Integrating Trait and Neurocognitive Mechanisms of Externalizing Psychopathology: A Joint Modeling Framework for Measuring Impulsive Behavior

Dissertation

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

By

Nathaniel Haines, B.A., M.A.

Graduate Program in Department of Psychology

The Ohio State University

2021

Dissertation Committee:

Theodore P. Beauchaine, Ph.D., Advisor

Brandon M. Turner, Ph.D., Co-Advisor

Trisha Van Zandt, Ph.D., Committee Member

Abstract

Trait impulsivity, defined by actions taken without forethought and a consistent preference for immediate over delayed rewards, confers vulnerability to all externalizing spectrum disorders. This includes all disorders along the common developmental progression of attention-deficit/hyperactivity disorder (ADHD) in early childhood to conduct disorder (CD) and delinquency in later childhood and adolescence, to substance use disorders (SUDs) and antisocial personality disorder (ASPD) in adulthood. Such externalizing progression derives from complex interactions among individual-level vulnerabilities and environmental risk factors over time. Specifying how such mechanisms interact across development is a burgeoning area of research. Although trait-level mechanisms have long been studied, research linking trait-level to behavioral mechanisms is more limited. Furthermore, most existing research uses standard inferential approaches, which are not well suited for modeling complex relations among causal influences at different levels of analysis. In this dissertation, I describe how both (1) the methods used to make inference on individual difference correlations across levels of analysis, and (2) the statistical models used to infer how data within levels of analysis arise often fail to fully embody the substantive theories that researchers aim to test. I use my prior work on the "Reliability Paradox" (Haines et al., 2020a) to demonstrate (1), and my work on the Iowa Gambling Task (Haines, Vassileva, & Ahn, 2018) to demonstrate (2). I then discuss a third study (Haines et al., 2020b) that shows how joint generative models across levels of analysis (between behavioral and trait mechanisms, behavioral and neural mechanisms, etc.) can be used to better capture individual differences of theoretical interest.

Acknowledgements

First, I would like to thank my family members for supporting me throughout this process, including my parents Blake and Anita Haines, and my siblings Nicole, Noah, Natalie, Nehemiah, and Nadia (yes, all N's!). Without you, I never would have been able to achieve something like this. I would also like to thank my amazing fiancé, Lydia Simon, who has helped me stay grounded throughout graduate school, the pandemic of 2020, and life more generally. Additionally, I am incredibly grateful for the amazing mentorship that I have been provided at The Ohio State University. Trisha Van Zandt and Robert Gore provided me with an environment to learn what exactly mathematical psychology and Bayesian statistics were all about, Jay Myung opened my eyes to the usefulness of cognitive/computational modeling in psychology and helped me realize that I had something to contribute to our understanding of the mind, Woo-Young Ahn gave me an opportunity and the resources necessary for me to refine my computational skills (despite a less-than-stellar undergraduate GPA...), Theodore Beauchaine passed on to me the importance of thinking both developmentally and in terms of "the big picture" (and also greatly influenced/improved my writing skills!), Brandon Turner both sharpened my technical and theoretical knowledge and guided me through tough career decisions, and Mark Pitt for helping be understand the importance of conveying complicated statistical methods to generalist audiences. Last but not least, I would like to thank everyone in my cohort and all the other graduate students who made graduate school an enjoyable process along the way.

Vita

2007 to 2011 ..... Jonathan Alder High School

2011 to 2015 ..... B.A. Psychology, The Ohio State University

2016 to 2017 ..... M.A. Psychology, The Ohio State University

2017 to 2020 ..... Graduate Teaching Associate, Department of Psychology, The Ohio State University

2020 to 2021 ..... Presidential Fellow, The Ohio State University

Publications

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., Pitt, M. A., Myung, J. I. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports, 10*, 12091. doi:10.1038/s41598-020-68587-x

Haines, N., Beauchaine, T. P. (2020). Moving beyond ordinary factor analysis in studies of personality and personality disorder: A computational modeling perspective. *Psychopathology*. doi:10.1159/000508539

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., Myung, J. I., Turner, B. M., Ahn, W.-Y. (2020). Anxiety modulates preference for immediate rewards in trait-impulsive individuals: A hierarchical Bayesian analysis. *Clinical Psychological Science, 8*, 1017-1036. doi:10.1177/2167702620929636

Romeu-Kelly*, Haines, N.*, R.J., Busemeyer, J.R., Ahn, W.-Y., Vassileva, J. (2020). Computational modeling of decision making performance on the Cambridge gambling task with applications to samples of drug users. *Drug and Alcohol Dependence, 206*, 107711. doi:10.1016/j.drugalcdep.2019.107711. *equal contribution

Hahn, H., Kalnitsky, S., Haines, N., Thamotharan, S., Beauchaine, T.P., & Ahn, W.-Y. (2019). Delay discounting of protected sex: Relationship type and sexual orientation influence sexual risk behavior. *Archives of Sexual Behavior, 48*, 2089-2102. doi:10.1007/s10508-019-1450-5

Haines, N., Bell, Z., Crowell, S. E., Hahn, H., Kamara, D., McDonough-Caplan, H., Shader, T., & Beauchaine, T. P. (2019). Using automated computer vision and machine learning to code facial expressions of affective valence and arousal: Implications for emotion dysregulation research. *Development and Psychopathology, 31*, 871-886. doi:10.1017/S0954579419000312

Beauchaine, T. P., & Haines, N. (2019). Functionalist and constructionist perspectives on emotion dysregulation. In T. P. Beauchaine & S. E. Crowell (Eds.), *The Oxford handbook of emotion dysregulation*. New York, NY: Oxford University Press. epublished ahead of print. ISBN:9780190689285

Haines, N., Southward, M. W., Cheavens, J. S., Beauchaine, T., & Ahn, W.-Y. (2019). Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity. *PLoS ONE*, *14*, e0211735. doi:10.1371/journal.pone.0211735

Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). The Outcome-Representation Learning model: A novel reinforcement learning model of the Iowa Gambling Task. *Cognitive Science, 47*, 1-28. doi:10.1111/cogs.12688

Rogers, A.H., Seager, I., Haines, N., Hahn, H., Aldao, A., Ahn, W.Y. (2017). The indirect effect of emotion regulation on minority stress and problematic substance use in lesbian, gay, and bisexual individuals. *Frontiers in Psychology, 8,* 1881. doi:10.3389/fpsyg.2017.01881

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry, 1*, 24-57. doi:10.1162/CPSY_a_00002

Fields of Study

Major Field: Psychology

Table of Contents

ABSTRACT II

ACKNOWLEDGEMENTS III

VITA IV

TABLE OF CONTENTS VI

LIST OF TABLES VIII

LIST OF FIGURES IX

CHAPTER 1: INTRODUCTION 1

1.1 Introduction 1
1.2 General Organization 3

CHAPTER 2: DEVELOPMENT OF EXTERNALIZING PSYCHOPATHOLOGY 4

2.1 Trait Impulsivity as the Core Vulnerability to ESDs 4
2.2 Neural and Neurocognitive Mechanisms of Trait Impulsivity 5
2.3 Moderators of Impulsivity 8
2.3.1 Trait Anxiety/Punishment Sensitivity 10
2.3.2 Executive Function and Top-Down Self-Control 13
2.3.3 Environmental Moderators of Trait Impulsivity 14
2.4 Interim Summary of ESD Progression 17

CHAPTER 3: PITFALLS OF TRADITIONAL BEHAVIORAL AND NEURAL MEASUREMENT. 20

3.1 Problems with Reliability 20
3.2 Problems with Theory 23
3.3 Summary of Problems with Traditional Behavioral Measures 25

CHAPTER 4: A FRAMEWORK FOR MEASURING NEUROCOGNITIVE MECHANISMS OF IMPULSIVITY 27

4.1 Joint Bayesian Modeling to Improve Reliability 27
4.2 Theoretically Motivated Generative Models to Improve Measurement 29

CHAPTER 5: THEORETICALLY INFORMED GENERATIVE MODELS CAN ADVANCE PSYCHOLOGICAL AND BRAIN SCIENCES: LESSONS FROM THE RELIABILITY PARADOX 32

5.1 Introduction 33
5.2 Method 43
5.2.1 Datasets and Behavioral Paradigms 43
5.2.2 Data Analysis 44
5.3 Results 57
5.3.1 Comparing Summary Statistics to Generative Model Parameters 64

5.4 Discussion 65

CHAPTER 6: THE OUTCOME-REPRESENTATION LEARNING MODEL: A NOVEL REINFORCEMENT LEARNING MODEL OF THE IOWA GAMBLING TASK 80

6.1 Introduction 81
6.1.1 Expected value 84
6.1.2 Win frequency 85
6.1.3 Perseveration 87
6.1.4 Reversal learning 88
6.1.5 The current study 89
6.2. Methods 90
6.2.1 Participants 90
6.2.2 Tasks 91
6.2.3 Reinforcement learning models 91
6.2.4 Hierarchical Bayesian analysis 99
6.2.5 Model comparison: Leave-one-out information criterion 101
6.2.6 Model comparison: Choice simulation 101
6.2.7 Model comparison: Parameter recovery 102
6.3. Results 104
6.3.1 Model comparison: Leave-one-out information criterion 104
6.3.2 Model comparison: Choice simulation 104
6.3.3 Model comparison: Parameter recovery 105
6.3.4 Applications to substance users 106
6.4. Discussion 107

CHAPTER 7: ANXIETY MODULATES PREFERENCE FOR IMMEDIATE REWARDS AMONG TRAIT-IMPULSIVE INDIVIDUALS: A HIERARCHICAL BAYESIAN ANALYSIS 120

7.1 Introduction 121
7.1.1 Approaches to Measuring Impulsivity 123
7.1.2 Modeling Functional Dependencies and Etiological Complexity 128
7.2 Objectives of The Current Study 130
7.3 Method 131
7.3.1 Participants 131
7.3.2 Measures 133
7.3.3 Behavioral Task 135
7.3.3 Procedure 137
7.3.4 Data Analysis 137
7.4 Results 146
7.4.1 State, Trait, and Behavioral Differences 146
7.4.2 Descriptive Models 147
7.4.3 Explanatory Models 149
7.5 Discussion 150

CHAPTER 8: CONCLUSIONS 161

BIBLIOGRAPHY 165

List of Tables

Table 6.1. Breakdown of datasets used in the current study. 111

Table 6.2. Mean squared deviations of true from simulated choice probabilities. 112

Table 7.1. Demographic Characteristics by Group. 155

List of Figures

Figure 2.1. Dopamine signaling as a mechanism of delay discounting..... 19
Figure 4.1. Bayesian joint modeling of behavioral and neural data..... 31
Figure 5.1. Pathway from theory to inference with behavioral data. 69
Figure 5.2. Qualitatively different distributions with the same mean. 71
Figure 5.3. Lognormal and shifted lognormal generative distributions..... 72
Figure 5.4. Building generative models consistent with theory. 73
Figure 5.5. Test-retest correlations and model misfit for the Stroop task. 75
Figure 5.6. Test-retest correlations for all tasks and models. 76
Figure 5.7. Relationship between two-stage estimates and generative model parameters. 78
Figure 6.1. Structure of the original and modified versions of the IGT..... 113
Figure 6.2. Post-hoc model fits across models and datasets. 114
Figure 6.3. True versus simulated choice proportions across time..... 115
Figure 6.4. Parameter recovery results across models and versions of the IGT..... 117
Figure 6.5. Group-level ORL parameters across healthy and substance using groups. 118
Figure 6.6. Differences in group-level ORL parameters between healthy and substance using groups. 119
Figure 7.1. Graphical depiction of the Explanatory model described in the main text. 156
Figure 7.2. Trait impulsivity, state anxiety, and behavioral impulsivity across groups. 158
Figure 7.3. Interaction of BIS-NP and STAI-S in predicting discounting rates for both Trait models. 159

Chapter 1: Introduction

1.1 Introduction

Progression of externalizing spectrum disorders (ESDs)—including attention-deficit/hyperactivity disorder (ADHD) in early childhood to oppositional defiant disorder (ODD), conduct disorder (CD), substance use disorders (SUDs), and antisocial personality disorder (ASPD) across the lifespan—is tremendously detrimental to individuals, families, and communities. This developmental pathway portends low academic achievement, underemployment, criminal behavior, incarceration, social rejection, and other negative outcomes such as suicidal ideation and suicide attempts (e.g., Barkley, Fischer, Smallish, & Fletcher, 2006; Beauchaine, Zisner, & Sauder, 2017; Biederman et al., 2008; Dyck et al., 1988; Fletcher, 2014; Hinshaw et al., 2012; Huemer et al., 2016; Loe & Feldman, 2007). In addition to interpersonal costs associated with such impairment, ESDs generate substantial economic burden. In 2010 alone, the US education system spent close to $25 billion on problems related to ADHD (Ruland, 2010), and state and federal corrections costs, which are driven disproportionately by

males with externalizing disorders (Teplin, 1994), exceed $70 billion annually (Schmitt, Warner, & Gupta, 2010).

Given tremendous interpersonal and societal costs of ESDs, prevention and early intervention are crucial. Targeting children who are most vulnerable to externalizing progression—typically those diagnosed with ADHD very early in life—reduces current symptoms, with benefits that accrue into adulthood (e.g., Conduct Problems Prevention Research Group, 2011; Jones, Daley, Hutchings, Bywater, & Eames, 2007; 2008; Webster-Stratton, Reid, & Beauchaine, 2011, 2013). Such benefits are observed 10-15 years after effective prevention programs, and include better reading literacy, fewer police contacts, less criminal justice system involvement, lower rates of ASPD, and fewer high-risk sexual behaviors, among other outcomes (e.g., Dodge et al., 2015; Scott, Briskman, & O'Connor, 2014).

Even without intervention, however, many children with ADHD—especially those reared in protective environments—do not progress to increasingly severe ESDs (see *Development of Externalizing Psychopathology* below). At the individual level, both traditional (e.g., regression-based) and contemporary (e.g., machine learning) models have limited prospective predictive utility given (1) inherent difficulties of specifying future events for low base rate phenomena (see Fusar-Poli, Hijazi, Stahl, & Steyerberg, 2018); (2) difficulties knowing which potentiating environment risk factors (e.g., trauma, deviant peer affiliations) any specific person will face as they mature; and (3) the importance of developmental timings of such exposures, which are also difficult to anticipate (Beauchaine & Hinshaw, 2020). In short, it is extraordinarily difficult to

model complex person × environment mechanisms over time, and predictive models are typically constructed atheoretically with regard to causal mechanisms. Consistent with broader perspectives in psychological science literature (e.g., Muthukrishna & Henrich, 2019), my aim is to develop a more formalized theory of the individual- and environmental-level mechanisms of ESDs, following the assumption that more precise, mathematical characterizations of such mechanisms provide insights into impulsive behaviors and their treatment across development.

1.2 General Organization

This proposal is organized as follows. First, in Chapter 2, I review current developmental models of ADHD within the broader context of ESDs summarized above. In Chapter 3, I review methods traditionally used when drawing inferences about impulsive behavior, and describe their limitations. I separate these methods into a discussion regarding: (1) the behavioral models used to make inferences about neurocognitive mechanisms, and (2) the statistical models used to make inferences about individual differences. In Chapter 4, I propose combining generative (or computational) models with hierarchical Bayesian analysis (HBA) as a joint framework to better understand relations among trait and neurocognitive mechanisms underlying ESDs. Chapters 5-7 then present three journal articles I have published or submitted that provide detailed theoretical and empirical arguments in favor of the proposed framework in Chapter 4. Finally, Chapter 8 provides a general conclusion and discusses future directions of the work I present in this dissertation.

Chapter 2: Development of Externalizing Psychopathology

2.1 Trait Impulsivity as the Core Vulnerability to ESDs

For many years, ESDs were characterized as distinct syndromes, as specified in the DSM-5 (APA, 2013). More recently, a growing consensus has emerged for a spectrum interpretation (see Beauchaine & Hinshaw, 2016; Krueger, Markon, Patrick, & Iacono, 2005). ESDs share genetic vulnerability (Gizer, Otto, & Ellingson, 2017; Krueger et al., 2002), and, as described above, follow a predictable sequence for those who develop more serious externalizing behavior as they mature (see Beauchaine et al., 2017). This heterotypic sequence typically begins with ADHD in early childhood, which progresses to oppositional defiant disorder (ODD), CD, delinquency, SUDs, and ASPD across development (Loeber, Keenan, & Zhang, 1997; Moffitt, 1993; Robins, 1966; Storebø & Simonsen, 2016). As noted above, all of these syndromes are characterized by trait impulsivity. In clinical samples, about half of adolescents recruited for ADHD have comorbid ODD, with lifetime co-occurrence rates that increase to around 80% across development (Biederman et al., 2008). Prospectively, children with ADHD that persists throughout development are at greater risk to develop CD relative to age-matched peers

(e.g., Gau et al., 2010), and those who develop CD are more likely to develop problematic alcohol and substance use (e.g., Elkins, McGue, & Iacono, 2007; Molina & Pelham, 2003; Pardini, White, & Stouthamer-Loeber, 2007). Given such findings, it is unsurprising that in factor analytic studies, all DSM externalizing disorders load on a single, higher-order latent factor (Krueger,1999; Tuvblad, Zheng, Raine, & Baker, 2009; Wright et al., 2013), which we refer to as trait impulsivity (Beauchaine & McNulty, 2013; Beauchaine et al., 2017).

It is important to note that progression of ADHD to more serious externalizing behaviors is specific to the hyperactive-impulsive (HI) and combined (C) presentations—not the inattentive (IN) presentation (Ahmad & Hinshaw, 2017; Beauchaine, Hinshaw, & Pang, 2010; Beauchaine et al., 2017; Elkins, McGue, & Iacono, 2007; Milich, Balentine, & Lynam, 2001). Results are similar for antisocial traits, which are elevated specifically among those with ADHD-HI and ADHD-C (Sprafkin, Gadow, Weiss, Schneider, & Nolan, 2007).[1] Accordingly, my focus is on mechanisms of ADHD-HI/C. Henceforth, I use the term ADHD to refer only to ADHD-HI/C.

2.2 Neural and Neurocognitive Mechanisms of Trait Impulsivity

A particularly compelling neural model of ADHD is the dynamic developmental theory (DDT; Sagvolden, Johansen, Aase, & Russell, 2005; see also Quay, 1997), which

[1]Other evidence also supports a distinction between ADHD-HI/C vs. ADHD-I. Relative to children with ADHD-HI/C, those with ADHD-IN exhibit drowsiness and lethargy (i.e. "sluggish cognitive tempo") as opposed to the distractibility characteristic of impulsivity, poorer response to stimulant medications, deficits in fronto-parietal rather than fronto-striatal connectivity, and different patterns of functional connectivity (see Adams, Derefinko, Milich, & Fillmore, 2008; Diamond, 2005; Fair et al., 2013; Lee, Burns, Beauchaine, & Becker, 2016; Martel et al., 2017; Martel, Nigg, & Von Eye, 2009).

provides a good starting point for more general models of ESDs (e.g., Luijten, Schellekens, Kühn, Machielse, & Sescousse, 2017; Plichta & Scheres, 2014). Behaviorally, the DDT defines trait impulsivity as a strong preference for immediate rewards over delayed rewards, difficulties inhibiting prepotent behaviors, and failures to consider consequences of immediate actions—fundamental components of ADHD-HI/C that are readily quantified using validated laboratory tasks (see e.g., Neuhaus & Beauchaine, 2017; Patros et al., 2016; Scheres, Tontsch, & Thoeny, 2013; Scheres, Tontsch, Thoeny, & Kaczkurkin, 2010).

Based on extensive animal research that has since been extended to humans, the DDT proposes that impulsive behaviors arise from low tonic dopamine (DA) activity and blunted phasic DA reactivity in neural circuits that subserve associative learning, including reward learning, extinction learning, and maintenance of previously rewarded behaviors. Foundational animal research shows that DA neurons in the substantia nigra, ventral tegmental area, and nucleus accumbens exhibit short-duration phasic bursts in response to unpredicted reinforcers (e.g., unexpected food rewards; Schultz, Dayan, & Montague, 1997; Schultz, 1998). Such phasic DA responses are one form of *positive prediction error*; they are elicited from contexts in which reward was unexpected (see e.g. Schultz, Dayan, & Montague, 1997). In contrast, when reinforcers are predicted (e.g., by conditioned stimuli such as learned signal tones), phasic DA reactivity is elicited at the earliest predictive signal (e.g., a tone) rather than the reinforcer itself (e.g., food). If an expected reinforcer (food) is then withheld following a learned, predictive signal (tone), phasic reduction in DA levels occurs, a *negative prediction error*

(Schultz et al., 1997). Through these associative learning processes, behaviors become linked with contingencies in local environments.

According to the DDT, diminished tonic DA activity and blunted phasic DA reactivity among those with ADHD-HI/C attenuates prediction errors and yields a narrower temporal window during which associative learning can occur. Behaviorally, both reward and extinction learning are affected (i.e. both positive and negative prediction errors), resulting in steeper *temporal discounting* (see Figure 2.1). This process results in immediate rewards and punishments being overvalued relative to delayed rewards and punishments—even when delayed consequences are larger. Steeper temporal discounting gradients are a common finding in children, adolescents, and adults with ADHD (Douglas & Parry 1994; Sagvolden, Aase, Zeiner, & Berger, 1998; Sagvolden et al., 2005; Scheres, Tontsch, & Thoeny, 2013; Tripp & Alsop, 2001). They are corroborated by neuroimaging work showing altered structure and function of key mesolimbic and mesocortical pathways among children with ADHD (Cole et al., 2011; Gatzke-Kopp et al., 2009; Kelly et al., 2009; Luman, Tripp, & Scheres, 2010; Rieckmann, Karlsson, Fischer, & Bäckman, 2011; Shannon, Sauder, Beauchaine, & Gatzke-Kopp, 2009; Tomasi & Valkow, 2014).

Although the DDT was initially an explanatory model of ADHD-HI/C, its relevance to other externalizing disorders has long been noted (Beauchaine et al., 2010, 2017; Gatzke-Kopp & Beauchaine, 2007; Gatzke-Kopp et al., 2009; Zisner & Beauchaine, 2016a, 2016b). Considerable neurobiological and behaviors evidence—in addition to the observed heterotypic continuity described above—indicates a shared

etiology among ADHD-HI/C, ODD, CD, SUDs, and ASPD. Like children with ADHD-HI/C, those with ODD and CD show reduced sensitivity to rewards (especially small rewards), which is also thought to arise from attenuated dopaminergic function (for a review see Matthys, Vanderschuren, & Schutter, 2013). Indeed, low striatal reactivity in anticipation of incentives is observed across disorders of impulse control, including but not limited to ADHD, current and remitted SUDs, and ASPD (e.g., Luijten et al., Oberlin et al., 2012; Plichta & Scheres, 2014; Sauder, Derbidge, & Beauchaine, 2016). Furthermore, steep temporal discounting is observed across all ESDs (e.g., Baker et al., 2003; Barker et al., 2015; Barkley et al., 2001; Beauchaine, Ben-David, & Sela, 2017; Bobova et al., 2009; Bornovalova et al., 2005; Haines et al., 2020b; Heil et al., 2006; Johnson et al., 2015; Mitchell et al., 1999; 2006; Ohmura et al., 2005; Petry, 2002; Reynolds et al., 2004; Scheres et al., 2010; Takahashi et al., 2009; Wilson et al., 2010). In fact, delay discounting has been proposed as a transdiagnostic index of impulsivity given its association with a broad range of externalizing psychopathologies (Amlung et al., 2019). Taken together, these findings suggest that the key neural mechanism identified by the DDT may affect all ESDs, and that temporal discounting paradigms or those that measure reward/punishment learning mechanisms provide a means of measuring such processes.

2.3 Moderators of Impulsivity

Although the DDT provides a good starting point to understand ESD progression, it is limited when viewed in isolation. For example, an extensive literature shows that DA activity/reactivity is controlled by numerous other neuromodulators (Cools,

Nakamura, & Daw, 2011; Doya, 2002; 2008; Long, Kuhn, & Platt, 2009; Macoveanu et al., 2013). Such complex interdependencies obscure simple 1:1 correspondences between neural/neurobiological activity and impulsivity at the behavioral level (Beauchaine & Hinshaw, 2020). Therefore, a major focus of our research group has been on *interactions* among neural systems that subserve behavior, including neural systems of approach/reward responding (described above), neural systems of avoidance/aversive responding, and neural systems of self-/emotion regulation (e.g., Beauchaine, 2001; Beauchaine & Cicchetti, 2019; Beauchaine, Katkin, Strassberg, & Snarr, 2001; Crowell et al., 2006; Sauder, Beauchaine, Gatzke-Kopp, Shannon, Aylward, 2012). A major premise of this work is that human behaviors are influenced by complex *functional dependencies* among a limited number of neural systems (Haines & Beauchaine, 2020; Haines et al., 2020b). Functional dependencies refer to situations in which neural or neurocognitive systems modulate one another to affect behavior. Overwhelming evidence suggests that trait anxiety, which is subserved by a largely independent neural network, modulates behavioral expressions of trait impulsivity. In the next section, I describe such modulatory effects. Importantly, modulatory effects of neural systems on one another make it difficult (and in many cases impossible) to construct explanatory models from overt behavior alone because any single behavior can result from diverse combinations of neural inputs (Beauchaine & Hinshaw, 2020; Haines & Beauchaine, 2020).

2.3.1 Trait Anxiety/Punishment Sensitivity

As described above, a portion of children with ADHD-HI/C go on to develop conduct problems. Children and adolescents who do develop conduct problems, like others in the population, vary considerably on their levels of trait anxiety (see Beauchaine et al., 2017; Schatz & Rostain, 2006). Notably, those with low symptoms of anxiety show more aggression, get along worse with peers, have more police contacts, and show worse responses to treatment (Beauchaine, Webster-Stratton, & Reid, 2005; Jensen et al., 2001; Walker et al., 1991). Relatedly, children with ADHD who also have elevated callous-unemotional traits—which are characterized by low trait anxiety—tend to develop more conduct problems than their peers (e.g., Enebrink, Andershed, & Långström, 2009; Frick & White, 2008; Tremblay, Pihl, Vitaro, & Dobkin, 1994). Although longitudinal work is limited, better functional outcomes for individuals with comorbid anxiety suggest that trait anxiety is protective against externalizing progression.

Functional dependencies between trait impulsivity and trait anxiety are captured by the *joint subsystems hypothesis* of reinforcement sensitivity theory (RST; Corr, 2001; 2004; Corr & McNaughton, 2012; see also Gray, 1987). RST proposes that *state anxiety* is elicited during situations characterized by *conflicting motivational valence* (approach-avoidance, approach-approach, or avoidance-avoidance). In such situations, anxiety is functional because it suppresses ongoing behavior—including impulsive behavior—so organisms (in this case people) can evaluate the safest and/or most advantageous course of action. From this perspective, high *trait anxiety* is presumed to set a lower threshold

for goal conflict and suppression of approach behaviors in contexts where such behaviors might bring about consequences (competing motivational valence). Neurally, trait anxiety is modulated by a broad, coordinated neural network including the septo-hippocampal system, the amygdala, the locus coeruleus, the raphe nuclei, and the hypothalamus (e.g., Gray & McNaughton, 2000; Silva & McNaughton, 2019). This system modulates approach behaviors through striatal-amygdalar projections (see e.g., Dong, Li, & Kirouac, 2017). Neural systems of anxiety are mediated by multiple neurotransmitters, including serotonin, acetylcholine, norepinephrine, and GABA (e.g., Ntamati, Creed, Achargui, & Luscher, 2018; Teles-Grilo Ruivo & Mellor, 2013).

Children with both ADHD-HI/C and ODD/CD show especially dampened sensitivity to punishment relative to reward cues across behavioral (e.g., learning paradigms), physiological (e.g., electrodermal responding), and neural (e.g., BOLD responses using fMRI) levels of analysis (Beauchaine et al., 2001; Crowell et al., 2006; Gao, Raine, Venables, Dawson, & Mednick, 2010a, 2010b; Gatzke-Kopp et al., 2009; Jones, Laurens, Herba, & Viding, 2009; Marsh, 2008; Posthumus, Böcker, Raaijmakers, van Engeland, & Matthys, 2009; Sterzer, Stadler, Krebs, Kleinschmidt, & Poustka, 2005; van Bokhoven, Matthys, van Goozen, & van Engeland, 2005). Such findings are similar to those observed among adults with SUDs (e.g., Haines, Vassileva, & Ahn, 2018; Romeu, Haines, Ahn, Busemeyer, & Vassileva, 2019). Notably, externalizing males with comorbid anxiety show less severe structural compromises in the striatum and anterior cingulate cortex (ACC)—neural regions involved in generating reward responses and prediction errors during associative learning (Sauder et al., 2012). The ACC plays a

critical role in error monitoring and representing the likelihood of unexpected events (see Alexander & Brown, 2019). Boys with externalizing disorders also show blunted ACC reactivity when incentives are eliminated prior to previously rewarded behaviors (Gatzke-Kopp et al., 2009). Among adults, experimentally-induced anxiety attenuates reward value encoding in the ventromedial PFC (Engelmann, Meyer, Fehr, & Ruff, 2015). Neurobiologically, interactions between impulsivity and anxiety are likely affected by functional dependencies between neurotransmitters, where DA represents prediction errors (as described by the DDT) and serotonin, acetylcholine, and norepinephrine modulate other aspects of associative learning such as reward cost/risk/uncertainty, learning rate, and exploration/exploitation, respectively (Cools, Nakamura, & Daw, 2011; Doya, 2002; 2008; Long, Kuhn, & Platt, 2009; Macoveanu et al., 2013).

Taken together, findings discussed thus far provide a means to explain impulsivity-anxiety interactions in the DDT framework. Blunted mesolimbic (and mesocortical) DA activity induces impulsive behaviors through its effects on temporal discounting. However, impulsivity is modulated by neural mechanisms of trait anxiety and their effects on inducing goal conflict and uncertainty/risk representations (see also Frost & McNaughton, 2017; Silva & McNaughton, 2019). If high trait anxiety perpetually induces goal conflict or uncertainty, approach behaviors are suppressed (behavioral inhibition)[2]. Conversely, if low trait anxiety yields limited goal conflict or uncertainty, impulsive behaviors are potentiated, with limited regard for risk. My own work

[2]It is worth noting that strong empirical associations have been identified between behavioral indices of risk aversion and clinical levels of anxiety (e.g., Charpentier, Aylward, Roiser, & Robinson, 2017; Maner et al., 2007). In addition, those with anxiety disorders display chronic engagement of neural networks that represent risk including the ACC and amygdala (e.g., Robinson et al., 2014).

demonstrates how state anxiety can inhibit impulsive delay discounting among trait impulsive individuals through its effects on risk aversion (see Chapter 7).

2.3.2 Executive Function and Top-Down Self-Control

Executive function (EF) and related constructs including self-control and emotion regulation is a third factor relevant to ESD progression. Collectively, EF and emotion regulation comprise a set of are top-down processes (e.g., attention, working memory, response inhibition, problem solving, reappraisal, etc.) through which behavioral and affective responses are shaped in the service of goal-directed behaviors (see Thompson, 1990). Children with ADHD-HI/C score consistently lower than their peers on EF (e.g., Nigg, Blaskey, Huang-Pollock, & Rappley, 2002), and emotion dysregulation among children with ADHD is linked prospectively to delinquency, arrests, academic underachievement, and unemployment (Sasser et al., 2016).

EF and emotion regulation are subserved by highly overlapping prefrontal systems, including the ACC, the orbitofrontal cortex (OFC), and the dorsolateral, ventrolateral, and ventromedial PFCs (see Beauchaine, 2015; Zelazo, 2015). These neural systems are recruited during effortful downregulation of impulsive and/or anxious responses (see e.g., Beauchaine & Cicchetti, 2019; Beauchaine & Zisner, 2017; Mochcovitch, da Rocha Freire, Garcia, & Nardi, 2014, Peters & Büchel, 2011; Willcutt, Doyle, Nigg, Faraone, & Pennington, 2005). Furthermore, both the dlPFC and vmPFC are intricately involved in valuing reward and choosing between smaller immediate vs. larger delayed rewards (e.g., Turner et al., 2018).

For typically developing children, neural systems of approach (impulsivity) and avoidance (anxiety) are anatomically mature in early childhood, whereas frontal regions responsible for EF and emotion regulation mature into early adulthood (see Casey, Heller, Gee, & Cohen, 2019). For children with externalizing problems, frontal neuromaturation is more protracted, and in some cases, absent. Compared with typically developing children, those with ADHD show delayed cortical neuromaturation (e.g., Shaw et al., 2012), with volume deficiencies becoming even more pronounced in adolescence among males with comorbid ADHD and CD (De Brito et al., 2009), Further blunting of brain growth is observed in adolescence for those who initiate heavy drinking (Pfefferbaum et al., 2018). Thus, ESD progression is associated with increasingly severe structural compromises in key neural networks responsible for regulating impulsivity, with under-development predicting future progression of ESDs (e.g., Brumback et al., 2016).

2.3.3 Environmental Moderators of Trait Impulsivity

Findings discussed thus far suggest that progression of ESDs is most likely when individuals (1) have reduced sensitivity to both rewards and punishments compared to controls; (2) are less sensitive to punishments relative to rewards; and (3) have difficulties down-regulating impulsive responding to immediate rewards to facilitate better long-term decision-making. When neural models are considered, these mechanisms map roughly onto trait impulsivity, trait anxiety, and EF, respectively, although mapping is not 1:1 due to complex interactions among mechanisms, as described in previous sections.

In addition to individual-level trait and neural/neurocognitive factors, external factors including the *frequency*, *consistency*, and *intensity* of rewards and punishments in a child's environment also affect development of ESDs. A history of research dating back to the mid-1960s shows that heterotypic continuity among ESDs (and development of most any human behavioral trait; see Sameroff, 2010) occurs *transactionally* as individuals interact with their local environments over time (see Beauchaine & McNulty, 2013; Beauchaine et al., 2017). For example, coercive family dynamics including harsh/hostile and inconsistent parenting mediate prospective relations between ADHD-HI/C and later delinquent behaviors characteristic of ODD/CD (e.g., Ahmad & Hinshaw, 2017; Bell, Fristad, Youngstrom, Arnold, & Beauchaine, 2021; Lorber & Egeland, 2011; Patterson, DeGarmo, & Knutson, 2000). In later childhood and adolescence, affiliation with deviant peers mediates prospective relations between CD and both delinquency and substance use (e.g., Dishion, McCord, & Poulin, 1999; Dishion & Racer, 2013). Recent findings show that cumulative effects of childhood stress, including maltreatment and abuse, can lead to blunted reward-related responding in the ventral striatum into adulthood (Birn, Roeber, & Pollak, 2018; Hanson et al., 2016)—a potential mechanism through which impulse control problems are shaped and maintained (e.g., Guendelman, Owens, Galán, Gard, & Hinshaw, 2015).

As children following the ESD trajectory become more independent from caregivers, the decisions they make can select for environments with a higher likelihood of risk exposure, and pre-existing vulnerabilities can be potentiated (see Beauchaine et al., 2010; Beauchaine & McNulty, 2013). In the genetics literature, this type of interaction is

often referred to as a gene by environment correlation (i.e. where genetic liability can lead to selection of an environment that further potentiates certain traits or symptoms), but I focus on general person-level factors (e.g., neurocognitive mechanisms) rather than genes to retain generality. For example, one state-wide, population-based study (N=85,000) showed that delinquent behavior—including threatening, stealing, and harming others—increases with age from 10-19 years and is most pronounced among boys and girls who both (1) score high trait impulsivity and (2) live in neighborhoods that are less safe, allow easier access to drugs and alcohol, and are less communal (Meier, Slutske, Arndt, & Cadoret, 2008; see also Lynam et al., 2000). Similarly, affiliation with deviant peers mediates prospective relations between childhood ADHD and alcohol use during late adolescence (e.g., Marshal, Molina, & Pelham, 2003; Marshal & Molina, 2006). General availability of substances of abuse and peer influences also play key roles in initiation of substance use and later development of SUDs (see Iacono, Malone, & McGue, 2008). Once initiated, continued substance use compounds pre-existing neurocognitive vulnerabilities due to effects on brain structure and function (e.g., Pfefferbaum et al., 2018), which—like childhood maltreatment and abuse—can potentiate tendencies, thereby furthering externalizing progression (see Volkow & Morales, 2015). Heavy alcohol, which is linked to steeper declines in frontal gray matter volume throughout adolescence, provides one such example (e.g., Pfefferbaum et al., 2017), and is associated with dramatically higher risk for development of antisocial and borderline personality characteristics (see Moran, 1999; Sher & Trull, 2002). Thus, an important avenue for future research is to identify relations between macro-level

individual differences that take place over long time scales (e.g., traits, accumulated exposure to rewards/punishments, etc.) and more micro-level individual differences including changes in decision-making and neural function (e.g., sensitivity to rewards/punishments, aversion to delayed rewards, etc.). More generally, the role of environmental reward/punishment contingencies in shaping impulsive behaviors across the lifespan suggests that the models we use to make inference from behavioral data should account for how people interact with their environments in a way that modifies impulsive behavior across time.

2.4 Interim Summary of ESD Progression

The transactional processes between individuals and their environments described above are best instantiated by *Ontogenic Process Models*, which describe multidirectional links among genetic, neural, cognitive, behavioral, and environmental levels of analysis as they interact across development to eventuate in observable individual differences—including ESDs (see Beauchaine & McNulty, 2013; Beauchaine et al., 2017; Macdonald, Goines, Novacek, & Walker, 2016; Senner, Conklin, & Piersma, 2015). Ontogenic process accounts ESDs suggests that progression along the externalizing spectrum depends on individual factors including (1) sensitivity to rewards, (2) sensitivity to punishments relative to rewards, (3) executive function/emotion regulation, and (4) various and complex environmental-contextual factors—which interact over the course of development to eventuate in psychopathology (see Beauchaine & Hinshaw, 2020). However, designing studies and developing statistical models to test theories that span neural, behavioral, and environmental levels

remains challenging (see also Luman, Tripp, & Scheres, 2010). I describe some these challenges in detail below before offering a theoretical and computational framework as a partial solution.

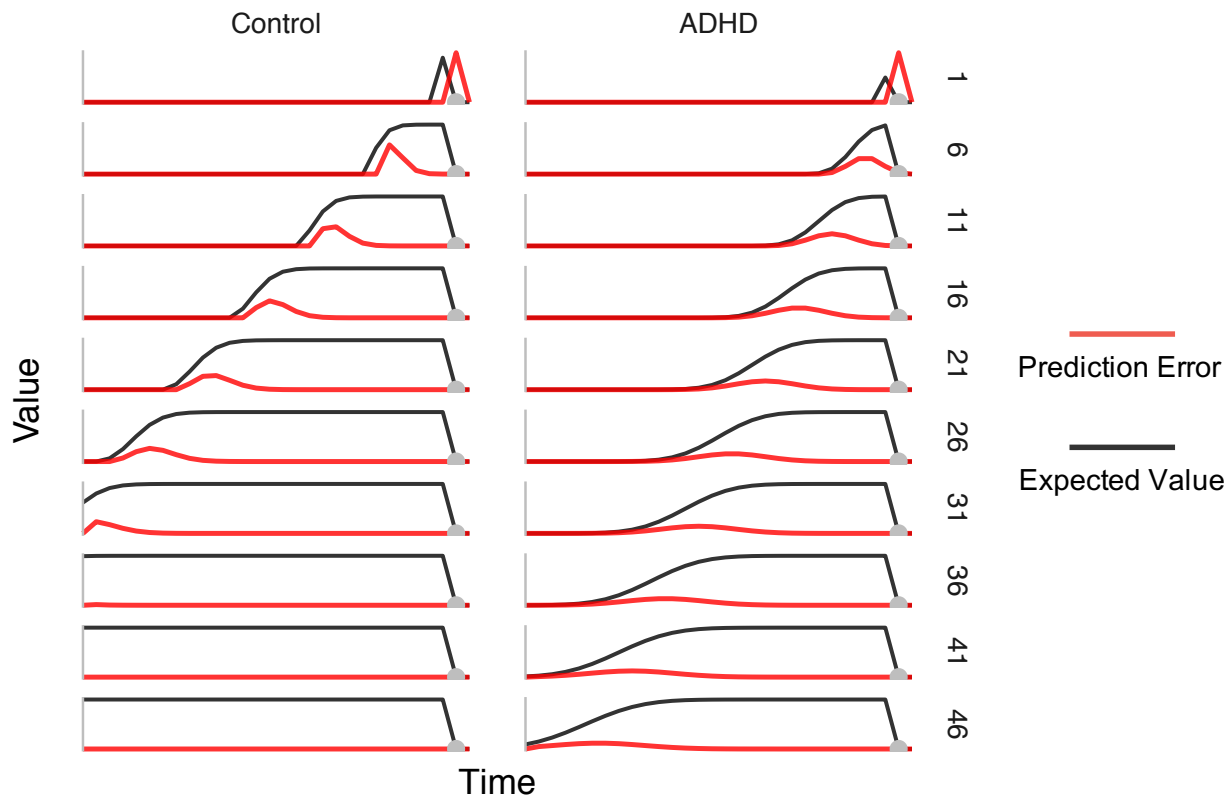

Figure 2.1. Dopamine signaling as a mechanism of delay discounting.

Here I illustrate how DA prediction errors (red) propagate backward in time from an initially unexpected reward (gray point toward the end of each panel) to the earliest stimulus that predicts receipt of reward (vertical gray line at the start of each panel) over repeated stimulus-reward pairings (trial numbers along the right side of the panels). I generated these curves using a simplified temporal discounting model (see Schultz, Dayan, & Montague, 1997) where the time dimension is treated as the "state", and the stimulus treated as an "action". The only difference between the simulated control and ADHD models is that the ADHD model has a lower learning rate (lower α parameter in the model), which theoretically corresponds to hypodopaminergic activity. In this way, hypoactivity at the level of DA signaling produces an effect where the control model learns to anticipate and assign value to the future reward given the stimulus at a much faster rate relative to the ADHD model (see also Sagvolden et al., 2005).

Chapter 3: Pitfalls of Traditional Behavioral and Neural Measurement

3.1 Problems with Reliability

As described above, the research community has long sought to link observable impulsive behaviors to underlying neural and cognitive mechanisms. One hope is that a better understanding of casual mechanisms will lead to more targeted treatments. However, the inferential approaches used in most mainstream research rely on strong yet often implicit assumptions that are usually not met in practice, and there is a dearth of applied work using more contemporary statistical methods that allow us to relax these assumptions. Here, I focus on relating behavioral data to neural data, but the same logic applies when linking data across any two (or more) levels of analysis. One example is described by Hedge, Powell, and Sumner (2017), who showed that behavioral tasks used across the cognitive and clinical neuroscience literatures show poor test-retest reliability and are therefore poorly suited for making inferences about individual differences. Specifically, behavioral measures of self-control and behavioral inhibition that can reliably differentiate at the group level (e.g., differentiating between those with

and without psychopathology) yet show unacceptably low test-retest intra-class correlations (ICCs) (e.g., Stop-Signal Reaction Time ICC $\approx .45$)[3]. Even the Stroop effect—one of the most robust, replicable phenomena in psychology—exhibits test-retest correlations as low as ICC $\approx .60$. These findings have since been replicated and extended, with a meta-analysis of self-control measures derived from behavioral tasks showing a median test-retest ICC $\approx .31$ (Enkavi et al., 2019). These test-retest reliabilities fall well below traditional acceptability thresholds often used in psychopathology research (e.g., recommended test-retest $r \geq .7$), which limits their utility for making inferences about trait-like individual differences relative to more reliable self-report measures (see Dang, King, & Inzlicht, 2020; Wennerhold & Friese, 2020).

Such problems are exacerbated by similarly low reliabilities for many neural measures used throughout clinical and cognitive science (e.g., task-based functional magnetic resonance imaging [fMRI], electroencephalography [EEG], etc.). Meta-analyses of graph-theoretic functional connectivity networks, for example, show test-retest ICCs $\approx .3$ (e.g., Chen et al., 2015; Noble, Scheinost, & Constable, 2019). Meta-analyses

[3]Note that ICCs as used here are meant to measure something different than Cronbach's α. In particular, ICCs are similar to Pearson's correlation (i.e. r) in that both can be used to determine the correlation between two paired, continuous variables, although ICCs also take into account the mean (as opposed to just variance/covariance) changes across timepoints (e.g., if all subjects perform better by a constant amount at the second timepoint, relative to no constant changes, the ICC will be attenuated whereas the Pearson's correlation would be unaffected). Therefore, the ICC is used to determine if there are consistent/temporally stable between-subject differences in task performance across timepoints (i.e. test-retest correlations). Conversely, Cronbach's α is primarily a measure of internal consistency, which is meant to estimate the amount of signal to noise (or true score to error score) attributed to a measure within a single timepoint given an assumed measurement model. We focus on test-retest reliability because we are interested in measuring individual differences over time, a form of reliability that Cronbach's α could easily underestimate (for a more in-depth discussion on reliability, see DeVellis, 1991).

of task-based fMRI measures (e.g., blood oxygen level dependent responses to task contrasts) show equally poor psychometric properties, with a mean ICC ≈ .40 (Elliot et al., 2019). Of particular relevance to the current study, task-based fMRI measures of amygdala response to emotional faces show test-retest reliabilities ranging from ICC = .0 to ICC = .68, depending on the specific stimuli and form of ICC used (e.g., those that take into account absolute agreement/mean changes show lower estimates than those that only account for consistency/between-subject variance).

Given such modest test-retest reliabilities across behavioral tasks and neural measures, making inferences about relations between neural functions and behaviors—an essential objective for evaluating complex ontogenic process models of psychopathology—requires highly powered studies using sample sizes that are much larger than those in most published research. For example, when behavioral and neural measures both have test-retest reliabilities of ICC = .6 and the true effect size is medium ($r = .30$), we need *239 participants* to reach 80% power with $\alpha = .05$ (Hedge et al., 2017). Much larger samples are required for subgroup analyses, tests of interactions, etc. Yet prior to 2018, of the 100 most cited articles using neuroimaging with clinical populations, the average sample size was 63 (Gong et al., 2019). This yields low signal-to-noise ratio situations (high measurement error, small sample size) that are conducive to type *S* and *M* errors, whereby effect sizes are estimated to have the wrong sign and to be of much larger magnitude than the "true" effects, are likely (Gelman & Carlin, 2014). This type of underpowered research plays a primary role in difficulties with

reproducibility in psychological research (see Loken & Gelman, 2017; Tackett, Brandes, King, & Markon, 2019).

3.2 Problems with Theory

A separate but related issue concerns widespread use of atheoretical behavioral models in which suboptimal statistical methods are applied to both behavioral and neural data. As I describe below these methods reduce precision (increase error) and fail to account for complex dynamics suggested by contemporary models of psychopathology. For example, many different behavioral paradigms measure different "facets" of impulsivity, risk-seeking, and otherwise maladaptive decision-making. Such paradigms include the Iowa Gambling Task (IGT) (Bechara, Damasio, Damasio, & Anderson, 1994), the Cambridge Gambling Task (CGT) (Rogers et al., 1999), the Balloon Analogue Risk Task (BART) (Lejuez et al., 2002), and delay and probability discounting tasks (e.g., Green & Myerson, 2004; McKerchar, & Renda, 2012), among others. Oftentimes, behavioral performance on these tasks is quantified using summary statistics of participants' choices (e.g., average number of risky choices, area under the discounting curve, etc.) I term the mathematical model (which is often implicit when using summary scores) used to infer underlying processes from behavioral data the *behavioral model*. Resulting summary scores are subsequently interpreted as reflecting some underlying decision-making process (e.g., "myopia" of future consequences on the IGT). Some of the largest studies to date on test-retest reliability have used the summary statistic approach (e.g., Hedge et al., 2017).

Despite their widespread use, summary measures do not disentangle possible neurocognitive processes/decision-making mechanisms that could lead to observed behavior. For example, when analyzing data collected from the IGT, participants are tasked with choosing among four "decks" with the goal of accruing as many points as possible. Unbeknownst to participants, two of the decks lead to long-term gains ("advantageous" decks) whereas two lead to long-term losses ("disadvantageous" decks). However, due to the structure of the task, consistent selection of disadvantageous decks could result from a number of different factors including (a) preference for decks with frequent gains, (b) a tendency to win-stay or lose-switch, (c) attentional biases toward positive or negative outcomes, (d) choice inconsistency, or (e) a failure to learn the long-term consequences of each choice (e.g., Busemeyer & Stout, 2002; Haines, Vassileva, & Ahn, 2018). As a result, two participants can be assigned the same behavioral summary score despite having significant differences in underlying neurocognitive mechanisms. In other words, even simple human behaviors can arise from complex combinations of causal factors, as described above in my discussion of RST. Behavioral summary scores therefore limit inferences about mechanisms (i.e., etiology). As a result, when a participant performs poorly on the IGT (or any other behavioral task), this information may be of limited value in predicting behavior on other laboratory tasks that recruit different neurocognitive mechanisms, and predicting real-world impulsive behavior is even more tenuous. Indeed, studies on the convergent validity of IGT and BART summary measures suggest they reflect independent latent factors, and may therefore index "different type[s] of decision-making" (Buelow & Blaine, 2015, p. 781) despite the

tasks sharing many features. Similar conclusions have been drawn from a host of other decision-making tasks developed as behavioral measures of impulsivity and self-control (see Cyders & Coskunpinar, 2011; Duckworth & Kern, 2011). However, it is important not to conflate the behavioral measurement model with the assumed underlying behavioral construct/mechanism when drawing such conclusions. Effective behavioral models isolate underlying decision-making mechanisms that we seek to measure such that tasks with similar features can be meaningfully compared—otherwise we risk making inappropriate comparisons across measures and drawing the wrong conclusion (see Navarro, 2020).

3.3 Summary of Problems with Traditional Behavioral Measures

In sum, many behavioral measures of impulsivity show both (1) low test-retest reliabilities, and (2) low convergent validity with other measures designed to measure the same decision-making mechanisms. These shortcomings limit the utility of behavioral measures for theory development and real-world applications pertaining to individual differences. They also make it unfeasible to conduct well-powered studies on individual differences in many settings. Nevertheless, most previous studies on trait-behavior or brain-behavior relations relevant to externalizing psychopathology have relied exclusively on summary measures of behavioral performance to index cognitive processes (i.e., the "summary statistic as behavioral model approach") as opposed to theoretically motivated models designed to represent how choice features relate to behavior. This overreliance on heuristic summary statistics artificially limits the utility

of behavioral data. Below I describe a framework that is based on my own and others' work which provides a partial solution to these problems.

Chapter 4: A Framework for Measuring Neurocognitive Mechanisms of Impulsivity

4.1 Joint Bayesian Modeling to Improve Reliability

Recent developments in computational and model-based neuroscience offer a partial solution to problems with estimating relations across behavioral and neural levels of analysis despite high measurement error within each level. Hierarchical Bayesian analysis (HBA) is a statistical framework that allows for *joint modeling* (e.g., joint trait-behavior, brain-behavior, or across any two or more levels of analysis). HBA uses both (1) the hierarchical structure of data (e.g., trials or repetition times nested within participants for behavioral or neural data), and (2) shared information between different sources of data to simultaneously estimate person- and group-specific model parameters across sub-models for each data source (e.g., Turner et al., 2013; Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016). When combined with models that capture theoretically important cognitive/behavioral processes, joint modeling provides a framework for developing and testing formal models of complex etiological processes (cf. Beauchaine & Hinshaw, 2020).

Figure 4.1 illustrates how joint modeling is used to simultaneously constrain parameter estimation of behavioral and neural data model parameters. From a theoretical perspective, joint modeling used in this manner allows us to determine if formal computational models of behavior are associated with activity in the brain. This is accomplished by making explicit assumptions about cognitive mechanisms that generative observed behavior, where the goal then shifts from estimating relations between summary measures (the traditional approach) to estimating relations between generative model parameters. For example, if we are interested in simple correlations between behavioral and neural measures, we assume that model parameters from both behavioral (θ) and neural (δ) models (captured with a generative behavioral model and a model of the hemodynamic response function, respectively) are drawn from a linking distribution (Ω, e.g., a multivariate normal distribution). This allows observations at behavioral and neural levels to mutually constrain one another by sharing (pooling) information across modalities (Turner et al., 2013; Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016). This pooling of information yields more precise model parameter estimates at the individual level, allowing us to identify individual differences in neurocognitive mechanisms with greater precision. For example, in a re-analysis of the Stroop task data from Hedge et al. (2017), we showed that joint Bayesian modeling produces Stroop effect test-retest estimates upwards of $r \approx .8$, contrasting the traditional summary statistic estimate of $r \approx .5$ (Haines et al., 2020a). This Improvement results from jointly estimating relations between model parameters rather

than estimating correlations between summary measures. I discuss these results in detail in Chapter 5 below.

4.2 Theoretically Motivated Generative Models to Improve Measurement

Despite these benefits, a critical challenge for implementing joint modeling is the requirement that use of formal generative models of behavior that can re-produce (or "simulate") patterns of trial-to-trial behavior consistent with observed data (sometimes referred to as computational models). Although such generative models are uncommon in the psychopathology literature, they are used widely in other areas of research, including mathematical psychology (Townsend, 2008), value-based decision-making and neuroeconomics (Rangel, Camerer, & Montague, 2008), and computational psychiatry (Hauser, Will, Dubois, & Dolan, 2019; Huys, Maia, & Frank, 2016; Wiecki, Poland, & Frank, 2015). Generative models have been developed that re-produce observed behavior based on assumptions regarding how people represent, evaluate, and learn the features relevant to engaging in particular behaviors (e.g., reward amount, delay, probability). As a result, generative models are being increasingly adopted by psychological scientists (see Guest & Martin, 2020), although they are still rather rare. In Chapter 6, I discuss a generative model of the Iowa Gambling Task that I developed based on principles of reward and punishment learning and show how such a model can facilitate mechanistic insights into impulsive behavior not afforded by traditional summary approaches.

Finally, although the joint modeling framework as described above has been mostly applied in the context of brain-behavior modeling, it is easily extendable beyond behavioral and neural measures (see Haines & Beauchaine, 2020). For example, a factor-analytic or item-response theory style model could be used to identify latent traits from self-report measures that can then be included in the linking distribution. Alternatively, traits can be related directly to specific behavioral model parameters to test theoretical trait-behavior relations (e.g., Haines et al., 2020b), which I demonstrate in Chapter 7. Because uncertainty across all measures is included in a single model, joint modeling alleviates problems with reliability that emerge when each level is analyzed in a non-hierarchical, multi-stage fashion (see Rouder & Haaf, 2019). Further, since joint modeling is inherently hierarchical, it yields both individual- and group-level parameters that can be used to test theoretical predictions (e.g., of correlations between neurocognitive mechanisms and traits, of group differences neurocognitive mechanisms, etc.). Using HBA to jointly model relationships across different data sources (each with their own generative sub-model) therefore offers a principled approach to identifying individual-level mechanisms that give rise to impulsive behaviors related to ESDs. It also facilitates inferences from smaller samples that traditional multi-stage approaches (i.e. when behavioral summary measures are entered into secondary statistical models) are underpowered for (e.g., Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017), which is important to consider for clinical research.

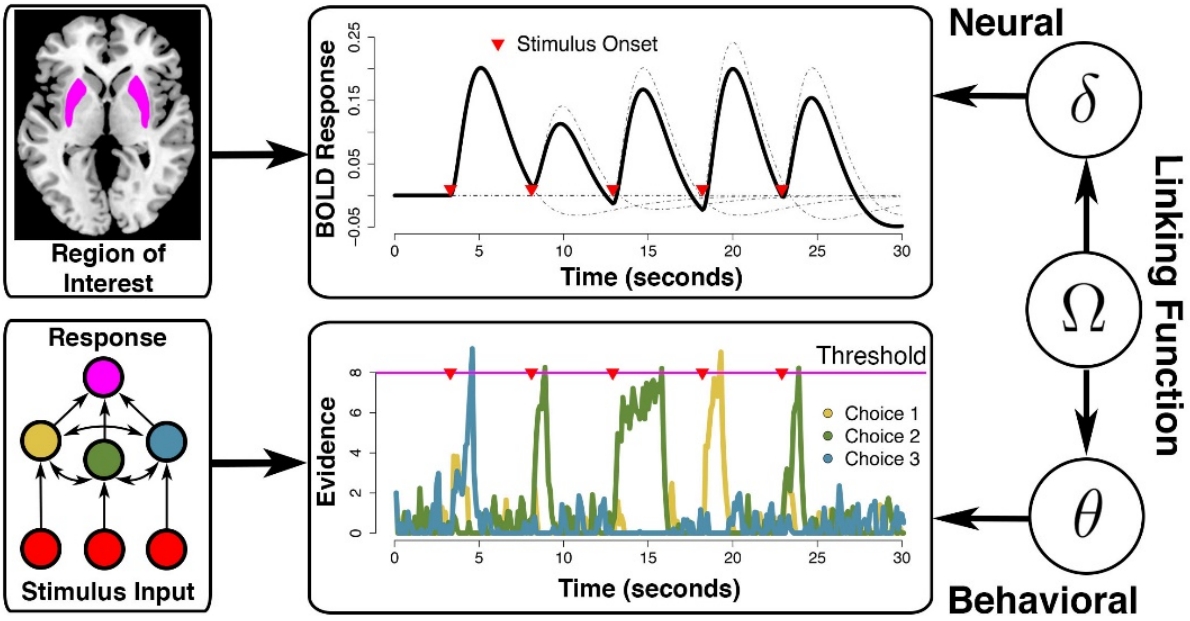

Figure 4.1. Bayesian joint modeling of behavioral and neural data.

Joint Bayesian modeling can be used to estimate both neural data parameters (e.g., neural contrasts from a general linear model) and computational model parameters (e.g., from a decision-theoretic or evidence accumulation model) within a single statistical model, which increases precision of parameter estimates at all levels. Although a joint model between neural and behavioral sub-models is shown here, in principle, joint Bayesian modeling can be conducted by linking any number of sub-models. For example, in Chapter 5, I demonstrate how a joint model with two behavioral sub-models can be used to better estimate test-retest correlations. See details in the section titled Joint Bayesian Modeling to Improve Reliability. Image reprinted from Turner, Palestro, Miletić, & Forstmann (2019) with permission.

Chapter 5: Theoretically Informed Generative Models Can Advance Psychological and Brain Sciences: Lessons from the Reliability Paradox

In this chapter, I present a manuscript submitted to Nature Communications as a general overview on how joint Bayesian modeling using generative models of behavior can reveal higher test-retest reliability in behavioral model parameters relative to traditional summary statistics approaches. Therefore, this chapter mainly demonstrates how joint modeling improves reliability as described briefly in Chapter 4.1, although it also touches on the importance of theoretically informed generative models of behavior described briefly in Chapter 4.2. Note that throughout the Chapter, Equations are numbered within the Chapter (as opposed to containing chapter indicators) to preserve the original text and its relationship to the online supplemental materials. This Chapter began as a blog post in late 2019 (http://haines-lab.com/post/thinking-generatively-why-do-we-use-atheoretical-statistical-models-to-test-substantive-psychological-theories/), which I then extended in collaboration with Peter D. Kvam, Louis Irving, Colin T. Smith, Theodore P. Beauchaine, Mark A. Pitt, Woo-Young Ahn, and Brandon M. Turner.

5.1 Introduction

Across the social, behavioral, and brain sciences, the researcher's primary agenda is to provide an explanation for *why* observable measures (i.e., data) exhibit systematic patterns (Hempel & Oppenheim, 1948). Typically, this process begins with the development of a theory or hypothesis about what should happen in specific situations (e.g., experiments) if our assumptions about the phenomena are correct. To test our theory, we design an environment that places the agent under consideration (e.g., human, rat) into a specific situation, and we collect data that describe the agent's experience (e.g., decisions, neural activity). To evaluate the fidelity of our hypothesis, we use statistical models such as *t*-tests, ANOVAs, multiple (logistic) regression, and factor analytic models, to connect our explanatory theories to the data observed in our experiment (Guest & Martin, 2020; Kellen, 2019; Suppes, 1966). We use these statistical models to make inferences, such as inferring population-level effects or estimating out-of-sample predictive accuracy using in-sample data, and relate these inferences to claims about theory. However, underlying our statistical models are causal or distributional assumptions that are too often mis-aligned with the substantive theories of interest, a situation we refer to as a *theory-description gap*. When assumptions are misaligned, theories become "divorced" from the statistical tests meant to validate or invalidate

them, and the disconnect impedes progress in science (Michell, 2008; Szollosi & Donkin, 2019; Yarkoni, 2019).

In this article, we demonstrate how *generative modeling*—a statistical modeling approach wherein model specification is intentionally aligned with theory—can alleviate the pressures associated with theory-description gaps. We illustrate how generative modeling resolves a vexing theory-description gap that is currently shaking the foundations of research on individual differences: the *reliability paradox*. The reliability paradox refers to the counterintuitive result that person-level measures of behavior across a broad range of tasks (e.g., the Implicit Association Test, Stroop, Flanker, and Posner Cueing tasks) and modalities (e.g., accuracy, response times, task-based and resting-state fMRI) show poor test-retest reliability despite task manipulations showing consistent and robust effects at the group level (Chen et al., 2015; Elliott et al., 2020; Enkavi et al., 2019; Gawronski, Morrison, Phills, & Galdi, 2017; Hedge, Powell, & Sumner, 2017; Noble, Scheinost, & Constable, 2019). The reliability paradox is important because low test-retest reliability implies that the psychological constructs on which theories are based are unstable across time, making the task or modality untrustworthy for validating theories based on temporally stable individual differences (Dang, King, & Inzlicht, 2020; Elliott et al., 2020; Schimmack, 2019; Wennerhold & Friese, 2020). As we demonstrate, however, the reliability paradox arises from the implicit, overly restrictive assumptions that researchers make when using standardized statistical models that fail to connect their theories to data in *meaningful* ways. To illustrate how *generative models* can overcome the reliability paradox, we use response

time measures as our observable data as an in-depth example, but the principles and benefits of generative modeling are much more general, applying to any measure extracted from data (e.g., accuracy). We include accuracy/preference data from the Delay Discounting task to demonstrate this generality, although we do not explore the task in detail here.

As a concrete example, consider the Stroop task (Stroop, 1935) which asks participants to indicate either the word or text color of a word presented on a screen. Words and colors can be either congruent ("red" in red text), neutral ("boat" in red text), or incongruent ("blue" in red text). In a typical experiment using the Stroop task, our observable measures are distributions of responses and response times for each person, each condition, and each perhaps session in the case of test-retest measurements. As illustrated in Figure 5.1, the "Stroop effect" is typically estimated for each person as the difference in mean response times (across trials) between incongruent and congruent trials, such that a positive value indicates that a person takes longer (on average) to respond to incongruent relative to congruent trials. This difference in mean response time serves as what we call the *behavioral model*, which is the mathematical model used to make inferences about person-level behavior. To estimate a test-retest correlation, the person-level Stroop effect values from one session are used as dependent variables in a secondary statistical model predicting the same participants' Stroop effect values from a later session. If a Stroop effect is stable, we should be able to use each person's Stroop effect at one point in time to accurately predict their Stroop effect at a later point in time. Yet if such an analysis shows that our behavioral measures have low test-retest

correlations—as revealed by the reliability paradox—it is difficult to imagine how we could use them to make inference on other individual differences relevant for testing our theories (e.g., estimating correlations between behavioral and neural or self-report measures). Altogether, we refer to the aforementioned approach to statistical inference as the "two-stage summary approach", given that summary statistics are estimated in one stage and then an individual difference correlation (i.e. test-retest) is estimated independently in a second stage.

The two-stage summary approach makes two strong yet inappropriate assumptions about within-person behavior that invalidate the test-retest correlations that are derived from it. Although we provide more substantive mathematical justification in the Methods section, the intuition is as follows: When we compute mean response times in task conditions, take their difference, and then input these difference scores into a secondary statistical model (*t*-test, etc.), we are making the (first) assumption that the first-stage estimates (difference scores in this case) are estimated with no measurement error, or equivalently that either (1) we have an *infinite number of trials* within each condition used to estimate the mean response time difference, or (2) response times within persons show no variation from trial-to-trial (Ly et al., 2017; Rouder & Haaf, 2019; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017a; Vandekerckhove, 2014). The same critique applies any time we summarize our data before entering the resulting summaries into a secondary model, irrespective of whether these summaries are sample means or point estimates derived from a classical, machine learning, or even Bayesian models. The key point is that when we ignore the fact that *summary statistics*

are estimates, and not observed measures, we ignore sources of uncertainty at one level of analysis that should be incorporated into our secondary statistical models (Davis-Stober, Park, Brown, & Regenwetter, 2016; Estes, 1956; Heathcote, Brown, & Mewhort, 2000; Liew, Howe, & Little, 2016; Pagan, 1984; Turner, Schley, Muller, & Tsetsos, 2018). In the case of test-retest analyses, ignoring uncertainty dramatically attenuates test-retest correlations, pushing them toward zero. In such cases, researchers understandably (yet incorrectly) conclude that the effect of interest is inconsistent across repetitions, even when the "true" effect is stable and consistent, as we demonstrate below.

The second problematic assumption is that the sample mean of a response time distribution (or choice proportions if measuring accuracy/preference) is an accurate way of connecting the substantive theory under investigation to observed data. In fact, the use of sample means in this way assumes a specific generative model of within-person behavior—namely, that response times within each person and condition arise from a normal (i.e., Gaussian) distribution (see Methods section). Researchers may be unaware that they assume this *implicit generative model* when using the two-stage summary approaches to assess test-retest reliability (or other individual difference correlations), and as a result they may erroneously infer that tasks or measurement modalities, rather than behavioral models, confer low test-retest reliability. To the extent that this assumption is inconsistent with empirical data, we will fail to detect individual differences that are otherwise readily apparent in empirical data. For example, Figure 5.2 shows that many different distributions, the shapes of which a theory must explain,

can all have the same mean. Without constructing generative models of how individual people behave, we fail to capture important aspects of observed data—including variance (Johnson & Busemeyer, 2005), bimodality (Kvam, 2019), and skew (Kvam & Busemeyer, 2020; Leth-Steensen, Elbaz, & Douglas, 2000)—which may otherwise invalidate the conclusions we draw from our mis-specified statistical models. Indeed, a history of examples ranging from basic research in cognitive science, to more policy-level research on eyewitness detection accuracy and racial bias in police shootings shows that *atheoretical* use of summary statistics (per the implicit generative models that they assume) can even lead to conclusions that are opposite of those derived from more theoretically-motivated generative models (Haines et al., 2020; Heathcote, Popiel, & Mewhort, 1991; Kellen, 2019; Romeu, Haines, Ahn, Busemeyer, & Vassileva, 2019; Ross, Winterhalder, & McElreath, 2020; Rotello, Heit, & Dubé, 2014). Notably, these shortcomings apply *irrespective to the replicability of the results* obtained using inappropriate or otherwise mis-specified statistical models (Devezer, Nardin, Baumgaertner, & Buzbas, 2019; Devezer, Navarro, Vandekerckhove, & Buzbas, 2020; Regenwetter & Robinson, 2017). In Supplementary Note 1, we provide a simulation that demonstrates the importance of distributional information on estimating stability in person-level behavior across time (i.e., test-retest reliability).

Generative modeling is a framework rooted in mathematical psychology and computational cognitive (neuro)science that resolves these issues by constructing statistical models that directly instantiate theoretical assumptions about how observed data are generated (Ahn & Busemeyer, 2016; Busemeyer, Gluth, Rieskamp, & Turner,

2019; Guest & Martin, 2020; Huys, Maia, & Frank, 2016; Montague, Dolan, Friston, & Dayan, 2012; Navarro, 2021; Rangel, Camerer, Camerer, & Montague, 2008; Townsend, 2008; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017a; Turner et al., 2013; Wiecki, Poland, & Frank, 2015; Wilson & Collins, 2019). To build a generative model, we begin with a *behavioral model*. The behavioral model captures our assumptions about how individual people generate data—in this case, of a person's distribution of response times across trials (and condition type, etc.). Our choice of behavioral model should be informed by theoretical assumptions and other background knowledge. For response time data, we know that response time distributions (1) must take on only positive real values, (2) have some central tendency with corresponding variability around this central tendency (e.g., a mean and variance), (3) are often right skewed, (4) are shifted away from 0 due to "non-decision" factors (e.g., visual encoding time, motor response time), and (5) typically show a linear relation the sample mean and standard deviation such that increases in mean response time are accompanied by increased variability of response times from trial-to-trial (the *law of response time*) (Wagenmakers & Brown, 2007).

There are many different distributions that can incorporate our knowledge about response times, and thus the process(es) that generated them. Here, we limit our discussion to the normal, lognormal, and shifted-lognormal distributions. We first show how the parameters of the lognormal models produce changes in predicted response time distributions (Figure 5.3). Figure 5.4 shows how the two-stage summary approach can be viewed as a highly constrained version of the normal generative model, in addition to

how the two lognormal models further improve upon shortcomings of the normal generative model. Given that these behavioral models can better characterize the full distribution of person-level response times, we can be more confident in using model parameters to infer meaningful consistencies (or inconsistencies) in behavior. In fact, it follows from the Fisher-Neyman factorization theorem that, if our assumed generative model is sufficient, variation in estimated model parameters will be *lower than or equal to* variation in summary statistics extracted from the same data across different conditions (in our case, test-retest sessions) (Casella & Berger, 2002; Fisher, 1922; Turner & Van Zandt, 2012). Intuitively, because behavior is inherently probabilistic, observed response times offer only limited information regarding each person's data-generating parameters (e.g., μ and σ). So long as we use summary statistics or generative models that are not sufficient to fully capture the data-generating process (e.g., only sample means or the normal generative model when response times are highly skewed), we ignore information that could otherwise inform or constrain parameter estimation. An implication is that, to the extent that we have more mechanistic behavioral models informed by the theory under investigation (e.g., computational models; see Guest & Martin, 2020; Jarecki, Tan, & Jenny, 2020; Wilson & Collins, 2019), they can be used to help us constrain inference in a way not afforded by the traditional two-stage summary approach.

The two-stage summary approach also makes a strong implicit assumption about between-person variation in behavior: That person-level parameters (those depicted in Figure 5.4) are distributed *uniformly* across an interval that spans far beyond the

reasonable range of parameter values, such that knowing the value of one person's parameters offers no information at all about the value of another person's parameters. For example, imagine that we collected data and estimated Stroop effects for *all but one* undergraduate psychology student at a large university, and we are then tasked with making an informed guess on this out-of-sample student's Stroop effect. Following the logic of the two-stage summary approach, we assume that the known distribution of undergraduate Stroop effects offers no information at all to help us make an informed guess! To the extent that the distribution of person-level parameters follow a shape that is not uniform (e.g., if they are normally distributed across people), the two-stage summary approach ignores information about between-person variation that can help us obtain more precise estimates of person-level parameters (in addition to test-retest or other individual difference correlations). A history of work in both mathematical statistics and psychometrics shows that we can obtain more precise estimates of the "true" person-level parameters by pooling information across people using either hierarchical models (which assume a group-level distribution over person-level parameters) or related methods that "shrink" person-level parameter estimates toward the group-level mean in proportion to how uncertain the person-level parameter estimates are (see the Methods section for details) (Efron & Morris, 1977; Gelman, 2006; Williams, Martin, DeBolt, Oakes, & Rast, 2020). In our case, uncertainty arises from both the inherent probabilistic nature of behavior (i.e. variation in response times within persons, or σ), and the limited number of trials that we observe within persons/conditions. Therefore, by defining an explicit group-level generative model of

the person-level parameters themselves (as illustrated in the "Generative Model" row of Figure 5.1), the group-level model can pool information across people while accounting for uncertainty at the person level that would otherwise produce biased, attenuated individual differences correlations using the two-stage summary approach. In fact, the test-retest correlation obtained from a full generative model spanning from within- to between-person variation is akin to the "true" or "dis-attenuated" correlation that one obtains when using classical corrections for attenuation due to unreliability (Rouder & Haaf, 2019; Williams et al., 2020), although the generative formulation allows for much more flexibility in our choice of both group- and person-level models. Here, we assume that person-level parameters follow correlated (i.e. multivariate) normal distributions, and we use this same specification for all three behavioral models (see the Methods section for details and visual depictions).

To demonstrate the power of generative modeling in closing the theory-description gap in the context of the reliability paradox, we re-analyzed data from three prior studies. We compare results obtained from the two-stage summary approach with those using the three generative models (Figure 5.4) across five common behavioral tasks. We analyze data from the Stroop, Flanker, and Posner Cueing tasks (Hedge et al., 2017), the Self-Concept (introversion/extraversion) and Race (Black/White) versions of the Implicit Association Test (IAT) (Gawronski et al., 2017), and the Delay Discounting task (Ahn et al., 2020). We include the Delay Discounting dataset to demonstrate that benefits of generative modeling extend beyond response time data. Individually, each of these tasks has yielded a deep body of literature—the Stroop, Flanker, and Posner

Cueing tasks have been used extensively to develop theories of attention and inhibitory control, the IAT has been used to develop theories of implicit cognition and evaluations with deep societal implications, and the Delay Discounting task has been used to develop theories of impulsivity and self-control. On Google Scholar alone (as of April 2021), the collective citation count of the original papers pertaining to these tasks is over 57,000 (B. A. Eriksen & Eriksen, 1974; Green & Myerson, 2004; Greenwald, McGhee, & Schwartz, 1998; Mazur, 1987; Posner, 1980; Stroop, 1935). Moreover, the five tasks cover areas of research spanning from psychology to neuroscience to behavioral economics.

5.2 Method

5.2.1 Datasets and Behavioral Paradigms

Given that the Stroop task is the running example throughout this article, we describe the details of the Stroop task from Hedge et al. (Hedge et al., 2017) below. We provide details of all other tasks and datasets in Supplementary Note 5.

For the Stroop task, two sets of participants ($n = 47$, $n = 60$ for Studies 1 and 2) performed the task in two separate sessions three weeks apart. The main effect of interest is the contrast between congruent and incongruent conditions. Specifically, participants responded to the color of a word (either red, blue, green, or yellow). The word could be the same as the font color (e.g., the word "red" in red font; congruent condition or $c = 1$), a non-color word (e.g., "ship"; neutral condition), or a color word mapping onto another response option (e.g., the word "red" colored blue, green, or

yellow; incongruent condition or $c = 2$). Participants completed 240 trials in each of the three conditions.

5.2.2 Data Analysis

5.2.2.1 Data Preprocessing

For all tasks involving response times, we removed trials with response times recorded as < 0, assuming that such trials could not be part of the data-generating process. For the delay discounting task, we did not remove trials. Keeping all trials (except negative response times) can help identify regions of model misfit that offer insights into cognitive mechanisms that get obscured by oversimplified preprocessing choices (e.g., removing trials with response times less than 100 milliseconds). Such heuristic preprocessing choices tend to have strong, unpredictable effects on inference (Parsons, 2020). Therefore, liberal inclusion criteria not only keep our models consistent with the goals of generative modeling but also demonstrate the utility of hierarchical modeling.

5.2.2.2 Two-Stage Summary Approach

The two-stage approach proceeds by reducing behavior within each participant to a point estimate before entering the resulting point estimates into a secondary statistical model to make inference. Below, we describe its implementation for each task.

5.2.2.2.1 Response Time Tasks

For the IAT, Stroop, Flanker, and Posner Cueing tasks, our first analysis followed the two-stage summary approach described in the introduction. We computed mean

contrasts across task conditions by taking the mean in each condition and then subtracting the congruent from incongruent condition mean response time (see Figure 5.1). In addition, we computed standard deviation contrasts for comparison with the generative models (i.e., standard deviations of incongruent condition response times minus standard deviations of congruent condition response times). To estimate test-retest reliabilities, we computed Pearson correlations across participants for the mean and standard deviation contrasts.

5.2.2.2.2 Delay Discounting Task

We used maximum likelihood estimation to estimate discounting rates (k) and choice sensitivity parameters (c) from a hyperbolic model for each participant and session, followed by Pearson correlations across participant to estimate test-retest reliabilities of model parameter point estimates (see Supplementary Notes 2 and 3 for details). We compare these estimates to a hierarchical Bayesian estimation approach described below.

5.2.2.3 Generative Modeling Approach

If the goal is to make group-level inferences, hierarchical models can appropriately account for person-level uncertainty. Further, hierarchical models can increase precision of person-level parameter estimates. Below, we extend the concept of generative modeling from person- to group-level model parameters.

5.2.2.3.1 Person-level Response Time Models

To characterize response time data, a generative model must obey some of the very simple properties outlined in the introduction. The normal (Gaussian) distribution is

perhaps the simplest behavioral model that can generate a full distribution of response times. Despite not capturing skew and allowing negative values, it can still be useful for shifting away from the summary statistic approach. At the minimum, the normal distribution characterizes both the central tendency and the variance or spread of the response time distribution.

For example, each person's response times in the Stroop task can be conceptualized as arising from a separate normal distribution. Parameters from each distribution (e.g., means/standard deviations) are specific to each person within each task condition. The Stroop effect can be characterized by within-participant changes in the shape of response time distributions across trials within conditions. When using a normal generative distribution, the shape of the response time distribution is characterized by changes in the mean and standard deviation parameters across congruent and incongruent condition trials for each participant. We can write the normal generative model as:

$$\mathbf{RT}_{i,c,t} \sim \mathcal{N}(\mu_{i,c,t}, \sigma_{i,c,t}) \tag{1}$$

where $\mathrm{RT}_{i,c,t}$ contains the set of response times for participant i in condition c during experimental session t. The notation $\mathrm{RT} \sim N(a, b)$ signifies that the response times are drawn from a generative process of a normal distribution (N) with mean a and standard deviation b. In Equation 1, the collection of response times in each block of our experiment are separately characterized by a specific mean ($\mu_{i,c,t}$) and standard deviation ($\sigma_{i,c,t}$). Hence, whereas the behavioral model implied by the two-stage summary approach reduces the response time data into a single summary statistic per

condition, the behavioral model in Equation 1 reduces the data into two parameters per condition—parameters which, as we discuss below, can be assessed in terms of their own mean and variance (Williams, Mulder, Rouder, & Rast, 2021).

The two-stage summary approach implicitly assumes Equation 1 as the data-generating model, because the sample mean is the analytical maximum likelihood estimator for the μ parameter in a normal distribution. However, the two-stage approach assumes the mean is estimated without measurement error. There are only two cases in which this assumption is valid. Specifically, recall that the standard error on the sample mean ($\sigma_{\hat{\mu}}$) (and correspondingly, the μ parameter of a normal distribution) is $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$. In our case, σ corresponds to the standard deviation of a person's response time distribution ($\sigma_{i,c,t}$ in Equation 1), and n to the number of trials they underwent in the given condition. Assuming no measurement error is equivalent to assuming that $\sigma_{\hat{\mu}} = 0$, which only occurs as either $\sigma \to 0$ or $n \to \infty$. Of course, we know *a priori* that each person's distribution of response time will have variability from trial-to-trial (i.e. $\sigma > 0$), and that we have a limited number of trials (i.e. $n \ll +\infty$). The normal generative model captures this knowledge, allowing joint estimation of μ and σ while appropriately accounting for uncertainty given the limited number of trials collected to estimate these parameters.

To facilitate interpretation, we will introduce a relabeling of the terms in Equation 1 based on the conditions they correspond to. First, we label the congruent condition (i.e., the first condition $c = 1$) as a baseline condition, where $\mathrm{RT}_{i,1,t} = \mathrm{RT}_{i,base,t}$, characterized by a baseline mean $\mu_{i,1,t} = \mu_{i,base,t}$ and baseline standard deviation

$\sigma_{i,1,t} = \exp(\sigma_{i,base,t})$. To isolate the effects of interference, or Stroop effects, we labeled a parameter Δ to signify the change from the baseline condition to the condition of interest (e.g., incongruent condition). This means that $\mathrm{RT}_{i,2,t}$ is characterized by a mean $\mu_{i,2,t} = \mu_{i,base,t} + \mu_{i,\Delta,t}$ and standard deviation $\sigma_{i,2,t} = \exp(\sigma_{i,base,t} + \sigma_{i,\Delta,t})$. Note that the *base* and Δ standard deviation parameters are estimated on the log scale and then exponentially transformed to ensure they are greater than 0. Therefore, the test-retest correlation for the Δ standard deviation parameters is on the log scale. See the Supplementary Note 3 for more details.

The normal generative model is better at characterizing distributional changes in response times across conditions and advancing theoretical knowledge relative to the two-stage summary approach. Nevertheless, the model is limited because it does not obey all the simple properties of response times outlined in the introduction. In particular, the normal model (1) can produce negative response times, and (2) cannot capture asymmetric variance with respect to the mean (i.e., right skew). One simple adjustment is to logarithmically transform the response time data and assume a normal model on the transformed data, a process equivalent to assuming the response time data come from a different generative model called the lognormal distribution. Given this equivalence, we can specify a more theoretically consistent generative model as:

$$\mathbf{RT}_{i,c,t} \sim \mathrm{Lognormal}(\mu_{i,c,t}, \sigma_{i,c,t}) \tag{2}$$

With this adjustment, parameters $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ will have very different abilities when characterizing the many shapes of response time distributions. In the lognormal model, the mean and standard deviation interact according to the *law of response time*. That

is, an increase in either parameter, holding the other constant, produces an increase in both the mean and standard deviation of the response times predicted by the model. Figures 5.3A and 5.3B show how changes in either parameter change the shape of the predicted response time data. Each possible distribution shape represents a prediction about how a participant's response time data should look, where the possible shapes are constrained by our commitments (or hypotheses) regarding the data-generating process (i.e. the lognormal model).

The lognormal model is an improvement over the normal model but still misses one important property of response time data. It is well-established that different response modalities (e.g., key press, mouse click, or verbal response) can produce shifts in response time distributions, even when the task demands and underlying evidence accumulation dynamics are identical (Gomez, Ratcliff, & Childers, 2015). Typically, extra time taken to interact with the stimuli and apparatus is not considered part of the decision process and is instead referred to as "non-decision" time to make the theoretical position clear. Although non-decision factors seem unimportant, they may compromise our ability to accurately characterize response time data. For example, suppose a person completes a Stroop task in two conditions where they respond either verbally or by pressing a key. Even assuming the person follows the same decision process in identifying the color of the word (i.e., they have the same $\mu_{i,c,t}$ parameter), differences in executing the response across conditions are still likely. For example, if responding verbally is quicker than via key press, we would expect the response times to be shifted higher relative to the verbal condition. In this case, fitting the lognormal distribution to

the observed response times would lead to different estimates for $\mu_{i,c,t}$ across the two conditions because the simple lognormal is not specified correctly according to the demands of the experiment. Consequently, having different estimates for $\mu_{i,c,t}$ might result in different interpretations about cognitive factors across the two contexts when non-decision factors actually caused the differences.

A simple solution is to adjust the lognormal distribution by introducing an additional parameter δ to move the distribution a distance of δ away from zero. Figure 5.3C illustrates the effect of δ on a specific lognormal distribution. To impose some theoretical constraints, we could assume δ is specific to each person and is unlikely to change between conditions within a behavioral task. This assumption would be inappropriate in our example above, but it would be justified for the analyses we perform in later sections due to the method of data collection. With this new shift parameter and imposed theoretical constraints, we can now specify a shifted-lognormal model as:

$$\mathbf{RT}_{i,c,t} \sim \text{Shifted-Lognormal}(\delta_{i,t}, \mu_{i,c,t}, \sigma_{i,c,t}) \tag{3}$$

where $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ have the same interpretations as described for Equation 2, and $\delta_{i,t}$ indicates the amount of shift or "non-decision time" specific to each person at each of the two experimental sessions.

5.2.2.3.2 Group-level Response Time Models

We have now defined generative models of person-level behavior for the response time tasks (normal, lognormal, and shifted lognormal models). The next step toward building full generative models of test-retest reliability is to define group-level

probability distributions for person-level parameters. Starting with the three response time models, we assume that all i person-level parameters in the congruent task condition at each of the two sessions t are drawn from a normal group-level distributions with unknown means and standard deviations:

$$\begin{aligned} \mu_{i,\text{base},t} &\sim \mathcal{N}(\mu_{\text{mean,base},t}, \mu_{\text{sd,base},t}) \\ \sigma_{i,\text{base},t} &\sim \mathcal{N}(\sigma_{\text{mean,base},t}, \sigma_{\text{sd,base},t}) \end{aligned} \tag{4}$$

The group-level normal distributions here are considered prior models (or prior distributions) on the person-level parameters. Estimating group-level parameters from prior models allows for information to be pooled across participants such that each person-level estimate influences its corresponding group-level mean and standard deviation estimates, which in turn influence all other person-level estimates. This interplay between person- and group-level parameters produces regression of person-level estimates toward the group mean (also referred to as *hierarchical pooling*, *shrinkage*, or *regularization*), which increases precision of person-level estimates (Gelman, 2006). The specific amount that person-level parameters are pooled toward the group-level mean is proportional to the uncertainty or reliability when estimating the person-level parameters. In fact, from the perspective of classical test theory, the pooling incurred through the hierarchical model is a form of "true score estimation", wherein the hierarchical pooling factor produces person-level estimates equivalent to reliability-based corrections under certain conditions (Rouder & Haaf, 2019; Williams et al., 2020). Note that the two-stage summary approach can be viewed as a special case of Equation 4 in combination with the normal behavioral model, wherein we assume that $\mu_{\text{sd,base,t}} \approx +\infty$, $\sigma_{\text{mean,base},t} \approx -\infty$, and $\sigma_{\text{sd,base},t} \approx 0$ (keeping in mind that $\sigma_{i,\text{base},t}$ is on the log scale).

In this highly restricted case, the group-level distribution across $\mu_{i,\text{base},t}$ is essentially uniform, and there is subsequently no pooling of $\mu_{i,\text{base},t}$ toward $\mu_{\text{mean},\text{base},t}$. Further, the person-level standard deviation parameters ($\sigma_{i,\text{base},t}$) are all assumed to be approximately 0, for reasons given in the text surrounding Equation 1 (see also 5.4). Therefore, in the same way that the normal behavioral model can be viewed as a relaxation of the overly restrictive (and implicit) assumptions underlying the two-stage summary approach, the normal group-level model (Equation 4) relaxes the assumption that person-level parameters follow a uniform distribution.

Beyond the relation between hierarchical pooling and reliability-based corrections, the normal distributions function similarly for person-level latent parameters in Equation 4 as they do for observed response times in Equation 1. Both cases assume that a normal distribution at one level of analysis generates observed or unobserved data at another level (e.g., observed response times are generated by normal distributions within participants, with unobserved means and standard deviations generated from normal group-level distributions). *This joint specification of relations between parameters over all levels of analysis embodies the generative perspective*. It allows for group- and person-level parameters to be estimated simultaneously, thus avoiding the pitfalls of the two-stage approach entirely (we illustrate the effect of these generative assumptions on person-level parameters in Figure 5.5). Although we do not demonstrate it here, the group-level model (i.e., Equation 4) itself is of theoretical interest—it can be extended to estimate relations between personality traits and

decision mechanisms (Haines et al., 2020b), or to generalize parameter estimates beyond non-representative samples (Kennedy & Gelman, 2019).

To estimate test-retest reliability, we can assume that person-level change parameters (e.g., $\mu_{i,\Delta,t}$ and $\sigma_{i,\Delta,t}$) are correlated across sessions. Staying true to the generative perspective, we can estimate this correlation by assuming scores are drawn from a multivariate normal distributions rather than independent normal distributions as in Equation 4:

$$\begin{bmatrix}\mu_{i,\Delta,1}\\ \mu_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\mu_{\text{mean},\Delta,1}\\ \mu_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\mu}\right)$$
$$\begin{bmatrix}\sigma_{i,\Delta,1}\\ \sigma_{i,\Delta,2}\end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix}\sigma_{\text{mean},\Delta,1}\\ \sigma_{\text{mean},\Delta,2}\end{bmatrix}, \mathbf{S}_{\sigma}\right) \quad (5)$$

Using a multivariate normal distribution allows us to estimate covariances ($\mathbf{S}_{\mu}$ and $\mathbf{S}_{\sigma}$ matrices) between person-level parameters across sessions that can be decomposed into group-level parameter variances and the correlation between person-level parameters across sessions—this correlation represents the test-retest reliability of the generative model parameters (see Supplementary Note 3 for mathematical details). If the correlation is zero, then Equation 5 is equivalent to Equation 4 (i.e. the normal distributions are independent). As the correlations become stronger, however, person-level parameters become increasingly pooled toward a common regression line, which in our case represents the test-retest correlation. In the same way that Equation 4 produces person-level estimates that are "corrected" for unreliability, Equation 5 produces a test-retest correlation that is corrected for attenuation due to unreliability (or uncertainty) in person-level estimates. Therefore, correlations extracted from the full

generative model indicate the "true" (where true indicates the statistical expectation) temporal stability of the person-level parameters given the assumed generative model (Rouder & Haaf, 2019). Similar to the discussion surrounding Equation 4, the two-stage summary approach can be viewed as a special case of Equation 5, wherein $\mu_{\mathrm{sd},\Delta,\mathrm{t}} \approx +\infty$, $\sigma_{\mathrm{mean},\Delta,t} \approx -\infty$, and $\sigma_{\mathrm{sd},\Delta,t} \approx 0$. In this highly restricted case, the test-retest correlation obtained from the normal generative model is equivalent to what is obtained with the two-stage summary approach.

For the shifted lognormal model, we estimated a single shift parameter for each participant at each timepoint (assuming that shift is equivalent between task conditions). Details about the shift parameter specification and prior distributions for group-level parameters in Equations 4-5 are available in Supplementary Note 3.

5.2.2.3.3 Delay Discounting Model

Delay discounting data are often modeled using a hyperbolic discounting model, which includes a discounting rate (k) that indicates how steeply a person discounts delayed rewards and a choice sensitivity parameter (c) that indicates how randomly versus deterministically a person chooses between smaller-sooner versus larger-later rewards based on their valuations for each. The mathematical details of this model are available in Supplementary Note 3. Extending the person-level hyperbolic delay discounting model to a full generative model that can estimate test-retest reliability follows the same logic as outlined for response time models. We used the same multivariate normal distribution parameterization to estimate test-retest correlations between discounting rate (k) and choices sensitivity (c) parameters.

5.2.2.4 Parameter Estimation

A benefit of Bayesian estimation is that after specifying a joint probability model (i.e. the full group- and person-level generative model), it is possible to compute conditional probabilities that determine which parameter values are most credible given the observed data. This results in *posterior distributions* over model parameters that express the probability of a specific value for the parameter given the model and data. Because computing conditional probabilities analytically requires solving complex and often intractable integrals, Bayesian model parameters are typically estimated using numerical integration methods. We estimated parameters from all models using Stan (version 2.19.2), a probabilistic programming language that uses a variant of Markov Chain Monte Carlo to estimate posterior distributions for parameters within Bayesian models(Carpenter et al., 2017). Details are described in Supplementary Note 3.

5.2.2.5 Sensitivity Analyses of Group-level Model Specification

We tested multiple different group-level models to determine how sensitive our results were to changes in generative assumptions. We report details in Supplementary Information, but offer a brief overview of results here. First, we used parameter recovery simulations to show that we can accurately recover the "true" underlying test-retest correlation using the full generative models as described in the main text (see Supplementary Note 4 and Figure S2). Second, we tested an alternative group-level model wherein both the person-level base and change parameters were drawn from separate multivariate normal distributions, as opposed to only the change parameters (the "Joint Separate" model in Supplementary Note 7 and Figure S8). Finally, we tested

another group-level model wherein we directly estimated the $\mu_{i,c,t}$ and $\sigma_{i,c,t}$ parameters as opposed to estimating baseline and change parameters. For this model, we assumed that person-level parameters were drawn from a single multivariate normal distribution (but separate for μ versus σ) across conditions and sessions (the "Joint Single" model in Supplementary Note 7 and Figure S9). Both models produced very similar results to the model specification presented throughout the main text.

5.3 Results

We discuss analysis of the Stroop data in-depth, followed by a summary of the results from the other tasks (see Supplementary Note 3 for an expanded discussion of each)

Results for the Stroop task in Study 1 of Hedge et al. (Hedge et al., 2017) are shown in Figure 5.5. Panel A compares the estimated test-retest correlation for the two-stage approach versus each of the normal, lognormal, and shifted lognormal generative models. For the two-stage mean and standard deviation contrasts, test-retest correlations were $r = .50$ (95% CI = [.25, .69]) and $r = .07$ (95% CI = [-.22, .35]), respectively. Note that previous studies did not include standard deviation contrasts, but we include them here for comparison to the generative models. These estimates are consistent with results obtained originally by Hedge et al. (Hedge et al., 2017). It is clear that the Stroop effect has less than ideal reliability when estimated using the two-stage approach: with a test-retest reliability of $r = .50$ to $r = .60$, we would need well over 200 participants to detect (with adequate power) a simple correlation between the Stroop effect and an alternative individual difference measure with similar reliability (Hedge et al., 2017). Given constraints on the number of participants available for any one design, low reliability limits the utility of the Stroop effect as a measure to advance theories that make predictions on the basis of individual differences.

Figure 5.5A shows *posterior probability distributions* of the model parameters obtained using Bayesian updating as opposed to the point estimates and confidence intervals obtained using the two-stage summary approach. Note that the posterior

distribution can be interpreted in a variety of ways depending on our goals. For example, if one is interested in the probability that the test-retest correlation of the normal generative model is greater than the two-stage estimate of $r = .50$, this quantity can be easily computed as the proportion of the posterior distribution greater than $r = .50$. Alternatively, if we are interested in the single most likely test-retest estimate, we can simply locate the mode (or peak) of the posterior distribution. However, we are typically interested not only in a single value, such as the mode, but a range of likely values that help us convey uncertainty. Therefore, to facilitate interpretability of posterior distributions, we report the posterior mean (sometimes referred to as the posterior "expectation") along with the 95% *highest density interval* (HDI). An HDI is a generalization of the concept of the mode, but it is an interval rather than a single value. For example, a 20% HDI would contain 20% of the area of the entire posterior distribution, where every value within the interval is more likely than every value outside of the interval. We report 95% HDIs to maintain consistency with the 95% CIs reported for the two-stage approach, although we caution readers that HDIs and CIs are different concepts that have different interpretations. As has been a focus throughout this article, a mean and CI alone may do a poor job of summarizing a skewed distribution, so we recommend that readers interpret the posterior distributions holistically to fully appreciate the generative model estimates.

For the generative models, the posterior distributions for the mean/difficulty contrast parameters ($\mu_{i,\Delta}$) across models were concentrated above the two-stage estimates (posterior mean test-retest ranging from $r = .76$ to $r = .81$). Furthermore, the

95% HDIs for the difficulty parameter in the normal (95% HDI = [.46, 1.00]), lognormal (95% HDI = [.47, 1.00]), and shifted-lognormal (95% HDI = [.53, 1.00]) models included $r = 1.00$, indicating that we cannot rule out the possibility that there is in fact a perfect correlation in the mean/difficulty parameter contrast between retest sessions. This can be observed in the posterior distributions, which are concentrated against the upper limit of the range at $r = 1.00$. Posterior distributions for the standard deviation/dispersion parameters ($\sigma_{i,\Delta}$) were also concentrated above the two-stage estimates, although primarily for the lognormal and shifted lognormal models (posterior mean test-retest ranging from $r = .23$ to $r = .62$). In fact, the test-retest estimate for the standard deviation/dispersion parameters were much higher for the lognormal (95% HDI = [.26, .89]) and shifted-lognormal (95% HDI = [.25, .96]) models relative to the normal model (95% HDI = [-.05, .50]), which demonstrates the importance of our data-generating (distributional) assumptions when making inference on individual differences; it is clear that the assumption of within-person variation in response times is a crucial factor for obtaining more reliable estimates of the mean/central tendency of response times (refer back to Figure 5.3). Furthermore, this variation itself can be a stable psychological property depending on its assumed functional form (i.e. normal, lognormal, shifted lognormal). Our theories of inhibitory control should therefore account for not only the mean, but also the variability in response times across trials. More generally, our behavioral models should account for the full distribution of observed data. Generative modeling is a principled method of closing these theory-description gaps.

We can also compare the person-level parameters across models to determine if the models produce different mechanistic inferences. For example, we may be interested in the proportion of participants who show a "Stroop effect" for each model. For demonstration, here we define an effect as when the 95% HDI of the person-level posterior distribution on the contrast parameter of interest is entirely above 0. We can then identify the proportion of participants meeting this criterion for each of the $\mu_{i,\Delta}$ and $\sigma_{i,\Delta}$ parameters. For Session 1, across all generative models, all 47 participants showed evidence for an increase in $\mu_{i,\Delta}$ in the incongruent condition. However, for $\sigma_{i,\Delta}$, 36, 27, and 20 participants showed evidence for an increase in the incongruent condition according to the normal, lognormal, and shifted-lognormal models, respectively. Results were similar for Session 2: across all models $\mu_{i,\Delta}$ showed evidence for an increase across conditions for all participants. Similarly, for $\sigma_{i,\Delta}$, 27, 24, and 18 participants showed an increase in dispersion across conditions for the normal, lognormal, and shifted lognormal models, respectively. This pattern of results suggests that changes in response times across conditions within participants may be attributable primarily to changes in $\mu_{i,\Delta}$ (difficulty) rather than $\sigma_{i,\Delta}$ (dispersion)—an inference facilitated by the lognormal models that provides clear motivation for extending the model (e.g., developing a more explicit mechanism of how $\mu_{i,\Delta}$ varies as a function of stimulus properties). Note that the addition of a shift parameter led to fewer participants showing evidence for a change in dispersion ($\sigma_{i,\Delta}$) across conditions, which emphasizes the importance of explicitly accounting for nuisance factors (i.e. non-decision time) when our goal is to interpret model parameters in psychologically meaningful ways. Without the shift parameter, we

would have incorrectly concluded that the majority of participants showed an increase in dispersion within both Sessions 1 and 2, which has implications for our theories of stimulus interference and inhibitory control.

Figure 5.5B shows fitted model predictions compared to the observed response times for a random, representative participant. The two-stage approach is represented simply as the mean response time within each of the congruent and incongruent conditions, whereas the generative model predictions are represented by the light red curves, which are response time distributions simulated from this participant's estimated person-level normal, lognormal, and shifted-lognormal model parameters, where variation between lines indicates uncertainty in underlying parameters. With these simulated response times, we can compare how well each model reproduces the observed response times. For this particular participant, the normal generative model reveals many shortcomings, the most obvious being failure to capture right-skew, and over-prediction of rapid response times. By contrast, the lognormal model in the middle panel provides a much better reproduction of the observed data, capturing both right-skew and the concentration of response times around the mean. Improvement offered by the shifted-lognormal model is more subtle in this example—it better captures the onset of the response time distribution (i.e. the most rapid response times) relative to the lognormal model due to the small shift, but otherwise performs similarly (Supplementary Note 6, Figures S3-S6, contains examples of where the shift makes a more noticeable difference). Note that improvement in model fit is accompanied by an increase in expected test-retest reliability for the lognormal models over the normal

model, particularly for dispersion parameters. Indeed, as we described above, the normal model suggests that a large majority of participants showed increased dispersion ($\sigma_{i,\Delta}$) across conditions in Session 1, and then much fewer in Session 2. Given that the normal model poorly characterizes observed data, we may attribute this inconsistency (which was not observed to such an extent for the lognormal models) to a theory-description gap rather than to an actual change in how the experimental manipulation affected underlying cognitive processes between sessions (refer back to Figure 5.3). Overall, scrutinizing person-level model predictions versus observed data contextualizes group-level results and provides insight into the properties of cognitive mechanisms involved in inhibitory control.

Figure 5.6 visualizes test-retest correlations for the remaining tasks. We include detailed results and figures (akin to Figure 5.5) for each of these tasks in Supplementary Note 6, Figures S3-S6, along with a table with the descriptive statistics pertaining to Figure 5.6 (Supplementary Table 1). However, we emphasize that each of these tasks can be explored in the same (or greater) depth as we explored Stroop Study 1 above—in all cases, generative models readily unveil richness underlying observed data that is otherwise ignored by the coarse nature of the two-stage summary approach.

There are three main take-aways from the results presented in Figure 5.6. First, generative models consistently infer higher test-retest correlations relative to the two-stage approach, and in some cases these changes are quite substantial. For example, in Study 2 of the Flanker task, the test-retest correlation of the sample mean obtained from the two-stage summary approach was quite low, r = -.13, whereas the normal

generative model inferred $r = .64$. For the IAT Race version, the two-stage sample mean contrast test-retest correlation was $r = .45$, whereas the normal generative model inferred $r = .83$. Such large differences have considerable implications for testing and developing theories of individual differences within each paradigm. Indeed, low test-retest correlations at the person level in the face of high group-level stability is the central paradox behind a recent influential theoretical advance within social psychology known as the "bias of crowds" (Payne, Vuletich, & Lundberg, 2017; Rivers, Rees, Calanchini, & Sherman, 2017), which leads to the argument that IAT scores could be caused by contexts but do not exist within individuals' minds, absent eliciting contexts (Jost, 2019). However, others have argued that measurement error in the IAT D-score is a more parsimonious solution to the apparent puzzle (Connor & Evers, 2020), which our generative modeling results partially corroborate.

Second, generative model estimates are highly consistent across replications of the same task, whereas the two-stage approach estimates often vary considerably (e.g., compare the two-stage and generative model estimates for Flanker Study 1 versus Study 2). For example, for the Stroop task, the two-stage standard deviation contrast is significant in Study 2 but not in Study 1. Similarly, for the Flanker task, the two-stage mean contrast is significant in Study 1 but not in Study 2. In contrast, the more theoretically informed generative model (i.e. the lognormal models) parameters replicated consistently across studies.

Third, there is variation among generative models themselves, indicating that test-retest reliability varies—sometimes substantially (e.g., compare the normal versus

lognormal models for the Stroop task and IAT Race version)—depending on the assumed behavioral model. Variability across models suggests that we should make efforts not to overgeneralize the failings (or successes) of a single behavioral model to attributes of the behavioral task itself. In other words, we should be explicit in acknowledging that inferences are conditional on a data-generating model and not the task per se. It is important to remember that the traditional two-stage summary approach does not escape this limitation. All statistical models, from *t*-tests to factor analysis, carry assumptions that bias inference. With this in mind, our results run counter to mounting claims that behavioral tasks are poorly suited for developing theories of individual differences (Dang et al., 2020; Enkavi et al., 2019; Gawronski et al., 2017; Hedge et al., 2017; Wennerhold & Friese, 2020); the central issue is not that behavior in these tasks is unreliable, but rather that outdated and overly restrictive statistical models fail to characterize the reliable characteristics of the data-generating process underlying behavior.

5.3.1 Comparing Summary Statistics to Generative Model Parameters

Why are test-retest correlations often higher in generative than two-stage models? Figure 5.7 illustrates the main reason*: hierarchical pooling*, which refers to regression of person-level parameters toward group-level means (or the regression line in this multivariate case). We chose examples to demonstrate how hierarchical pooling within generative models affects person-level parameter estimates. In Study 1 of the Stroop task, mean contrast estimates ($\mu_{i,\Delta}$) that would ordinarily be considered outliers in the two-stage approach are pooled toward the group-mean, producing higher expected test-

retest correlations. Generative model parameter estimates also reveal potential practice effects in which almost every participant's expected mean contrast (i.e., Stroop effect, $\mu_{i,\Delta}$) is lower at Session 2 relative to Session 1. In contrast, standard deviation contrast estimates ($\sigma_{i,\Delta}$) show weak pooling, as well as poor expected test-retest correlations. The same general pattern holds in the IAT (Black/White Race version), where $\mu_{i,\Delta}$ and $\sigma_{i,\Delta}$ exhibit strong and weak pooling, respectively. However, pooled $\mu_{i,\Delta}$ estimates for the IAT show regression toward the mean but without indication of potential practice effects. For the Delay Discounting task, both discounting rate (k_i) and choice sensitivity (c_i) parameters show moderate pooling (Figure 5.7). Together, these examples also show that hierarchical models do not automatically confer higher test-retest correlations. Instead, pooling only occurs to the extent it is warranted by data (see also the test-retest parameter recovery simulation results in the Supplementary Note 4).

5.4 Discussion

Generative modeling is a framework that allows researchers to use theoretical knowledge to inform their statistical models. Comparisons across five popular tasks demonstrate how incorporating assumptions about data-generating processes allows for a more precise characterization of individual differences in behavior better isolating sources of variability across experimental conditions. By attending to data-generating processes underlying behavior, generative modeling offers a solution not only to problems of low reliability (and by extension predictive validity), but also to problems with theory-description gaps arising from use and overinterpretation of statistical

models that fail to instantiate reasonable assumptions, or worse, carry misleading assumptions. In contrast, traditional two-stage summary approaches used to analyze behavioral data are largely atheoretical mechanical exercises, and researchers may be unaware of implicit data-generating assumptions they make when using such methods (refer back to Figure 5.4). Such implicit assumptions can lead to attenuated individual difference correlations and overall impoverished inferences with behavioral data. The explicitness of generative models ensures that assumptions and consequences of model design can be easily tracked, evaluated, and iteratively expanded on in future research.

It is important to emphasize that generative modeling is an iterative process, and throughout model development, model parameters can be assessed to determine their psychometric properties. Although we focused on test-retest reliability, there are many other properties worth exploring including parameter identifiability (Spektor & Kellen, 2018), parameter recovery (Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011; Haines, Vassileva, & Ahn, 2018; Miletić, Turner, Forstmann, & Van Maanen, 2017), tests of selective influence (Criss, 2010) (a form of construct validity where experimental manipulations cause expected changes in parameter values), as well as parameter convergence between behavioral models and models derived at other levels of analysis (Haines et al., 2020b; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017b) (e.g., with trait or neural models). Bayesian analysis facilitates joint estimation of all model parameters and their hypothesized relations, thus allowing for proper calibration of uncertainty in key parameters, such as test-retest correlations. We discuss extensions to the generative models we developed in the current study, including using more

sophisticated evidence accumulation models of choice and response time behavior, in the Supplementary Note 8, in addition to future directions that take advantage of adaptive experimental designs to further maximize informativeness of behavioral data in Supplementary Note 9 (Cavagnaro, Pitt, & Myung, 2011; Myung, Cavagnaro, & Pitt, 2013; Yang, Pitt, Ahn, & Myung, 2020).

Advances in computational statistics have only recently made generative modeling widely accessible. We anticipate that generative modeling approaches will proliferate as scientists from all backgrounds recognize their utility for rigorous theory development and testing. There are now many accessible resources and software packages available to help researchers gain a deeper understanding of generative modeling, so they can apply these techniques to their own work. Resources include introductions to the philosophy and utility of generative or computational modeling for theory development (Guest & Martin, 2020; van Rooij & Baggio, 2020), tutorials on building generative models from first principles (van Rooij & Blokpoel, 2020; Wilson & Collins, 2019), practical textbooks that combine introductions to both behavioral model development and hierarchical Bayesian modeling (Farrell & Lewandowsky, 2018; Lee & Wagenmakers, 2014), tutorials and case examples on developing joint generative models of behavior and brain activity (Palestro et al., 2018; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017b), and open source R and Python software packages that allow both beginners and advanced users to apply popular generative models of behavioral to their own data using hierarchical Bayesian modeling (Ahn, Haines, & Zhang, 2017; Mathys et al., 2014; Wiecki, Sofer, & Frank, 2013). These and related sources make it clear that

there is much more to building sound generative models than just capturing the shape of empirical data distributions. Computational models are a good example (Guest & Martin, 2020; Jarecki et al., 2020; Wilson & Collins, 2019), which facilitate a level of mechanistic inference not provided by the simple behavioral models used in the current study. We hope our message has come through clear—that the landscape for generative model development is vast, whereas traditional two-stage summary approaches are inherently limited in scope and application.

We end with a cautionary yet hopeful note: As history has revealed, heuristic use of summary statistics absent generative models impedes scientific progress, leading us down paths we would have never explored had we been made aware of the implicit assumptions that directed us there. Although our generative models may be wrong or mis-specified, their explicitness forces us to specify theoretical assumptions regarding how behavior arises, thereby requiring us to spend time thinking about the mechanisms that underly the brain, behavior, and their inter-relations. By embracing their incompleteness, we can strive to build generative models that are precise and thus meaningfully incorrect, rather than relying on vague, heuristic theories whose verisimilitude can be deceptive because of the theory-description gap. Identifying and knowing where our assumptions are wrong provides a natural path toward deepened understanding of the mechanisms underlying behavior.

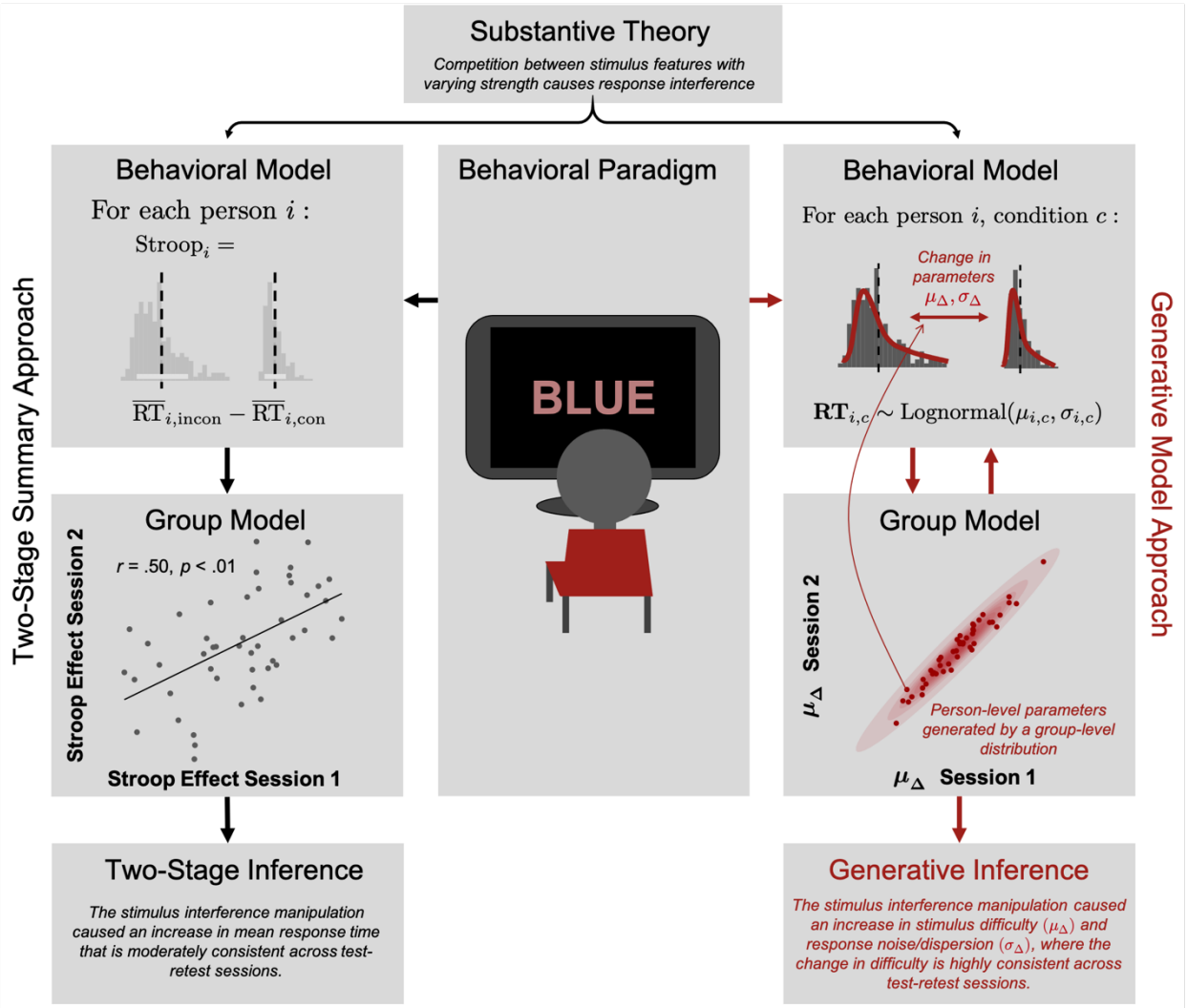

Figure 5.1. Pathway from theory to inference with behavioral data.

Behavioral tasks are designed to elicit behaviors that test the substantive theory. Behavioral models formally relate the theory to features of observed behavior within individual participants. Here we show the "behavioral model" often assumed when analyzing Stroop data in the "Two-Stage Approach" column, which makes implicit assumptions about how response times arise that we describe in the next section. The group-level model captures variation in person-level behavioral model parameters (e.g., mean response time). Here, we illustrate a group-level model estimating test-retest reliability. The traditional two-stage summary approach treats these models in a contiguous way, estimating person-level behavioral model parameters (e.g., mean response time contrasts) and then entering the resulting point estimates into a secondary model to assess group differences, individual difference correlations (i.e. test re-test), etc. Conversely, with generative modeling we construct a single model that integrates our assumptions and hypotheses about the data generating process, spanning distributions of trial-by-trial response times within persons to the distributions of

individual differences across people. This example depicts a lognormal behavioral model, which is comprised of two parameters that capture the shape of a person's response time distribution (described in detail in the next section). Person-level parameters then follow a group-level distribution that captures dependence in person-level parameters across test-retest sessions. This approach allows information to flow between the person- and group-level models when estimating parameters, which leads to more accurate estimation of person- and group-level parameters relative to the two-stage summary approach (e.g., mean response times, individual difference correlations). Better parameter estimation allows for us to test theories with more fidelity.

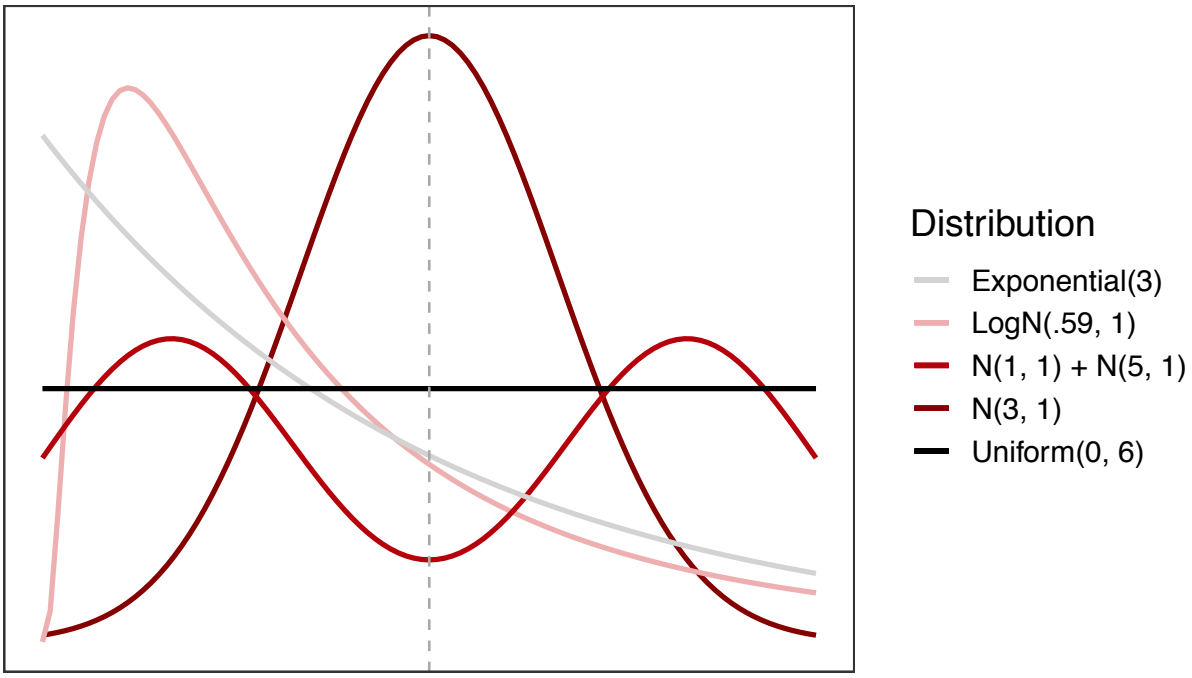

Figure 5.2. Qualitatively different distributions with the same mean.

All five distributions have the same mean (indicated by the dotted gray line) and would therefore produce the same conclusions if analyzed using only the sample mean summary statistic, regardless of how different their data-generating processes are. Statistical models that fail to fully account for the information contained in the full distribution of observed data fall victim to the theory-description gap.

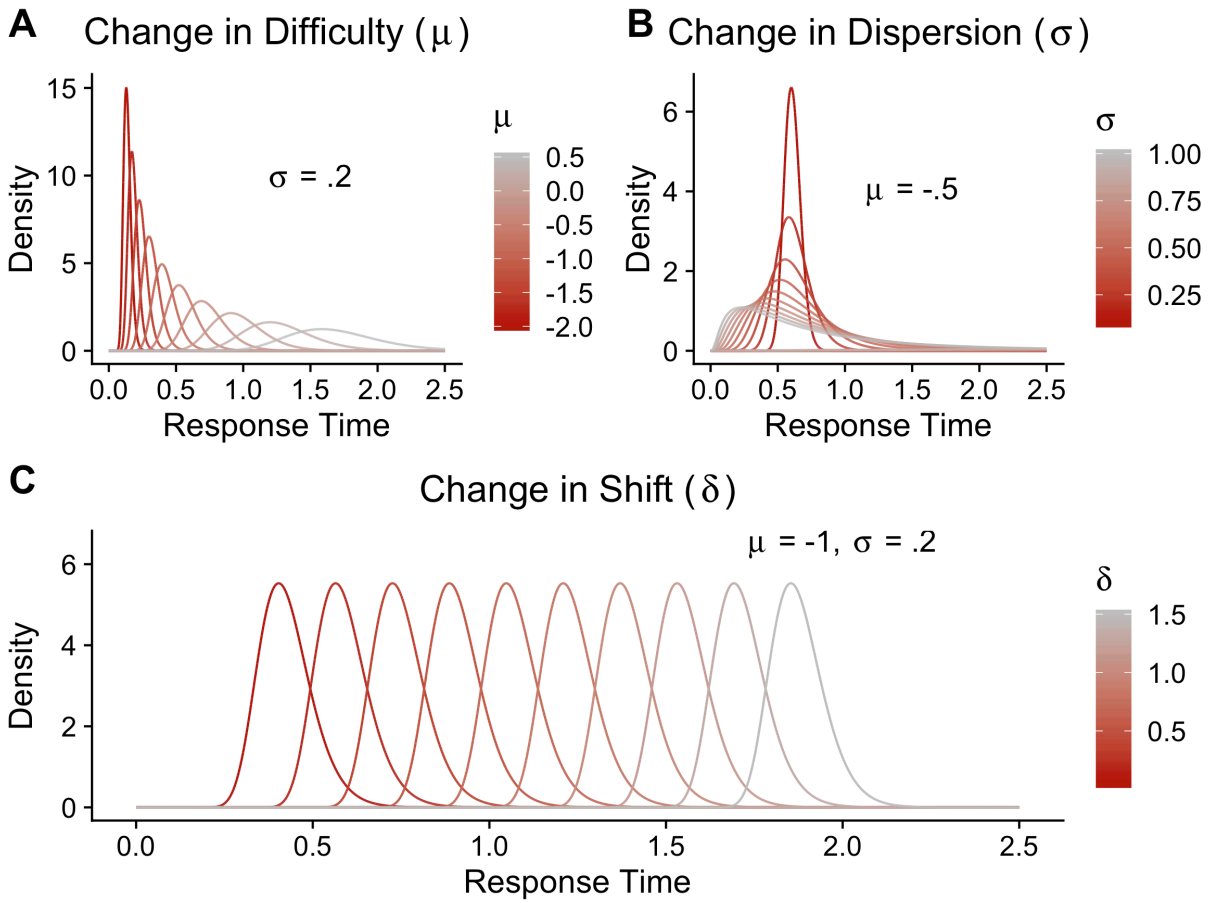

Figure 5.3. Lognormal and shifted lognormal generative distributions.

(A) For the lognormal and shifted lognormal distributions, changes in the μ parameter (interpreted as "stimulus difficulty") produce changes in both means and variances of response time distributions(Heathcote & Love, 2012; Rouder, Province, Morey, Gomez, & Heathcote, 2014). (B) The σ parameter controls dispersion (interpreted as "decision noise"); changes in σ affect means and ranges of likely response times, but medians remain constant. (C) For the shifted lognormal distribution, the shift parameter δ translates the lognormal distribution forward in time without changing the shape of the response time distribution. The lognormal distribution is a special case of the shifted lognormal distribution wherein the shift parameter (δ) is set to 0. The shift parameter is interpreted as "non-decision time", as it can capture individual differences in components of the data-generating process that are not relevant to decision-making but nevertheless produce variation in mean response times (e.g., visual encoding time, motor responses).

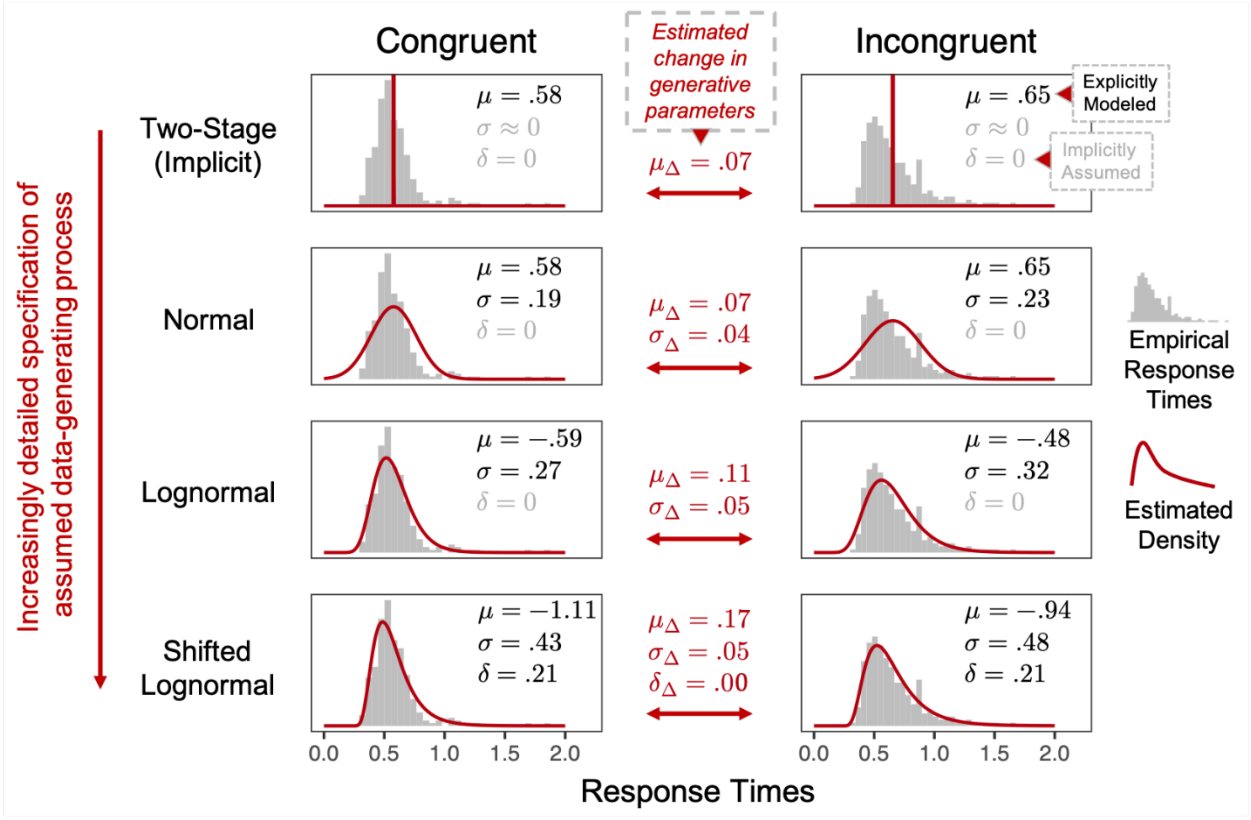

Figure 5.4. Building generative models consistent with theory.

The two-stage summary statistic approach is often used by default and is chosen without reference to an underlying theory. However, it is not a "safe" default—the implicit generative model it assumes is highly constrained and most likely inappropriate, assuming that either the within-person sample size (number of trials) is infinite/arbitrarily large, or that the within-person standard deviation of response times is approximately zero (the latter is depicted here; see the Methods section and Supplementary Note 2 for details). In the case of the Stroop task, these assumptions respectively correspond to us collecting an infinite number of responses in both conditions before taking the mean difference in response times, or alternatively, that every response time is exactly the same within conditions. Of course, we know a priori that neither of these assumptions are met—the generative approach begins by relaxing these overly restrictive assumptions. We begin with a simple normal model of response times, wherein we estimate both a mean and standard deviation within persons and conditions. However, the normal model predicts negative response times and does not capture the skew readily apparent in empirical response times distributions. The lognormal model improves upon these shortcomings, yet it cannot capture "non-decision" factors (e.g., stimulus encoding or motor response time) that could bias parameters of interest. The shifted lognormal model retains key properties of the lognormal model, but can also capture non-decision factors with the shift parameter (which is constrained to be equal across conditions within persons; see the Methods

section for an explanation). In constructing these models, we reduce the theory-description gap by creating models that are more consistent with our theoretical and background knowledge.

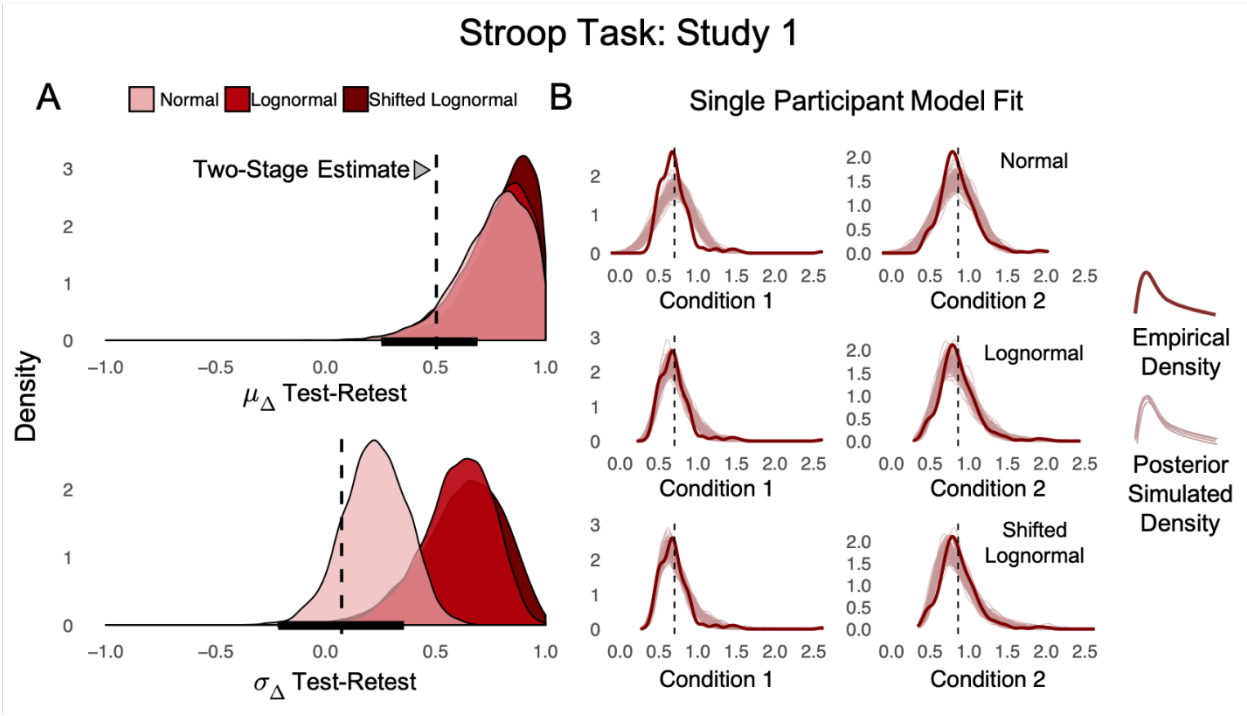

Figure 5.5. Test-retest correlations and model misfit for the Stroop task.

(A) Posterior distributions (red) for the test-retest correlations of each of the three generative models versus the two-stage sample mean/standard deviation approach (vertical dotted black line with corresponding horizontal 95% confidence interval) for the Stroop task in Study 1 of Hedge et al.(Hedge et al., 2017). (B) Posterior predictive simulations and sample means (vertical dotted black lines) for each of the generative models for a representative participant. The lognormal models can re-produce person-level response time distributions with much more fidelity than the normal model (and subsequently, the two-stage summary approach).

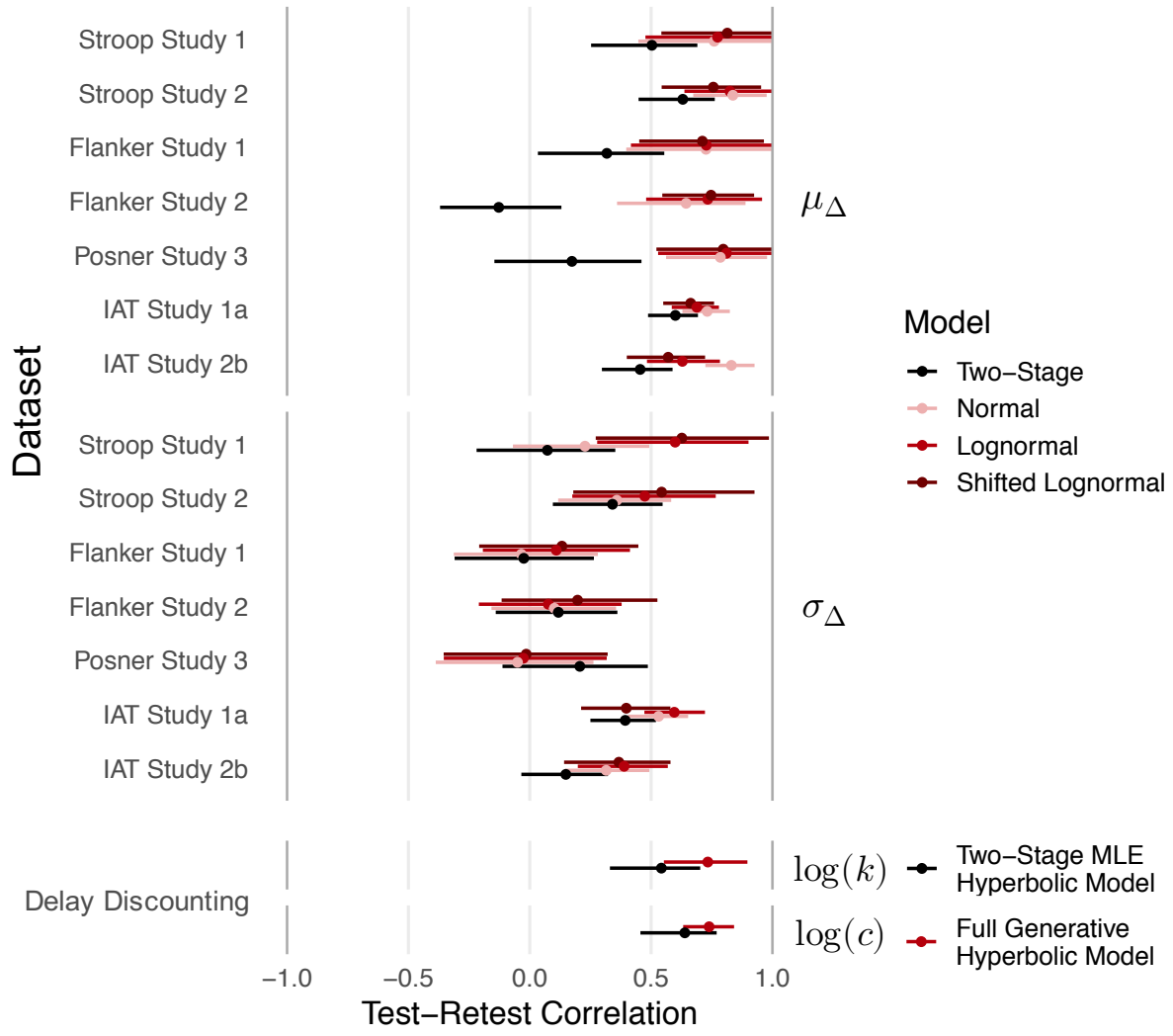

Figure 5.6. Test-retest correlations for all tasks and models.

Here we show means and 95% confidence intervals for the two-stage summary approach (for both sample mean and standard deviation contrasts) in black, along with the posterior means and 95% highest density intervals for the generative model parameter estimates (in various shades of red). The Implicit Association Test (IAT) datasets are from the Self-Concept (introversion/extraversion; Study 1a) and Race (Black/White; Study 2b) versions. The most striking increase in test-retest reliability is observed for μ, which is consistently higher than two-stage summary approach estimates across tasks (with an expected test-retest correlation of approximately .75 or greater for most tasks/models). In contrast, the test-retest reliabilities for σ are higher for some, but not all tasks (but note that estimation of σ is in part what is responsible for better estimation of μ; see Methods section and Figure 5.4). Test-retest windows for each task were as follows: (1) 3 weeks for the Stroop, Flanker, and Posner Cueing tasks, (2) 8 weeks for both versions of the IAT, and (3) 4 weeks for the Delay Discounting task. See

Supplementary Table 1 and the Supplementary Notes 5 and 6 for more detailed figures and full descriptions of each task.

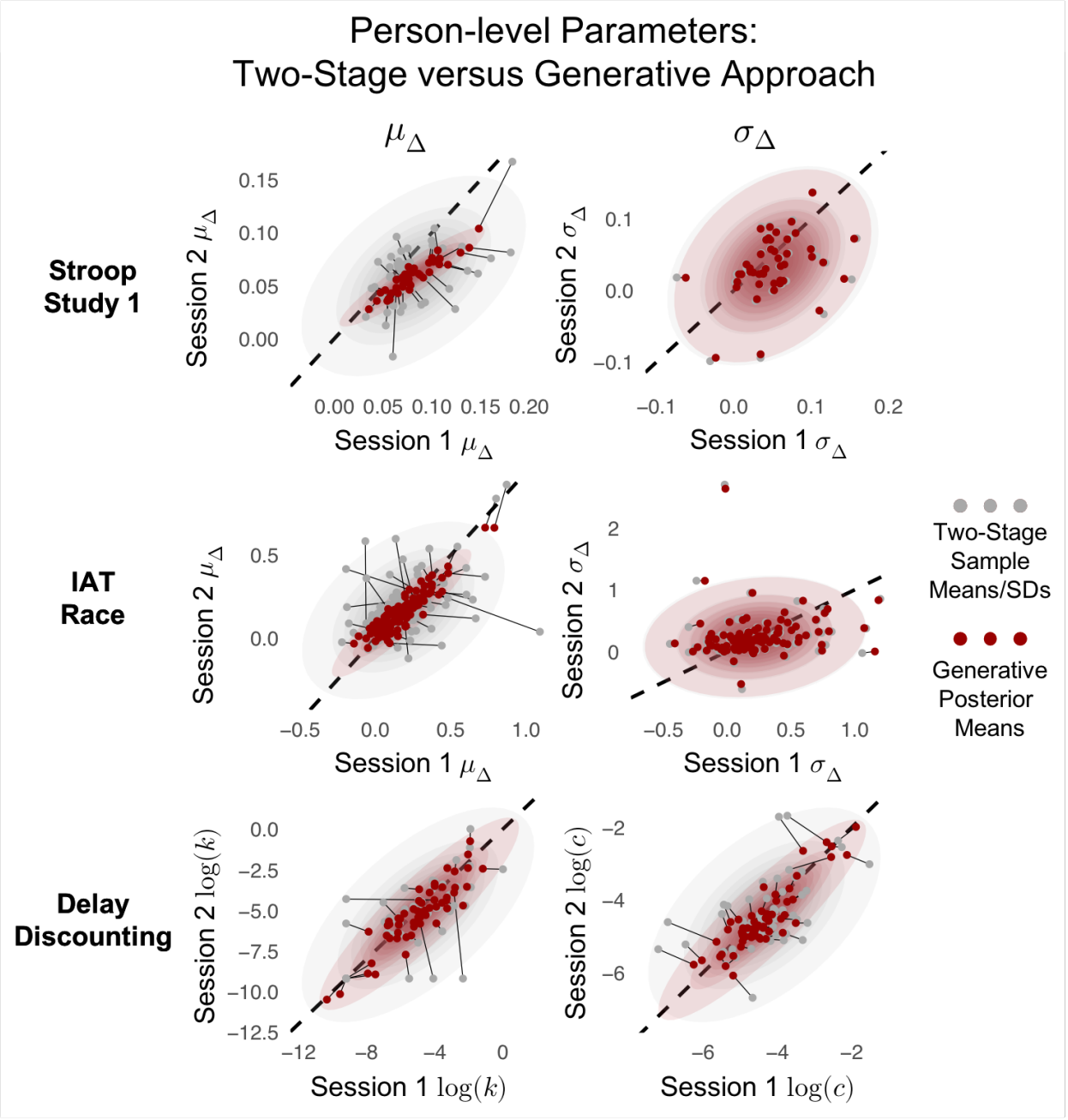

Figure 5.7. Relationship between two-stage estimates and generative model parameters.

For the response time models (Stroop and IAT tasks), two-stage estimates are the sample mean and standard deviation contrasts for each participant and retest session (i.e., estimates from the summary statistic approach). Generative model parameters are means of the person-level posterior distributions (i.e., posterior expectations) for each participant. Gray lines connect two-stage estimates and generative model parameters for each participant, demonstrating how hierarchical models induce regression to the group-level mean/regression line. To help visualize the low correlation for the Stroop study,

the standard deviation panel is zoomed in and two participants are not shown. For the Delay Discounting task, two-stage estimates reflect maximum likelihood estimates for each participant's discounting rate (k_i) and choice sensitivity (c_i) parameters; generative model parameters are means of person-level posterior distributions for each participant given by the full generative hyperbolic model. Gray and red ellipses indicate normal distribution quantiles (10%, 20%, ..., 90%, 99%) for the two-stage and generative model estimates, respectively.

Chapter 6: The Outcome-Representation Learning model: A novel reinforcement learning model of the Iowa Gambling Task

The previous chapter detailed why the traditional summary scores that researchers use to analyze behavioral data throughout the psychological and brain sciences are often inadequate for make theoretically useful inferences on individual differences. However, the models discussed in Chapter 5 were relatively simple and fail to capture some of the complex person × environment interactions that characterize externalizing psychopathology. For example, as discussed throughout Chapters 1-3, current theories of externalizing progression suggest that sensitivity to rewards versus punishments play a strong role in both the development and continuity of impulsive behavior across the lifespan. To capture such processes, our generative models need to capture dynamics in how people behave over time, in addition to capturing how people change their behavior in response to the consequences of their behaviors (i.e. in response to environmental feedback). I tackle this problem in Chapter 6, wherein I present a journal article that I published in Cognitive Science (in collaboration with Woo-Young Ahn and Jasmin Vassileva) on: (1) developing a reinforcement learning model of probabilistic learning from rewards versus punishments, and (2) use the model to investigate differences in

reward and punishment learning in adults with substance use disorders.

6.1 Introduction

There is a growing interest among researchers to develop and apply computational (i.e. cognitive) models to classical assessment tools to help guide clinical decision making (e.g., Ahn & Busemeyer, 2016; Batchelder, 1998; McFall & Townsend, 1998; Neufeld, Vollick, Carter, Boksman, & Jetté, 2002; Ratcliff, Spieler, & Mckoon, 2000; Treat, McFall, Viken, & Kruschke, 2001; Wallsten, Pleskac, & Lejuez, 2005). Despite this interest, clinical assessment has yet to be influenced by the many computational assays available today (see, Ahn & Busemeyer, 2016). There are many potential reasons for this, but two important factors are the lack of both: (1) precise characterizations of neurocognitive processes, and (2) optimal, externally valid paradigms for assessing psychiatric conditions.

The Iowa Gambling Task (IGT) is an example, which was successfully used to classify various clinical populations from healthy populations (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; Bechara et al., 2001). Originally developed to detect damage in ventromedial prefrontal brain regions, the IGT has since been used to identify a variety of decision making deficits across a wide range of clinical populations (e.g., Grant, Contoreggi, & London, 2000; Shurman, Horan, & Nuechterlein, 2005; Stout, Rodawalt, & Siemers, 2001; Whitlow et al., 2004). While the IGT is highly

sensitive to decision making deficits, the specific underlying neurocognitive processes that are responsible for these observed deficits are difficult to identify using only behavioral performance data.

To address the lack of specificity provided by the IGT, multiple computational models have been proposed which aim to break down the decision making process into its component parts (Ahn, Busemeyer, Wagenmakers, & Stout, 2008; Busemeyer & Stout, 2002; d'Acremont, Lu, Li, Van der Linden, & Bechara, 2009; Worthy, Pang, & Byrne, 2013b), and the modeling approach has been applied to several clinical populations (for a review, see Ahn, Dai, Vassileva, Busemeyer, & Stout, 2016). In particular, the first cognitive model proposed for the IGT—termed the Expectancy-Valence Learning (EVL) model (Busemeyer & Stout, 2002)—was used to identify differences in cognitive mechanisms between healthy controls and multiple clinical populations ranging from those with substance use to neuropsychiatric disorders (Yechiam, Busemeyer, Stout, & Bechara, 2005). The EVL led to several new competing models, which capture participants' decision making behavior more accurately. Specifically, two models show excellent performance: (1) the Prospect Valence Learning model with Delta rule (PVL-Delta) shows excellent long-term prediction accuracy and parameter recovery (Ahn et al., 2008; 2014; Steingroever, Wetzels, & Wagenmakers, 2013; 2014), and (2) the Value-Plus-Perseverance model (VPP) shows excellent short-term prediction accuracy (Ahn et al., 2014; Worthy et al., 2013b). Long-term prediction accuracy (a.k.a., absolute performance; Steingroever, Wetzels, & Wagenmakers, 2014) is defined as how well a model can generate the whole choice patterns when only the fitted

parameters are used, and short-term prediction accuracy is defined as a measure of model prediction accuracy on one-step-ahead trials using fitted parameters and a history of choices while penalizing model complexity. Parameter recovery performance indicates how well "true" model parameters can be estimated (i.e. recovered) after they are used to simulate behavior, which is essential for making valid inference with model parameters (Donkin, Brown, Heathcote, & Wagenmakers, 2011; Wagenmakers, van der Maas, & Grasman, 2007). Because all three of these metrics are important in understanding how well model parameters capture the true cognitive processes underlying decision making (see Heathcote, Brown, & Wagenmakers, 2015) and there is no single model that shows good performance in all three metrics, it is unclear which model should be used to make inference on the IGT.

Additionally, no studies to our knowledge have explicitly assessed different models' performance across the multiple versions of the IGT. While many studies to date have employed the original version of the task developed in 1994 (Bechara et al., 1994), the modified version has a non-stationary payoff structure (see section 6.2.2) and is widely used in practical applications involving populations with severe decision making impairments (e.g., Ahn et al., 2014; Bechara & Damasio, 2002). Importantly, a model that performs well across both versions of the task would be more generalizable to other experience-based cognitive tasks which are used extensively in the decision making and cognitive science literature.

To develop a new and improved computational model for the IGT, it is necessary to first identify the cognitive strategies that decision makers may engage in during IGT

administration. In the sections that follow, we describe four separable cognitive strategies/effects that are consistently observed in IGT behavioral data including: (1) maximizing long-term expected value, (2) maximizing win frequency, (3) choice perseveration, and (4) reversal learning. As mentioned previously, the IGT falls under the umbrella of more general experience-based cognitive tasks, so a model that accurately captures these multiple strategies has broad implications for models of decisions from experience.

6.1.1 Expected value

In experience-based cognitive tasks, people typically learn the long-term expected value of choice alternatives across trials and make choices appropriately. The IGT is a specific instantiation of an experienced-based task in which people make decisions based on expected value (e.g., Bechara, Damasio, Damasio, & Anderson, 1994; Beitz, Salthouse, & Davis, 2014). In fact, the most common metric used to summarize IGT behavioral performance is the difference between the number of "good" versus "bad" decks selected, where good and bad decks are those with positive and negative expected values, respectively. For example, in Bechara et al.'s (1994) original work, the net good minus bad deck selections was used to successfully differentiate healthy controls from individuals with ventromedial prefrontal cortex damage. However, it has since become clear that healthy subjects do not always learn to make optimal selections (see Steingroever, Wetzels, Horstmann, Neumann, & Wagenmakers, 2013b), which is consistent with extant literature on experience-based tasks (e.g., Erev & Barron, 2005). In extreme cases, healthy controls make decisions similar to that of severely impaired

decision makers when evaluated using expected value criterion alone (e.g., Caroselli, Hiscock, Scheibel, & Ingram, 2006).

The PVL-Delta and VPP models both assume that decision makers first value the outcomes according to the Prospect Theory utility function (Kahneman & Tversky, 1979), and the resulting subjective utilities are then used to update decision makers' trial-by-trial expectations using the delta rule (i.e. the simplified Rescorla-Wagner updating rule; see Rescorla & Wagner, 1972). Together, the Prospect Theory utility shape and loss aversion parameters determine which decks decision makers learn to prefer—holding other parameters constant, low loss aversion can lead to a preference for disadvantageous decks (i.e. decks A and B) because large losses become discounted, while a shape parameter closer to 0 (and below 1) makes decks with frequent gains more valuable than those with infrequent gains despite having the same objective expected value (see section 2.3; Ahn et al., 2008). Notably, reduced loss aversion on the IGT, but not a difference in utility shape, has been linked to decision making deficits in multiple clinical populations (Ahn et al., 2014; Vassileva et al., 2013), suggesting that differential valuation of gains versus losses is an individual difference with potential real-world implications. Therefore, a new IGT model should capture differential valuation of gains versus losses.

6.1.2 Win frequency

In experience-based paradigms like the IGT, it is well known that a majority of individuals have strong preferences for choices (i.e. decks) that win frequently, irrespective to long-term expected value (e.g., Barron & Erev, 2003; Chiu & Lin, 2007;

Chiu et al., 2008; Yechiam, Stout, Busemeyer, Rock, & Finn, 2005). For example, across studies using the IGT, deck B (win frequency=90%) is often more preferred than deck A (win frequency=50%) despite the long-term value of the two decks being equivalent (Lin, Chiu, Lee, & Hsieh, 2007; Steingroever et al., 2013b). In fact, this preference is so strong that most healthy subjects fail to make optimal decisions when the IGT task structure is altered so that good and bad decks have low and high win frequency, respectively (Chiu et al. 2008).

In principle, decision makers may prefer deck B over more advantageous options because they do not accurately account for rare events (i.e. 1 large loss per 10 trials; see Figure 6.1). Barron & Erev (2003) describe this general tendency as an underweighting of rare events that may be attributable to multiple cognitive mechanisms including recency effects, estimation error, and/or reliance on cognitive heuristics (see Hertwig & Erev, 2009). However, it is clear from the IGT literature that recency effects alone cannot account for the observed preferences for decks with high win frequency. For example, Steingroever et al. (2013a) showed that the Expectancy Valence Learning model (EVL; Busemeyer & Stout, 2002)—despite capturing recency effects using the delta learning rule—cannot account for the win frequency effect in the IGT. Conversely, the concave downwards Prospect Theory utility function utilized by the PVL-Delta and VPP allows for both models to implicitly account for win frequency (see section 2.3; Ahn et al., 2008). Further, the structure of the IGT is such that the high win frequency decks (i.e. B and D) each have a single loss, so the loss aversion parameter in both the PVL-Delta and VPP models may directly underweight the rare, negative outcomes in

these decks. Therefore, the PVL-Delta and VPP implicitly capture win frequency effects and underweighting of rare events through the Prospect Theory utility function, but their parameters do not dissociate the effects of loss aversion or valuation (i.e. the utility shape) from that of win frequency. Relatedly, the individual posterior distributions of the utility shape parameter are sometimes not well estimated (e.g., confined around a boundary value), which is problematic from a modeling perspective. This is a potentially important oversight given the centrality of win frequency to healthy participants' IGT performance, which may differentiate healthy from clinical samples (see Steingroever, et al., 2013b). Moreover, a model that explicitly accounts for win frequency may offer insight into experience-based underweighting of rare events.

6.1.3 Perseveration

A series of studies shows that IGT choice preferences can be explained well by heuristic models of choice perseveration—the tendency to continue selecting an option regardless of the choice value. In particular, Worthy et al. (2013a) showed that win-stay/lose-switch choice strategies exhibit good short-term prediction accuracy relative to typical reinforcement learning models, indicating that many decision makers may engage in simple stay/switch strategies that obfuscate inferences made on their learning processes. Furthermore, decay learning rules (Erev & Roth, 1998) provide better short-term prediction accuracy than typical updating rules (i.e. the delta rule), which may be because they can mimic choice perseveration heuristics by increasing the probability that recently selected decks are chosen again (Ahn et al., 2008). Finally, despite the IGT being designed to capture the exploration-exploitation trade-off (Bechara et al.,

1994), recent studies show that healthy participants fail to show evidence of progressing from a state of exploration to exploitation across trials (Steingroever et al., 2013b). Instead, participants' individual tendencies to perseverate on or frequently switch choices remain relatively stable over time. Therefore, a new IGT model should capture decision makers' tendencies to stay versus switch decks. Otherwise, other model parameters of theoretical interest (e.g., learning rates, loss aversion, etc.) may become conflated with perseverative tendencies.

6.1.4 Reversal learning

Due to the structure of both the original (Bechara et al., 1994) and modified (Bechara et al., 2001) versions of the IGT (see section 2.2 for the details of the task structure), reversal learning plays a critical role in some people's decision making process. For example, deck B appears optimal after its first 8 selections (+100 point rewards on each selection), but the expected value becomes negative after a large loss (-1,150 points) on the ninth selection. Because many decision makers begin the IGT with a pronounced preference for deck B, which rapidly declines over the first 20-30 trials (see, Steingroever et al., 2014), it is crucial that models can quickly reverse the preference for deck B after a large loss is encountered. In fact, participants who show performance deficits on the original version of the IGT become indistinguishable from healthy controls when the deck structure is altered to make the bad decks less appealing during the first few draws, and this increase in performance is strongly predictive of reversal learning abilities (Fellows & Farah, 2005).

Neither the PVL-Delta nor the VPP models were developed to account for reversal learning. However, the perseverance heuristic in the VPP can potentially mimic short-term effects of reversal learning by increasing the probability of selecting the same choice after a gain while increasing the probability of switching choices after a loss (see section 2.3; Worthy et al., 2013b). Both reversal learning and counter-factual (i.e. fictive) updating models can exhibit this behavior by updating the unchosen option utilities in reference to the chosen option outcome (e.g., Gläscher, Hampton, & O'Doherty, 2009; Lohrenz, McCabe, Camerer, & Montague, 2007). Unlike the VPP's perseverance heuristic, counter-factual updating can speed the learning process itself, which can lead to more rapid, long-term preference reversals. Importantly, reversal learning/counter-factual reasoning is a well-replicated behavioral phenomenon (see Roese & Summerville, 2005) and has strong support in the model-based cognitive neuroscience literature in application to reinforcement learning tasks (i.e. experience-based tasks; Gläscher, Hampton, & O'Doherty, 2009; Hampton, Bossaerts, & O'Doherty, 2006).

6.1.5 The current study

In summary, current state-of-the-art computational models of the IGT do not (1) explicitly account for the various effects observed in behavioral data, or (2) provide a compromise between the multiple different model comparison metrics used for model selection (i.e. short- and long-term prediction accuracy and parameter recovery). Here, we present the Outcome-Representation Learning model (ORL), a novel reinforcement learning model which explicitly accounts for the effects of expected value, gain-loss

frequency, choice perseveration, and reversal-learning with only 5 free parameters. By fitting 393 subjects' IGT choice data, we show that the ORL model provides good short- and long-term prediction accuracy and parameter recovery in comparison to the PVL-Delta and VPP models. Furthermore, the ORL performs consistently well for both the original and modified version of the IGT and on data collected across multiple different research sites. Finally, we apply the ORL to IGT data collected from amphetamine, heroin, and cannabis users (Ahn et al., 2014; Fridberg et al., 2010), and we show that the ORL identifies theoretically meaningful differences in decision making between substance using groups which are supported by prior studies.

6.2. Methods

6.2.1 Participants

We used IGT data collected from multiple studies to validate the ORL model including: (1) an openly-accessible, "many labs" collaboration dataset containing IGT data from 247 healthy participants across 8 independent studies (Steingroever et al., 2015)[4]; (2) data from Ahn et al. (2014) where 48 healthy controls, and 43 pure heroin and 38 pure amphetamine users in protracted abstinence completed a modified version of the IGT; and (3) data from Fridberg et al. (2010), where 17 chronic cannabis users

[4] We only included data from Steingroever et al. (2015) where participants underwent either the original or modified versions of the IGT as described in Fig. 1. This criterion excluded any datasets where the order of cards in each deck was randomized or where participants were required to complete other tasks (i.e. introspective judgements) throughout IGT administration.

completed the original version of the IGT[5]. **Table 1** summarizes the multiple datasets used in the current study. In total, our study includes data from 393 participants. See the cited studies for specific details on the participants included in each dataset.

6.2.2 Tasks

In both versions of the IGT, decks A and B are considered "bad" decks because they have a negative expected value, and decks C and D are "good" decks because they have a positive expected value (Figures 6.1a and 6.1c). The order of cards within each deck (for both versions) is predetermined so that each subject will experience the same sequence of outcomes when drawing from a given deck (e.g., Figures 6.1b and 6.1d). The original version of the IGT maintains a stationary payoff distribution throughout the task (Bechara et al., 1994), whereas the payoff distribution of the modified version changes over trials (Bechara et al., 2001)—the net losses in good and bad decks become less and more extreme, respectively, after every 10 selections made from a given deck (c.f. Figure 6.1b to 6.1d).

6.2.3 Reinforcement learning models

6.2.3.1 Prospect Valence Learning model with delta rule (PVL-Delta)

The PVL-Delta model (Ahn et al., 2008) uses a prospect theory utility function (Kahneman & Tversky, 1979) to transform realized, objective monetary outcomes into subjective utilities:

[5] Healthy controls from Fridberg et al. (2010) are included in the many labs dataset from Steingroever et al. (2015).

$$u(t) = \begin{cases} x(t)^{\alpha}, & \text{if } x(t) \geq 0 \\ -\lambda|x(t)|^{\alpha}, & \text{otherwise} \end{cases} \quad (1)$$

Above, t denotes the trial number, $u(t)$ is the subjective utility of the experienced outcome, $x(t)$ is the experienced net outcome (i.e. the amount won minus amount lost on trial t), and α ($0 < \alpha < 2$) and λ ($0 < \lambda < 10$) are free parameters which govern the shape of the utility function and sensitivity to losses relative to gains, respectively. The α parameter in the Prospect Theory utility function can account for the win frequency effect (e.g., Chiu et al., 2008). For example, when $\alpha < 1$, the summed subjective utility of receiving \$1 five times is greater than receiving \$5 once (i.e. the utility curve is concave for positive outcomes and convex for negative ones), so decision makers with an α below 1 would be expected to prefer decks with high win frequency over objectively equivalent decks which win less often (Ahn et al., 2008). Likewise, if $\lambda > 1$, the subjective experience of a given loss is greater in magnitude than an equivalent gain, which captures the idea that "losses loom larger than equivalent gains" (Kahneman & Tversky, 1979) when being subjectively evaluated. Note that when making decisions from experience—as in the IGT—the modal participant does not typically show loss aversion (Erev, Ert & Yechiam, 008); instead, participants tend to underweight rare events (e.g., Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004). Previous modeling analyses with the IGT have exhibited a similar pattern, where group-level loss aversion parameters are mostly below 1 (e.g., Ahn et al., 2014).

The PVL-Delta model assumes that decision makers update their expected values for each deck using a simplified variant of the Rescorla-Wagner rule (i.e. the delta rule; Rescorla & Wagner, 1972):

$$E_j(t+1) = E_j(t) + A \cdot (u(t) - E_j(t)) \tag{2}$$

Here, $E_j(t)$ is the expected value of chosen deck j on trial t, and A $(0 < A < 1)$ is a learning rate controlling how quickly decision makers integrate recent outcomes into their expected value for a given deck. Expected values are entered into a softmax function to generate choice probabilities:

$$Pr[D(t+1) = j] = \frac{e^{\theta \cdot E_j(t+1)}}{\sum_{k=1}^{4} e^{\theta \cdot E_k(t+1)}} \tag{3}$$

where $D(t)$ is the chosen deck on trail t, and θ is determined by:

$$\theta = 3^c - 1 \tag{4}$$

Here, c $(0 < c < 5)$ is a free parameter which represents trial-independent choice consistency (Yechiam & Ert, 2007). If c is close to 0 or 5, it indicates that decision makers are responding randomly or (near)deterministically, respectively, with respect to their expected values for each deck. Altogether, the PVL-Delta model contains 4 free parameters (A, α, c, λ).

6.2.3.2 Value-Plus-Perseverance model (VPP)

The VPP model expands upon the PVL-Delta model by adding an additional term for choice perseverance (Worthy et al., 2013b):

$$P_j(t+1) = \begin{cases} K \cdot P_j(t) + \epsilon_P, & \text{if } x(t) \geq 0 \\ K \cdot P_j(t) + \epsilon_N, & \text{otherwise} \end{cases} \tag{5}$$

$P_j(t)$ indicates the perseveration value for chosen deck j on trial t, which decays by K $(0 < K < 1)$ on each trial. When chosen, the perseveration value for deck j is updated by ϵ_P $(-\infty < \epsilon_P < \infty)$ or ϵ_N $(-\infty < \epsilon_N < \infty)$ based on the sign of outcome. Positive values for ϵ_P and ϵ_N indicate tendencies for decision makers to "perseverate" the deck chosen on the previous trial, whereas negative values indicate a switching tendency.

The VPP assumes that the expected value (from the PVL-Delta model) and perseveration terms are integrated into a single value signal:

$$V_j(t+1) = \omega \cdot E_j(t+1) + (1-\omega) \cdot P_j(t+1) \tag{6}$$

where ω $(0 < \omega < 1)$ is a parameter that controls the weight given to the expected value and perseveration signals. As ω approaches 0 or 1, the VPP reduces to the perseveration model or the PVL-Delta model alone, respectively. The VPP uses the same softmax function as the PVL-Delta to generate choice probabilities, except that $E_j(t+1)$ is replaced with $V_j(t+1)$. Altogether, the VPP contains 8 free parameters (A, α, c, λ, ϵ_P, ϵ_N, K, ω).

6.2.3.3 Outcome-Representation Learning model (ORL)

Here, we propose the ORL as a novel learning model for the IGT. Unlike the PVL-Delta and VPP models, the ORL assumes that the expected value and win frequency for each deck are tracked separately as opposed to implicitly within the Prospect Theory utility function (Pang, Blanco, Maddox, & Worthy, 2016).[6] Note that separate tracking

[6] Pang, B., Byrne, K., A., Worthy, D., A. (unpublished). When More is Less: Working Memory Load Reduces Reliance on a Frequency Heuristic During Decision-Making.

of expected value and win frequency makes the ORL similar to the class of risk-sensitive reinforcement learning models which forgo maximizing expected value to minimize potential risks (e.g., Mihatsch & Neuneier, 2002). The expected value of a deck is updated with separate learning rates for positive and negative outcomes:

$$EV_j(t+1) = \begin{cases} EV_j(t) + A_{rew} \cdot (x(t) - EV_j(t)), & \text{if } x(t) \geq 0 \\ EV_j(t) + A_{pun} \cdot (x(t) - EV_j(t)), & \text{otherwise} \end{cases} \quad (7)$$

where $EV_j(t)$ denotes the expected value of chosen deck j on trial t, and A_{rew} ($0 < A_{rew} < 1$) and A_{pun} ($0 < A_{pun} < 1$) are learning rates which are used to update expectations after reward (i.e. positive) and punishment (i.e. negative) outcomes, respectively. Unlike the PVL-Delta and VPP models, the ORL is updating expected values using the objective outcome $x(t)$, not the subjective utility $u(t)$.

The use of separate learning rates for positive versus negative outcomes allows for the ORL model to account for over- and under-sensitivity to losses and gains, similar to the loss aversion parameter shared by the PVL-Delta and VPP. Specifically, the larger the difference is between the positive and negative learning rates, the more learning is dominated by either positive or negative outcomes. We used separate learning rates, as opposed to a loss-aversion parameterization, because there is strong neurobiological and behavioral evidence for learning models with separate learning rates for positive versus negative outcomes (e.g., Doll, Jacobs, Sanfey, & Frank, 2009; Gershman, 2015). For example, Parkinson's patients learn more quickly from negative compared to positive outcomes, and dopamine medication reverses this bias (Frank, Seeberger, & O'Reilly, 2004). Additionally, positive and negative learning rates are modulated by genes that

are partially responsible for striatal dopamine functioning (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007), and more recent evidence implicates striatal D1 and D2 receptor stimulation in learning from positive and negative outcomes, respectively (Cox et al., 2015).

To account for the win frequency effect, the ORL separately tracks win frequency as follows:

$$EF_j(t+1) = \begin{cases} EF_j(t) + A_{rew} \cdot (sgn(x(t)) - EF_j(t)), & \text{if } x(t) \geq 0 \\ EF_j(t) + A_{pun} \cdot (sgn(x(t)) - EF_j(t)), & \text{otherwise} \end{cases} \quad (8)$$

where $EF_j(t)$ denotes the "expected outcome frequency", A_{rew} $(0 < A_{rew} < 1)$ and A_{pun} $(0 < A_{pun} < 1)$ are learning rates shared with the expected value learning rule, and $sgn(x(t))$ returns 1, 0, or -1 for positive, 0, or negative outcome values on trial t, respectively. The ORL model also includes a reversal-learning component for $EF_j(t)$. $EF_{j'}(t)$ refers to the expected outcome frequency of all unchosen decks j' on trial t:

$$EF_{j'}(t+1) = \begin{cases} EF_{j'}(t) + A_{pun} \cdot (\frac{-sgn(x(t))}{C} - EF_{j'}(t)), & \text{if } x(t) \geq 0 \\ EF_{j'}(t) + A_{rew} \cdot (\frac{-sgn(x(t))}{C} - EF_{j'}(t)), & \text{otherwise} \end{cases} \quad (9)$$

Here, the learning rates are shared from the expected value learning rule, and C is the number of possible alternative choices to chosen deck j. Note that when updating unchosen decks j', the reward learning rate is used if the chosen outcome was negative and the punishment learning rate is used if the chosen outcome was positive. Because there are 4 possible choices in both versions of the IGT, there are always 3 possible alternative choices. Therefore, C is set to 3 in the current study. Note that if there were only a single alternative choice (e.g. simple two-choice tasks), C would be set to 1 and

the frequency heuristic would reduce to a "double-updating" rule often used to model choice behavior in probabilistic reversal learning tasks (e.g., Gläscher, Hampton, & O'Doherty, 2009).[7]

The ORL model also employs a simple choice perseverance model to capture decision makers' tendencies to stay or switch decks, irrespective to the outcome:

$$PS_j(t+1) = \begin{cases} \frac{1}{1+K}, & \text{if } D(t) = j \\ \frac{PS_j(t)}{1+K}, & \text{otherwise} \end{cases} \tag{10}$$

where K is determined by:

$$K = 3^{K'} - 1 \tag{11}$$

Here, $PS_j(t)$ is the perseverance weight of deck j on trial t, and K is a decay parameter controlling how quickly decision makers forget their past deck choices. K' is estimated $\in$ [0,5], therefore $K \in$ [0,242] (see equation 11). The above model implies that the perseverance weight of the chosen deck is set to 1 on each trial, and subsequently all perseverance weights decay exponentially before a choice is made on the next trial. We used this parameterization because it showed the best performance for estimating K compared to other parameterizations (e.g., $PS_j(t+1) = PS_j(t) \times K$). Low or high values for K suggest that decision makers remember long or short histories of their own deck selections, respectively.

The ORL model assumes that value, frequency, and perseverance signals are integrated in a linear fashion to generate a single value signal for each deck:

[7] We tried various versions of the reversal learning process (e.g., reversal learning on $EV_j(t)$ or both $EV_j(t)$ and $EF_j(t)$) and versions of the model without the reversal learning component, but the version we report in this paper showed the best model fit.

$$V_j(t+1) = EV_j(t+1) + EF_j(t+1) \cdot \beta_F + PS_j(t+1) \cdot \beta_P \quad (12)$$

Here, β_F ($-\infty < \beta_F < \infty$) and β_P ($-\infty < \beta_P < \infty$) are weights which reflect the effect of outcome frequency and perseverance on total value with respect to the expected value of each deck. Therefore, values for β_F less than or greater than 0 indicate that decision makers prefer decks with low or high win frequency, respectively. Additionally, values for β_P less than or greater than 0 indicate that decision makers prefer to switch or stay with recently chosen decks, respectively. Note that the expected value (EV) is a reference point which frequency and perseverance effects are evaluated against, so the ORL assumes that the "weight" of EV is equal to 1.

The ORL uses the same softmax function as the VPP to generate choice probabilities, except that the choice consistency/inverse temperature parameter (θ) is set to 1. We do not estimate choice consistency for the ORL due to parameter identifiability problems between θ, β_F, and β_P. Altogether, the ORL contains 5 free parameters (A_{rew}, A_{pun}, K, β_F, β_P).[8]

The ORL model will be added to *hBayesDM*, an easy-to-use R toolbox for computational modeling of a variety of different reinforcement learning and decision making models using hierarchical Bayesian analysis (Ahn, Haines, & Zhang, 2017). Additionally, all R codes used to preprocess, fit, simulate, and plot our results will be uploaded to our GitHub repository upon publication of this manuscript (https://github.com/CCS-Lab).

[8] Note that we tried various other models from the reinforcement learning literature, including: variants with the Pearce-Hall updating rule (Pearce & Hall, 1980), working memory models (Collins, Albrecht, Waltz, Gold & Frank, 2017), and risk aversion models (d'Acremont et al., 2009). However, none of these models provided an improved fit of the data and we do not report them for brevity.

6.2.4 Hierarchical Bayesian analysis

We used hierarchical Bayesian analysis (HBA) to estimate free parameters for each model (Kruschke, 2015; M. D. Lee, 2011; M. D. Lee & Wagenmakers, 2011; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008). HBA offers many benefits over more conventional approaches (i.e. maximum likelihood estimation) including: (1) modeling of individual differences with shrinkage (i.e. pooling) across subjects, (2) computation of posterior distributions as opposed to point estimates. Previous studies show that HBA leads to more accurate individual-level parameter recovery than the individual MLE approach (e.g., Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011).

HBA was conducted using Stan (version 2.15.1), a probabilistic programming language which uses Hamiltonian Monte Carlo (HMC), a variant of Markov Chain Monte Carlo (MCMC), to efficiently sample from high-dimensional probabilistic models as specified by the user (Carpenter, Gelman, Hoffman, & Lee, 2016). For each dataset used in the current study, we assumed that individual-level parameters were drawn from group-level distributions. Group-level distributions were assumed to be normally distributed, where the priors for locations (i.e. means) and scales (i.e. standard deviations) were assigned normal distributions. Additionally, we used non-centered parameterizations to minimize the dependence between group-level location and scale parameters (Betancourt & Girolami, 2013). Bounded parameters (e.g. learning rates $\in$ (0,1)) were estimated in an unconstrained space and then probit-transformed to the constrained space-and scaled if necessary-to maximize MCMC efficiency within the parameter space (Ahn et al., 2014; 2017; Wetzels et al., 2010). Using the reward

learning rate A_{rew} from the ORL model as an example, formal specification of the bounded parameters followed the form:

$$\begin{aligned} \mu_{A_{rew}} &\sim \text{Normal}(0, 1) \\ \sigma_{A_{rew}} &\sim \text{Normal}(0, 0.2) \\ \mathbf{A_{rew}}' &\sim \text{Normal}(0, 1) \\ \mathbf{A_{rew}} &= \text{Probit}(\mu_{A_{rew}} + \sigma_{A_{rew}} \cdot \mathbf{A_{rew}}') \end{aligned} \tag{13}$$

where μ_{Arew} and σ_{Arew} are the location and scale parameters for the group-level distribution, $\mathbf{A_{rew}}'$ is a vector of individual-level parameters on the unconstrained space, $\mathbf{A_{rew}}$ is a vector of individual-level parameters after they have been probit-transformed back to the constrained space, and Probit(x) is the inverse cumulative distribution function of the standard normal distribution. This parameterization ensures that after being probit-transformed, the hyper-prior distribution over the subject-level parameters is (near)uniform between the parameter bounds. For parameters bounded $\in (0, \mathit{upper})$ (e.g. K), we used the same parameterization as above but scaled to the upper bound accordingly:

$$\mathbf{K} = \text{Probit}(\mu_K + \sigma_K \cdot \mathbf{K}') \cdot 5 \tag{14}$$

For unbounded parameters (e.g., β_F), we used the same parameterization outline in equation 13 except we set the hyper-standard deviation to a half-Cacuhy$(0, 1)$. All models were sampled for 4,000 iterations, with the first 1,500 as warmup (i.e. burn-in), across 4 sampling chains for a total of 10,000 posterior samples for each parameter. Convergence to target distributions was checked visually by observing trace-plots and numerically by computing Gelman-Rubin—also known as $\widehat{R}$—statistics for each

parameter (Gelman & Rubin, 1992). $\widehat{R}$ values for all models were below 1.1, suggesting that the variance between chains did not outweigh variance within chains.

6.2.5 Model comparison: Leave-one-out information criterion

We used the leave-one-out information criterion (LOOIC) to compare one-step-ahead prediction accuracy across models. LOOIC is an approximation to full leave-one-out prediction accuracy that can be computed using the log pointwise posterior predictive density (lpd) of observed data (Vehtari, Gelman, & Gabry, 2017). Here, we computed the lpd by taking the log likelihood of each subject's actual choice on trial $t + 1$ conditional on their parameter estimates and choices from trials $\in \{1,2, \dots, t\}$. This procedure is iterated for all trials and for each posterior sample. Log likelihoods are then summed across trials within subjects. This summation results in an $N \times S$ lpd matrix, where N is the number of subjects and S is the number of posterior samples. We used the *loo* R package (Vehtari et al., 2017) to estimate the LOOIC from the lpd matrix. LOOIC is on the deviance scale, where lower values indicate better model fits.

6.2.6 Model comparison: Choice simulation

We used the simulation method to compare long-term prediction accuracy across models (Ahn et al., 2008; Steingroever et al., 2014). The simulation method involves two steps: (1) models are fit to each group's data, and (2) fitted model parameters from step 1 are used to simulate subjects' choice behavior given the task payoff structure. Simulated and true choice patterns are then compared to determine how well the model parameters capture subjects' choice behavior. In the current study, we employ a fully

Bayesian simulation method, which takes random draws from each subject's joint posterior distribution across fitted model parameters to simulate choice data (Steingroever et al., 2014; Steingroever, Wetzels, & Wagenmakers, 2013a). We iterated this procedure 1,000 times for each subject (i.e. 1,000 draws from individual-level, joint posteriors), and choice probabilities for each deck were stored for each iteration. We then averaged the choice probabilities for each deck across iterations and then subjects. Finally, we computed the mean squared deviation (MSD) between the experimental and simulated choice probabilities as follows:

$$\text{MSD} = \frac{1}{4 \cdot n} \sum_{t=1}^{n} \sum_{j=1}^{4} (\bar{D}_{exp_j}(t) - \bar{D}_{sim_j}(t))^2, \tag{15}$$

where n is the number of trials, t is the trial number, j is the deck number, $\overline{D}_{exp}(t)$ is the average across-subject probability of choosing deck j on trial t, and $\overline{D}_{sim}(t)$ is the average across-subject simulated probability (across 1,000 iterations as described above) of selecting deck j on trial t. Before computing MSD scores, we smoothed the experimental data (i.e. $\overline{D}_{exp}(t)$) with a moving average of window size 7 (Ahn et al., 2008). Additionally, this method is different from a posterior predictive check because it does not condition on observed response data (Gelman, Hwang, & Vehtari, 2013).

6.2.7 Model comparison: Parameter recovery

Parameter recovery is a method used to determine how well a model can estimate (i.e. recover) known parameter values, and it typically follows two steps: (1) choice data are simulated using a set of true parameters for a given model and task structure, and (2) the model is fit to the simulated choice data and the recovered parameter estimates

are compared to the true parameters (e.g., Ahn et al., 2011; Donkin, Brown, Heathcote, & Wagenmakers, 2011; Wagenmakers, van der Maas, & Grasman, 2007). We used the same set of parameters to simulate choices from the modified and original IGT task structure. We generated the parameter set by taking the means of the individual-level posterior distributions of each model fit to the 48 control subjects' data from Ahn et al. (2014) to ensure that the true parameter values were reasonably distributed and representative of human decision makers for each model.

We used two different parameter recovery methods. First, we compared the means of the posterior distributions for each individual-level parameter, and for each model, to the true parameters by plotting all the parameter values in a standardized space. We transformed parameters by z-scoring the recovered posterior means of each parameter by the mean and standard deviation of true parameters (i.e. the parameter set used to simulate choices) across individual-level parameters, which allowed us to determine how well the location of true parameters was recovered for each parameter and model. Second, we compared each of the true parameters to the entire posterior distribution of the respective recovered parameter by computing rank-ordered (i.e. *Spearman's*) correlations between the true and recovered parameter values across individual-level parameters. We iterated this procedure over each sample from the joint posterior distribution to estimate how well the rank-order between true parameters could be recovered for each parameter and model. The rank-order is particularly important for making inferences on relative parameter differences between subjects. Together, the

parameter recovery methods we used here allowed us to infer how well each model could recover parameters in an absolute and relative sense.

6.3. Results

6.3.1 Model comparison: Leave-one-out information criterion

Figure 6.2 shows the one-step-ahead leave-one-out information criterion (LOOIC) performance for each model and datasets used in the current study. As seen in the graphs, while the ORL and VPP outperform the PVL-Delta, they show similar performance to one another. Notably, the ORL outperformed the VPP in all three substance using groups, albeit by only a negligible amount in heroin users. Altogether, the LOOIC comparisons suggest that the ORL shows similar short-term prediction performance to the VPP (i.e. better than the PVL-Delta) across both versions of the IGT and across multiple populations with different decision making strategies despite the fact that the ORL has three fewer parameters than the VPP (5 vs. 8).

6.3.2 Model comparison: Choice simulation

The raw choice data and choice simulations for each dataset are depicted in Figure 6.3, and the mean squared deviations (MSDs) are shown in Table 6.2. Similarly to previous analyses (Ahn et al., 2014; Steingroever et al., 2013a), the PVL-Delta showed good simulation performance for both modified and original IGT versions in both healthy control and substance using groups. Unlike previous analyses (Ahn et al., 2014; but see Worthy et al., 2013b), the VPP showed similar performance to the PVL-Delta

across datasets.[9] Altogether, the simulation results are less clear on which of the models performs best for long-term prediction accuracy. In fact, the variation in performance between datasets is much greater than the variation in performance between models within each dataset (see Table 6.2).

6.3.3 Model comparison: Parameter recovery

Parameter recovery results for both versions of the IGT are shown in Figure 6.4. For the modified IGT, the PVL-Delta and ORL both show good parameter recovery across model parameters while the VPP performs poorly. For the VPP, the recovered posterior means were systematically higher than the true parameters for the learning rate (A), and systematically lower for the choice consistency (c) and reinforcement weight (ω). For the PVL-Delta and ORL, recovered posterior means were well-distributed around the true parameter means. Additionally, the full posterior recovery results for the VPP showed much more variable correlations between true parameters and the recovered posteriors compared to the PVL-Delta and ORL, suggesting that the PVL-Delta and ORL provide more precise posterior estimates and better capture the variance between individual-level parameter estimates (i.e. "subjects") compared to the VPP. For the original IGT, parameter recovery results were similar. While the VPP showed slightly better performance in the original IGT, still the posterior means for ω and c were systematically lower and posterior means for A were systematically higher than their true values. Together, the parameter recovery results suggest that both the

[9] Note that an error was discovered in simulation code used for the VPP in Ahn et al. (2014), which may partially account for the previous finding that the VPP exhibited poor simulation performance.

PVL-Delta and ORL provide more accurate and precise parameter estimates than the VPP for both versions of the IGT.

6.3.4 Applications to substance users

Because the ORL consistently performed as well or better than competing models across all groups in the current study, we used the ORL to examine group differences in model parameters. Note that we only compared substance using groups to the healthy control groups within the same studies to minimize any potential between-study effects. Figures 6.5 and 6.6 show the posterior estimates and differences in posterior estimates for each group, respectively. Below, we use the term "strong evidence" to refer to group differences where the 95% highest density interval (HDI) excludes 0 (Kruschke, 2015). We do not endorse binary interpretations of significant differences using this threshold, and we refer readers to the graphical comparisons (Figure 6.6) to judge parameters for meaningful differences. Within the dataset from Ahn et al. (2014), the heroin using group showed strong evidence of lower punishment learning rates than healthy controls (95% HDI = [0.003, 0.04]). A low punishment learning rate indicates less updating of expectations after experiencing a loss, a finding which is consistent with prior studies showing that heroin users have lower loss-aversion than controls (Ahn et al., 2014). We did not find strong evidence of differences between amphetamine and heroin users. However, there was some evidence (see Figure 6.6) that amphetamine users had more negative perseverance weights than heroin users (95% HDI = [-2.67, 0.79]). Within the dataset from Fridberg et al. (2010), chronic cannabis users showed strong evidence of greater reward learning rates (95% HDI = [-0.23, -0.05]) and some evidence of lower

punishment learning rates (95% HDI = [-0.001, 0.04]) compared to healthy controls, which is consistent with a previous analysis of this dataset using the PVL-Delta model showing that cannabis users were more sensitive to rewards and less sensitive to losses compared to healthy controls (Fridberg et al., 2010). Lastly, cannabis users showed strong evidence for more negative perseverance weights than healthy controls (95% HDI = [0.004, 4.09]), indicating a strong preference toward switching, as opposed to perseverating on, choices irrespective to the expected value of each deck.

6.4. Discussion

We present a novel cognitive model (the ORL) for the IGT which shows excellent short- and long-term prediction accuracy across both versions of the task and across an array of different clinical populations. The ORL explicitly models the four most consistent trends found in IGT behavioral data including long-term expected value, gain-loss frequency, perseverance, and reversal-learning. Overall, we showed that the ORL outperformed or showed comparable performance to competing models in all three model comparison indices including: post-hoc test (LOOIC), simulation performance, and parameter recovery. The results suggest that future research using the IGT should consider the ORL a top choice for cognitive modeling analyses.

Consistent with prior studies, our model comparison results suggest that any single measure used to compare models might not be sufficient (Ahn et al., 2008; 2014; Steingroever et al., 2014; Yechiam & Ert, 2007). For example, we found that the ORL consistently outperformed the VPP using parameter recovery metrics yet performed similarly to the VPP in short- and long-term prediction accuracy. Our results

underscore the importance of using many model comparison metrics in deciding between competing cognitive models (Heathcote, Brown, & Wagenmakers, 2015; Palminteri, Wyart, & Koechlin, 2017). Many studies use only information criteria such as LOOIC (e.g. Akaike or Bayesian information criteria) when choosing one among many cognitive models, and our results suggest that this may lead to imprecise inferences. Indeed, despite the VPP performing excellently when assessed using information criteria alone (i.e. LOOIC), the parameter recovery results indicate that multiple VPP model parameters might be imprecise at the subject-level and biased at the group-level (see Figure 6.4). For cognitive models to be useful in identifying individual differences (e.g., for clinical decision making), it is crucial that future studies conduct parameter recovery tests to ensure that parameter interpretations are valid.

When applied to IGT performance of pure substance users, the ORL revealed that heroin users in protracted abstinence were less sensitive to punishments (i.e. lower punishment learning rates) compared to healthy controls. The finding of lower punishment sensitivity in the heroin-using group is consistent with Ahn et al. (2014), where heroin users showed lower loss aversion (i.e. λ from the VPP) than healthy controls. We also found some evidence that amphetamine users engaged in more switching behavior than heroin users (see β_P in Figures 6.5 and 6.6). Although weak in comparison to other reported differences, this finding is consistent with a previous study showing that high levels of experience seeking traits are positively and negatively predictive of amphetamine and heroin users, respectively (Ahn & Vassileva, 2016). Notably, behavioral summaries of the amphetamine and heroin user's choice preferences

were indistinguishable (see Ahn et al., 2014). Additionally, the ORL revealed that chronic cannabis users were more sensitive to rewards (i.e. higher reward learning rates) and more likely to engage in exploratory behavior (i.e. more negative perseveration weight) than healthy controls. These findings converge with previous modeling results using the PVL-Delta (Fridberg et al., 2010) and with pharmacological studies showing that cannabis administration can increase sensitivity to rewards (and not punishments) which in turn may lead to more risk-taking behaviors (Lane, 2002; Lane, Cherek, Tcheremissine, Lieving, & Pietras, 2005). Importantly, our finding that chronic cannabis users tend to engage in exploratory behavior—irrespective to the value of each deck—suggests that the high levels of risk-taking induced by acute cannabis consumption may have long-lasting effects that influence not only sensitivity to rewards, but also the tendency to seek out novel stimuli. Future studies may further clarify the temporal relationship between reward sensitivity and sensation seeking in cannabis users by applying the ORL to cross-sectional or longitudinal samples. Finally, research by our own and other groups consistently reveals that computational model parameters are more sensitive to dissociating substance-specific and disorder-specific neurocognitive profiles than standard neurobehavioral performance indices (see Ahn et al., 2016 for a review). Such parameters show significant potential as novel computational markers for addiction and other forms of psychopathology, which could help refine neurocognitive phenotypes and develop more rigorous mechanistic models of psychiatric disorders (Ahn & Busemeyer, 2016).

Our results have implications for a wide range of cognitive tasks that involve

learning from experience. In particular, our finding that differential learning rates for positive and negative outcomes can capture the same behavioral patterns that have previously been attributed to a loss aversion parameter (cf. controls versus heroin users in Figure 6.5 to findings published in Ahn et al. [2014]) suggests that the underweighting of rare events that is observed in experience-based tasks may arise from learning, rather than valuation mechanisms (e.g., Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004). While the ORL limits this underweighting to tasks including outcomes in both gain and loss domains, future studies may extend the model to capture decisions in purely gain or loss domains by modifying the function that codes outcomes as gains versus losses (see equations 7-9). One potential solution could be to code outcomes as gains versus losses based on the sign of the prediction error rather than the objective outcome; in fact, cognitive models utilizing separate learning rates for positive versus negative prediction errors are gaining popularity in the decision sciences due to their theoretical and empirical support (e.g., Gershman, 2015).

Dataset	N	Population	IGT Version	Study Citation
Kjome	19	Healthy	Modified	Kjome et al. (2010)
Premkumar	25	Healthy	Modified	Premkumar et al. (2008)
Wood	153	Healthy	Modified	Wood et al. (2005)
Worthy	35	Healthy	Original	Worthy et al. (2013b)
Ahn	48	Healthy	Modified	Ahn et al. (2014)
Ahn	38	Amphetamine	Modified	Ahn et al. (2014)
Ahn	43	Heroin	Modified	Ahn et al. (2014)
Fridberg	15	Healthy	Original	Fridberg et al. (2010)
Fridberg	17	Cannabis	Original	Fridberg et al. (2010)

Table 6.1. Breakdown of datasets used in the current study.

	Dataset								
Model	1	2	3	4	5	6	7	8	9
ORL	**41.6**	**20.3**	6.9	23.4	**7.4**	15.4	9.7	81.5	25.1
PVL-Delta	44.9	20.6	**4.4**	17.3	8.5	**12.9**	**7.7**	**72.8**	**18.8**
VPP	44.7	20.9	6.0	**16.9**	8.8	15.0	9.0	85.5	20.6

Table 6.2. Mean squared deviations of true from simulated choice probabilities.

1 = Kjome; 2 = Premkumar; 3 = Wood; 4 = Worthy; 5 = Ahn (Healthy); 6 = Ahn (Amphetamine); 7 = Ahn (Heroin); 8 = Fridberg (Healthy); 9 = Fridberg (Cannabis). The lowest mean squared deviation (MSD) is bolded within each dataset.

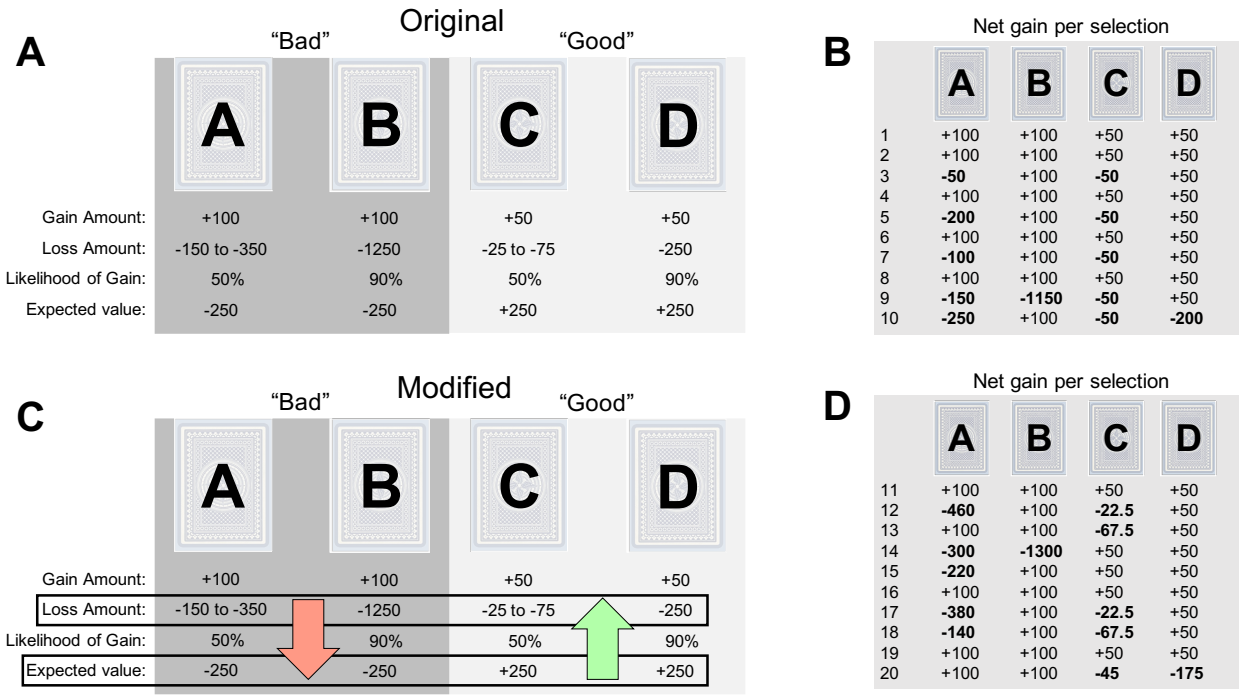

Figure 6.1. Structure of the original and modified versions of the IGT.

(A) The original version of the Iowa Gambling Task (IGT) maintains a stationary payoff distribution for all 100 trials. Decks A and B are both "bad" decks, each with an expected value of -250 points. In contrast, decks C and D are both "good" decks, each with an expected value of +250 points. Additionally, decks B and D both have a 90% chance of gaining points when chosen, whereas decks A and C have only a 50% chance. We present net outcomes here, but during the actual task, participants will see a gain and loss after each selection. Actual gains presented are +100 and +50 for the bad and good decks, respectively. Actual losses range in value depending on the deck. (B) Net gains (i.e. sum of actual gain and loss) for the first ten draws from each deck. (C) The modified version of the IGT is equivalent to the original version in all respects but one: the losses in the modified version become more and less severe in the bad and good decks, respectively, resulting in a drifting payoff distribution that makes the good decks easier to identify over time. The loss values change in a stepwise manner, where they are incremented after every ten draws from a given deck. (D) Net gains for the second set of ten draws (i.e. draws 11-20) from the modified IGT. Note that the first ten draws are identical to the original version, and that the bad decks have decreased in expected value while the good decks have increased.

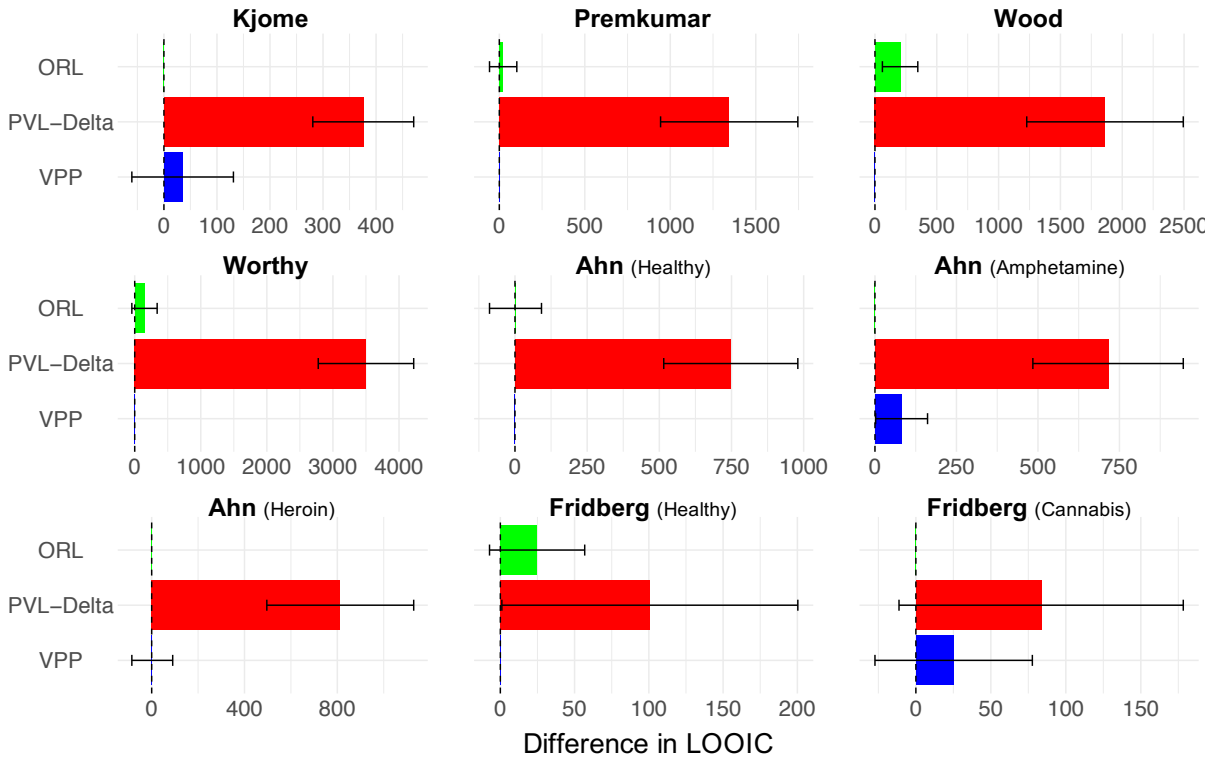

Figure 6.2. Post-hoc model fits across models and datasets.

Results of the leave-one-out information criterion (LOOIC) model comparison on one-step-ahead (i.e. short-term) prediction accuracy for each of the datasets analyzed in the current study. Lower LOOIC values indicate better model performance. LOOIC values were baselined by the best model in each comparison. The dashed line represents the zero point (i.e. best model LOOIC = 0), and any deviations from the zero point represent competing model LOOIC values. Error bars represent 2 standard errors on the difference between the best model and the respective competing model.

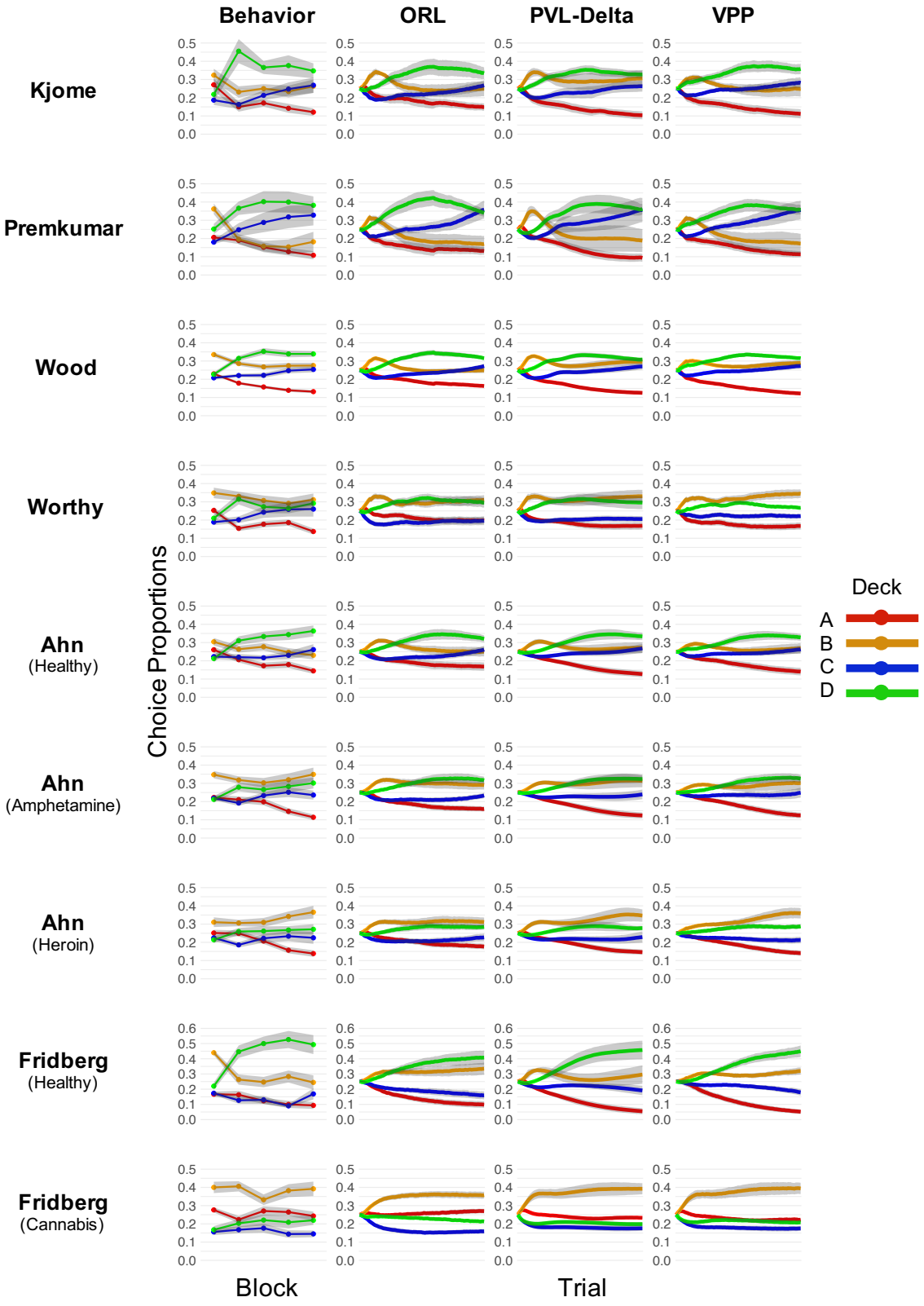

Figure 6.3. True versus simulated choice proportions across time.

Behavioral and simulation performance for the healthy control data for each of the datasets in the current study. Choice behavior is summarized per block, where blocks were constructed by calculating the proportion of choices made from each deck, across subjects, in 20-trial increments (i.e. block 1 = trials 1-20, block 2 = trials 21-40, etc.). Choice proportions across subjects are represented by points, and grey ribbons indicate 1 standard error. In general, subjects begin with a preference for deck B, but learn to prefer deck D as they progress through the task. Additionally, subjects show a clear preference for decks with high win frequency (B and D) over alternatives. Simulation performance is summarized per trial, across subjects within each dataset. The grey ribbons represent 1 standard error across subjects' averaged simulated choice probabilities.

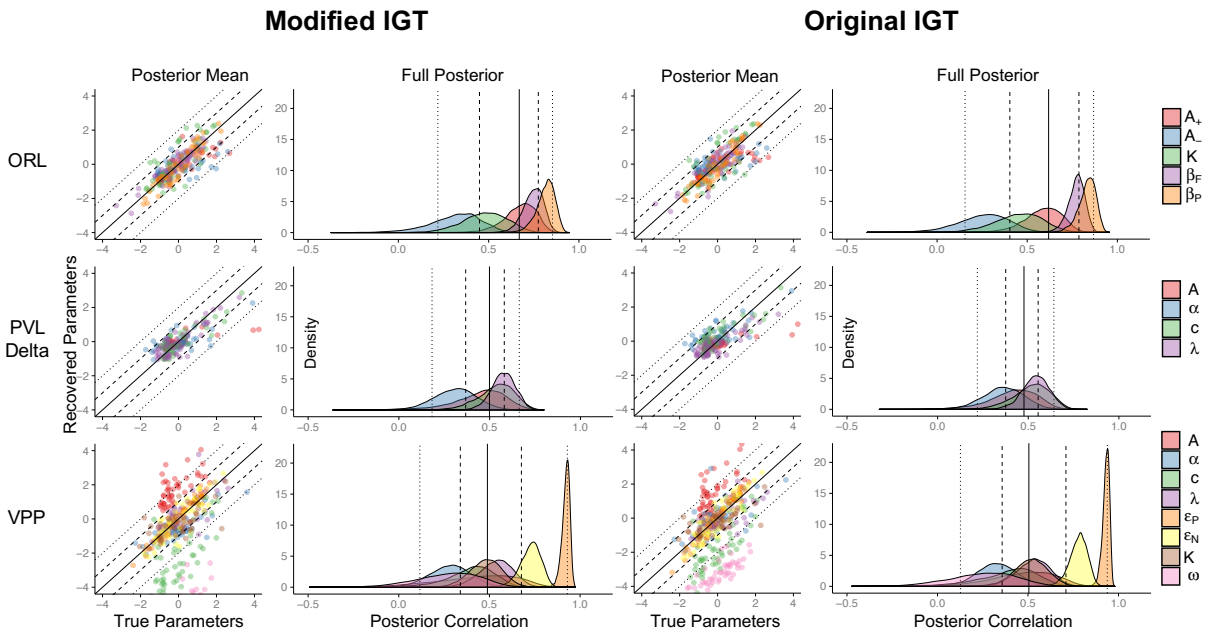

Figure 6.4. Parameter recovery results across models and versions of the IGT.

Parameter recovery results for the modified and original IGT tasks. Each task structure was simulated for each model using the same set of 48 individual-level parameter sets across modified and original task structures. Posterior mean results show comparisons of the true parameters with the means of the posterior distributions of the recovered parameters after being standardized. We standardized parameters by z-scoring the true and recovered posterior means by the mean and standard deviation of each of the 48 true parameter sets. This method allowed us to visualize the bias in recovered posterior means, where any values falling above or below the solid diagonal line indicate higher or lower recovered means in reference to the true parameters, respectively. Dashed and dotted lines reflect 1 and 2 standard deviations in the standardized space, respectively. Note that some parameter values fell outside of the graphs (particularly for the VPP), but zooming out further obfuscates the results. Full posterior recovery results were generated by computing a Spearman's rank-order correlation between each set of individual-level true parameters and the respective set of individual-level recovered parameters for each sample in the recovered posterior distribution. Full posterior recovery results therefore represent the uncertainty in recovering the relative positions of the true parameters across all individual-level parameters (i.e. across all "subjects"). Distributions with mass closer to 1 indicate that the order between true parameters is recovered well for a given parameter and model. Dotted lines represent 2.5% and 97.5% quantiles, dashed lines represent 25% and 75% quantiles, and the solid line represents the median. Quantiles were calculated across all parameters.

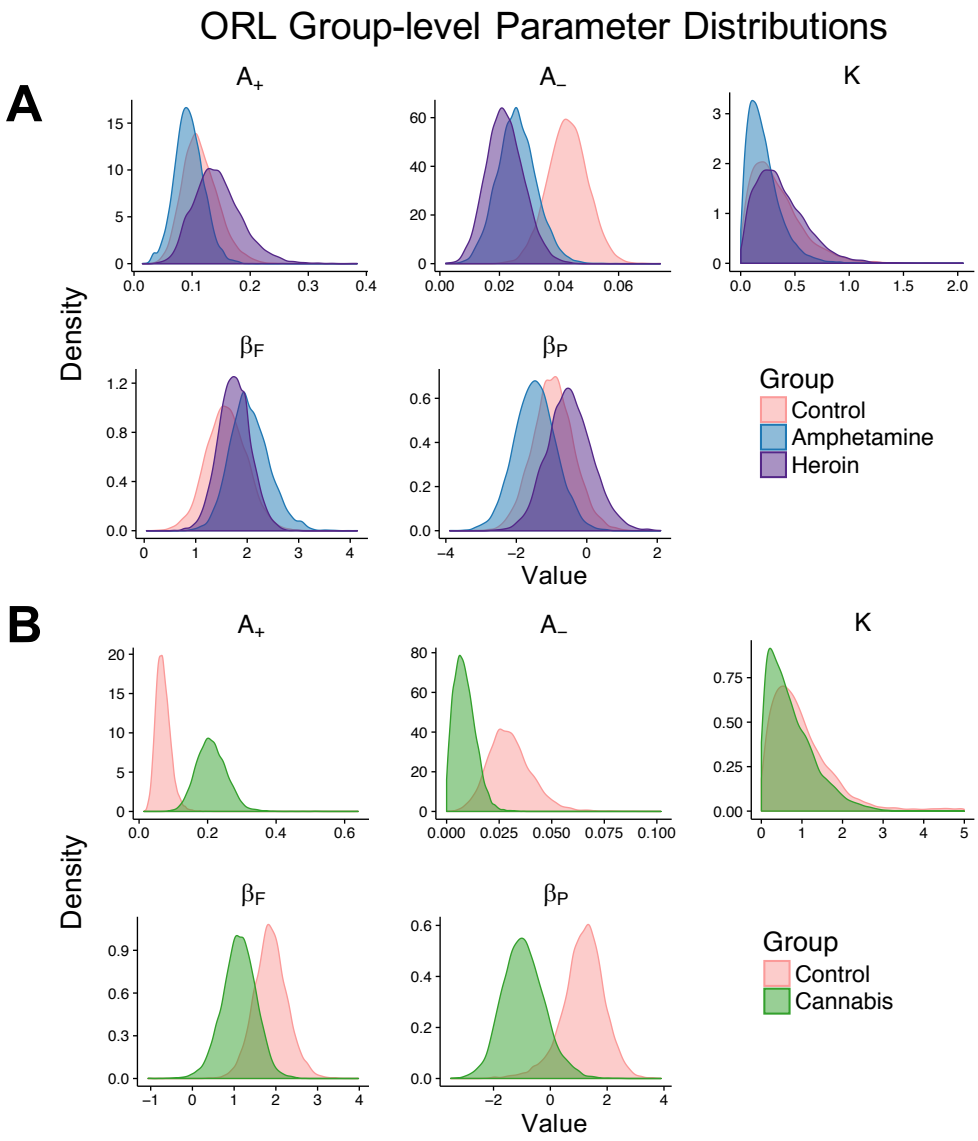

Figure 6.5. Group-level ORL parameters across healthy and substance using groups.

(A) Group-level parameter distributions for the healthy controls, amphetamine users, and heroin users who underwent the modified IGT. (B) Group-level parameter distributions for the healthy controls and chronic cannabis users who underwent the original IGT.

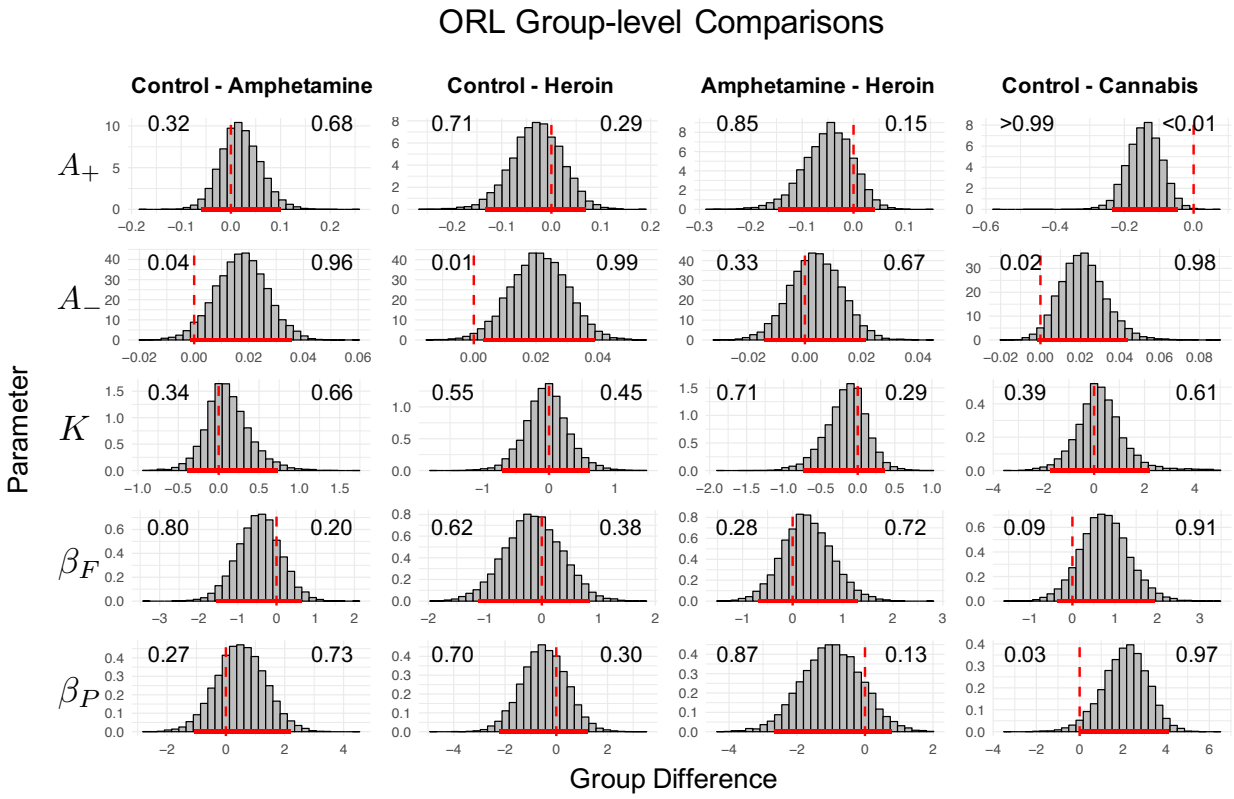

Figure 6.6. Differences in group-level ORL parameters between healthy and substance using groups.

Differences in group-level parameter distributions (for the ORL) between healthy controls and substance using groups. Solid red lines highlight the 95% highest posterior density interval (HDI), and dashed red lines reflect the 0 point. Values on the left and right sides of each graph represent the proportion of each distribution falling below and above the 0 point, respectively. Note that groups were compared within studies to minimize any confounding effects of task implementation, study design, and other site-specific experimental details.

Chapter 7: Anxiety Modulates Preference for Immediate Rewards among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis

Chapter 6 demonstrated how generative models can be developed to capture complex person × environment interactions relevant to the development of externalizing psychopathology. However, the focus was on group differences between controls versus those with substance use disorders, and not on individual differences per se. Therefore, Chapter 7 combines the power of Bayesian modeling to more precisely estimate individual difference relationships (discussed in Chapter 5) with more sophisticated generative models of behavior (discussed in Chapter 6) to test a conceptual theory of externalizing psychopathology that makes predictions involving individual differences. Specifically, I present a paper that I published in Clinical Psychological Science (Haines et al., 2020b) wherein I test predictions derived from Reinforcement Sensitivity Theory—a conceptual theory of approach/avoidance behavior which assumes that individual differences in sensitivity to punishments/rewards interact to give rise to impulsive or "self-controlled" behavior observed data level. This paper was published in collaboration with Theodore P. Beauchaine, Matthew Galdo, Andrew H. Rogers, Hunter

Hahn, Mark A. Pitt, Jay I. Myung, Brandon M. Turner, and Woo-Young Ahn.

7.1 Introduction

Impulsivity, defined behaviorally as a preference for immediate over delayed rewards, actions taken without forethought, and difficulties inhibiting prepotent behaviors (Neuhaus & Beauchaine, 2017; Sagvolden, Johansen, Aase, & Russell, 2005), is a highly heritable trait that confers vulnerability to all externalizing spectrum disorders (Beauchaine, Zisner, & Sauder, 2017), including attention-deficit/hyperactivity disorder (ADHD), oppositional defiant disorder, conduct disorder (CD), substance use disorders (SUDs), and antisocial personality disorder (ASPD). In structural models of adult psychopathology, all of these disorders load on a single, highly heritable latent vulnerability trait (see e.g., Krueger et al., 2002). A similar heritable trait emerges in structural models of child psychopathology, with the exceptions of ASPD and SUDs given limited opportunity for children to engage in criterion behaviors (Tuvblad, Zheng, Raine, & Baker, 2009). This shared latent vulnerability is often characterized as *trait impulsivity* based on common genetic, neural, cognitive, and behavioral processes observed across disorders (e.g., Beauchaine, Zisner et al., 2017; Gatzke-Kopp et al., 2009; Gizer, Otto, & Ellingson, 2017). Notably, those who are highly impulsive early in life—as manifested in the hyperactive-impulsive and combined presentations of ADHD—are at considerable risk for developing more severe forms of externalizing

conduct across development (Beauchaine & McNulty, 2013; Beauchaine, Zisner et al., 2017). Such progression is most likely in contexts of adversity, including family dysfunction (Patterson, Degarmo, & Knutson, 2000), child maltreatment (e.g., Shin, Cook, Morris, McDougle, & Groves, 2016), delinquent peer affiliations (e.g., McGloin & O'Neill Shermer, 2008), and exposure to neighborhood violence and criminality (Lynam et al., 2000; Meier, Slutske, Arndt, & Cadoret, 2008).

Given the high heritability of impulsivity and its associations with concurrent and future externalizing outcomes, many candidate biomarkers and endophenotypes of externalizing liability have been proposed including neural functions, autonomic responses, and laboratory task performance (e.g., Ersche, Turton, Pradhan, Bullmore, & Robbins, 2010; Foell et al., 2016; Ortiz & Raine, 2004; Patrick et al., 2006). As reviewed elsewhere, biological and behavioral markers could be useful for early identification of vulnerability given sufficient measurement precision (e.g., Beauchaine & Constantino, 2017). Such efforts are challenging, however, because like most human behavioral traits, impulsivity is distributed continuously in the population, and becomes impairing only when expressed at extremes. Accordingly, impulsivity and related constructs, such as self-control, figure prominently in theories of personality (e.g., Corr, 2004; Hampson, 2012). Other literatures link excessive impulsivity to certain mood disorders (Lombardo et al., 2012), personality disorders other than ASPD (McCloskey et al., 2009), and vulnerability to psychopathology more broadly (e.g., Beauchaine, Hinshaw, & Bridge, 2019; Carver & Johnson, 2018). These conceptualizations are consistent with burgeoning efforts to identify transdiagnostic features of mental illness (e.g., Beauchaine,

Constantino, & Hayden, 2018; Beauchaine & Hinshaw, 2020; Beauchaine & Thayer, 2015; Robbins, Gillan, Smith, de Wit, & Ersche, 2012). Impulsivity is therefore a construct of considerable interest both as an individual difference and as a marker of vulnerability to psychopathology. In this article, we consider complexities of measuring impulsivity, including possible explanations for low correspondences between self-reports and lab tasks.

7.1.1 Approaches to Measuring Impulsivity

Historically, impulsivity has been measured in many ways, often at different levels of analysis, including self-reports, informant reports, and assorted behavioral/cognitive tasks (for reviews see Neuhaus & Beauchaine, 2017; Oas, 1985; Rung & Madden, 2018; Vassileva & Conrod, 2019). For example, when assessing clinical levels of impulsivity among children and adolescents, informant-reports are commonly used. Such reports show high reliability and strong predictive validity to concurrent and future psychological function (see e.g., Achenbach & Edelbrock, 1991; Beauchaine, Zisner et al., 2017). Among adults, self-reports are commonly used given ease-of-administration and similarly strong reliability and predictive validity (e.g., Patton, Stanford, & Barratt, 1995). Notably, many adult measures assess multiple facets of impulsivity (Sharma, Markon, & Clark, 2014; Whiteside & Lynam, 2001). For example, the Barratt Impulsiveness Scale (BIS-11) assesses non-planning, motor, and attentional impulsivity (Patton et al., 1995). High scores on non-planning (BIS-NP), which captures preferences for immediate over delayed rewards, are observed consistently among those who abuse

substances, including alcohol, nicotine, stimulants, and heroin (Dom, Hulstijn, & Sabbe, 2006).

Self-reports aside, behavioral and cognitive approaches used to assess impulsivity include set-shifting tasks (e.g., Avila, Cuenca, Félix, Parcet, & Miranda, 2004), continuous performance tasks (e.g., Conners & MHS Staff, 2000), and go/no-go tasks (e.g., Bezdjian, Baker, Lozano, & Raine, 2009). More recently, monetary delay discounting tasks (DDTs) have gained popularity. DDTs, which we use here, assess how individuals assign value to delayed rewards by presenting them with sequences of choices between smaller magnitude, sooner (SS) rewards and larger magnitude, later (LL) rewards (e.g., Green & Myerson, 2004). Performance is quantified by individuals' *discounting rates*, which describe how precipitously they discount rewards as a function of increasing time delay to receipt of reward. Steeper discounting rates are observed among those with ADHD, CD, and ASPD, and among those who abuse alcohol, nicotine, heroin, and cocaine (e.g., Beauchaine, Ben-David, & Sela, 2017; Bickel & Marsch, 2001; Bobova, Finn, Rickert, & Lucas, 2009; Bornovalova, Daughters, Hernandez, Richards, & Lejuez, 2005; Petry, 2001; Wilson, Mitchell, Musser, Schmitt, & Nigg, 2010).

Despite frequent use of both self-report and task measures of impulsivity, correspondences between the approaches are usually weak (see Sharma et al., 2014). Meta-analyses show average correlations between multidimensional self-reports and behavioral measures of $r \approx .10$ (Cyders & Coskunpinar, 2011). These low correspondences are attributed to several sources, including low test-retest reliability of

behavioral tasks (Cyders & Coskunpinar, 2011; Hedge, Powell, & Sumner, 2017); state-dependence of behavioral tasks relative to self-reports (Cyders & Coskunpinar, 2011; Koff & Lucas, 2011); and failures of behavioral tasks to capture the multidimensional nature of impulsivity (Duckworth & Kern, 2011).

An additional possibility, which we examine here, is that impulsivity is determined in part by *functional dependencies* among different neurobehavioral substrates of behavior (e.g., Beauchaine & Constantino, 2017; Beauchaine & Hinshaw, 2020). Such perspectives date at least to the mid-20th Century, when Gray (1970, 1987) proposed that propensities toward approach behaviors derive from *competing effects* of individual differences in sensitivity to reward cues (trait impulsivity) vs. frustrative non-reward/punishment cues (trait anxiety). Gray's perspective (see also Gray & McNaughton, 2000), which generated a large body of research on psychophysiological correlates of impulsivity (e.g., Beauchaine, Katkin, Strassberg, & Snarr, 2001; Fowles, 2000), is currently instantiated in Reinforcement Sensitivity Theory (RST; Corr, 2001; 2004). RST specifies neural substrates of and functional interactions among cognitive-emotional valuation systems of activation and inhibition (Corr, 2008), including implications for externalizing behavior (Corr & McNaughton, 2016). Although full articulation of RST is beyond the scope of this article, it suggests that concurrently assessed dimensions of impulsivity (approach) and anxiety (avoidance), rather than measures of impulsivity alone, might better account for performance on specific tasks.

RST and similar perspectives are supported behaviorally by consistent evidence that trait anxiety mollifies externalizing risk among vulnerable children and adolescents (see

Beauchaine, Zisner, et al., 2017; Schatz & Rostain, 2006). For example, anxiety symptoms predict better responses to certain treatments among externalizing children (Jensen et al., 2001). Furthermore, youth with CD and comorbid anxiety are less aggressive, experience less peer rejection, and face fewer police contacts than youth with CD alone (Walker et al., 1991). In contrast, low trait anxiety is a hallmark of callous unemotional traits—which predict clinical severity of conduct problems (e.g., Enebrink, Andershed, & Långström, 2009; Frick & White, 2008; Tremblay, Pihl, Vitaro, & Dobkin, 1994). Thus, externalizing behaviors are often *potentiated* by low levels of anxiety, consistent with RST.

To date however, few studies have examined mechanisms through which anxiety moderates impulsive behaviors. At the neurobiological level of analysis, computational models of reward learning and delay discounting suggest that impulsivity-anxiety interactions may emerge from opponent dopaminergic and serotonergic systems, where dopamine facilitates learning from reward prediction errors across time and serotonin modulates cost and risk valuation of potential rewards (Cools, Nakamura, & Daw, 2011; Doya, 2002; 2008; Long, Kuhn, & Platt, 2009; Macoveanu et al., 2013). Among healthy controls, tryptophan (a serotonin precursor) depletion induces steeper delay discounting and stronger memory decay of previously experienced negative outcomes (Schweighofer et al., 2008; Tanaka et al., 2009).

At the neural level, experimentally induced anxiety attenuates value signals generated by the ventromedial prefrontal cortex (vmPFC) when encoding rewards, yielding more risk-averse decision-making (Engelmann, Meyer, Fehr, & Ruff, 2015).

Furthermore, comorbid anxiety among externalizing males is associated with less severe structural compromises in several brain regions implicated in impulsive decision-making, including the ventral striatum and the anterior cingulate cortex (Sauder, Beauchaine, Gatzke-Kopp, Shannon, & Aylward, 2012). Behaviorally, both typically developing children and children with ADHD show better response inhibition on stop-signal tasks if they experience symptoms of anxiety (Bloemsma et al., 2012; Manassis, Tannock, & Barbosa, 2000; Zinbarg & Revelle, 1989). Additionally, computational models derived from prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) reveal that those who meet criteria for generalized anxiety disorder show stronger risk aversion relative to healthy controls when making choices among both certain and probabilistic rewards/punishments (Charpentier, Aylward, Roiser, & Robinson, 2017). Similarly, individual differences in social anxiety, trait anxiety, and worry in both clinical and non-clinical samples are associated with risk aversion in the Balloon Analogue Risk Task, which mixes reward and punishment cues (Maner et al., 2007). Collectively, such findings are captured by RST through the *joint subsystem hypothesis*, which postulates a positive relation between anxiety and indecision (e.g., arising from goal conflict among reward magnitude and delay). Thus, anxiety and associated indecision allows for more thorough risk assessment, attenuating subjective valuations of reward relative to risk (see Corr, 2004; 2008).

Despite the relevance of RST to decision-making, to our knowledge no studies have tested interactive mechanisms through which impulsivity and anxiety affect impulsive decision-making, even though main effects of both are well characterized (e.g., Avila &

Parcet, 2001; Bloemsma et al., 2012; Duckworth & Kern, 2011; Manassis et al., 2000; Xia, Gu, Zhang, & Luo, 2017; Zhao, Cheng, Harris, & Vigo, 2015). Dependence of impulsive decisions on both trait impulsivity and anxiety may help to explain why self-report and behavioral measures of impulsivity show low correspondence (Cyders & Coskunpinar, 2011). Indeed, we would expect any 1:1 correspondence between trait and behavioral measures of impulsivity to be diminished to the extent that impulsive and anxious tendencies interact to affect decision-making (see Beauchaine & Hinshaw, 2020). More importantly, a fuller understanding of interactive effects between impulsivity and anxiety may help to explain mixed findings regarding differential effects of anxiety across different forms of impulsive decision-making and different groups of participants. Indeed, some studies find that anxiety decreases impulsive decision-making through increased risk-aversion, whereas others show increased impulsivity through steeper delay discounting (e.g., Charpentier et al., 2017; Schweighofer et al., 2008; Tanaka et al., 2009).

7.1.2 Modeling Functional Dependencies and Etiological Complexity

Quantifying complex functional dependencies among biobehavioral systems, such as those described above, presents significant barriers to testing theories of personality and psychopathology (Beauchaine & Constantino, 2017). In the present example, multiple neural mechanisms affect behavior in ways that are not well accounted for by traditional main effects regression models used in psychology. Instead, statistical models that account for functional dependencies among predictors across levels of analysis are needed. Traditional approaches linking DDT performance to personality traits first

quantify behavioral summary statistics separately for each participant (e.g., discounting rates) then use regression to estimate relations between those summary statistics and outcomes of interest (e.g., personality measures). This *two-stage* approach—as it is often termed in the cognitive neuroscience literature—does not allow for statistical constraint across levels of analysis (see Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017). In summarizing behavioral data before entering it into a secondary statistical model for hypothesis testing, the two-stage approach assumes implicitly that participants *share no group-level information* (e.g., knowing the average discounting rate across participants does not inform estimates at the individual-level), and that behavioral summary statistics are estimated with *infinite precision* (i.e., discounting rates are estimated without error)[10]. When these assumptions are not met, the two-stage method inflates measurement error. In turn, inflated measurement error leads to overconfident, biased estimates of model parameters/effects, particularly when numbers of observations for a measure are not fixed across participants and/or within conditions. In classic test theory terms, such estimates are *non-portable* (see Rouder & Haaf, 2019). Of note, self-report measures are often constructed using stringent criteria to help enforce portable estimates (e.g., ensuring high test-retest reliability, requiring all participants to answers the same questions, etc.). Summary measures from behavioral rarely meet these standards (e.g., Hedge, Powell, & Sumner, 2017).

A solution to these problems is to construct a single model that simultaneously pools

[10]We explain mathematical details underlying these assumptions in the Supplementary Text (see Model Parameterizations and Fitting Procedures *Base Descriptive model*).

behavioral data within and across participants to estimate both individual- and group-level summary statistics, and assumes theoretically relevant relations between behavioral (e.g., discounting rate) and external (e.g., personality traits) measures (e.g., Rouder & Haaf, 2019; Turner et al., 2017). Hierarchical Bayesian analysis (HBA; Craigmile, Peruggia, & Van Zandt, 2010; Kruschke, 2015; Lee & Wagenmakers, 2013; Rouder & Lu, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008) is a framework that can jointly estimate relations between task performance measures and individual-level personality measures (or any other combination of levels). HBA produces posterior distributions that convey how much *certainty* we have in parameter estimates given the data. Such information is not readily derived from traditional (frequentist) hierarchical modeling approaches that rely on maximum likelihood estimation. As we also demonstrate, HBA allows us to construct *competing models*, and to use formal Bayesian comparison techniques to determine which model best accounts for observed data while penalizing model complexity (for more information on benefits of Bayesian modeling, see Ahn, Krawitz, Kim, Busemeyer, & Brown, 2011; Craigmile et al., 2010; Rouder & Lu, 2005; Wagenmakers, 2007).

7.2 Objectives of The Current Study

Here we use an adaptive version of the delay discounting task (DDT), HBA, and Bayesian model comparison to show that current levels of anxiety moderate effects of trait-impulsivity on decision-making. We present data from three groups of participants (total N=967) with low to severe substance use patterns. The descriptive models we developed reveal that high state anxiety decreases rates at which trait-impulsive

individuals discount future rewards while performing the DDT. However, such findings appear to apply only to those who report concurrently high trait impulsivity *and* state anxiety. To better explain our pattern of findings, we develop a more mechanistic model that assumes anxiety and impulsivity are linked to cognitive mechanisms of reward/risk valuation and delay valuation, respectively. Given formal correspondence between our explanatory model and other models used in the decision-making literature, we can offer testable predictions regarding anxiety-impulsivity effects in alternative forms of impulsive decision-making (e.g., risky decision-making paradigms).

Results offer potential insight into mechanisms through which anxiety serves a protective role among impulsive individuals, yet *potentiates* impulsive decision-making among those without elevated trait impulsivity. We conclude that (1) main effects of single biobehavioral systems are often insufficient to describe task performance among those with psychopathology (see Beauchaine & Hinshaw, 2020; Beauchaine et al., 2018); (2) methods such as HBA offer principled means of testing complex theories of psychopathology that span levels of analysis; and (3) future research should gravitate away from searching for 1:1 correspondences between traits and task performance toward constructing statistical models that link levels of analysis in theoretically motivated ways (see Beauchaine & Constantino, 2017).

7.3 Method

7.3.1 Participants

Date were collected from three independent samples. Demographic characteristics of

each sample appear in Table 7.1. The first sample comprised adult undergraduates ($n_{\text{student}} = 132$) who participated for credit in an introductory psychology course. Students were recruited from a general pool, so we anticipated lower scores on both trait impulsivity and state anxiety than among the other groups, described below, who were selected for substance use behaviors. There were no exclusion criteria for students. Including the student group was important so we could determine whether or not state anxiety shows moderating effects on trait impulsivity when both are within normal ranges (cf. Corr, 2004, 2008; Corr & McNaughton, 2017).

The second group ($n_{\text{MTURK}} = 800$) was recruited through Amazon Mechanical Turk (MTURK), an online platform through which people participate in various tasks and/or surveys for money. Prior research demonstrates the utility of MTURK for rapid and large-scale collection of valid and reliable data for clinical and behavioral research (Mason & Suri, 2011; Shapiro, Chandler, & Mueller, 2013). MTURK participants were eligible if they lived in the United States, had approval ratings of 90% or above on past work (Mason & Suri, 2011), and reported problematic use of cigarettes, alcohol, marijuana, stimulants, or opioids during pre-screening. Only those who (1) believed they had a problem, or (2) reported having a relative or friend who was concerned with their substance use were enrolled. After pre-screening, MTURK participants were excluded if they failed more than 1 of 4 attention check questions randomly dispersed among questionnaires (e.g., "*Most people would rather lose than win*" is failed if a participant selects *True*). Additionally, we excluded MTURK participants who completed the DDT but failed to complete the trait impulsivity and/or state anxiety questionnaires

described below (8 total). MTURK participants were paid $10/hr. We anticipated this group would show higher levels of trait impulsivity than students given pre-screening criteria.

The third group ($n_{SUD} = 35$) comprised current patients at a local inpatient alcohol and drug treatment clinic (SUDs group). Participants were eligible if they met *DSM-5* (American Psychiatric Association, 2013) criteria for any alcohol or substance use disorder according to the Structured Clinical Interview for *DSM-5* (First, Williams, Karg, & Spitzer, 2015). Exclusion criteria included any history of head trauma with loss of consciousness for more than 5 min, a history of psychotic disorders, eight or more seizures, electroconvulsive therapy, or any neurological disorder. Participants were offered gift cards to a local grocery store at a rate of $10/hr. We expected SUDs participants would show the highest levels of trait impulsivity.

7.3.2 Measures

7.3.2.1 Barratt Impulsiveness Scale

The BIS-11 is a 30-item self-report questionnaire that assesses three facets of impulsivity including non-planning, motor, attentional impulsivity (Patton, Stanford, & Barratt, 1995). We used the non-planning subscale (BIS-NP), which comprises 11 questions and (a) is most closely aligned with conceptualizations of trait impulsivity reviewed above and (b) is a consistent correlate of DDT performance (e.g., Koff & Lucas, 2011). Internal consistency (Cronbach's α) and one-month test-retest reliability (r) of the BIS-NP both exceed .7 (Stanford et al., 2009).

7.3.2.2 State-Trait Anxiety Inventory

The STAI is a 40-item self-report measure that assesses state and trait anxiety (Spielberger, 1983). We used the state anxiety measure (STAI-S), as we hypothesized that current levels of anxiety, although affected by trait levels, would more potently moderate effects of trait impulsivity (i.e., BIS-NP) on discounting behavior—this hypothesis is based on the known casual effect that state anxiety has on risk sensitivity/reward valuation (e.g., Engelmann, Meyer, Fehr, & Ruff, 2015). Test-retest reliability of the STAI-S ranges from $r = .16$ to .83 for time periods spanning one week to many months (Barker, Wadsworth, & Wilson, 1976; Spielberger, 1983). Internal consistency (Cronbach's α) exceeds .80 (Spielberger, 1983). See Table S1 in the Supplementary Text for the bivariate correlations between all impulsivity and anxiety subscales.

7.3.2.3 Alcohol Use Disorder Identification Test

The AUDIT comprises 10 items that are used to assess risk for alcohol use disorder (Bohn, Babor, & Kranzler, 1995). We included the AUDIT to measure ranges of alcohol use across groups. A score of 8 or more among men (7 among women) indicates a strong likelihood of hazardous/harmful alcohol use. A score above 20 suggests alcohol use disorder. The AUDIT is both reliable ($r > .80$) and internally consistent ($\alpha > .80$) (Daeppen, Yersin, Landry, Pécoud, & Decrey, 2000; Hays, Merz, & Nicholas, 1995).

7.3.2.4 Drug Abuse Screening Test

The DAST-10 is a 10-item brief version of the 28-item DAST, which is used to assess past 12-month problematic substance use (Skinner, 1982). As with the AUDIT, we included the DAST-10 to measure variation in problematic substance use across groups.

A DAST-10 score > 2 indicates problematic substance use (Cocco & Carey, 1998). The DAST-10 shows acceptable test-retest reliability ($r > .70$) and good internal consistency ($\alpha > .80$) across validation studies (see Yudko, Lozhkina, & Fouts, 2007).

7.3.2.5 Structured Clinical Interview for the DSM-5, Research Version (SCID)

The SCID (First et al., 2015) was used to assess eligibility for the substance use treatment clinic group, primarily to assess which substances caused the most dysfunction for participants. All SCIDs were conducted by either: (1) trained graduate students in a clinical psychology Ph.D. program, or (2) by trained research assistants. Final diagnostic decisions were rendered by W.-Y. A. using a combination of SCID assessments and patient medical records to ensure patients did not meet exclusion criteria.

7.3.3 Behavioral Task

7.3.3.1 Delay Discounting Task

The monetary DDT comprises a sequence of binary choices between rewards varying in magnitude (dollars) and time of delivery (days, weeks, months, years). Each DDT trial consists of a choice between a smaller-sooner (SS) or larger-later (LL) reward (e.g., would you rather have *$10 now* or *$20 in one week*). After collecting choice data, impulsivity is captured by participants' discounting rates—a model parameter that measures how steeply they discount values of temporally-delayed rewards. A hyperbolic model (Mazur, 1987) is often used to describe discounting rates because it is simple and

fits choice patterns better than many similar alternatives (e.g., exponential, power) (but see Cavagnaro, Aranovich, McClure, Pitt, & Myung, 2016). Steeper discounting rates are observed among those with a wide range of externalizing conditions (ADHD, CD, ASPD), and among those who abuse various substances (Beauchaine, Ben-David et al., 2017; Bickel & Marsch, 2001; Bobova et al., 2009; Bornovalova et al., 2005; Petry, 2001; Wilson et al., 2010).

We used a DDT (Ahn et al., 2020) that uses a version of Bayesian active learning, adaptive design optimization to improve task efficiency and the precision of parameter estimation (ADO, see Myung, Cavagnaro, & Pitt, 2013). Trial-by-trial, ADO selects dollar-day pairs that are expected to improve parameter estimation the most. Participant-level parameters (discounting rate [k], and choice sensitivity [c]) are updated between trials using Bayesian updating, and delays and monetary values are then selected using a grid search over potential dollar-day pairs such that participants' choices minimize uncertainty in parameter estimates. This DDT version makes it possible to collect data 3-8 times more rapidly and 3-5 times more precisely than traditional staircase approaches (Ahn et al., 2020). Although each participant's parameters were estimated as they progressed through the task, modeling was conducted on raw choice data to facilitate hierarchical modeling. All three groups underwent two sessions of ADO-DDT separated by a 5 min break. Data from both sessions were combined to fit models described below. Student and SUDs groups both underwent 42 trials per session, whereas the MTURK group underwent 20 trials per session. We used fewer trials for MTURK participants because analyses of data from the

other groups, who were tested first, showed that additional trials rarely improved parameter estimation (test-retest reliability of delay discounting estimates exceeds $r =$ .95 after 20 or fewer ADO trials) and to minimize off-task behavior (Ahn et al., 2020).

7.3.3 Procedure

All participants provided informed consent before completing questionnaires (including the BIS-11 and STAI). They then completed two sessions of ADO for the DDT. Following the DDT, participants were debriefed and either given course credit or paid.

7.3.4 Data Analysis

First, we conducted Bayesian *t-test*s to determine whether trait impulsivity and state anxiety varied across groups in predicted directions (i.e., Students < MTURK < SUD). We used the R package *BEST*, which conducts Bayesian estimation of mean differences between groups as described by Kruschke (2015). *BEST* estimates parameters for means, *SD*s, and normality within groups, and differences between estimated means are used to infer group differences. We used the default, non-informative prior distributions for all parameters. We then interpreted each distribution using highest density intervals, which we describe in detail under *Interpreting Bayesian Models*.

Next, we developed two classes of competing models to test the hypothesis that state anxiety moderates trait impulsivity to predict discounting rates on the DDT. We term the first class of models *Descriptive*, in that they take the form of traditional

interaction models used throughout psychology (albeit within a hierarchical Bayesian framework). This allowed us to determine general relations between state anxiety, trait impulsivity, and delay discounting. We term the second class of models *Explanatory*, in that they make specific assumptions about how people value both rewards and delays in a way that gives rise to the interactive effect between impulsivity and anxiety we found with the *Descriptive* model[11]. Below, we describe *Base* and *Trait* versions of both classes of models, which assume that personality measures have either no relation to or are linear related to delay discounting model parameters, respectively.

7.3.4.1 Base Descriptive Model

The *Base Descriptive* model assumes that each participant discounts delayed rewards according to a hyperbolic function (Mazur, 1987) of the following form:

$$V = \frac{A}{1 + kt} \tag{1}$$

where V is the value of the delayed reward, A is the actual (objective) amount of the reward, k ($0 < k < +\infty$) is the discounting rate, and t is the time delay measured in weeks. With this parameterization, as k increases, the time delay (t) leads to greater decreases in the value of delayed rewards (V), which indicates steeper discounting of decision-making. V is computed for both the immediate and delayed options on each trial, and the subsequent values are then entered into a logistic equation to produce the probability of selecting the LL option:

[11]Note that we use the term *explanatory* because the model offers a specific explanation for how anxiety and impulsivity interact through their relations with different cognitive processes. We note, however, that the model is still descriptive because it does not identify a direct, causal mechanism.

$$Pr(LL) = \frac{1}{1 + e^{-c(V_{LL} - V_{SS})}} \quad (2)$$

Here, V_{LL} and V_{SS} reflect values of the LL and SS choice options after being discounted in Equation 1, and c ($0 < c < 5$) is a choice sensitivity (i.e. inverse temperature) parameter that captures how deterministically (c closer to 5) versus randomly (c closer to 0) participants make choices according to differences in V_{LL} and V_{SS}.

We used hierarchical Bayesian analysis (HBA) to simultaneously estimate group- and participant-level parameters separately for each of the three groups (Kruschke, 2015; Lee & Wagenmakers, 2013; Rouder & Lu, 2005; Shiffrin et al., 2008). HBA estimates posterior distributions that quantify uncertainty for each parameter, which makes it ideal for drawing reliable inferences on parameters in complex hierarchical models (e.g., Ahn et al., 2011). Details on the prior distributions and on the detailed fitting procedures (including all the models overviewed below) are in the Supplementary Text.

7.3.4.2 Trait Descriptive Model

To test our hypothesis of an impulsivity-anxiety dependency in affecting discounting, we implemented Bayesian regression by re-parameterizing k so it was determined by a linear combination of BIS-NP, STAI-S, and the interaction of BIS-NP and STAI-S (Boehm, Steingroever, & Wagenmakers, 2018). To do so, we first standardized each measure by mean-centering and rescaling by the *SD* separately within each group. Standardizing measures within each group allowed us to test if within-participant competing effects of trait impulsivity and state anxiety varied across groups. We then estimated deviations in the group-level discounting rate attributable to anxiety and

impulsivity using the following regression (see Equation S1 in the Supplementary Text for more details):

$$\mu_k = \beta_0 + \beta_1 \cdot \text{BIS-NP} + \beta_2 \cdot \text{STAI-S} + \beta_3 \cdot \text{BIS-NP} \cdot \text{STAI-S} \quad (3)$$

Here, β weights are interpreted similarly as in a standard multiple regression. Intuitively, β_0 is now interpreted as the group average discounting rate (i.e., μ_k from Eq. S1 in the Supplementary Text), and other β weights account for participant-level variance in k that is attributable to their respective BIS-NP and STAI-S scores. Note that we omitted participant-level subscripts in Eq. 3 for simplicity. Use of personality/trait measures to statistically constrain individual-level delay discounting estimates allows for the *Trait Descriptive* model to account for uncertainty in behavioral data when estimating personality-behavior relations. This contrasts with the traditional two-stage method, described above, which reduces behavioral summary statistics to single point (infinitely precise) estimates before probing personality-behavior relations[12].

7.3.4.3 Base Explanatory Model

Given our pattern of findings across groups from the *Trait Descriptive* model, we developed a more explanatory, mechanistic model of the interaction between impulsivity and anxiety using models derived from computational neuroscience, decision-making, and translational research on delay discounting (e.g., Cools et al., 2011; Doya, 2002; 2008; Ho, Mobini, Chiang, Bradshaw, & Szabadi, 1999; Luckman, Donkin, & Newell, 2017). Specifically, we made a simple extension to the traditional hyperbolic model

[12]We also tested the traditional frequentist version of the two-stage approach, which showed evidence for an interaction only in the SUDs group. We discuss these results in detail in the Supplementary Text (see Traditional Two-stage Approach from the Supplementary Text).

which assumes that reward magnitudes (e.g., $10) and delays (e.g., in two weeks) are valued independently and then combined in a way that naturally gives rise to an interactive effect:

$$V = \frac{A^{\alpha}}{1 + kt} \tag{4}$$

In equation 5, α ($0 < \alpha < +\infty$) is a reward magnitude valuation parameter that controls how sensitive people are to differences in reward (independent of delay) across choices on each trial. Importantly, changes in α can lead to similar behaviors compared to changes in the traditional discounting rate k. Specifically, as $\alpha \to 0$, rewards are valued more for their frequency than for their objective values, which leads to indifference between either reward offered on each trial (e.g., receiving $10 once is equivalent to receiving $1 once). Conversely, as $\alpha \to +\infty$, people become very sensitive to even small differences between rewards (e.g., receiving $10.25 once is strongly preferred over receiving $10 once). This extended model can be viewed as a variant of the multiplicative hyperbolic discounting model used in animal research (e.g., Ho et al., 1999), with the major difference being that we assume a power function for reward valuation as opposed to a hyperbolic saturating function.

As defined mathematically in Eq. 4, α corresponds to the "risk sensitivity/aversion" parameter from prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), as is leads to an increase in risk aversion at the behavioral level of analysis when $\alpha < 1$. Although our DDT does not involve risky decision-making, model comparison studies offer strong evidence that the risk aversion parameter (i.e. α) is preserved within-participants across risky- and inter-temporal choice paradigms (Luckman,

Donkin, & Newell, 2017). Therefore, although we do not interpret α as risk-aversion *per se*, it is a useful theoretical correspondence that leads to specific predictions regarding how anxiety may influence impulsive decisions in the DDT (see *Trait Explanatory model* below for details). Specifically, RST predicts that anxiety leads to risk assessment (see Corr, 2004, pg. 324), and we can encode this prediction in the model by assuming that state anxiety is linked to α. Therefore, we refer to α as "reward sensitivity" due to its direct interpretation, but emphasize that it produces risk aversion at the level of observed behavioral data, consistent with RST.

Finally, unlike in the *Descriptive* models, we did not estimate c (choice sensitivity) as a free parameter and instead set c to 1 for all participants when fitting the *Explanatory* models. We made this decision because α and c have similar functions in the model, which results in co-linearity between parameters[13]. More importantly, when $c = 1$ for all participants, the model described by Eq. 4 produces better interactive effects between α and k, which are described in more detail below (see *Trait Explanatory* model). See Figure 7.1 for graphical depiction of independent and interactive effects of α and k.

7.3.4.4 Trait Explanatory Model

Evidence suggests that temporal valuation of rewards (i.e., discounting rate, k) is related to impulsivity/excessive approach, whereas reward valuation/risk aversion (i.e.,

[13]We conducted an additional sensitivity analysis to determine if setting $c = 1$ affected our inference, as describes in the Supplementary Text (see Sensitivity Analysis). In Brief, we fit a model that estimated a single value for c across all participants (akin to the group-level parameters for α and k). Effects of state anxiety and trait impulsivity on α and k, respectively, were consistent with the reported model where $c = 1$ (Fig. S5).

α) is related to anxiety/excessive avoidance. Although k has traditionally been thought to capture impulsivity, correlational and experimental studies reveal a correspondence between trait and state measures of anxiety and behavioral/computational model parameters reflecting risk aversion, which is captured by α, as described above (see *Approaches to Measuring Impulsivity*) (e.g., Charpentier et al., 2017 Engelmann et al., 2015; Maner et al., 2007). Therefore, we assume that individual-level α and k parameters are systematically related to individual differences in state anxiety and trait impulsivity across participants, respectively:

$$\begin{aligned} \mu_\alpha &= \beta_{\alpha_0} + \beta_{\alpha_1} \cdot \text{STAI-S} \\ \mu_k &= \beta_{k_0} + \beta_{k_1} \cdot \text{BIS-NP} \end{aligned} \tag{5}$$

As in Eq. 3 (for the *Trait Descriptive* model), μ_α and μ_k indicate group-level means for reward (α) and delay (k) valuation parameters, which are estimated as a linear combination of a group-level "intercept" (β_{α_0}) and an "effect" (β_{α_1}) of individual differences in state anxiety (and similarly for impulsivity). Because this is the first empirical test of a model of this kind, we also tested the opposite model in which BIS-NP and STAI-S were assumed to relate to α and k, respectively (termed the *Trait Explanatory Incongruent* model; we use the term *Incongruent* for clarity, although it is possible that impulsivity and anxiety do in fact relate to α and k in this way despite empirical evidence suggesting otherwise). We also conducted a sensitivity analysis to determine whether our choice of BIS and STAI subscales appreciably affected our inference (see Supplementary Text). In general, results held across subscales, with the model presented in main text showing the strongest hypothesized relations (see Fig. S6).

By setting the choice sensitivity (c) parameter for the *Explanatory* models to 1, "competition" between reward valuation (α) and delay discounting (k) can lead to patterns of impulsive decision-making that explain likely anxiety-impulsivity interactions[14]. As $\alpha \to 0$, the effect of the discounting rate becomes increasingly attenuated, which leads to (near) indifference between the SS and LL options, irrespective of the magnitude of k. Conversely, as $\alpha \to +\infty$, the effect of k becomes increasingly strong, such that having a high k leads to consistent choices of the SS option and *vice-versa*. Therefore, if state anxiety and trait impulsivity are negatively and positively associated with α and k (through Eq. 5), respectively, then the *Trait Explanatory* model offers a more formal explanation of how state anxiety may interact with trait impulsivity to lead to impulsive decision-making (see Figure 7.1B for a graphical depiction).

7.3.4.5 Model Comparison

To compare *Descriptive* models in a fully Bayesian manner, we used the leave-one-out information criterion (LOOIC), which approximates how well a model should generalize to new data (Vehtari, Gelman, & Gabry, 2017). Because we fit *Descriptive* models separately to each group, we used LOOIC to estimate how well the models should perform on new participants sampled from the same groups (i.e., within student, MTURK, SUDs). In contrast, to compare *Explanatory* models, which were fit to all groups simultaneously, we used a leave-one-group-out measure (termed LPPD). We fit

[14]We use the term *competition* to refer broadly to the interactive nature of parameters in the model. We use this term instead of *interaction*, which could be misinterpreted to mean a traditional interaction as in Eq. 3.

Explanatory models simultaneously to the student and MTURK groups, and then made predictions on individual-level choices for each participant in the SUDs group using their state anxiety and trait impulsivity scores alone. Further details on the model comparison measures are included in the Supplementary Text.

7.3.4.6 Interpreting Bayesian Models

To interpret Bayesian models, we report highest density intervals (HDIs) to summarize posterior distributions, which are analogous but not equivalent to frequentist confidence intervals. An $x\%$ HDI covers the range of parameter values comprising $x\%$ of the area of the posterior distribution, where every value falling inside the interval is more probable than any value falling outside of the interval. Using the *Trait Descriptive* model as an example, a 95% HDI = [0.15, 0.3] on β_1 would indicate that the most probable 95% of values for β_1 fall between .15 and .3. Intuitively, it is useful to imagine the behavior of the HDI as we use a smaller and smaller $x\%$. As $x \to 0$, the interval converges to the single most probable parameter value (i.e., the mode of the distribution). As $x \to 100$, HDI continues to highlight the $x\%$ of most probable parameter values until covering the entire range of the distribution. In this way, HDI extends the concept of a mode from a point estimate to a range of values. Therefore, HDIs differ from frequentist confidence intervals in that they make direct assertions about which parameter values are most probable, whereas frequentist confidence intervals only make probability statements about the proportion of confidence intervals containing a given value under repeated sampling. Note that we do not endorse binary interpretations of "significant differences" using HDIs, but instead use them as a general

measure of evidence (e.g., "*Which discounting rate estimates are most probable?*", "*Which values best represent the effect of trait impulsivity on discounting rates?*", etc.). Again using the *Trait* model as an example, a 95% HDI = [0.15, 0.30] on β_1 would indicate strong evidence for a positive effect, given that the range of 95% most probable values are well above 0, and the 95% range is itself relatively narrow (i.e., the estimate is precise). Conversely, a 95% HDI = [-0.3, 0.4] on β_1 would indicate weak evidence for no effect, given that the range is both centered around 0 and relatively wide (i.e., the estimate is not precise). For detailed discussion of HDIs, their uses, and their similarities/differences with respect to frequentist confidence intervals, see chapter 11 of Kruschke (2015).

7.4 Results

7.4.1 State, Trait, and Behavioral Differences

Here, we report highest density intervals (HDIs) on estimated differences in mean trait impulsivity (BIS-NP) and state anxiety (STAI-S) scores between groups, in addition to estimated group-level discounting rates for each group. As depicted in Figure 7.2A, trait impulsivity varied across groups in the anticipated direction. Students had lower BIS-NP scores than both the MTURK, 95% $\text{HDI}_{\text{student-MTURK}}$ = [-2.67, -0.87], and SUDs groups, 95% $\text{HDI}_{\text{student-SUD}}$ = [-11.36, -7.61]. The MTURK group also had lower BIS-NP scores than the SUDs group, 95% $\text{HDI}_{\text{MTURK-SUD}}$ = [-9.42, -5.92]. Results were similar for state anxiety (see Figure 7.2A), where students had lower STAI-S scores than both the MTURK, 95% $\text{HDI}_{\text{student-MTURK}}$ = [-6.16, -2.23], and SUDs groups,

95% $HDI_{student\text{-}SUD}$ = [-19.82, -11.36]. The MTURK group also had lower STAI-S scores than the SUDs group, 95% $HDI_{MTURK\text{-}SUD}$ = [-15.29, -7.31]. Frequentist *t-test*s offered the same conclusions[15]. In the *Base Descriptive* model, discounting rates varied as predicted across groups (Figure 7.2B). The SUDs group showed the steepest discounting, followed by the MTURK group, then students. Taken together, results indicate that our selection criteria effectively produced three different groups with varying levels of trait impulsivity, state anxiety, and impulsive decision-making during the DDT.

7.4.2 Descriptive Models

Model comparison of the *Base* versus *Trait Descriptive* models showed that the *Trait Descriptive* model more effectively accounted for student ($LOOIC_{Base}$ - $LOOIC_{Trait}$ = 2.5, $SE_{Difference}$ = 6.9), MTURK ($LOOIC_{Base}$ - $LOOIC_{Trait}$ = 24.8, $SE_{Difference}$ = 25.9), and SUDs ($LOOIC_{Base}$ - $LOOIC_{Trait}$ = 68.4, $SE_{Difference}$ = 74.5) participants' DDT performance[16]. This suggests that main and/or dependent effects of trait impulsivity and state anxiety accounted for meaningful variance in individual-level decision-making[17]. The difference in LOOIC between models for students was lowest relative to

[15]Traditional frequentist *t-test*s showed that students had lower BIS-NP scores than both the MTURK, $t(196.7) = -3.91$, $p < .001$, $d = -0.56$, and SUDs groups, $t(51.5) = -10.02$, $p < .001$, $d = -2.79$, and the MTURK group had lower BIS-NP scores than the SUDs group, $t(37.9) = -8.79$, $p < .001$, $d = -2.86$. In addition, students had lower STAI-S scores than both the MTURK, $t(220.0) = -4.22$, $p < .001$, $d = -0.57$, and SUDs groups, $t(49.5) = -7.40$, $p < .001$, $d = -2.10$, and the MTURK group had lower STAI-S scores than the SUDs group, $t(38.6) = -5.74$, $p < .001$, $d = -1.85$.

[16]Because lower LOOIC values indicate better model performance, positive values for the difference of $LOOIC_{Base}$ - $LOOIC_{Trait}$ indicate better performance for the *Trait Descriptive* model.

[17]We fit a main effects only Trait Descriptive model (i.e., no impulsivity-anxiety interaction term) in addition to the full interaction model, which we describe in the Sensitivity Analysis section of the Supplemental Text. Results were consistent with those reported in text.

the *SE* of the difference, which may be due to a lack of dependency between BIS-NP and STAI-S among students. In fact, the 95% HDI on β_3, the interaction term, for students indicates weak evidence for no moderating effects of BIS-NP and STAI-S on discounting rates (95% HDI_{β_3} = [-0.34, 0.29]), whereas both the MTURK (95% HDI_{β_3} = [-0.25, 0.00]) and SUDs (95% HDI_{β_3} = [-0.98, -0.27]) samples showed evidence of moderating effects (see Fig. S1).

Additionally, both student (95% HDI_{β_1} = [0.16, 0.78]) and MTURK (95% HDI_{β_1} = [0.11, 0.40]) groups showed strong correspondences between non-planning impulsivity (BIS-NP) and discounting rates, conditioned on state anxiety and their interaction (Fig. S1). Conversely, the SUDs (95% HDI_{β_1} = [-0.66, 0.45]) group showed weak evidence for no conditional effect of BIS-NP on delay discounting. Conditional effects of state anxiety (STAI-S) on discounting rates were weaker, with some evidence for a negative association among students (95% HDI_{β_2} = [-0.57, 0.04]), and some evidence for a positive relationship in the MTURK (95% HDI_{β_1} = [-0.03, 0.26]) and SUDs (95% HDI_{β_1} = [-0.05, 0.89]) groups. Figure 7.3A shows *Descriptive Trait* model-predicted discounting rates for each group at varying levels of BIS-NP and STAI-S, which makes the moderating effect of anxiety on the association between impulsivity and discounting more clear. In the figure, interactions are evident for both the MTURK and SUDs groups, such that discounting rates are highest when individuals endorse both low levels of state anxiety and high levels of trait impulsivity. In contrast, the student group showed no interaction, with discounting rates best characterized by independent main effects of trait impulsivity and state anxiety. These results suggest that impulsive

decision-making is multiply determined by both trait impulsivity and state anxiety, although there is some apparent discrepancy between groups (i.e., main effects with no interaction in the student sample). Below, we expand to provide an explanatory account of the impulsivity-anxiety interactions, and develop a more robust model across all groups.

7.4.3 Explanatory Models

Model comparison of the *Base* versus *Trait Explanatory* models showed that the *Trait Explanatory* model—where α and k are assumed to relate to state anxiety and trait impulsivity, respectively—provided the best out-of-sample predictions (LPPD = -4098) compared to the *Base* (LPPD = -5011) and *Trait Incongruent* (LPPD = -4441) *Explanatory* models[18]. In addition to best predicting performance across the whole SUD group, the *Trait Explanatory* model outperformed competing models for individual participants in the SUDs group (see Fig. S3). Such results provide relatively strong evidence that state anxiety and trait impulsivity are linked to mechanisms of reward/risk and delay valuation (captured by α and k, respectively) in a way that generalizes across qualitatively different groups.

Posterior distributions for parameters of the *Trait Explanatory* model are shown in Fig. S4. We found strong evidence for a negative relationship between the STAI-S and α, such that increases in state anxiety predict attenuated reward valuation (95% $\text{HDI}_{\beta_{\alpha 1}}$ = [-0.045, -0.011]), consistent with both (1) relations between anxiety and

[18]LPPD closer to 0 indicates better predictive performance within the out-of-sample SUD group. See the Supplementary Text for further details on interpretation of LPPD.

indecision/increased risk sensitivity predicted by RST (Corr, 2004, 2008), and (2) previous studies showing that state anxiety increases risk aversion. Additionally, we found a positive association between BIS-NP and k, such that increases in trait impulsivity predicted increases in delay discounting (95% $\text{HDI}_{\beta_{k1}} = [0.20, 0.40]$). These results corroborate the interaction revealed by the *Trait Descriptive* model, and offer an explanation for how state anxiety and trait impulsivity interact to produce impulsive decisions. For example, Figure 7.3B demonstrates the estimated group-level effects of state anxiety and trait impulsivity on four different example choices from the DDT.

7.5 Discussion

Psychopathology research continues to shift from discrete syndromal conceptualizations of mental illness toward transdiagnostic trait approaches that specify complex interactions among multiple vulnerabilities (e.g., Beauchaine & Cicchetti, 2019; Beauchaine & Constantino, 2017; Beauchaine & Hinshaw, 2020; Robbins et al., 2012). Trait impulsivity is one such vulnerability (Beauchaine & McNulty, 2013; Beauchaine, Zisner et al., 2017; Ersche et al., 2010; Lombardo et al., 2012; McCloskey et al., 2009). Our findings demonstrate a clear functional dependency between trait impulsivity and state anxiety, such that high state anxiety decreases rates at which trait-impulsive individuals discount delayed rewards. Our *Trait Explanatory* model suggests that this pattern of behavior is better explained by a delay discounting model assuming that impulsivity and anxiety reflect delay/time and reward/risk valuation, respectively. Furthermore, given evidence that reward/risk sensitivity is preserved across intertemporal and risk decision-making paradigms within participants (Luckman,

Donkin, & Newell, 2017), and that anxiety inductions increase risk aversion (Engelmann, Meyer, Fehr, & Ruff, 2015), our model provides an explanation for why anxiety has differential effects on impulsive decisions across both different paradigms and levels of trait impulsivity. Decreases in reward/risk sensitivity (α) in response to anxiety lead to more random responding during delay discounting paradigms, which can be interpreted as either an increase or decrease in impulsivity depending on the individual's discounting rate (see Figure 7.1). However, in risky decision-making paradigms (e.g., \$3 with certainty or \$4 with probability .8), the same decrease in α leads to a higher likelihood of choosing the safe (i.e., "non-impulsive") option. Future studies might manipulate state anxiety experimentally among those who are low vs. high on trait impulsivity. Experimental manipulations, combined with alternative forms of delay discounting (e.g., cigarette discounting), may reveal novel strategies for decreasing reward values of drug cues among those with substance use disorders.

Our findings also have broader implications for traditional methods used to test hypotheses in psychopathology research. For example, psychopathology research efforts continue to shift away from single-level analyses and toward multiple-level analysis in development and validation of theories of mental illness (e.g., Beauchaine & McNulty, 2013; Cicchetti & Dawson, 2002; Cicchetti, Ackerman, & Izard, 1995). Oftentimes, researchers assume 1:1 links between constructs across levels of analysis. As in the two-stage approach, this assumes that behavioral measures are unidimensional and portable. However, behavior observed on seemingly single-dimension tasks (the DDT here) is often determined by multiple, competing mechanisms (see also Ahn et al., 2014;

Beauchaine & Constantino, 2017; Beauchaine & Hinshaw, 2020; Finucane, Challman, Martin, & Ledbetter, 2016; Haines et al., 2018). We demonstrated this across self-report and behavioral measures, but similar effects are observed when linking behavior to neural data (Turner et al., 2018). Consequently, main effects analyses using summary statistics derived from behavioral data alone are insufficient for identifying latent cognitive, emotional, and neural mechanisms underlying complex behaviors. Furthermore, the assumption of portability is rarely considered for data collected from anything other than self-report measures (e.g., behavioral, physiological, and neural data), which can lead to biased inferences and overconfidence in the wrong parameter values (e.g., *beta* weights from a multiple regression). HBA offers a flexible statistical framework to solve such problems and construct interpretable, complex models of psychopathology that can be formally compared (see Boehm et al., 2018; Rouder & Haaf, 2019).

Several limitations should be considered. First, the student sample and especially the SUDs sample were smaller than the MTURK sample. Smaller samples are underpowered relative to larger samples, and may also be influenced more by outliers and/or sample-specific characteristics. Of note, however, HBA is less sensitive to small sample sizes than traditional methods, which often do not pool information across participants in a principled manner to estimate effects (Kruschke, 2015; Lee & Wagenmakers, 2013; Rouder & Lu, 2005; Shiffrin et al., 2008). Additionally, our *Explanatory* model was fit simultaneously to all 967 participants, and we identified patterns consistent with the *Descriptive* models that were fit to each sample. Still, findings specific to the SUDs

sample in particular should be replicated in future research.

Second, the student, MTURK, and SUDs groups likely differ in other ways not measured, which may have contributed to finding no impulsivity-anxiety interaction within the student group using the *Trait Descriptive* models. Executive function/self-control is one possible explanation. In theory, strong executive control could modulate competition between impulsivity and anxiety, consistent with RST (see Figure 7.2A and section *State, Trait, and Behavioral Differences*; Beauchaine & Hinshaw, in press; Corr, 2004, 2008). However, there is significant overlap among posterior distributions for the interaction terms in the student and MTURK models (see Panel 3 of Fig. S1), and it is possible that a larger student sample could reveal an interaction similar to the MTURK group. Therefore, we caution over-interpretation of the interaction from the *Trait Descriptive* model in the student group given large uncertainty intervals. Future studies may address these points by incorporating additional relevant measures such as executive function into the *Trait Descriptive* model we developed. Furthermore, our study design was cross-sectional, and we are not claiming that links between trait impulsivity and state anxiety are causal. Use of anxiety manipulations in future studies may identify potential causal effects. Finally, although our *Explanatory* model takes a step in this direction, use of neurally-inspired computational models that account for dynamics among choices, response times, and neural activation might allow for more precise inferences on the joint effects of state anxiety and trait impulsivity on impulsive decision-making (cf. Turner et al., 2018; Turner, Van Maanen, & Forstmann, 2015). Although we did not collect reaction time measures in this study, future studies may

leverage such models to more precisely determine separable effects of impulsivity, anxiety, and executive function on impulsive decision-making and behavior.

In sum, state anxiety moderates the association between trait impulsivity and impulsive decision-making, such that high trait-impulsive individuals show reduced discounting of delayed rewards when they endorse high concurrent levels of anxiety. Such reduced discounting leads to more optimal, future-oriented decisions in the DDT. Further, our findings from the *Trait Explanatory* model reveal a mechanism through which anxiety may serve as a protective factor against impulsive behavior in those with externalizing spectrum disorders, yet lead to relatively more impulsive behavior for those with low trait impulsivity. Future research may use experimental manipulations to determine if within-subject anxiety inductions can decrease the value of drug cues in high trait-impulsive individuals with substance use disorders. More broadly, hierarchical Bayesian analysis offers a principled way to explore how mechanisms at one level of analysis interact to produce observations at another level, which can shed light on the dimensional neural, cognitive, and/or trait-level constructs that underlie traditionally discretized behavioral syndromes.

	Group			F	η^2
	Student (n=132)	MTURK (n=800)	SUD (n=35)		
Age (*SD*)	20.1 (4.6)	35.1 (10.8)	35.8 (10.3)	124.8	.2
Sex (male/female)	61/71	363/437	25/10	-	-
AUDIT score	4.9 (3.3)	9.6 (7.3)	14.9 (11.7)	25.4	.06
DAST-10 score	0.6 (0.9)	2.4 (2.1)	7.7 (2.9)	141.6	.23

Table 7.1. Demographic Characteristics by Group.

Due to experimenter error, participants recruited near the beginning of the study were not shown a portion of the AUDIT questionnaire. Summary statistics/statistical tests for the AUDIT were therefore computed on data collected from participants who completed the full questionnaire. Reduced sample sizes were 91, 674, and 27 for the student, MTURK, and SUD groups, respectively. AUDIT scores ≥ 7 in women (8 in men) indicate harmful/hazardous alcohol use. DAST-10 scores > 2 indicate problematic substance use. On average, the MTURK and SUD groups—but not the student group—reported problematic alcohol and substance use. For sex across groups, $\chi^2 = 9.1$.

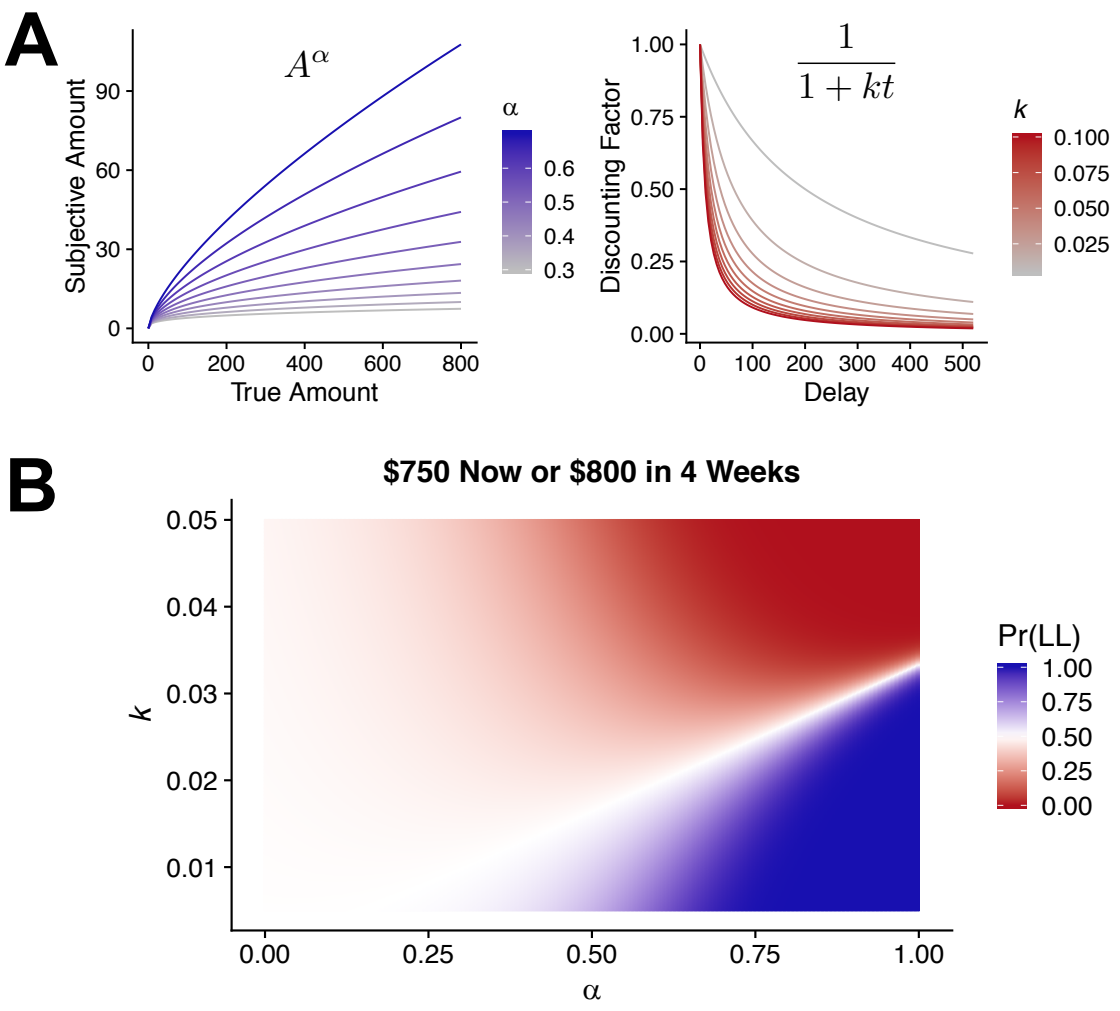

Figure 7.1. Graphical depiction of the Explanatory model described in the main text.

(A) The Explanatory model consists of two separate valuation mechanisms: one capturing reward magnitude sensitivity (α), and another capturing the traditional reward delay discounting rate (k). As α decreases toward 0, the subjective difference between two rewards of different magnitudes becomes increasingly small, and vice-versa. As k increases toward $+\infty$, rewards become increasingly discounted with time, and vice-versa. (B) The Explanatory model assumes that both valuation mechanisms described shown in panel A are combined such that they give rise to interactive effects (we constrained the parameter ranges for visualization purposes). Specifically, when reward sensitivity is low (i.e. as $\alpha \to 0$), the discounting rate (k) has a dampened effect on the resulting preference, and both choices become more equally preferred. Conversely, when reward sensitivity is high (i.e. as $\alpha \to +\infty$), the effect of k becomes increasingly strong, such that the larger later (LL) or shorter sooner (SS) choice becomes strongly preferred dependent on the specific choices and discounting rate. Assuming that reward magnitude and delay sensitivity are related to state anxiety and trait impulsivity,

respectively (see Trait Explanatory model), the model offers a more formal account of how anxiety and impulsivity may interact to produce (non)impulsive decisions.

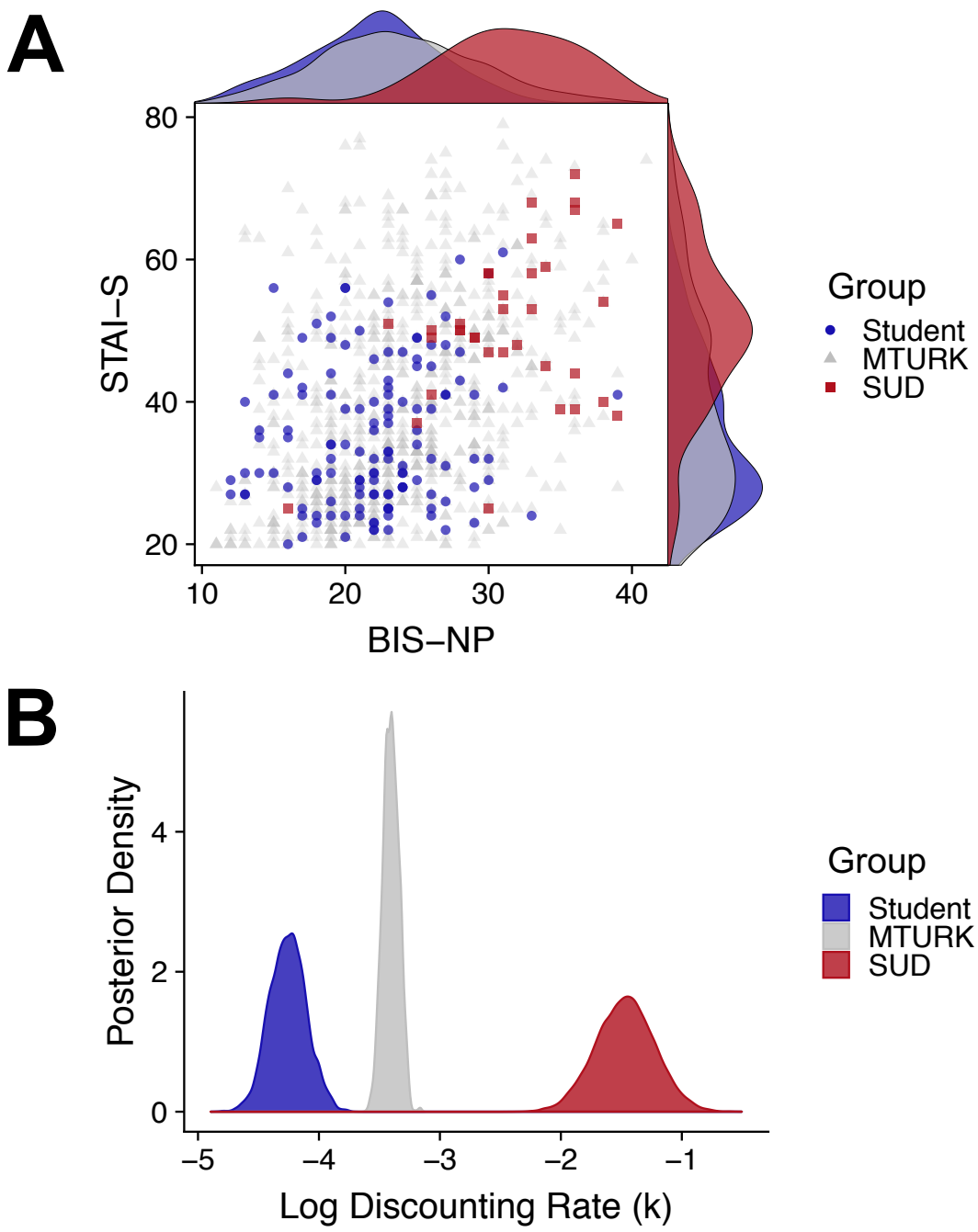

Figure 7.2. Trait impulsivity, state anxiety, and behavioral impulsivity across groups.

(A) Scatterplot with marginal distributions for summed scores of trait impulsivity (BIS-NP) and state anxiety (STAI-S) across groups. Pearson's correlations between BIS-NP and STAI-S scores for each group were $r_{student} = .17$, $r_{MTURK} = .38$, and $r_{SUD} = .39$. (B) Posterior distributions over group-level delay discounting rates estimated using the Base Descriptive model. Note that the distributions contain uncertainty in parameter estimates and can therefore be directly compared across groups.

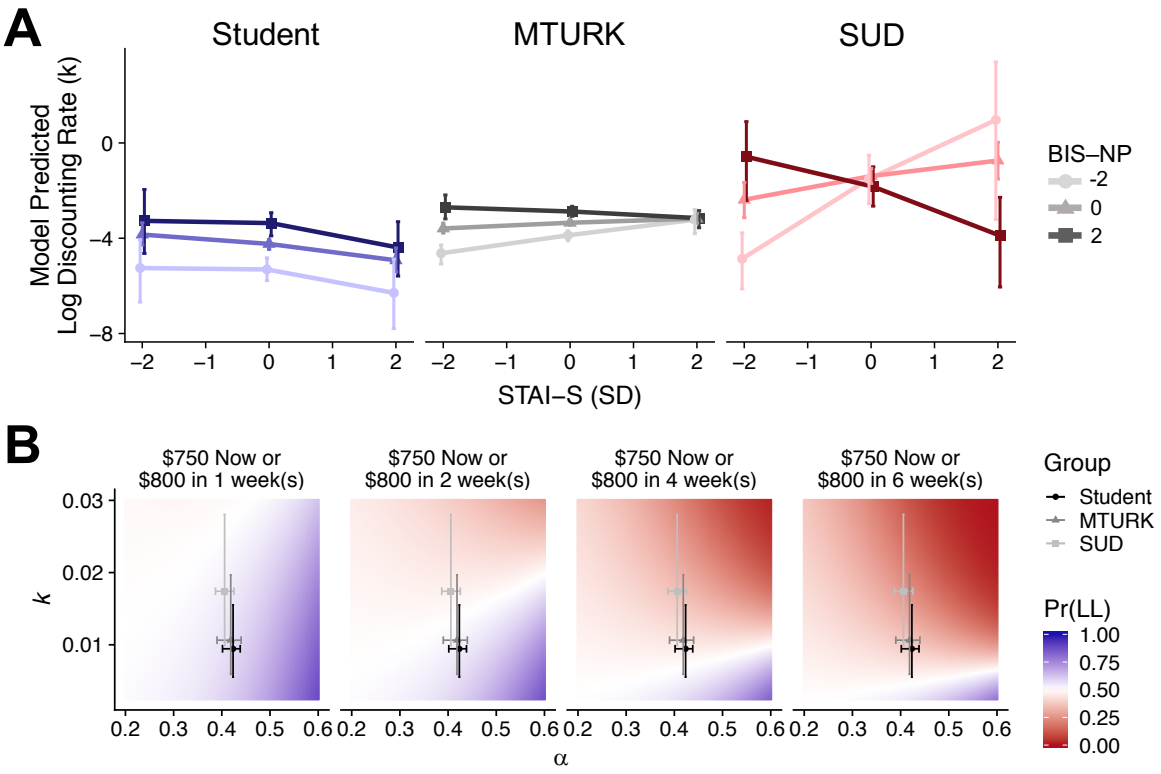

Figure 7.3. Interaction of BIS-NP and STAI-S in predicting discounting rates for both Trait models.

(A) Model-predicted discounting rates for different combinations (i.e., standard deviations from mean) of trait impulsivity (BIS-NP) and state anxiety (STAI-S) within each group given parameter estimates from the Trait Descriptive model. Points indicate modes of model predictions, and uncertainty intervals (vertical bars) reflect 80% HDIs of model predicted discounting rates (i.e., posterior predictive distributions over group-level discounting rates), which help to visualize how uncertainty in the Trait Descriptive model parameters affects estimates of discounting rate. Of note, Bayesian intervals indicate probabilities. Although low BIS-NP, high anxious participants in the SUD group appear to have steeper discounting rates relative to others on average, there is a non-negligible probability that they also have lower rates (i.e., the HDI spans both above and below others). When accounting for such uncertainty, MTURK and SUD groups show a very similar pattern. (B) Model-estimated effects of STAI-S and BIS-NP on parameters of the Trait Explanatory model for different example choices from the DDT (LL = larger later choice). Individual plots are "zoomed-in" versions of the same plot from Figure 7.1B, which we chose for interpretative purposes. Points indicate predicted group-level estimates (i.e., μ_k and μ_α from Eq. 5; see also Eq. S3 for more details) for individuals with sample-average levels of STAI-S and BIS-NP within each group. Uncertainty intervals highlight the same estimates, but for individuals at the 5^{th} and 95^{th} in-sample quantiles of STAI-S and BIS-NP within each group. Therefore, uncertainty intervals represent variation in α and k across participants that is

attributable to individual differences in state/trait measures, where α and k are negatively and positively associated with STAI-S and BIS-NP scores, respectively.

Chapter 8: Conclusions

In this dissertation, I reviewed theoretical models of how trait impulsivity and externalizing psychopathology develop across the lifespan, in addition to the non-trivial problems that researchers encounter when trying to measure and make inference on impulsivity—and psychological constructs more generally—across different levels of analysis (including trait, neural, and behavioral levels). Additionally, I present a framework that offers partial solutions to these measurement issues. I summarize the Chapters below.

To begin, Chapters 1-2 provide an overview of the Ontogenic Process Model of externalizing psychopathology, which suggests that individual differences in reward and punishment sensitivity interact with people's environmental contingencies (e.g., parenting style, peer influence, etc.) across the lifespan to give rise to individual differences in observed impulsive approach and anxious avoidance behavior that characterize ADHD, ODD, CD, SUDs, ASPD, and BPD.

Chapter 3 then discusses the central issues researchers face when operationalizing and testing predictions derived from such complex theories, including: (1) the poor reliability of many behavioral, neural, and other individual difference measures; and (2)

the lack of theoretical insight provided by the summary statistics that are often used to make inference on mechanisms underlying observed data. Chapter 4 briefly overviews solutions to issues (1) and (2), including the use of joint Bayesian modeling to simultaneously estimate parameters across all levels of analysis, thus accounting for measurement error that otherwise attenuates individual difference correlations; and the use of theoretically informed generative models of observed data to create a tighter link between quantities assumed by our conceptual theories and the statistical measures (i.e. parameters) that we estimate from data.

Next, Chapters 5-7 present three different journal articles that I published or have submitted that provide in-depth theoretical arguments and empirical demonstrations of the issues I raise and solutions I argue for in Chapters 3 and 4. Specifically, Chapter 5 begins by introducing the "Reliability Paradox", a phenomenon whereby many robust group-level effects derived from behavioral and neural data (e.g., the "Stroop effect") nevertheless show low test-retest reliability. The Reliability Paradox has led to many overly general conclusions regarding the utility of behavioral and neural data for individual difference research. Chapter 5 details the formal assumptions made by researchers when analyzing behavioral data that lead to the apparent Reliability Paradox, and shows how behavioral effects can have quite good test-retest reliability when using generative models that make more reasonable assumptions about behavior.

Chapter 6 then develops a generative model of reward and punishment learning in the context of the Iowa Gambling Task—a probabilistic learning task that has historically been used to identify "deficits" in decision-making between clinical and non-

clinical groups. I show that traditional summary measures (i.e. the proportion of "optimal" versus "non-optimal" choices) used to analyze data from such learning tasks conflate many different cognitive processes, and that computational models that make explicit assumptions about reward and punishment learning are better suited to infer differences in decision-making mechanisms relevant to learning. I then use the model to identify group differences in punishment learning between control participants and those with substance use disorders, which revealed that substance users had lower punishment learning rates (i.e. less sensitivity to punishment) relative to controls.

Chapter 7 then brings together insights from Chapters 5 and 6, using hierarchical Bayesian modeling to identify relationships between individual differences in self-reported trait impulsivity and state anxiety relate to impulsive behavior in the context of a Delay Discounting task. As described in detail throughout Chapter's 1 and 2 (see also Figure 2.1), diminished learning from rewards and/or punishments can lead to a preference for immediate over delayed rewards at the level of observed behavior (i.e. steeper delay discounting). In Chapter 7, I found a positive relationship between self-reported trait impulsivity and task-inferred discounting rates in people reporting low state anxiety, but no relationship between impulsivity and discounting in those reporting high state anxiety. This finding supports predictions derived from Reinforcement Sensitivity theory, suggesting that state anxiety inhibits impulsive approach behavior. In the same study, I developed a more detailed model that revealed a positive correlation between trait impulsivity and discounting rate and a negative correlation between state anxiety and reward/risk sensitivity. This "explanatory model"

suggests that anxiety may inhibit impulsive tendencies by modulating the subjective value of reward, such that higher state anxiety leads to diminished reward valuation. In the future, I hope to explore these results in more detail, linking them back to dynamic learning models as discussed in Chapter 6.

In summary, it is my hope that the theoretical and empirical findings I presented in this dissertation will facilitate both: (1) the appropriate treatment of uncertainty when estimating correlations or relationships between different levels of analysis, and (2) the development and use of theoretically informative generative models when making inference on mechanisms underlying observed data. In the context of externalizing psychopathology, the results I have presented so far show that adherence to both (1) and (2) gives way to strong forms of inference that traditional methods (e.g., computing point estimate summary statistics for different levels and using such point estimates to infer relationships between or among levels of analysis) are either underpowered (e.g., in the case of attenuation due to unreliability) or in-principle incapable (e.g., in the case of being used to measure quantities that conflate multiple cognitive processes) of providing.

To close: Making strong inference in the face of complexity requires that our statistical models both (1) take advantage of all the information available in data, and (2) adequately embody the conceptual theories that we aim to test or explore. Otherwise, our statistical methods become divorced from their intended purpose—at best, we risk failing to see important aspects of our data; at worse, we mischaracterize data in ways that lead our theories astray.

Bibliography

Achenbach, T.M., & Edelbrock, C.S. (1991). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University Associates in Psychiatry.

Adams, Z. W., Derefinko, K. J., Milich, R., & Fillmore, M. T. (2008). Inhibitory functioning across ADHD subtypes: Recent findings, clinical implications, and future directions. *Developmental Disabilities Research Reviews, 14*, 268-275. doi:10.1002/ddrr.37

Ahmad, S. I., & Hinshaw, S. P. (2017). Attention-Deficit/Hyperactivity Disorder, trait impulsivity, and externalizing behavior in a longitudinal sample. *Journal of Abnormal Child Psychology*, *45*, 1077-1089. doi:10.1007/s10802-016-0226-9

Ahn, W.-Y., & Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences*, *11*, 1-7. doi:10.1016/j.cobeha.2016.02.001

Ahn, W.-Y., Busemeyer, J., Wagenmakers, E.-J., & Stout, J. (2008). Comparison of Decision Learning Models Using the Generalization Criterion Method. *Cognitive Science*, *32* (8), 1376-1402. doi:10.1080/03640210802352992

Ahn, W.-Y., Dai, J., Vassileva, J., Busemeyer, J. R., & Stout, J. C. (2016). Computational modeling for addiction medicine: From cognitive models to clinical applications. *Progress in Brain Research*, *224*, 53-65. doi:10.1016/bs.pbr.2015.07.032

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., et al. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian

learning algorithm. *Scientific Reports*, *10*(1), 1-10. doi:10.1038/s41598-020-68587-x

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *bioRxiv*, *1*, 24-57. doi:10.1162/CPSY_a_00002

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J. R., & Brown, J. W. (2011). A model-based fMRI analysis with hierarchical Bayesian parameter estimation. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 95-110. doi:10.1037/a0020684

Ahn, W.-Y., Vasilev, G., Lee, S.-H., Busemeyer, J. R., Kruschke, J. K., Bechara, A., & Vassileva, J. (2014). Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. *Frontiers in Psychology*, *5*, 1376. doi:10.3389/fpsyg.2014.00849

Ahn, W.-Y., & Vassileva, J. (2016). Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. *Drug and Alcohol Dependence*, *161*, 247-257. doi:10.1016/j.drugalcdep.2016.02.008

Alexander, W. H., & Brown, J. W. (2019). The role of the anterior cingulate cortex in prediction error and signaling surprise. *Topics in Cognitive Science*, *11*, 119-135. doi:10.1111/tops.12307

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA. American Psychiatric Press.

Amlung, M., Marsden, E., Holshausen, K., Morris, V., Patel, H., Vedelago, L., et al. (2019). Delay discounting as a transdiagnostic process in psychiatric disorders. *JAMA Psychiatry*, *76*, 1176-1186. doi:10.1001/jamapsychiatry.2019.2102

Avila, C., & Parcet, M.A. (2001). Personality and inhibitory deficits in the stop-signal task: The mediating role of Gray's anxiety and impulsivity. *Personality and Individual Differences*, *31*, 975-986. doi:10.1016/S0191-8869(00)00199-9

Avila, C., Cuenca, I., Félix, V., Parcet, M.A., & Miranda, A. (2004). Measuring impulsivity in school-aged boys and examining its relationship with ADHD and ODD ratings. *Journal of Abnormal Child Psychology*, *32*, 295-304. doi:10.1023/B:JACP.0000026143.70832.4b

Barker, V., Romaniuk, L., Cardinal, R. N., Pope, M., Nicol, K., & Hall, J. (2015). Impulsivity in borderline personality disorder. *Psychological Medicine*, *45*, 1955-1964. doi:10.1017/S0033291714003079

Barker, H.R., Wadsworth, A.P., & Wilson, W. (1976). Factor structure of the state-trait anxiety inventory in a nonstressful situation. *Journal of Clinical Psychology*, *32*, 595-598. doi:10.1002/1097-4679(197607)32:3<595::AID-JCLP2270320322>3.0.CO;2-7

Barkley, R. A., Edwards, G., Laneri, M., Fletcher, K., & Metevia, L. (2001). Executive functioning, temporal discounting, and sense of time in adolescents with Attention Deficit Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD). *Journal of Abnormal Child Psychology*, *29*, 541-556. doi:10.1023/A:1012233310098

Barkley, R. A., Fischer, M., Smallish, L., & Fletcher, K. (2006). Young adult outcome of hyperactive children: Adaptive functioning in major life activities. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*, 192-202. doi:10.1097/01.chi.0000189134.97436.e2

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16* (3), 215-233. doi:10.1002/bdm.443

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10* (4), 331-344. doi:10.1037/1040-3590.10.4.331

Beauchaine, T. P. (2001). Vagal tone, development, and Gray's motivational theory: Toward an integrated model of autonomic nervous system functioning in psychopathology. *Development and Psychopathology, 13*, 183-214. doi:10.1017/S0954579401002012

Beauchaine, T. P. (2015). Future Directions in Emotion Dysregulation and Youth Psychopathology. *Journal of Clinical Child and Adolescent Psychology, 44*, 875-896. doi:10.1080/15374416.2015.1038827

Beauchaine, T. P., Ben-David, I., & Sela, A. (2017). Attention-deficit/hyperactivity disorder, delay discounting, and risky financial behaviors: A preliminary analysis of self-report data. *PLoS ONE*, *12*, e0176933. doi:10.1371/journal.pone.0176933

Beauchaine, T. P., & Cicchetti, D. (2019). Emotion dysregulation and emerging psychopathology: A transdiagnostic, transdisciplinary perspective. *Development and Psychopathology, 31*, 799-804. doi:10.1017/S0954579419000671

Beauchaine, T.P., & Constantino, J. (2017). Redefining the endophenotype concept to accommodate transdiagnostic vulnerabilities and etiological complexity. *Biomarkers in Medicine*, *11*, 769-780. doi:10.2217/bmm-2017-0002

Beauchaine, T.P., Constantino, J.N., & Hayden, E.P. (2018). Psychiatry and developmental psychopathology: Unifying themes and future directions. *Comprehensive Psychiatry, 87*, 143-152. doi:10.1016/j.comppsych.2018.10.014

Beauchaine T. P., Hinshaw S.P. (Eds) (2016). The Oxford handbook of externalizing spectrum disorders. New York, NY: Oxford University Press.

Beauchaine T. P., Hinshaw S. P. (2020). RDoC and psychopathology among youth: Misplaced assumptions and an agenda for future research. *Journal of Clinical Child and Adolescent Psychology*. epublished ahead of print. doi:10.1080/15374416.2020.1750022

Beauchaine, T.P., Hinshaw, S.P., & Bridge, J.A. (2019). Nonsuicidal self-injury and suicidal behaviors in girls: The case for targeted prevention in preadolescence. *Clinical Psychological Science, 7*, 643-667. doi:10.1177/2167702618818474

Beauchaine, T. P., Hinshaw, S. P., & Pang, K. L. (2010). Comorbidity of attention-deficit/hyperactivity disorder and early-onset conduct disorder: Biological, environmental, and developmental mechanisms. *Clinical Psychology: Science and Practice, 17*, 327-336. doi:10.1111/j.1468-2850.2010.01224.x

Beauchaine, T. P., Katkin, E. S., Strassberg, Z., & Snarr, J. (2001). Disinhibitory psychopathology in male adolescents: Discriminating conduct disorder from attention-deficit/hyperactivity disorder through concurrent assessment of multiple autonomic states. *Journal of Abnormal Psychology, 110*, 610-624. doi:10.1037/0021-843X.110.4.610

Beauchaine, T. P., & McNulty, T. (2013). Comorbidities and continuities as ontogenic

processes: Toward a developmental spectrum model of externalizing behavior. *Development and Psychopathology, 25*, 1505-1528. doi:10.1017/S0954579413000746

Beauchaine, T.P., & Thayer, J.F. (2015). Heart rate variability as a transdiagnostic biomarker of psychopathology. *International Journal of Psychophysiology*, *98*, 338-350. doi:10.1016/j.ijpsycho.2015.08.004

Beauchaine, T. P., Webster-Stratton, C., & Reid, M. J. (2005). Mediators, Moderators, and Predictors of 1-Year Outcomes Among Children Treated for Early-Onset Conduct Problems: A Latent Growth Curve Analysis. *Journal of Consulting and Clinical Psychology, 73*, 371-388. doi:10.1037/0022-006x.73.3.371

Beauchaine, T. P., & Zisner, A. (2017). Motivation, emotion regulation, and the latent structure of psychopathology: An integrative and convergent historical perspective. *International Journal of Psychophysiology, 119*, 108-118. doi:10.1016/j.ijpsycho.2016.12.014

Beauchaine, T. P., Zisner, A., & Sauder, C. L. (2017). Trait impulsivity and the externalizing spectrum. *Annual Review of Clinical Psychology, 13*, 343-368. doi:10.1146/annurev-clinpsy-021815-093253

Bechara, A., & Damasio, H. (2002). Decision-making and addiction (part I): impaired activation of somatic states in substance dependent individuals when pondering decisions with negative future consequences. *Neuropsychologia*, *40* (10), 1675-1689. doi:10.1016/S0028-3932(02)00015-5

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

3), 7-15. doi:10.1016/0010-0277(94)90018-3

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. doi:10.1016/S0028-3932(00)00136-6

Beitz, K. M., Salthouse, T. A., & Davis, H. P. (2014). Performance on the Iowa Gambling Task: From 5 to 89 years of age. *Journal of Experimental Psychology: General*, *143* (4), 1677-1689. doi:10.1037/a0035823

Bell ZE, Fristad MA, Youngstrom EA, Arnold LE, Beauchaine TP (2021). Prospective prediction of externalizing psychopathology by hyperactivity-impulsivity versus inattention: An empirical test of trait impulsivity theory. Journal of the American Academy of Child and Adolescent Psychiatry. Manuscript in press.

Betancourt, M. J., & Girolami, M. (2013). Hamiltonian Monte Carlo for Hierarchical Models. *arXiv Preprint*.

Bezdjian, S., Baker, L.A., Lozano, D.I., & Raine, A. (2009). Assessing inattention and impulsivity in children during the Go/NoGo task. *British Journal of Developmental Psychology*, *27*, 365-383. doi:10.1348/026151008X314919

Bickel, W.K.E., & Marsch, L.A. (2001). Toward a behavioral economic understanding of drug dependence: delay discounting processes. *Addiction*, *96*, 73-86. doi:10.1046/j.1360-0443.2001.961736.x

Biederman, J., Petty, C. R., Dolan, C., Hughes, S., Mick, E., Monuteaux, M. C., & Faraone, S. V. (2008). The long-term longitudinal course of oppositional defiant

disorder and conduct disorder in ADHD boys: findings from a controlled 10-year prospective longitudinal follow-up study. *Psychological Medicine*, *38*, 1027-1036. doi:10.1017/S0033291707002668

Birn, R. M., Roeber, B. J., & Pollak, S. D. (2017). Early childhood stress exposure, reward pathways, and adult decision making. *Proceedings of the National Academy of Sciences, 114*, 13549-13554. doi:10.1073/pnas.1708791114

Bloemsma, J.M., Boer, F., Arnold, R., Banaschewski, T., Faraone, S.V., Buitelaar, J.K., ...Oosterlaan, J. (2012). Comorbid anxiety and neurocognitive dysfunctions in children with ADHD. *European Child and Adolescent Psychiatry*, *22*, 225-234. doi:10.1007/s00787-012-0339-9

Bobova, L., Finn, P.R., Rickert, M.E., & Lucas, J. (2009). Disinhibitory psychopathology and delay discounting in alcohol dependence: Personality and cognitive correlates. *Experimental and Clinical Psychopharmacology*, *17*, 51-61. doi:10.1037/a0014503

Boehm, U., Steingroever, H., & Wagenmakers, E.-J. (2018). Using Bayesian regression to test hypotheses about relationships between parameters and covariates in cognitive models. *Behavior Research Methods*, *50*, 1248-1269. doi:10.3758/s13428-017-0940-4

Bohn, M.J., Babor, T.F., & Kranzler, H.R. (1995). The alcohol use disorders identification test (AUDIT): Validation of a screening instrument for use in medical settings. *Journal of Studies on Alcohol, 56*, 423-432. doi:10.15288/jsa.1995.56.423

Bornovalova, M.A., Daughters, S.B., Hernandez, G.D., Richards, J.B., & Lejuez, C.W.

(2005). Differences in impulsivity and risk-taking propensity between primary users of crack cocaine and primary users of heroin in a residential substance-use program. *Experimental and Clinical Psychopharmacology*, *13*, 311-318. doi:10.1037/1064-1297.13.4.311

Brumback, T. Y., Worley, M., Nguyen-Louie, T. T., Squeglia, L. M., Jacobus, J., & Tapert, S. F. (2016). Neural predictors of alcohol use and psychopathology symptoms in adolescents. *Development and Psychopathology*, *28*, 1209-1216. doi:10.1017/S0954579416000766

Buelow, M. T., & Blaine, A. L. (2015). The assessment of risky decision making: A factor analysis of performance on the Iowa Gambling Task, Balloon Analogue Risk Task, and Columbia Card Task. *Psychological Assessment*, *27*, 777-785. doi:10.1037/a0038622

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, *23*(3), 251-263. doi:10.1016/j.tics.2018.12.003

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*, 253-262. doi:10.1037/1040-3590.14.3.253

Caroselli, J. S., Hiscock, M., Scheibel, R. S., & Ingram, F. (2006). The Simulated Gambling Paradigm Applied to Young Adults: An Examination of University Students' Performance. *Applied Neuropsychology*, *13* (4), 203-212. doi:10.1207/s15324826an1304_1

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1-32. doi:10.18637/jss.v076.i01

Carver, C.S., & Johnson, S.L. (2018). Impulsive reactivity to emotion and vulnerability to psychopathology. *American Psychologist*, *73*, 1067-1078. doi:10.1037/amp0000387

Casella, G., & Berger, R. L. (2002). Statistical Inference. Duxbury Press.

Casey, B. J., Heller, A. S., Gee, D. G., & Cohen, A. O. (2019). Development of the emotional brain. *Neuroscience Letters*, *693*, 29-34. doi:10.1016/j.neulet.2017.11.055

Cavagnaro, D.R., Aranovich, G.J., McClure, S.M., Pitt, M.A., & Myung, J.I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, *52*, 233-254. doi:10.1007/s11166-016-9242-y

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review*, *18*(1), 204-210. doi:10.3758/s13423-010-0030-4

Charpentier, C.J., Aylward, J., Roiser, J.P., & Robinson, O.J. (2017). Enhanced risk aversion, but not loss aversion, in unmedicated pathological anxiety. *Biological Psychiatry*, *81*, 1014-1022. doi:10.1016/j.biopsych.2016.12.010

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., et al. (2015). Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS ONE*, *10*, e0144963. doi:10.1371/journal.pone.0144963

Chiu, Y.-C., & Lin, C.-H. (2007). Is deck C an advantageous deck in the Iowa Gambling Task? *Behavioral and Brain Functions*, *3* (1), 37. doi:10.1186/1744-9081-3-37

Chiu, Y.-C., Lin, C.-H., Huang, J.-T., Lin, S., Lee, P.-L., & Hsieh, J.-C. (2008). Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa Gambling Task? *Behavioral and Brain Functions*, *4* (1), 13. doi:10.1186/1744-9081-4-13

Cicchetti, D., Ackerman, B.P., & Izard, C.E. (1995). Emotions and emotion regulation in developmental psychopathology. *Development and Psychopathology*, *7*, 1-10. doi:10.1017/S0954579400006301

Cicchetti, D., & Dawson, G. (2002). Editorial: Multiple levels of analysis. *Development and Psychopathology*, *14*, 417-420. doi:10.1017/S0954579402003012

Cocco, K.M., & Carey, K.B. (1998). Psychometric properties of the Drug Abuse Screening Test in psychiatric outpatients. *Psychological Assessment*, *10*, 408-414. doi:10.1037/1040-3590.10.4.408

Cole, D. M., Beckmann, C. F., Searle, G. E., Plisson, C., Tziortzi, A. C., Nichols, T. E., et al. (2011). Orbitofrontal connectivity with resting-state networks is associated with midbrain dopamine D3 receptor availability. *Cerebral Cortex*, *22*, 2784-2793. doi:10.1093/cercor/bhr354

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biological Psychiatry*, *82* (6), 431-439. doi:10.1016/j.biopsych.2017.05.017

Conduct Problems Prevention Research Group. (2011). The effects of the fast track preventive intervention on the development of conduct disorder across childhood.

Child Development, 82, 331-345. doi:10.1111/j.1467-8624.2010.01558.x

Conners, C.K., & MHS Staff. (2000). *Conners' Continuous Performance Test II* (CPT II). North Tonawanda, NY: Multi-Health Systems.

Connor, P., & Evers, E. R. K. (2020). The Bias of Individuals (in Crowds): Why Implicit Bias Is Probably a Noisily Measured Individual-Level Construct:. *Perspectives on Psychological Science*, *15*, 1329-1345. doi:10.1177/1745691620931492

Cools, R., Nakamura, K., & Daw, N.D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology, 36*, 98-113. doi:10.1038/npp.2010.121

Corr, P.J. (2001). Testing problems in J. A. Gray's personality theory: A commentary on Matthews and Gilland (1999). *Personality and Individual Differences*, *30*, 333-352. doi:10.1016/S0191-8869(00)00028-3

Corr, P.J. (2004). Reinforcement sensitivity theory and personality. *Neuroscience and Biobehavioral Reviews*, *28*, 317-332. doi:10.1016/j.neubiorev.2004.01.005

Corr, P.J. (2008). *Reinforcement Sensitivity Theory of personality*. Cambridge, UK: Cambridge University Press.

Corr, P.J., & McNaughton, N. (2012). Neuroscience and approach/avoidance personality traits: A two stage (valuation-motivation) approach. *Neuroscience and Biobehavioral Reviews, 36*, 2339-2354. doi:10.1016/j.neubiorev.2012.09.013

Corr, P.J., & McNaughton, N. (2016). Neural mechanisms of low trait anxiety and risk for externalizing behavior. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *The Oxford handbook of externalizing spectrum disorders* (pp. 220-238). New York: Oxford

University Press.

Cox, S. M. L., Frank, M. J., Larcher, K., Fellows, L. K., Clark, C. A., Leyton, M., & Dagher, A. (2015). Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *Neuroimage*, *109*, 95-101. doi:10.1016/j.neuroimage.2014.12.070

Coyle, J.T., & Campochiaro, P. (1976). Ontogenesis of dopaminergic-cholinergic interactions in the rat striatum: A neurochemical study. *Journal of Neurochemistry, 27*, 673-678. doi:10.1111/j.1471-4159.1976.tb10393.x

Craigmile, P.F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika*, *75*, 613-632. doi:10.1007/s11336-010-9172-6

Criss, A. H. (2010). Differentiation and response bias in episodic memory: evidence from reaction time distributions. *Psychological Assessment*, *36*, 484-499. doi:10.1037/a0018435

Crowell, S. E., Beauchaine, T. P., Gatzke-Kopp, L., Sylvers, P., Mead, H., & Chipman-Chacon, J. (2006). Autonomic correlates of attention-deficit/hyperactivity disorder and oppositional defiant disorder in preschool children. *Journal of Abnormal Psychology, 115*, 174-178. doi:10.1037/0021-843x.115.1.174

Cyders, M., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clinical Psychology Review*, *31*, 965-982. doi:10.1016/j.cpr.2011.06.001

d'Acremont, M., Lu, Z.-L., Li, X., Van der Linden, M., & Bechara, A. (2009). Neural

correlates of risk prediction error during reinforcement learning in humans. *Neuroimage*, *47* (4), 1929-1939. doi:10.1016/j.neuroimage.2009.04.096

Daeppen, J.B., Yersin, B., Landry, U., Pécoud, A., & Decrey, H. (2000). Reliability and validity of the Alcohol Use Disorders Identification Test (AUDIT) embedded within a general health risk screening questionnaire: Results of a survey in 332 primary care patients. *Alcoholism, Clinical and Experimental Research*, *24*, 659-665. doi:10.1111/j.1530-0277.2000.tb02037.x

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, 1-3. doi:10.1016/j.tics.2020.01.007

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

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE*, *14*(5), e0216125. doi:10.1371/journal.pone.0216125

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020, April 28). The case for formal methodology in scientific reform. *bioRxiv*. Cold Spring Harbor Laboratory. doi:10.1101/2020.04.26.048306

Diamond, A. (2005). Attention-deficit disorder (attention-deficit/hyperactivity disorder without hyperactivity): A neurobiologically and behaviorally distinct disorder from attention-deficit/hyperactivity disorder (with hyperactivity). *Development and*

Psychopathology, 17, 807-825. doi:10.1017/S0954579405050388

Dingle, H. (1950). A theory of measurement. *British Journal for the Philosophy of Science, 1,* 5-26. doi:10.1093/bjps/I.1.5

Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm. *American Psychologist, 54*, 755-764. doi:10.1037/0003-066X.54.9.755

Dishion, T. J., & Racer, K. H. (2013). Development of adult antisocial behavior. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *Child and adolescent psychopathology* (2nd ed., pp. 453-487). Hoboken, NJ: Wiley.

De Brito, S. A., Mechelli, A., Wilke, M., Laurens, K. R., Jones, A. P., Barker, G. J., et al. (2009). Size matters: Increased grey matter in boys with conduct problems and callous-unemotional traits. *Brain, 132*, 843-852. doi:10.1093/brain/awp011

DeVellis, R. F. (1991). *Scale Development: Theory and Applications*. Newbury Park, California: Sage.

Dodge, K.A., Bierman, K.L., Coie, J.D., Greenberg, M.T., Lochman, J.E., McMahon, R.J., Pinderhughes, E.E., for the Conduct Problems Prevention Research Group (2015). Impact of early intervention on psychopathology, crime, and well-being at age 25. *American Journal of Psychiatry, 172*, 59-70. doi:10.1176/appi.ajp.2014.13060786

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74-94. doi:10.1016/j.brainres.2009.07.007

Dom, G., Hulstijn, W., & Sabbe, B. (2006). Differences in impulsivity and sensation

seeking between early- and late-onset alcoholics. *Addictive Behaviors*, *31*, 298-308. doi:10.1016/j.addbeh.2005.05.009

Dong, X., Li, S., & Kirouac, G. J. (2017). Collateralization of projections from the paraventricular nucleus of the thalamus to the nucleus accumbens, bed nucleus of the stria terminalis, and central nucleus of the amygdala. *Brain Structure and Function, 222*, 3927-3943. doi:10.1007/s00429-017-1445-8

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2010). Diffusion versus linear ballistic accumulation: different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18* (1), 61-69. doi:10.3758/s13423-010-0022-4

Douglas, V. I., & Parry, P. A. (1994). Effects of reward and nonreward on frustration and attention in attention deficit disorder. *Journal of Abnormal Child Psychology*, *22*, 281-302. doi:10.1007/BF02168075

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*, 495-506. doi:10.1016/S0893-6080(02)00044-8

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*, 410-416. doi:10.1038/nn2077

Duckworth, A.L., & Kern, M.L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*, 259-268. doi:10.1016/j.jrp.2011.02.004

Dyck, R. J,, Bland, R. C., Newman, S. C., & Orn, H. (1988). Suicide attempts in psychiatric disorders in Edmonton. *Acta Psychiatrica Scandinavia, 77*, 64-71.

doi:10.1111/j.1600-0447.1988.tb08549.x

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119-127. doi:10.2307/24954030

Elkins, I. J., McGue, M., & Iacono, W. G. (2007). Prospective effects of attention-deficit/hyperactivity disorder, conduct disorder, and sex on adolescent substance use and abuse. *Archives of General Psychiatry*, *64*, 1145-1152. doi:10.1001/archpsyc.64.10.1145

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., et al. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792-806. doi:10.1177/0956797620916786

Enebrink, P., Andershed, H., & Långström, N. (2005). Callous-unemotional traits are associated with clinical severity in referred boys with conduct problems. *Nordic journal of psychiatry*, *59*, 431-440. doi:10.1080/08039480500360690

Engelmann, J.B., Meyer, F., Fehr, E., & Ruff, C. C. (2015). Anticipatory anxiety disrupts neural valuation during risky choice. *Journal of Neuroscience*, *35*, 3085-3099. doi:10.1523/JNEUROSCI.2880-14.2015

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

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement

Learning Among Cognitive Strategies. *Psychological Review*, *112* (4), 912-931. doi:10.1037/0033-295X.112.4.912

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*, 369-409. doi:10.1037/rev0000062

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, *21* (5), 575-597. doi:10.1002/bdm.602

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, *88* (4), 848-881. doi:10.2307/117009

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149. doi:10.3758/BF03203267

Ersche, K.D., Turton, A.J., Pradhan, S., Bullmore, E.T., & Robbins, T.W. (2010). Drug addiction endophenotypes: Impulsive versus sensation-seeking personality traits. *Biological Psychiatry*, *68*, 770-773. doi:10.1016/j.biopsych.2010.06.015

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Assessment*, *53*, 134-140. doi:10.1037/h0045156

Fair, D. A., Nigg, J. T., Iyer, S., Bathula, D., Mills, K. L., Dosenbach, N. U. F., ... Milham, M. P. (2013). Distinct neural signatures detected for ADHD subtypes after controlling for mico-movements in resting state functional connectivity MRI data.

Frontiers in Systems Neuroscience, 6, 80. doi:10.3389/fnsys.2012.00080

Farrell, S., & Lewandowsky, S. (2018). Computational Modeling of Cognition and Behavior. New York, NY: Cambridge University Press.

Finucane, B., Challman, T.D., Martin, C.L., & Ledbetter, D.H. (2016). Shift happens: Family background influences clinical variability in genetic neurodevelopmental disorders. *Genetics in Medicine*, *18*, 302-304. doi:10.1038/gim.2015.92

First, M.B., Williams, J.B. W., Karg, R.S., & Spitzer, R.L. (2015). *Structured Clinical Interview for DSM-5—Research Version* (SCID-5). Arlington, VA: American Psychiatric Association.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series a, Containing Papers of a Mathematical or Physical Character*, *222*, 309-368. doi:10.1098/rsta.1922.0009

Fletcher, J. M. (2014). The effects of childhood ADHD on adult labor market outcomes. *Health Economics,* 23, 159-181. doi:10.3386/w18689

Foell, J., Brislin, S.J., Strickland, C.M., Seo, D., Sabatinelli, D., & Patrick, C.J. (2016). Externalizing proneness and brain response during pre-cuing and viewing of emotional pictures. *Social Cognitive and Affective Neuroscience*, *11*, 1102-1110. doi:10.1093/scan/nsv080

Fowles, D.C. (2000). Electrodermal hyporeactivity and antisocial behavior: Does anxiety mediate the relationship? *Journal of Affective Disorders*, *61*, 177-189. doi:10.1016/S0165-0327(00)00336-0

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104* (41), 16311-16316. doi:10.1073/pnas.0706111104

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, *306* (5703), 1940-1943. doi:10.1126/science.1102941

Frick, P. J., & White, S. F. (2008). Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*, 359-375. doi:10.1111/j.1469-7610.2007.01862.x

Fridberg, D. J., Queller, S., Ahn, W.-Y., Kim, W., Bishara, A. J., Busemeyer, J. R., et al. (2010). Cognitive mechanisms underlying risky decision-making in chronic cannabis users. *Journal of Mathematical Psychology*, *54* (1), 28-38. doi:10.1016/j.jmp.2009.10.002

Frost, R., & McNaughton, N. (2017). The neural basis of delay discounting: A review and preliminary model. *Neuroscience and Biobehavioral Reviews, 79*, 48-65. doi:10.1016/j.neubiorev.2017.04.022

Fusar-Poli, P., Hijazi, Z., Stahl, D., & Steyerberg, E. W. (2018). The science of prognosis in psychiatry: A review. *JAMA Psychiatry*, *75*, 1289-1297. doi:10.1001/jamapsychiatry.2018.2530

Gao, Y., Raine, A., Venables, P. H., Dawson, M. E., & Mednick, S. A. (2010a). Reduced

electrodermal fear conditioning from ages 3 to 8 years is associated with aggressive behaviour at age 8 years. *Journal of Child Psychology and Psychiatry, 51*, 550-558. doi:10.1111/j.1469-7610.2009.02176.x

Gao, Y., Raine, A, Venables, P. H., Dawson, M. E., & Mednick, S. A. (2010b). Association of poor childhood fear conditioning and adult crime. *American Journal of Psychiatry, 167*, 56-60. doi:10.1176/appi.ajp.2009.09040499

Gatzke-Kopp, L. M., Beauchaine, T. P., Shannon, K. E., Chipman, J., Fleming, A. P., Crowell, S. E., . . . Aylward, E. (2009). Neurological correlates of reward responding in adolescents with and without externalizing behavior disorders. *Journal of Abnormal Psychology, 118*, 203-213. doi:10.1037/a0014378

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures. *Personality and Social Psychology Bulletin*, *43*(3), 300-312. doi:10.1177/0146167216684131

Gelman, A. (2006). Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *Technometrics*, *48*, 432-435. doi:10.1198/004017005000000661

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24* (6), 997-1016. doi:10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7* (4), 457-472. doi:10.2307/2246093

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22* (5), 1320-1327. doi:10.3758/s13423-014-0790-3

Gizer, I. R., Otto, J. M., & Ellingson, J. M. (2017). Molecular genetics of the externalizing spectrum. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *The Oxford handbook of externalizing spectrum disorders* (pp. 149-169). New York: Oxford University Press.

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. *Cerebral Cortex*, *19* (2), 483-495. doi:10.1093/cercor/bhn098

Gomez, P., Ratcliff, R., & Childers, R. (2015). Pointing, looking at, and pressing keys: A diffusion model account of response modality. *Psychological Assessment*, *41*, 1515-1523. doi:10.1037/a0039653

Gong, B., Naveed, S., Hafeez, D. M., Afzal, K. I., Majeed, S., Abele, J., et al. (2019). Neuroimaging in psychiatric disorders: A bibliometric analysis of the 100 most highly cited articles. *Journal of Neuroimaging*, *29*, 14-33. doi:10.1111/jon.12570

Grant, S., Contoreggi, C., & London, E. D. (2000). Drug abusers show impaired performance in a laboratory test of decision making. *Neuropsychologia*, *38* (8), 1180-1187. doi:10.1016/S0028-3932(99)00158-X

Gray, J.A. (1970). The psychophysiological basis of introversion-extraversion. *Behaviour Research and Therapy*, *8*, 249-266. doi:10.1016/0005-7967(70)90069-0

Gray, J.A. (1987). Perspectives on anxiety and impulsivity: A commentary. *Journal of Research in Personality*, *21*, 493-509. doi:10.1016/0092-6566(87)90036-5

Gray, J.A., & McNaughton, N. (2000). The neuropsychology of anxiety: An enquiry into

the functions of the septo-hippocampal system (2nd ed.). Oxford, UK: Oxford University Press.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin, 130,* 769-792. doi:10.1037/0033-2909.130.5.769

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Psychological Assessment*, *74*, 1464-1480. doi:10.1037/0022-3514.74.6.1464

Guendelman, M. D., Owens, E. B., Galán, C., Gard, A., & Hinshaw, S. P. (2015). Early-adult correlates of maltreatment in girls with attention-deficit/hyperactivity disorder: Increased risk for internalizing symptoms and suicidality. *Development and Psychopathology*, *28*, 1-14. doi:10.1017/S0954579414001485

Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *PsyArXiv Preprint*, 1-13. doi:10.31234/osf.io/rybh9

Haines, N., Beauchaine, T. P. (2020). Moving beyond ordinary factor analysis in studies of personality and personality disorder: A computational modeling perspective. *In press at Psychopathology*.

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., et al. (2020b). Anxiety Modulates Preference for Immediate Rewards Among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis. *Clinical Psychological Science*. doi:10.1177/2167702620929636

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., ...

Turner, B. (2020a, August 24). Learning from the Reliability Paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences. doi:10.31234/osf.io/xr7y3

Haines, N., Rass, O., Shin, Y. W., Busemeyer, J. R., Brown, J. W., O'Donnell, B. F., & Ahn, W. Y. (under review). Negative affect induces rapid formation of counterfactual representations: A model-based facial expression analysis approach.

Haines, N., Vassileva, J., & Ahn, W.-Y. (2018). The Outcome-Representation Learning Model: A novel reinforcement learning model of the Iowa Gambling Task. *Cognitive Science, 47*, 1-28. doi:10.1111/cogs.12688

Hampson, S.E. (2012). Personality processes: Mechanisms by which personality traits "Get outside the skin." *Annual Review of Psychology*, *63*, 315-339. doi:10.1146/annurev-psych-120710-100419

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. *Journal of Neuroscience*, *26* (32), 8360-8367. doi:10.1523/JNEUROSCI.1010-06.2006

Hand, D. J. (1996). Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society. Series a (Statistics in Society)*, *159*, 445-473. doi:10.2307/2983326

Hanson, J. L., Albert, D., Iselin, A.-M. R., Carré, J. M., Dodge, K. A., & Hariri, A. R. (2016). Cumulative stress in childhood is associated with blunted reward-related brain activity in adulthood. *Social Cognitive and Affective Neuroscience*, *11*, 405-412. doi:10.1093/scan/nsv124

Hauser, T. U., Will, G. J., Dubois, M., & Dolan, R. J. (2019). Annual research review: Developmental computational psychiatry. *Journal of Child Psychology and Psychiatry*, *60*, 412-426. doi:10.1111/jcpp.12964

Hays, R.D., Merz, J.F., & Nicholas, R. (1995). Response burden, reliability, and validity of the CAGE, Short MAST, and AUDIT alcohol screening measures. *Behavior Research Methods, Instruments, and Computers*, *27*, 277-280. doi:10.3758/BF03204745

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An Introduction to Good Practices in Cognitive Modeling. In *An Introduction to Model-Based Cognitive Neuroscience* (pp. 25-48). New York, NY: Springer, New York, NY. doi:10.1007/978-1-4939-2236-9_2

Heathcote, A., & Love, J. (2012). Linear Deterministic Accumulator Models of Simple Choice. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00292

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185-207. doi:10.3758/BF03212979

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

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *103*, 1-21. doi:10.3758/s13428-017-0935-1

Heil, S. H., Johnson, M. W., Higgins, S. T., & Bickel, W. K. (2006). Delay discounting in currently using and currently abstinent cocaine-dependent outpatients and non-drug-using matched controls. *Addictive Behaviors*, *31*, 1290-1294. doi:10.1016/j.addbeh.2005.09.005

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*, 135-175. doi:10.1086/286983

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, *15* (8), 534-539. doi:10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13* (12), 517-523. doi:10.1016/j.tics.2009.09.004

Hinshaw, S. P., Owens, E. B., Zalecki, C., Huggins, S. P., Montenegro-Nevado, A. J., Schrodek, E., & Swanson, E. N. (2012). Prospective follow-up of girls with ADHD into early adulthood: Continuing impairment includes elevated risk for suicide attempts and self-injury. *Journal of Consulting and Clinical Psychology, 80*, 1041-1051. doi:10.1037/a0029451

Ho, M.Y., Mobini, S., Chiang, T.J., Bradshaw, C.M., & Szabadi, E. (1999). Theory and method in the quantitative analysis of" impulsive choice" behaviour: implications for psychopharmacology. *Psychopharmacology*, *146*, 362-372. doi:10.1007/PL00005482

Huemer, J., Riegler, A., Völkl-Kernstock, S., Wascher, A., Lesch, O. M., Walter, H., & Skala, K. (2016). The influence of reported ADHD and substance abuse on suicidal ideation in a non-clinical sample of young men. *Neuropsychiatrie, 30*, 131-137.

doi:10.1007/s4021

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*, 404-413. doi:10.1038/nn.4238

Iacono, W. G., Malone, S. M., & McGue, M. (2008). Behavioral disinhibition and the development of early-onset addiction: common and specific influences. *Annual Review of Clinical Psychology*, *4*, 325-348. doi:10.1146/annurev.clinpsy.4.022007.141157

Jensen, P.S., Hinshaw, S.P., Kraemer, H.C., Lenora, N., Newcorn, J.H., Abikoff, H.B., ...Vitiello, B. (2001). ADHD comorbidity findings from the MTA Study: Comparing comorbid subgroups. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 147-158. doi:10.1097/00004583-200102000-00009

Jarecki, J. B., Tan, J. H., & Jenny, M. A. (2020). A framework for building cognitive process models. *Psychonomic Bulletin & Review*, *27*(6), 1218-1229. doi:10.3758/s13423-020-01747-2

Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Assessment*, *112*, 841-861. doi:10.1037/0033-295X.112.4.841

Johnson, M. W., Johnson, P. S., Herrmann, E. S., & Sweeney, M. M. (2015). Delay and probability discounting of sexual and monetary outcomes in individuals with cocaine use disorders and matched controls. *PLoS ONE*, *10*, e0128641. doi:10.1371/journal.pone.0128641

Jones, A. P., Laurens, K. R., Herba, C. J., & Viding, E. (2009). Amygdala hypoactivity to fearful faces in boys with conduct problems and callous-unemotional traits. *American Journal of Psychiatry, 166*, 95-102. doi:10.1176/appi.ajp.2008.07071050

Jones, K., Daley, D., Hutchings, J., Bywater, T., & Eames, C. (2007). Efficacy of the Incredible Years Basic parent training programme as an early intervention for children with conduct problems and ADHD. *Child: care, health and development*, *33*, 749-756. doi:10.1111/j.1365-2214.2007.00747.x

Jones, K., Daley, D., Hutchings, J., Bywater, T., & Eames, C. (2008). Efficacy of the Incredible Years Programme as an early intervention for children with conduct problems and ADHD: long-term follow-up. *Child: care, health and development*, *34*, 380-390. doi:10.1111/j.1365-2214.2008.00817.x

Jost, J. T. (2019). The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology:. *Current Directions in Psychological Science*, *28*, 10-19. doi:10.1177/0963721418797309

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-292. doi:10.2307/1914185

Kellen, D. (2019). A Model Hierarchy for Psychological Science. *Computational Brain & Behavior*, *2*(3-4), 160-165. doi:10.1007/s42113-019-00037-y

Kelly, C., de Zubicaray, G., Di Martino, A., Copland, D. A., Reiss, P. T., Klein, D. F., et al. (2009). l-Dopa modulates functional connectivity in striatal cognitive and motor networks: A double-blind placebo-controlled study. *The Journal of Neuroscience*, *29*, 7364-7378. doi:10.1523/JNEUROSCI.0810-09.2009

Kennedy, L., & Gelman, A. (2019, June 26). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *arXiv*.

Kirby, K. N. (2009). One-year temporal stability of delay-discount rates. *Psychonomic Bulletin & Review*, *16*, 457-462. doi:10.3758/PBR.16.3.457

Kjome, K. L., Lane, S. D., Schmitz, J. M., Green, C., Ma, L., Prasla, I., et al. (2010). Relationship between impulsivity and decision making in cocaine dependence. *Psychiatry Research*, *178* (2), 299-304. doi:10.1016/j.psychres.2009.11.024

Koff, E., & Lucas, M. (2011). Mood moderates the relationship between impulsiveness and delay discounting. *Personality and Individual Differences*, *50*, 1018-1022. doi:10.1016/j.paid.2011.01.016

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*: *Vol. 1*. New York, NY: Academic Press.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*, 921-926. doi:10.1001/archpsyc.56.10.921.

Krueger, R. F., Hicks, B. M., Patrick, C. J., Carlson, S. R., Iacono, W. G., & McGue, M. (2002). Etiologic connections among substance dependence, antisocial behavior and personality: Modeling the externalizing spectrum. *Journal of Abnormal Psychology, 111*, 411-424. doi:10.1037/0021-843X.111.3.411

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: a dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology*, *114*, 537-550.

doi:10.1037/0021-843X.114.4.537

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

Kvam, P. D. (2019). Modeling accuracy, response time, and bias in continuous orientation judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 301-318. doi:10.1037/xhp0000606

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Assessment*, *127*, 1053-1078. doi:10.1037/rev0000215

Lane, S. (2002). Marijuana Effects on Sensitivity to Reinforcement in Humans. *Neuropsychopharmacology*, *26* (4), 520-529. doi:10.1016/S0893-133X(01)00375-X

Lane, S. D., Cherek, D. R., Tcheremissine, O. V., Lieving, L. M., & Pietras, C. J. (2005). Acute Marijuana Effects on Human Risk Taking. *Neuropsychopharmacology*, *30* (4), 800-809. doi:10.1038/sj.npp.1300620

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55* (1), 1-7. doi:10.1016/j.jmp.2010.08.013

Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian Cognitive Modeling. Cambridge, UK: Cambridge University Press.

Lee, S., Burns, G. L., Beauchaine, T. P., & Becker, S. P. (2016). Bifactor latent structure of attention-deficit/hyperactivity disorder (ADHD)/oppositional defiant disorder (ODD) symptoms and first-order latent structure of sluggish cognitive

tempo symptoms. *Psychological Assessment*, *28*, 917-928. doi:10.1037/pas0000232

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75-84. doi:10.1037/1076-898X.8.2.75

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: a response time distributional approach. *Acta Psychologica*, *104*(2), 167-190. doi:10.1016/S0001-6918(00)00019-6

Lin, C.-H., Chiu, Y.-C., Lee, P.-L., & Hsieh, J.-C. (2007). Is deck B a disadvantageous deck in the Iowa Gambling Task? *Behavioral and Brain Functions*, *3* (1), 16. doi:10.1186/1744-9081-3-16

Loe, I. M., & Feldman, H. M. (2007). Academic and educational outcomes of children with ADHD. *Journal of Pediatric Psychology, 32*, 643-654. doi:10.1016/j.ambp.2006.05.005

Loeber, R., Keenan, K., & Zhang, Q. (1997). Boys' experimentation and persistence in developmental pathways toward serious delinquency. *Journal of Child and Family Studies, 6*, 321-357. doi:10.1023/A:1025004303603

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, *104* (22), 9493-9498. doi:10.1073/pnas.0608842104

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*,

355, 584-585. doi:10.1126/science.aal3618

Lombardo, L.E., Bearden, C.E., Barrett, J., Brumbaugh, M.S., Pittman, B., Frangou, S., & Glahn, D.C. (2012). Trait impulsivity as an endophenotype for bipolar I disorder. *Bipolar Disorders*, *14*, 565-570. doi:10.1111/j.1399-5618.2012.01035.x

Long, A.B., Kuhn, C.M., & Platt, M.L. (2009). Serotonin shapes risky decision making in monkeys. *Social Cognitive and Affective Neuroscience*, *4*, 346-356. doi:10.1093/scan/nsp020

Luckman, A., Donkin, C., & Newell, B.R. (2017). Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice. *Psychonomic Bulletin and Review*, *25*, 785-792. doi:10.3758/s13423-017-1330-8

Luijten, M., Schellekens, A. F., Kühn, S., Machielse, M. W. J., Sescousse, G. (2017). Disruption of reward processing in addiction. *JAMA Psychiatry, 74*, 387-398. doi:10.1001/jamapsychiatry.2016.3084

Luman, M., Tripp, G., & Scheres, A. (2010). Identifying the neurobiology of altered reinforcement sensitivity in ADHD: a review and research agenda. *Neuroscience & Biobehavioral Reviews*, *34*, 744-754. doi:10.1016/j.neubiorev.2009.11.021

Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B. U., Marsman, M., & Matzke, D. (2017). A Flexible and Efficient Hierarchical Bayesian Approach to the Exploration of Individual Differences in Cognitive-model-based Neuroscience. In A. A. Moustafa (Ed.), *Computational Models of Brain and Behavior* (pp. 467-480). Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9781119159193

Lynam, D. R., Caspi, A., Moffit, T. E., Wikström, P.-O., Loeber, R., & Novak, S. (2000). The interaction between impulsivity and neighborhood context on offending: The effects of impulsivity are stronger in poorer neighborhoods. *Journal of Abnormal Psychology, 109*, 563-574. doi:10.1037/0021-843x.109.4.563

Macdonald, A. N., Goines, K. B., Novacek, D. M., & Walker, E. F. (2016). Prefrontal mechanisms of comorbidity from a transdiagnostic and ontogenic perspective. *Development and Psychopathology*, *28*, 1147-1175. doi:10.1017/S0954579416000742

Macoveanu, J., Rowe, J.B., Hornboll, B., Elliott, R., Paulson, O.B., Knudsen, G.M., & Siebner, H.R. (2013). Serotonin 2A receptors contribute to the regulation of risk-averse decisions. *Neuroimage*, *83*, 35-44. doi:10.1016/j.neuroimage.2013.06.063

Manassis, K., Tannock, R., & Barbosa, J. (2000). Dichotic listening and response inhibition in children with comorbid anxiety disorders and ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *39*, 1152-1159. doi:10.1097/00004583-200009000-00015

Maner, J.K., Richey, J.A., Cromer, K., Mallott, M., Lejuez, C.W., Joiner, T.E., & Schmidt, N.B. (2007). Dispositional anxiety and risk-avoidant decision-making. *Personality and Individual Differences*, *42*, 665-675. doi:10.1016/j.paid.2006.08.016

Marsh, A. A., Finger, E. C., Mitchell, D. G. V., Reid, M., Sims, C., Kosson, D. S., et al. (2008). Reduced amygdala response to fearful expressions in children and adolescents with callous-unemotional traits and disruptive behaviour disorders. *American Journal of Psychiatry, 165*, 712-720. doi:10.1176/appi.ajp.2007.07071145

Marshal, M. P., & Molina, B. S. (2006). Antisocial behaviors moderate the deviant peer pathway to substance use in children with ADHD. *Journal of Clinical Child & Adolescent Psychology*, *35*, 216-226. doi:10.1207/s15374424jccp3502_5

Marshal, M. P., Molina, B. S. G., & Pelham, W. E. (2003). Childhood ADHD and adolescent substance use: an examination of deviant peer group affiliation as a risk factor. *Psychology of Addictive Behavior, 17*, 293-302. doi:10.1037/0893-164X.17.4.293

Martel, M. M., Levinson, C. A., Lee, C. A., & Smith, T. E. (2017). Impulsivity symptoms as core to the developmental externalizing spectrum. *Journal of Abnormal Child Psychology, 45*, 83-90. doi:10.1007/s10802-016-0148-6

Martel, M. M., Nigg, J. T., & Von Eye, A. (2009). How do trait dimensions map onto ADHD symptom domains? *Journal of abnormal child psychology*, *37*, 337. doi:10.1007/s10802-008-9255-3

Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1-23. doi:10.3758/s13428-011-0124-6

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*. doi:10.3389/fnhum.2014.00825

Mazur, J. E. (1987). An adjustment procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior* (Vol. 5, pp. 55-73). Hillsdale, NJ: Lawrence Erlbaum Associates.

McCloskey, M.S., New, A.S., Siever, L.J., Goodman, M., Koenigsberg, H.W., Flory,

J.D., & Coccaro, E.F. (2009). Evaluation of behavioral impulsivity and aggression tasks as endophenotypes for borderline personality disorder. *Journal of Psychiatric Research*, *43*, 1036-1048. doi:10.1016/j.jpsychires.2009.01.002

McFall, R. M., & Townsend, J. T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment*, *10* (4), 316-330. doi:10.1037/1040-3590.10.4.316

McGloin, J.M., & O'Neill Shermer, L. (2008). Self-control and deviant peer network structure. *Journal of Research in Crime and Delinquency*, *46*, 35-72. doi:10.1177/0022427808326585

McKerchar, T. L., & Renda, C. R. (2012). Delay and Probability Discounting in Humans: An Overview. *The Psychological Record*, *62*, 817-834. doi:10.1007/BF03395837

Meier, M. H., Slutske, W. S., Arndt, S., & Cadoret, R. J. (2008). Impulsive and callous traits are more strongly associated with delinquent behavior in higher risk neighborhoods among boys and girls. *Journal of Abnormal Psychology, 117*, 377-385. doi:10.1037/0021-843X.117.2.377

Mihatsch, O., & Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, *49* (2-3), 267-290. doi:10.1023/A:1017940631555

Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research & Perspective*, *6*(1-2), 7-24. doi:10.1080/15366360802035489

Miletić, S., Turner, B. M., Forstmann, B. U., & Van Maanen, L. (2017). Parameter

recovery for the Leaky Competing Accumulator model. *Journal of Mathematical Psychology*, *76*, 25-50. doi:10.1016/j.jmp.2016.12.001

Milich, R., Balentine, A. C., & Lynam, D. R. (2001). ADHD combined type and ADHD predominantly inattentive type are distinct and unrelated disorders. *Clinical Psychology: Science and Practice*, *8*, 463-488. doi:10.1093/clipsy.8.4.463

Mitchell, S. H. (1999). Measures of impulsivity in cigarette smokers and non-smokers. *Psychopharmacology*, *146*, 455-464. doi:10.1007/pl00005491

Mitchell, J. M., Fields, H. L., D'Esposito, M., & Boettiger, C. A. (2006). Impulsive responding in alcoholics. *Alcoholism, Clinical and Experimental Research*, *29*, 2158-2169. doi:10.1097/01.alc.0000191755.63639.4a

Mochcovitch, M. D., da Rocha Freire, R. C., Garcia, R. F., & Nardi, A. E. (2014). A systematic review of fMRI studies in generalized anxiety disorder: Evaluating its neural and cognitive basis. *Journal of Affective Disorders*, *167*, 336-342. doi:10.1016/j.jad.2014.06.041

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701. doi:10.1037/0033-295X.100.4.674

Molina, B. S. G., Pelham, W. E., Jr., Cheong, J., Marshal, M. P., Gnagy, E. M., & Curran, P. J. (2012). Childhood attention-deficit/hyperactivity disorder (ADHD) and growth in adolescent alcohol use: The roles of functional impairments, ADHD symptom persistence, and parental knowledge. *Journal of Abnormal Psychology, 121*, 922-935. doi:10.1037/a0028260

Montague, R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72-80. doi:10.1016/j.tics.2011.11.018

Moran, P. (1999). The epidemiology of antisocial personality disorder. *Social Psychiatry and Psychiatric Epidemiology*, *34*, 231-242. doi:10.1007/s001270050138

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221-229. doi:10.1038/s41562-018-0522-1

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3-4), 53-67. doi:10.1016/j.jmp.2013.05.005

Naneix, F., Marchand, A.R., Di Scala, G., Pape, J.-R., & Coutureau, E. (2012). Parallel maturation of goal-directed behavior and dopaminergic systems during adolescence. *The Journal of Neuroscience, 32*, 16223-16232. doi:10.1523/JNEUROSCI.3080-12.2012

Navarro, D. J. (2020, March 16). If mathematical psychology did not exist we would need to invent it: A case study in cumulative theoretical development. *PsyArXiv*. doi:10.31234/osf.io/ygbjp

Neufeld, R. W. J., Vollick, D., Carter, J. R., Boksman, K., & Jetté, J. (2002). Application of stochastic modeling to the assessment of group and individual differences in cognitive functioning. *Psychological Assessment*, *14* (3), 279-298. doi:10.1037/1040-3590.14.3.279

Neuhaus, E., & Beauchaine, T.P. (2017). Impulsivity and vulnerability to psychopathology. In T.P. Beauchaine & S.P. Hinshaw (Eds.), *Child and adolescent*

psychopathology (pp. 178-212). Hoboken, NJ: Wiley.

Nigg, J. T., Blaskey, L. G., Huang-Pollock, C. L., & Rappley, M. D. (2002). Neuropsychological executive functions and DSM-IV ADHD subtypes. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*, 59-66. doi:10.1097/00004583-200201000-00012

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, *203*, 116157. doi:10.1016/j.neuroimage.2019.116157

Ntamati, N. R., Creed, M., Achargui, R., & Lüscher, C. (2018). Periaqueductal efferents to dopamine and GABA neurons of the VTA. *PLoS ONE, 13*, e0190297. doi:10.1371/journal.pone.0190297

Oas, P. (1985). The psychological assessment of impulsivity: A review. *Journal of Psychoeducational Assessment*, *3*, 141-156. doi:10.1177/073428298500300205

Oberlin, B. G., Dzemidzic, M., Bragulat, V., Lehigh, C. A., Talavage, T., O'Connor, S. J., & Kareken, D. A. (2012). Limbic responses to reward cues correlate with antisocial trait density in heavy drinkers. *NeuroImage, 60*, 644-652. doi:10.1016/j.neuroimage.2011.12.043

Ohmura, Y., Takahashi, T., & Kitamura, N. (2005). Discounting delayed and probabilistic monetary gains and losses by smokers of cigarettes. *Psychopharmacology*, *182*, 508-515. doi:10.1007/s00213-005-0110-8

Ortiz, J., & Raine, A. (2004). Heart rate level and antisocial behavior in children and adolescents: A meta-analysis. *Journal of the American Academy of Child and*

Adolescent Psychiatry, 43, 154-162. doi:10.1097/00004583-200402000-00010

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *Scientific American*, *25*, 221-247. doi:10.2307/2648877

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*(13), 20-48. doi:10.1016/j.jmp.2018.03.003

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21* (6), 425-433. doi:10.1016/j.tics.2017.03.011

Pang, B., Blanco, N. J., Maddox, W. T., & Worthy, D. A. (2016). To not settle for small losses: evidence for an ecological aspiration level of zero in dynamic decision-making. *Psychonomic Bulletin & Review*, *24* (2), 536-546. doi:10.3758/s13423-016-1080-z

Pardini, D., White, H. R., & Stouthamer-Loeber, M. (2007). Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood. *Drug and Alcohol Dependence, 88*, S38-S49. doi:10.1016/j.drugalcdep.2006.12.014

Parsons, S. (2020). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *PsyArXiv Preprint*, 1-35. doi:10.31234/osf.io/y6tcz

Patrick, C.J., Bernat, E.M., Malone, S.M., Iacono, W.G., Krueger, R.F., & Mcgue, M.K. (2006). P300 amplitude as an indicator of externalizing in adolescent males. *Psychophysiology, 43*, 84-92. doi:10.1111/j.1469-8986.2006.00376.x

Patros, C. H. G., Alderson, M., Kasper, L. J., Tarle, S. J., Lea, S. E., & Hudec, K. L. (2016) Choice-impulsivity in children and adolescents with ADHD: A meta-analytic review. *Clinical Psychology Review, 43*, 162-174. doi:10.1016/j.cpr.2015.11.001

Patterson, G.R., Degarmo, D.S., & Knutson, N. (2000). Hyperactive and antisocial behaviors: Comorbid or two points in the same process? *Development and Psychopathology*, *12*, 91-106. doi:10.1017/S0954579400001061

Patton, J.H., Stanford, M.S., & Barratt, E.S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*, 768-774. doi:10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. *Psychological Inquiry*, *28*, 233-248. doi:10.1080/1047840X.2017.1335568

Pearce, J.M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532-552. doi:10.1037/0033-295X.87.6.532

Peters, J., & Büchel, C. (2011). The neural mechanisms of inter-temporal decision-making: understanding variability. *Trends in Cognitive Sciences*, *15*, 227-239. doi:10.1016/j.tics.2011.03.002

Petry, N.M. (2001). Delay discounting of money and alcohol in actively using alcoholics, currently abstinent alcoholics, and controls. *Psychopharmacology*, *154*, 243-250. doi:10.1007/s002130000638

Petry, N. M. (2002). Discounting of delayed rewards in substance abusers: relationship

to antisocial personality disorder. *Psychopharmacology, 162*, 425-432. doi:10.1007/s00213-002-1115-1

Pfefferbaum, A., Kwon, D., Brumback, T., Thompson, W. K., Cummins, K., Tapert, S. F., ... Sullivan, E. V. (2018). Altered Brain Developmental Trajectories in Adolescents After Initiating Drinking. *American Journal of Psychiatry, 175*, 370-380. doi:10.1176/appi.ajp.2017.17040469

Plichta, M. M., & Scheres, A. (2014). Ventral-striatal responsiveness during reward anticipation in ADHD and its relation to trait impulsivity in the healthy population: A meta-analytic review of the fMRI literature. *Neuroscience and Biobehavioral Reviews, 38*, 125-134. doi:10.1016/j.neubiorev.2013.07.012

Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology*, *32*, 3-25. doi:10.1080/00335558008248231

Posthumus, J. A., Böcker, K. B. E., Raaijmakers, M. A. J., van Engeland, H., & Matthys, W. (2009). Heart rate and skin conductance in 4-year old children with aggressive behavior. *Biological Psychology, 82*, 164-168. doi:10.1016/j.biopsycho.2009.07.003

Quay, H. C. (1965). Psychopathic personality as pathological stimulation-seeking. *American Journal of Psychiatry*, *122*, 180-183. doi:10.1176/ajp.122.2.180

Quay, H. C. (1997). Inhibition and Attention Deficit Hyperactivity Disorder. *Journal of Abnormal Child Psychology*, *25*, 7-13. doi:10.1023/A:1025799122529

Rangel, A., Camerer, C. F., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews*

Neuroscience, *9*, 545-556. doi:10.1038/nrn2357

Ratcliff, R., Spieler, D., & Mckoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin & Review*, *7* (1), 1-25. doi:10.3758/BF03210723

Regenwetter, M., & Robinson, M. M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*(5), 533-550. doi:10.1037/rev0000067

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp.64-99). New York: Appleton Century Crofts.

Reynolds, B., Richards, J. B., Horn, K., & Karraker, K. (2004). Delay discounting and probability discounting as related to cigarette smoking status in adults. *Behavioural Processes*, *65*, 35-42. doi:10.1016/S0376-6357(03)00109-8

Rivers, A. M., Rees, H. R., Calanchini, J., & Sherman, J. W. (2017). Implicit Bias Reflects the Personal and the Social. *Psychological Inquiry*, *28*, 301-305. doi:10.1080/1047840X.2017.1373549

Robbins, T.W., Gillan, C.M., Smith, D.G., de Wit, S., & Ersche, K.D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, *16*, 81-91. doi:10.1016/j.tics.2011.11.009

Robins, L. N. (1966). *Deviant children grown up.* Baltimore, MD: Williams & Wilkins.

Roese, N. J., & Summerville, A. (2005). What we regret most... and why. *Personality and Social Psychology Bulletin*, *31* (9), 1273-1285. doi:10.1177/0146167205274693

Rogers, R., Everitt, B. J., Baldacchino, A., Blackshaw, A. J., Swainson, R., Wynne, K., et al. (1999). Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: Evidence for monoaminergic mechanisms. *Neuropsychopharmacology*, *20*, 322-339. doi:10.1016/S0893-133X(98)00091-8

Romeu, R. J., Haines, N., Ahn, W.-Y., Busemeyer, J. R., & Vassileva, J. (2019). A computational model of the Cambridge gambling task with applications to substance use disorders. *Drug and Alcohol Dependence*, 107711. doi:10.1016/j.drugalcdep.2019.107711

Ross, C. T., Winterhalder, B., & McElreath, R. (2020). Racial Disparities in Police Use of Deadly Force Against Unarmed Individuals Persist After Appropriately Benchmarking Shooting Data on Violent Crime Rates:. *Social Psychological and Personality Science*, *16*, 194855062091607. doi:10.1177/1948550620916071

Rotello, C. M., Heit, E., & Dubé, C. (2014). When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944-954. doi:10.3758/s13423-014-0759-2

Rouder, J.N., & Haaf, J.M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*, 452-467. doi:10.3758/s13423-018-1558-y

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12* (4), 573-604. doi:10.3758/BF03196750

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The Lognormal Race: A Cognitive-Process Model of Choice and Latency with Desirable Psychometric Properties. *Psychometrika*, *80*(2), 491-513. doi:10.1007/s11336-013-9396-3

Ruland, T. (2012). *The "hidden" costs of ADHD*. neaToday, Jan. 9. neatoday:2012/01/09/the-hid- den-costs-of-adhd/

Rung, J.M., & Madden, G.J. (2018). Experimental reductions of delay discounting and impulsive choice: A systematic review and meta-analysis. *Journal of Experimental Psychology: General*, *147*, 1349-1381. doi:10.1037/xge0000462

Sagvolden, T., Aase, H., Zeiner, P., & Berger, D. (1998). Altered reinforcement mechanisms in attention-deficit/hyperactivity disorder. *Behavioural Brain Research*, *94*, 61-71. doi:10.1016/S0166-4328(97)00170-8

Sagvolden, T., Johansen, E. B., Aase, H., & Russell, V. A. (2005). A dynamic developmental theory of attention-deficit/hyperactivity disorder (ADHD) predominantly hyperactive/impulsive and combined subtypes. *Behavioral and Brain Sciences, 28*, 397-468. doi:10.1017/S0140525X05000075

Sameroff, A. (2010). A unified theory of development: A dialectic integration of nature and nurture. *Child development*, *81*, 6-22. doi:10.1111/j.1467-8624.2009.01378.x

Sasser, T. R., Kalvin, C. B., & Bierman, K. L. (2016). Developmental trajectories of

clinically significant attention-deficit/hyperactivity disorder (ADHD) symptoms from grade 3 through 12 in a high-risk sample: Predictors and outcomes. *Journal of Abnormal Psychology*, *125*, 207-219. doi:10.1037/abn0000112

Sauder, C.L., Beauchaine, T.P., Gatzke-Kopp, L.M., Shannon, K.E., & Aylward, E. (2012). Neuroanatomical correlates of heterotypic comorbidity in externalizing male adolescents. *Journal of Clinical Child and Adolescent Psychology*, *41*, 346-352. doi:10.1080/15374416.2012.658612

Sauder, C. L., Derbidge, C. M., & Beauchaine, T. P. (2016). Neural responses to monetary incentives among self-injuring adolescent girls. *Development and Psychopathology, 28*, 277-291. doi:10.1017/S0954579415000449

Sauder, C. L., Hajcak, G., Angstadt, M., & Phan, K. L. (2013). Test-retest reliability of amygdala response to emotional faces. *Psychophysiology, 50*, 1147-1156. doi:10.1111/psyp.12129

Schatz, D.B., & Rostain, A.L. (2006). ADHD with comorbid anxiety. *Journal of Attention Disorders*, *10*, 141-149. doi:10.1177/1087054706286698

Scheres, A., Dijkstra, M., Ainslie, E., Balkan, J., Reynolds, B., Sonuga-Barke, E., & Castellanos, F. X. (2006). Temporal and probabilistic discounting of rewards in children and adolescents: Effects of age and ADHD symptoms. *Neuropsychologia*, *44*, 2092-2103. doi:10.1016/j.neuropsychologia.2005.10.012

Scheres, A., Tontsch, C., & Thoeny, A. L. (2013). Steep temporal reward discounting in ADHD-Combined type: Acting upon feelings. *Psychiatry Research*, *209*, 207-213. doi:10.1016/j.psychres.2012.12.007

Scheres, A., Tontsch, C., Thoeny, A. L., & Kaczkurkin, A. (2010). Temporal reward discounting in Attention-Deficit/Hyperactivity Disorder: The contribution of symptom domains, reward magnitude, and session length. *Biological Psychiatry*, *67*, 641-648. doi:10.1016/j.biopsych.2009.10.033

Schimmack, U. (2019). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, *117*, 174569161986379. doi:10.1177/1745691619863798

Schmitt, J., Warner, K., & Gupta, S. (2010). *The high budgetary cost of incarceration*. Center for Economic Policy Research. Retrieved November 6, 2018, from www.cepr.net/documents/publications/incarcera- tion-2010-06.pdf

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S.C., Yamawaki, S., & Doya, K. (2008). Low serotonin levels increase delayed reward discounting in humans. *Journal of Neuroscience*, *28*, 4528-4532. doi:10.1523/JNEUROSCI.4982-07.2008

Scott, S., Briskman, J., and O'Connor, T.G. (2014). Early prevention of antisocial personality: Long-term follow-up of two randomized controlled trials comparing indicated and selective approaches. *American Journal of Psychiatry, 171*, 649-657. doi:10.1176/appi.ajp.2014.13050697

Senner, N. R., Conklin, J. R., & Piersma, T. (2015). An ontogenetic perspective on individual differences. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20151050. doi:10.1098/rspb.2015.1050

Shannon, K. E., Sauder, C., Beauchaine, T. P., & Gatzke-Kopp, L. M. (2009).

Disrupted effective connectivity between the medial frontal cortex and the caudate in adolescent boys with externalizing behavior disorders. *Criminal Justice and Behavior, 36*, 1141-1157. doi:10.1177/0093854809342856

Shapiro, D.N., Chandler, J., & Mueller,P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*, 213-220. doi:10.1177/2167702612469015

Sharma, L., Markon, K.E., & Clark, L.A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*, 374-408. doi:10.1037/a0034418

Shaw, P., Malek, M., Watson, B., Sharp, W., Evans, A., & Greenstein, D. (2012). Development of cortical surface area and gyrification in attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *72*, 191-197. doi:10.1016/j.biopsych.2012.01.031

Sher, K. J., & Trull, T. J. (2002). Substance use disorder and personality disorder. *Current Psychiatry Reports*, *4*, 25-29. doi:10.1007/s11920-002-0008-7

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32* (8), 1248-1284. doi:10.1080/03640210802414826

Shin, S.H., Cook, A.K., Morris, N.A., McDougle, R., & Groves, L.P. (2016). Different faces of impulsivity as links between childhood maltreatment and young adult crime. *Preventive Medicine*, *88*, 210-217. doi:10.1016/j.ypmed.2016.03.022

Shurman, B., Horan, W. P., & Nuechterlein, 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. doi:10.1016/j.schres.2004.03.020

Silva, C., & McNaughton, N. (2019). Are periaqueductal grey and dorsal raphe the foundation of appetitive and aversive control? A comprehensive review. *Progress in Neurobiology, 177*, 33-72. doi:10.1016/j.pneurobio.2019.02.001

Skinner, H.A. (1982). The drug abuse screening test. *Addictive Behaviors*, *7*, 363-371. doi:10.1016/0306-4603(82)90005-3

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin & Review*, *4*, 1-22. doi:10.3758/s13423-018-1446-5

Spielberger, C.D. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Sprafkin, J., Gadow, K. D., Weiss, M. D., Schneider, J., & Nolan, E. E. (2007). Psychiatric comorbidity in ADHD symptom subtypes in clinic and community adults. *Journal of Attention Disorders*, *11*, 114-124. doi:10.1177/1087054707299402

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*, 385-395. doi:10.1016/j.paid.2009.04.008

Steingroever, H., Fridberg, D., Horstmann, A., Kjome, K., Kumari, V., Lane, S. D., et al. (2015). Data from 617 healthy participants performing the Iowa Gambling Task: A "Many Labs" collaboration. *Journal of Open Psychology Data*, *3* (1), 7.

doi:10.5334/jopd.ak

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013a). A comparison of reinforcement learning models for the Iowa Gambling Task using parameter space partitioning. *The Journal of Problem Solving*, *5* (2). doi:10.7771/1932-6246.1150

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *1* (3), 161-183. doi:10.1037/dec0000005

Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., & Wagenmakers, E.-J. (2013b). Performance of healthy participants on the Iowa Gambling Task. *Psychological Assessment*, *25* (1), 180-193. doi:10.1037/a0029929

Sterzer, P., Stadler, C., Krebs, A., Kleinschmidt, A., & Poustka, F. (2005). Abnormal neural responses to emotional stimuli in adolescents with conduct disorder. *Biological Psychiatry, 57*, 7-15. doi:10.1016/j.biopsych.2004.10.008

Storebø, O. J., & Simonsen, E. (2016). The association between ADHD and antisocial personality disorder (ASPD). *Journal of Attention Disorders, 20*, 815-824. doi:10.1177/1087054713512150

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. doi:10.1017/S1355617701711095

Stroop, R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *18*, 643-662. doi:10.1037/h0054651

Suppes, P. (1966). Models of Data. *Studies in Logic and the Foundations of*

Mathematics, *44*, 252-261. doi:10.1016/S0049-237X(09)70592-0

Szollosi, A., & Donkin, C. (2019). Neglected Sources of Flexibility in Psychological Theories: from Replicability to Good Explanations. *Computational Brain & Behavior*, *2*(3-4), 190-192. doi:10.1007/s42113-019-00045-y

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual review of clinical psychology*, *15*, 579-604. doi:10.1146/annurev-clinpsy-050718-095710

Takahashi, T., Ohmura, Y., Oono, H., Radford, M. (2009). Alcohol use and discounting of delayed and probabilistic gain and loss. *Neuroendocrinology Letters*, *30*, 749-752.

Tanaka, S.C., Shishida, K., Schweighofer, N., Okamoto, Y., Yamawaki, S., & Doya, K. (2009). Serotonin affects association of aversive outcomes to past actions. *Journal of Neuroscience*, *29*, 15669-15674. doi:10.1523/JNEUROSCI.2799-09.2009

Tarter, R., Vanyukov, M., Giancola, P., Dawes, M., Blackson, T., Mizzich, A., & Clark, D. B. (1999). Etiology of early age onset substance use disorder: A maturational perspective. *Development and Psychopathology*, *11*, 657-683. doi:10.1017/S0954579499002266

Teles-Grilo Ruivo, L. M., & Mellor, J. R. (2013). Cholinergic modulation of hippocampal network function. *Frontiers in Synaptic Neuroscience, 5*. doi:10.3389/fnsyn.2013.00002

Teplin, L. A. (1994). Psychiatric and substance abuse disorders among male urban jail detainees. *American Journal of Public Health, 84*, 290-293. doi:10.2105/ajph.84.2.290

Thompson, R. A. (1990). Emotion and self-regulation. In R. A. Thompson (Ed.),

Nebraska *Symposium on Motivation: Vol. 36. Socioemotional development* (pp. 383-483). Lincoln: University of Nebraska Press.

Tomasi, D., & Volkow, N. D. (2014). Functional connectivity of substantia nigra and ventral tegmental area: Maturation during adolescence and effects of ADHD. *Cerebral Cortex*, *24*, 935-944. doi:10.1093/cercor/bhs382

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of Mathematical Psychology*, *52*, 269-280. doi:10.1016/j.jmp.2008.05.001

Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, *13* (4), 549-565. doi:10.1037/1040-3590.13.4.549

Tremblay, R. E., Pihl, R. O., Vitaro, F., & Dobkin, P. L. (1994). Predicting early onset of male antisocial behavior from preschool behavior. *Archives of General Psychiatry, 51*, 732-739. doi:10.1001/archpsyc.1994.03950090064009

Tripp, G., & Alsop, B. (2001). Sensitivity to reward delay in children with attention deficit hyperactivity disorder (ADHD). *Journal of Child Psychiatry, 42*, 691-698. doi:10.1017/S0021963001007430

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65-79. doi:10.1016/j.jmp.2016.01.001

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B.,

& Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage, 72*, 193-206. doi:10.1016/j.neuroimage.2013.01.048

Turner, B. M., Palestro, J. J., Miletić, S., & Forstmann, B. U. (2019). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews, 102*, 327-336. doi:10.1016/j.neubiorev.2019.04.018

Turner, B. M., Rodriguez, C. A., Liu, Q., Molloy, M. F., Hoogendijk, M., & McClure, S. M. (2018). On the neural and mechanistic bases of self-control. *Cerebral Cortex*, *29*, 732-750. doi:10.1093/cercor/bhx355

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *Neuroimage, 128*, 96-115. doi:10.1016/j.neu- roimage.2015.12.030

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(3), 329-362. doi:10.1037/rev0000089

Turner, B.M., Van Maanen, L., & Forstmann, B.U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*, 312-336. doi:10.1037/a0038894

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69-85. doi:10.1016/j.jmp.2012.02.005

Tuvblad, C., Zheng, M., Raine, A., & Baker, L. A. (2009). A common genetic factor explains the covariation among ADHD, ODD, and CD symptoms in 9-10-year-old

boys and girls. *Journal of Abnormal Child Psychology, 37*, 153-167. doi:10.1007/s10802-008-9278-9

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323. doi:10.1007/BF00122574

van Bokhoven, I, Matthys, W., van Goozen, S. H. M., & van Engeland, H. (2005). Prediction of adolescent outcome in children with disruptive behaviour disorders: A study of neurobiological, psychological and family factors. *European Child and Adolescent Psychiatry, 14*, 153-163. doi:10.1007/s00787-005-0455-x

van Rooij, I., & Baggio, G. (2020, February 28). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *PsyArXiv*. doi:10.31234/osf.io/7qbpr

van Rooij, I., & Blokpoel, M. (2020). Formalizing Verbal Theories. *Social Psychology*, *51*, 285-298. doi:10.1027/1864-9335/a000428

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58-71. doi:10.1016/j.jmp.2014.06.004

Vassileva, J., Ahn, W.-Y., Weber, K. M., Busemeyer, J. R., Stout, J. C., Gonzalez, R., & Cohen, M. H. (2013). Computational modeling reveals distinct effects of HIV and history of drug use on decision-making processes in women. *PLoS ONE*, *8* (8), e68962. doi:10.1371/journal.pone.0068962

Vassileva, J., & Conrod, P.J. (2019). Impulsivities and addictions: a multidimensional

integrative framework informing assessment and interventions for substance use disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*, 20180137. doi:10.1098/rstb.2018.0137

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27 (5), 1413-1432. doi:10.1007/s11222-016-9696-4

Volkow, N. D., & Morales, M. (2015). The brain on drugs: From reward to addiction. *Cell*, *162*, 712-725. doi:10.1016/j.cell.2015.07.046

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

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830-841. doi:10.1037/0033-295X.114.3.830

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

Walker, J.L., Lahey, B.B., Russo, M.F., Frick, P.J., Christ, M.A.G., McBurnett, K., ...Green, S.M. (1991). Anxiety, inhibition, and conduct disorder in children: I. Relations to social impairment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *30*, 187-191. doi:10.1097/00004583-199103000-00004

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review*, *112* (4),

862-880. doi:10.1037/0033-295X.112.4.862

Webster-Stratton, C. H., Reid, M. J., & Beauchaine, T. P. (2011). Combining parent and child training for young children with ADHD. *Journal of Clinical Child & Adolescent Psychology*, *40*, 191-203, doi:10.1080/15374416.2011.546044

Webster-Stratton, C. H., Reid, M. J., & Beauchaine, T. P. (2013). One-year follow-up of combined parent and child intervention for young children with ADHD. *Journal of Clinical Child and Adolescent Psychology, 42*, 251-261. doi:10.1080/15374416.2012.723263

Wennerhold, L., & Friese, M. (2020). Why self-report measures of self-control and inhibition tasks do not substantially correlate. *Collabra: Psychology*, *6*, 9. doi:10.1525/collabra.276

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, *54* (1), 14-27. doi:10.1016/j.jmp.2008.12.001

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*, 669-689. doi:10.1016/S0191-8869(00)00064-7

Whitlow, C. T., Liguori, A., Brooke Livengood, L., Hart, S. L., Mussat-Whitlow, B. J., Lamborn, C. M., et al. (2004). Long-term heavy marijuana users make costly decisions on a gambling task. *Drug and Alcohol Dependence*, *76* (1), 107-111. doi:10.1016/j.drugalcdep.2004.04.009

Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry: Clustering and Classification. *Clinical Psychological Science*, *3*(3), 378-399. doi:10.1177/2167702614565359

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 1-10. doi:10.3389/fninf.2013.00014

Williams, D. R., Martin, S. R., DeBolt, M., Oakes, L., & Rast, P. (2020). A fine-tooth comb for measurement reliability: Predicting true score and error variance in hierarchical models. *PsyArXiv Preprint*. doi:10.31234/osf.io/2ux7t

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, *26*, 74-89. doi:10.1037/met0000270

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, 558. doi:10.7554/eLife.49547

Wilson, V.B., Mitchell, S.H., Musser, E.D., Schmitt, C.F., & Nigg, J.T. (2010). Delay discounting of reward in ADHD: Application in young children. *Journal of Child Psychology and Psychiatry*, *52*, 256-264. doi:10.1111/j.1469-7610.2010.02347.x

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013a). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20* (2), 364-371. doi:10.3758/s13423-012-0324-9

Worthy, D. A., Pang, B., & Byrne, K. A. (2013b). Decomposing the roles of

perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, 4:640. doi:10.3389/fpsyg.2013.00640

Wright, A. G. C., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: Toward an expanded quantitative empirical model. *Journal of Abnormal Psychology, 122*, 281-294. doi:10.1037/a0030133

Xia, L., Gu, R., Zhang, D., & Luo, Y. (2017). Anxious individuals are impulsive decision-makers in the delay discounting task: An ERP study. *Frontiers in Behavioral Neuroscience*, *11*, 1-11. doi:10.3389/fnbeh.2017.00005

Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). ADOpy: a python package for adaptive design optimization. *Behavior Research Methods*, 1-24. doi:10.3758/s13428-020-01386-4

Yarkoni, T. (2019, November). The Generalizability Crisis. *PsyArXiv Preprint*. doi:10.31234/osf.io/jqw35

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. doi:10.1111/j.1467-9280.2005.01646.x

Yechiam, E., & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology*, *51* (2), 75-84. doi:10.1016/j.jmp.2006.11.002

Yi, R., Chase, W. D., & Bickel, W. K. (2007). Probability discounting among cigarette

smokers and nonsmokers: molecular analysis discerns group differences. *Behavioural Pharmacology*, *18*, 633-639. doi:10.1097/FBP.0b013e3282effbd3

Yudko, E., Lozhkina, O., & Fouts, A. (2007). A comprehensive review of the psychometric properties of the Drug Abuse Screening Test. *Journal of Substance Abuse Treatment*, *32*, 189-198. doi:10.1016/j.jsat.2006.08.002

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review, 38*, 55-68. doi:10.1016/j.dr.2015.07.001

Zhao, J., Cheng, J., Harris, M., & Vigo, R. (2015). Anxiety and intertemporal decision making: The effect of the behavioral inhibition system and the moderation effects of trait anxiety on both state anxiety and socioeconomic status. *Personality and Individual Differences*, *87*, 236-241. doi:10.1016/j.paid.2015.08.018

Zinbarg, R., & Revelle, W. (1989). Personality and conditioning: A test of four models. *Journal of Personality and Social Psychology*, *57*, 301-314. doi:10.1037/0022-3514.57.2.301

Zisner, A., Beauchaine, T. P. (2016a). Midbrain neural mechanisms of trait impulsivity. In T. P. Beauchaine & S. P. Hinshaw (Eds), *The Oxford handbook of externalizing spectrum disorders* (184-200). New York, NY: Oxford University Press.

Zisner, A., Beauchaine, T. P. (2016b). Neural substrates of trait impulsivity, anhedonia, and irritability: Mechanisms of heterotypic comorbidity between externalizing disorders and unipolar depression. *Development and Psychopathology, 28*, 1179-1210. doi:10.1017/S0954579416000754