Computational Models Describe Individual Differences in Cognitive Function and Their Relationships to Mental Health Symptoms

Anahita Talwar

Prepared under the supervision of:

Professor Jonathan P. Roiser, Professor Quentin J. M. Huys and Dr Francesca Cormack

Submitted for the degree of Doctor of Philosophy

Institute of Cognitive Neuroscience

University College London

August 2022

Declaration

I, Anahita Talwar, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This thesis is dedicated to my late father, Dr Sanjiv Talwar.
He was there at the beginning, but sadly not the end of this journey. I know he would have been both proud and amused to know that I've followed exactly in his footsteps by taking on a new job and struggling to finish writing up my PhD at the same time.

I will always appreciate how hard you worked to give us a life full of opportunities and experiences, and that you showed us how to have a zest for life, make the most of every moment, and cherish the time spent with those closest to you.

Abstract

Cognitive alterations have long been reported in patients with mental health disorders, though with inconsistent results. These inconsistencies are likely due to highly heterogeneous diagnostic categories used for recruitment, and imprecise cognitive task measures. This thesis addresses the former by measuring symptoms with continuous questionnaire scales, and the latter by using theory-driven computational models that summarise participant behaviour using a small number of mechanistic parameters. This methodology is applied within the realm of attention set shifting and risky decision making to improve understanding of cognition in mental health, using large samples collected online. Following a general introduction (Chapter 1), Chapter 2 describes the computational approach employed in subsequent experimental chapters. In Chapter 3, we develop models of CANTAB IED (Intra-Extra Dimensional Set Shifting Task) to explore how learning and attention processes lead to differences in attention set shifting ability, and to investigate their relationship with symptoms of compulsivity. The second study (Chapter 4) applies the computational approach to risky decisions with CANTAB CGT (Cambridge Gamble Task) and explores the relationship between model parameters and symptoms of depression and anxiety. The final experimental chapter (Chapter 5) examines whether specific symptoms of anxiety are related to changes in risky decision making, focusing on the relationship between catastrophising and probability weighting. Overall, the computational approach offers increased precision when examining behavioural data. In several chapters we identify moderate relationships between model parameters and demographic variables such as age, gender, and level of education, which often exceed associations with traditional model-agnostic measures. However, relationships with mental health symptoms are minimal in the general population datasets tested here. The general discussion (Chapter 6) considers these findings in relation to the wider field of computational psychiatry, discussing both the limitations of the work presented and possible future directions.

Impact Statement

Mental health disorders include some of the most debilitating and prevalent conditions, affecting one in four people every year (Ginn & Horder, 2012). Despite this, the mechanisms underlying mental health disorders remain poorly understood, which limits the efficiency of treatment selection in clinical practice. In recent years, computational psychiatry has gained popularity as an approach to provide a more mechanistic understanding of mental health disorders that will ultimately lead to clinical improvements. This thesis makes several contributions to this field that are outlined below.

In Chapter 3, we validate a novel model of attention set shifting in a large general population sample. This model provides a mechanistic explanation for individual variation in set shifting ability as measured by CANTAB IED, in terms of the interaction between attention and learning processes. Examining the model parameters also provides a more precise explanation for the relationship between symptoms of compulsivity and difficulties in performing set shifts. In Chapter 4, we validate a novel model of risky decision making as assessed by CANTAB CGT that describes how individual differences in risk and loss aversion lead to individual differences in overall task behaviour, and are associated with demographic variables. We also report a noteworthy null finding with respect to the relationship between model parameters and mental health symptoms. Both of these chapters apply computational models to tasks from the widely used CANTAB task battery, and therefore offer a novel analysis approach for researchers using these tasks. In addition, due to the use of these tasks in clinical trials, these models offer potential opportunities to explore the impact of pharmacological compounds on model parameters, providing further mechanistic insights in the realm of mental health and its treatment. Chapter 5 attempts to clarify the relationship between risky decision making and symptoms of anxiety. We identify the importance of subjective probability weighting to individual differences in how risky decisions are made and report another noteworthy null finding with respect to mental health symptoms. These findings are preliminary and require replication, but nevertheless offer contributions to the field of computational psychiatry.

Acknowledgements

Thank you…

to my supervisors Professor Jonathan Roiser and Professor Quentin Huys-Gavric,
for giving me the opportunity to study and learn, for their ever-perceptive insights, for their constant support and guidance, for creating an incredible work environment, and for providing the best possible examples of great leadership.

to my colleagues Dr Alexandra Pike, Dr Vincent Valton, and Professor Oliver Robinson,
for always being kind, patient and ready to lend a helping hand.

to the friends I made along the way Jolanda, Lara, Maddy, Millie, Alex and Anahit,
for teaching me to take all difficulties in my stride, for always being up for a café debrief or a Franco Manca, and for all of the lockdown dance routines.

to my mum,
for exemplifying grace and strength in the most testing of times.

to my brother,
for putting up with the return of your annoying sister.

to my husband,
for being by my side through all of the highs and lows.

List of Abbreviations

ART	Abstract Reasoning Task
BIS-11	Barratt Impulsiveness Scale
CANTAB	Cambridge Neuropsychological Test Automated Battery
CAT	Catastrophising Questionnaire
CI	Confidence Interval
CGT	Cambridge Gamble Task
DSM	Diagnostic and Statistical Manual of Mental Disorders
ED	Extra Dimensional
EM	Expectation Maximisation
GAD	Generalised Anxiety Disorder
GAD-7	Generalised Anxiety Disorder 7-item Scale
HiToP	Hierarchical Taxonomy of Psychopathology
iBIC	Integrated Bayesian Information Criterion
ICD	International Classification of Diseases
ID	Intra Dimensional
IED	Intra-Extra Dimensional Set Shifting Task
MDD	Major Depressive Disorder
NIMH	National Institute of Mental Health
OCD	Obsessive-Compulsive Disorder
OCI-R	Obsessive Compulsive Inventory-Revised
PHQ	Patient Health Questionnaire Depression Scale
PSWQ	Penn State Worry Questionnaire
RDoC	Research Domain Criteria
SD	Standard Deviation
SRDS	Self-Rating Depression Scale
SSMS	Short Scales for Measuring Schizotypy
STAI-S	State Trait Anxiety Inventory – State
STAI-T	State Trait Anxiety Inventory – Trait
TEPS	Temporal Experience of Pleasure Scale

List of Tables and Figures

Contents

Notes to examiners

All findings from Chapter 3 have previously been published in a pre-print: Talwar A., Cormack F., Huys Q. J. M., Roiser J. P. (2021). A Hierarchical Reinforcement Learning Model Explains Individual Differences in Attention set shifting. bioRxiv 2021.10.05.463165; doi: https://doi.org/10.1101/2021.10.05.463165

All findings from Chapter 4 have previously been published in a pre-print: Talwar A., Cormack F., Huys Q. J. M., Roiser J. P. (2022). Individual Variation in Risky Decisions Is Related to Age and Gender but not to Mental Health Symptoms. bioRxiv 2022.07.11.499611; doi: https://doi.org/10.1101/2022.07.11.499611

The work reported in this thesis is entirely my own except for the contributions acknowledged below. My supervisor, Prof Jonathan Roiser, has contributed comments and guidance on all chapters of the thesis. In Chapters 3 & 4, the data was collected by Cambridge Cognition Ltd, who partly sponsored my MRC iCASE PhD. In Chapter 5, the original dataset was collected by Emily Bagley, an intern in our research group.

1 General Introduction

1.1 Categorical Diagnoses in Psychiatry

In current clinical practice, diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5, American Psychiatric Association, 2013) and International Classification of Diseases, Tenth Edition (ICD-10, World Health Organisation, 2016) operationalise a psychiatric nosology which classifies patients into diagnoses based on specific symptom criteria. For instance, to be diagnosed with major depressive disorder (MDD), according to the DSM-5, the patient must exhibit symptoms of dysphoria ('low or depressed mood') or anhedonia ('markedly diminished interest or pleasure in all, or almost all, activities') most of the day, nearly every day, and at least four (or three if both dysphoria and anhedonia are present) of the following symptoms: weight change, sleep pattern change, psychomotor change, fatigue or loss of energy, feelings of worthlessness or guilt, indecisiveness or diminished ability to think/concentrate, and suicide-related thoughts/behaviours. This conceptualisation of mental health conditions aimed to align psychiatric disorders with a medical model (akin to physical diseases), whereby discrete diagnoses are established based on specific observable symptoms. The relevant cluster of symptoms suggests a common underlying mechanistic basis for the disorder and is therefore of potential clinical value through guiding differential diagnosis and targeted treatment accordingly.

The differential efficacy of some medications for certain psychiatric disorders provided some affirmation for this model of psychopathology and for a categorical conceptualisation of mental health. For instance, the successful use of antidepressant medications for MDD (Cipriani et al., 2018), and the efficacy of lithium in patients with bipolar disorder (Baldessarini et al., 2006; Severus et al., 2014), but not schizophrenia (Bender & Dittmann-Balcar, 2004; Leucht et al., 2015), lent support to the idea that these diagnostic categories had some biological underpinning. Another key advantage of specifying diagnoses in this way that was particularly crucial to the design of the DSM-III (an earlier edition of the current DSM manual) is that it standardised psychiatry practice and improved consistency in diagnosis and treatment across clinicians (Spitzer et al., 1980). These outcomes somewhat corroborated the medical model framework, with the specified diagnostic categories solidifying their application not only in clinical practice and decision making, but also in scientific research where participants were recruited according to diagnostic criteria. The DSM in particular, having been translated into over twenty languages became *the* reference for psychiatric practice in the US, much of Europe and more recently in some Asian countries. As the gold-standard guideline, it had a notable influence on clinical practice, research, and society.

Despite their impact and clinical utility, these diagnostic nosologies have come under increasing scrutiny, largely due to growing scepticism that the symptomatically defined disorders clearly demarcate normality from disease, or one disorder from another (Fried, 2015; Hyman, 2010; Jablensky, 2016; Kendell & Jablensky, 2003). Whilst the impact of pharmacological therapeutics, particularly antidepressants, in psychiatry has been noteworthy, an important finding was that only about 30% of patients with major depression achieved symptom remission in response to first-line antidepressants in a large-scale trial of typical community patients, STAR*D (Rush et al., 2003). The remission rates reduced with each subsequent antidepressant treatment attempt, and around 40% of patients were non-remitters after attempting a range of different antidepressant drugs (Rush et al., 2003). Crucially, we have very little explanatory or predictive power regarding who will respond to antidepressant treatment, let alone to which one of the many available. Results such as these are widespread in psychiatry and cast doubt on the idea that DSM categories represent truly distinct natural entities (Cipriani et al., 2009; Loerinc et al., 2015; Miura et al., 2014; Seppälä et al., 2021).

In recent years, substantial attention has been drawn to the vast heterogeneity within psychiatric diagnostic groups to explain the high variation in treatment response. From a brief look at the aforementioned DSM-5 criteria for depression, for example, it is clear that many distinct combinations of symptoms can qualify for a diagnosis of MDD, and as a result the clinical presentation of patients with a particular diagnosis is rather diverse (Fried, 2017; Fried & Nesse, 2015). This clinical heterogeneity may be underpinned by mechanistic heterogeneity where patients with different mechanistic alterations may express similar symptoms; and conversely, closely related biological alterations may result in distinct symptom patterns in different patients. In addition to heterogeneity, there is a fair amount of overlap in the non-core diagnostic criteria between some disorders. For example, fatigue, difficulty concentrating, and difficulty sleeping are all DSM symptoms for both MDD and generalised anxiety disorder. Perhaps at least in part due to this imprecise nosology, but also likely due to common mechanistic pathways, a high level of comorbidity has been reported between psychiatric disorders, most notably between depressive and anxiety disorders (Kessler, 1994; Kessler et al., 2015). This degree of comorbidity is especially troubling when considering that most research studies recruit participants that have a single diagnosis only (or allow only limited comorbidities), and specifically exclude those with most psychiatric comorbidities. The result of this is that many research studies prohibit participants that represent the majority of psychiatric patients in real-world practice. Given the heterogeneity, and comorbidity in psychiatric diagnoses, is not surprising that in recent studies the reliability of diagnosing patients with some DSM-5 disorders including MDD and generalised anxiety disorder (GAD) was reported as being in the questionable range (intraclass kappa reliability of 0.2 – 0.39) (Regier et al., 2013). This is noteworthy as consistent practice between

clinicians was an important motivation for the development of this classification framework in the first place (Spitzer et al., 1980).

In summary, whilst the current categorical psychiatric framework has some utility in clinical practice and treatment, there is also little choice but to apply it until a superior alternative is proposed. However, in the realm of scientific research, clinging on to a framework rooted in a categorical and symptom-based nosology akin to diagnosis of discrete physical illness, which has little grounding in behavioural and neuroscientific data, will likely hamper progress in understanding mental health disorders. This has led to a momentum to turn to consider other models of psychopathology that might better capture its complexity.

1.2 'Symptomics' and Dimensional Psychiatry

The high heterogeneity, comorbidity, and lack of a mechanistic basis underlying the categorical diagnostic framework has motivated researchers to explore what a more appropriate representation of mental health disorders might look like (Hyman, 2010; Jablensky, 2016). This has led to breaking down the current diagnostic groups into their constituent symptoms and examining the latter more carefully, using a dimensional perspective. Ultimately, the aim is to reach a better understanding of the complex presentation and relationships between different symptoms of mental health disorders, and build back up to a novel organisation of psychopathology that overcomes the aforementioned limitations. This approach involves different kinds of studies, statistics, and analysis. Instead of summing the presence or absence of specific symptoms in the diagnostic criteria to allocate a binary label (e.g. depressed or not) to a person, symptom-based analysis involves questionnaires that measure the degree to which a person expresses a particular symptom (e.g. anhedonia) on a continuous scale (Mason et al., 2005; Meyer et al., 1990; Patton et al., 1995). Measuring symptoms of mental health disorders on a continuum more clearly acknowledges that most of these are present in the general population to some extent (Johns, 2005). Finally, by modelling the interactions between symptoms, we can identify those that tend to naturally cluster together and thus to some extent tackle heterogeneity and comorbidity, while also providing a better mechanistic understanding of psychopathology. Larger-scale datasets combined with data-driven approaches have ushered in an era of 'symptomics'.

These tools have been used to directly probe the composition of the current diagnostic framework, for instance by investigating which psychological symptoms are most connected to others and are therefore considered the most central in the development of certain syndromes or disorders. For example, Fried and colleagues used such an approach to investigate which symptoms of depression are most pivotal in depressive processes in a sample of over 3000 outpatients (Fried et al., 2016). They

found that some of the DSM symptoms (e.g. sad mood), but also some of the non-DSM symptoms (e.g. anxiety) were among the most central symptoms (Fried et al., 2016). While dysphoria is a core symptom of the criteria and therefore consistent with these results, anxiety does not appear on the diagnostic criteria for MDD *at all* suggesting once again that demarcation into discrete diagnoses does not reflect the high degree of overlap present in the real world. These results demonstrate further that the DSM criteria for MDD (and other disorders) could be better refined from such analyses of its component symptoms.

Further to the reorganisation of the current framework, researchers have turned to using symptoms as dimensional measurements of psychopathology in their own right (Jablensky, 2016; Kotov et al., 2017; Robbins et al., 2012; Widiger & Samuel, 2005). For instance, studies are increasingly examining associations between specific depressive symptoms and adverse life events (Keller et al., 2007), risk factors (Fried et al., 2014), or neurobiological markers (Fried et al., 2020), as opposed to the traditional approach of investigating associations with categorical diagnoses. In alcohol and substance use disorders, as well as problem gambling and eating disorders, the recognition of impulsivity and compulsivity as distinct components has also been encouraged, with a call to focus on transdiagnostic symptoms that may result in the development of transdiagnostic treatments (Robbins et al., 2012). Another approach has been the creation of new dimensions by analysing the shared variance across different symptoms using methods such as factor analysis. Korszun and colleagues used this method in a sample of over 1000 depressed patients who reported on 26 depression symptom items. They identified a four-factor solution suggesting the following dimensions best capture the variance in depressed patients: 'Mood Symptoms and Psychomotor Retardation', 'Anxiety', 'Psychomotor Agitation, Guilt, and Suicidality'; and 'Appetite Gain and Hypersomnia' (Korszun et al., 2004). Gillan and colleagues expanded this approach to include a greater variety of symptom types, asking participants in the general population to complete nine symptom questionnaires assessing compulsivity, depression, trait anxiety, alcohol use, eating attitudes, apathy, impulsivity, schizotypy, and social anxiety (Gillan et al., 2016). They reported the solution that best fit their data comprised three factors, which they labelled 'Anxious-Depression', 'Compulsive Behaviour and Intrusive Thought' and 'Social Withdrawal' (note that these different factors may result from differences in symptom covariance within clinical groups vs largely non-clinical samples recruited from the general population). This shift in focus from diagnoses to symptoms represents a promising start to understanding the interplay and overlap between diagnostic criteria.

At a more abstract level, the symptom-based approach reformulates the notion of psychiatric disorders as a complex web of causally connected symptoms with intricate dynamical patterns (Borsboom & Cramer, 2013; Boschloo et al., 2015). Whole new frameworks have been conceptualised

to account for this perspective, such as the Hierarchical Taxonomy of Psychopathology (HiToP), which has the aim of developing an empirically based organisation of mental disorders using multivariate factor analytic methods (Kotov et al., 2017). It posits a layered framework with symptoms, syndromes, factors and spectra making up the four levels of the hierarchy. Whilst the realisation of novel, comprehensive frameworks in clinical practice may seem far off, the shift in conceptualisation has encouraged a corresponding shift in the methodology and practice of scientific research, where researchers measure symptoms along a continuous scale rather than recruiting patients from one of the discrete diagnoses (Kendler et al., 2011; Robbins et al., 2012).

1.3 Cognition in Psychiatry

Cognitive neuroscience has long been used to investigate the mechanisms underlying mental health disorders. Whilst self-reported symptoms are easily influenced by recall bias and differences in personal experience, objectively measured cognition is thought to provide a more abstract and unbiased representation of how an individual acts and performs psychological operations under various situations. Cognitive neuroscientists use carefully designed tasks to assess specific processes, and they hold huge potential as cheap, rapid and easily administered measures of potential cognitive biomarkers. Many behavioural measures, such as those related to attention and decision making, have been reported to be associated with mental health symptoms (Paulus & Yu, 2012; Wilson et al., 2018). The importance of cognition in mental health is further emphasised by its inclusion in the Research Domain Criteria (RDoC; Cuthbert & Insel, 2013) framework which was introduced a decade ago by the National Institution of Mental Health (NIMH) whose disappointment in the biological validity of the current diagnostic framework led them to develop a system for linking biological markers at different levels to different mental health symptoms. However, perhaps unsurprisingly, until now no clear biomarker has been found that can differentiate between diagnostic groups or between patients and non-patients for a specific disorder (Berggren & Derakshan, 2013). This failure has been blamed, at least in part, on the aforementioned highly heterogeneous categories specified in diagnostic manuals and the resulting use of these groups in case-control studies of cognition. Further, a variety of task adaptations are used in cognitive research which might lead to slight differences in behaviour and inconsistent results between studies. Finally, the typical outcome measures used are coarse, and do not consider the full complexity of the data available, which may also lead to lack of precision and inconsistent results especially in small samples. Below I will highlight some of these inconsistencies in the research using the examples of attention set shifting and risky decision making literature, particularly using the Cambridge Neuropsychological Test Automated Battery (CANTAB) Intra-Extra Dimensional Set Shifting Task (IED), and Cambridge Gamble Task (CGT), which are the focus of two of the experimental chapters in this thesis.

1.3.1 Attention set shifting

Attention set shifting refers to learning and switching the focus of our attention. The ease with which we can generalise pre-learnt action-outcome associations, and the flexibility with which we can grasp previously irrelevant ones, depends on how and what we learn to attend to. These cognitive faculties known as attentional set formation, and attention set shifting, respectively, have commonly been assessed with the CANTAB IED (Owen et al., 1991). The task, originally designed as a computerised analogue of the Wisconsin Card Sort Task (Berg, 2010; Grant & Berg, 1948), consists of several stages requiring participants to use trial and error to learn which feature of a multidimensional stimulus signals the correct response for that stage of the task. The introduction of novel stimuli occurs twice, first to assess set formation by the reapplication of previously learned rules, and second to assess the crucial attentional set shift by switching attentional focus to the previously irrelevant stimulus dimension.

The IED has been used extensively to document cognitive impairments in patients with psychiatric diagnoses. Early research identified difficulties in performing attentional set shifts but not simpler reversals, as measured by increased errors, in patients such as those with frontal lobe excisions, suggesting that attentional set shifts represent a distinct and complex higher-level process (Owen et al., 1991). Difficulties in set shifting have also been considered a cognitive hallmark of obsessive-compulsive disorder (OCD) (Chamberlain et al., 2006, 2007; Purcell et al., 1998; Vaghi et al., 2017; Veale et al., 1996) exemplified by the use of CANTAB IED as an endpoint in clinical trials (Tyagi et al., 2019). Despite this, there have been conflicting reports regarding the nature of these difficulties (Gottwald et al., 2018). Furthermore, difficulties in attention set shifting have been reported in depression (Purcell et al., 1998; Purcell et al., 1997), schizophrenia (Elliott et al., 1995; Levaux et al., 2007; Liang et al., 2018) and anxiety (Kim et al., 2019), highlighting that these alterations in cognition are not specific to OCD patients.

As mentioned above, one reason for these differing results between studies is that traditional measures of task performance are blunt. Task performance on CANTAB IED is typically measured in terms of the errors made per stage, which provides an overview of how an individual has performed but is unable to provide a mechanistic account of this performance. More specific to this task, the multidimensional nature of CANTAB IED task stimuli means that we cannot easily interpret the reasons for participants' stimulus choices, and therefore what leads to variations in set shifting performance. Counting the errors made per stage is coarse - it does not consider the choices participants make at the trial-by-trial level and therefore what they learnt or attended to as the task progressed. Without a mechanistic account of task performance, it is not clear whether all participants failing the extra-

dimensional set shift stage are doing so for the same reasons, or whether various behavioural alterations can explain this performance.

1.3.2 Risky Decision Making

Mental illnesses are often characterised by differences in decision making, particularly in situations that involve maximising expected rewards or minimising punishments (Cáceda et al., 2014). This has been examined extensively, with different disorders such as depression, obsessive-compulsive disorder, and psychosis-related disorders showing some influence on various aspects of decision making (Deserno et al., 2016; Halahakoon et al., 2020; Pratt et al., 2021; Sachdev & Malhi, 2005). Gambling tasks have often been used to examine such differences in decision making, requiring participants to choose between options with uncertain payoffs. These tasks are useful as they simulate the risky decisions that we often face in daily life. Early studies suggested that such tasks are sensitive to brain damage (specifically to the ventromedial prefrontal cortex, involved in fear and planning - Bechara et al., 1994). Within the realm of mental health research, prior studies using gambling tasks have suggested that individuals with depressive and anxious disorders have heightened aversion to risks and losses (Baek et al., 2017; Charpentier et al., 2017; Smoski et al., 2008), and that patients with schizophrenia make more random choices leading to lower winnings overall (Pedersen et al., 2017; Woodrow et al., 2019).

One particularly informative set of studies has used the CGT from the CANTAB task battery, which requires participants to bet a proportion of their points on a simple decision. This set of studies is noteworthy because a large number of participants with a variety of diagnoses have performed the task (Ackerman et al., 2015; Deakin et al., 2004; Hutton et al., 2002; Rogers et al., 1999; Rubinsztein et al., 2001). The CGT was originally designed to remove some of the learning confounds present in previously popular gambling tasks, such as the Iowa gambling task (Bechara et al., 1994), and presents participants with explicit information about the values and probabilities of gambles (Rogers et al., 1999). One of the most consistent findings with the CGT in mental health research is that, relative to controls, depressed individuals choose to bet fewer points overall, particularly when the probability of winning is high (Mannie et al., 2015; Murphy et al., 2001; Rawal et al., 2013). This has usually been interpreted as reflecting a conservative, or risk-averse, decision making strategy. Studies observing this pattern have included diverse groups including adult patients with depression (Murphy et al., 2001), young people with a family history of depression but who had not been diagnosed themselves (Mannie et al., 2015), and adolescents with depression (Rawal et al., 2013). However, this was not observed in a smaller study including patients with bipolar depression (Rubinsztein et al., 2006), and the opposite pattern was found in a study that included adolescents with recent first episode

depression, with patients betting more overall than controls (Kyte et al., 2005). In addition to the small sample sizes in these studies, and the different groups examined, another possible explanation for these inconsistent results is that the dependent variables typically examined are multifactorial. For instance, the 'overall proportion bet' measure depends on the proportion of points that a participant chose to bet on trials over the entire task, and disregards the specific aspects of different trials, such as the probability of winning or the stake. This challenges the interpretation of results as the key underlying mechanistic processes cannot be examined.

1.4 Computational Models of Cognition

Computational models of cognitive tasks have gained popularity in recent years due to the solutions they offer for addressing a number of current issues in mental health research: their ability to account for trial-by-trial behaviours, dissect traditional measures into more precise components, and make concrete predictions about participants' choices at an individual level (Adams et al., 2016; Montague et al., 2012).

Traditional measures of overall task performance do not take into account patterns of trial-by-trial choices or the underlying mechanisms that are thought to produce them, and are therefore limited in the insight they can provide. Computational models overcome this by providing theory-driven explanations for how choices are generated on each trial. These models can then be tested against data from real participants to assess which model provides the most parsimonious account of the data. Further, these models can interrogate the vast heterogeneity present in mental health disorders by modelling behaviour at the individual level, rather than assessing group characteristics. Each participant's performance is captured by a small number of mechanistically relevant model parameters measured on a continuous scale. Finally, computational models are able to account for the level of randomness present in human performance on cognitive tasks, and by accounting for this, models can add precision. This is of huge value due to the high level of noise in human behaviour and samples. A major theme of this thesis is exploring the additional insights that models of cognitive tasks are able to provide.

1.5 Summary of Chapters

The overall aim of this thesis is to investigate the computational mechanisms of cognitive functioning and its relationship to mental health symptoms. The second chapter will outline the general computational methodology applied in all experimental chapters including model development, parameter estimation, recovery, and model comparison. The three experimental chapters focus on applying models to different cognitive tasks, specifically attention set shifting and risky decision

making tasks, and explore their relationships to demographic variables as well as symptoms of compulsivity, depression and catastrophising.

1.5.1 Chapter 2: Online Data Collection and Computational Analysis

This chapter outlines computational modelling methodology common to the three experimental chapters. The general model development procedure will be explained, along with the expectation-maximisation iterative algorithm that was used to estimate model parameters under a Bayesian hierarchical framework. Methods for parameter recovery, a key technique for assessing model validity, are described as well as qualitative and quantitative metrics for comparing model performance. Qualitative assessments of model quality include posterior predictive checks comparing model-simulated data with real data on typical task outcome measures, whilst quantitative checks include calculation of the integrated Bayesian Information Criterion (iBIC).

1.5.2 Chapter 3: Modelling CANTAB IED

This chapter focuses on attention set shifting, how the focus of attention is directed and switched. Cognitive tasks such as CANTAB IED reveal great variation in set shifting ability in the general population, with notable impairments in those with psychiatric diagnoses. The attentional and learning processes underlying this cognitive ability, and how they lead to the observed variation remain unknown. To directly examine this question, this chapter used a computational modelling approach on two independent large-scale (N > 700), general-population samples, tested online, performing the CANTAB IED, with one sample including additional psychiatric symptom assessment. This data showcases a new methodology to analyse data from the CANTAB IED task and suggests a possible mechanistic explanation for the variation in set shifting performance, and its relationship to compulsive symptoms.

1.5.3 Chapter 4: Modelling CANTAB CGT

Risky decisions involve choosing between options in which the outcomes are not certain. Cognitive tasks such as CANTAB CGT have revealed differences in risky decisions in patients with depression, but the mechanisms of choice evaluation underlying these cognitive decisions, and how they lead to the observed differences in depressed patients remain unknown. To directly test this, this chapter uses a computational modelling approach on a large-scale (N = 753), general-population sample, tested online, performing CANTAB CGT and completing additional psychiatric symptom assessment, including depression scales. We fit five different models, including two novel ones, inspired by Prospect Theory. This study showcases a new methodology to analyse data from the CANTAB CGT task, and the advantages that computation can offer in cognitive neuroscience.

1.5.4 Chapter 5: Probability Weighting in Catastrophising

Previous research has suggested that anxiety is associated with differences in risky decision making, but it remains unclear which specific facets or types of anxiety are most associated with these differences. Further, many existing studies of risky decisions have not been able to disentangle all of the components of these decisions, and in particular tend to neglect 'probability weighting' – how people's subjective weighting of probability differs from the true probability. The hypothesis motivating this study was that this component is highly relevant to catastrophising symptoms in particular. This was tested using a computational modelling approach in a broad general-population sample tested online (N = 212), who performed a novel gambling task and completed questionnaires assessing psychiatric symptoms, including catastrophising. This study highlights the importance of incorporating probability weighting parameters into studies of risky decision making.

2 Computational Modelling Methodology

This chapter outlines aspects of the methodology that are common to all experimental chapters. Where methods differ between experimental chapters, they will be introduced and detailed within each specific chapter.

2.1 Ethical Approval

All participants were presented with an online information sheet and subsequently provided informed consent online. They could leave the study at any time by closing their browser. All participants were paid at a rate of £7.50 per hour. All studies were approved by the University College London Research Ethics Committee (approval number 5253/001).

2.2 Model Development

All models were written in Python 3 (van Rossum, 1995). Initial models were based on previously published models from the literature; however, model development was an iterative process in which analysis of the parsimony of these models in explaining participant data, and identification of discrepancies between simulated and real data, were used to guide development of the next model.

2.3 Parameter Estimation

We used a hierarchical Bayesian parameter estimation approach, described previously (Huys et al., 2011), which finds the maximum a posteriori parameter estimates for each participant, given the model and the data, and sets the parameters of the prior distribution to the maximum likelihood estimates given all participants' data. The purpose of using hierarchical estimation here is primarily that priors over the parameters act to regularise the estimates such that unrealistic, extreme values are avoided. We used an expectation-maximisation (EM) approach which repeatedly iterates over two steps until convergence is reached. Briefly, in the E-step the model finds the best-fitting individual level parameter estimates for each participant given their data and the current parameters of the prior distribution; and in the M-step, the maximum likelihood group level prior parameters are updated to reflect the current individual parameter estimates.

All parameters were transformed to ensure that they were in the appropriate range (such as learning rates between 0 and 1). Recoverability of parameters was calculated by simulating 300 datasets with parameters drawn randomly from the estimated prior distribution. The best-fitting parameters for these data sets were found as above, and parameter recoverability is indicated by the correlation between the simulated and recovered parameters. Un-transformed parameter values were used for statistical inference as these are estimated to be distributed with a standard multivariate Gaussian distribution and therefore are more suitable for parametric analysis.

2.4 Model Comparison

A priori we did not assume that any of the models should be more likely. Therefore, they can be compared by examining the approximate model log likelihood with the iBIC (Huys et al., 2011). This procedure gives the model that fits the data most parsimoniously, whilst penalising for unnecessary added model complexity (additional parameters).

We carried out posterior predictive checks (i.e. qualitative model comparisons) in which we assessed model performance by comparing model-simulated data, using participants' estimated parameters, to real data. This involved comparing overall group level performance patterns, as well as correlating individual participant summary statistics with corresponding values calculated from simulated data. For the latter, ten simulated datasets were run for each participant and correlations with real data were calculated for each iteration. The mean correlation is reported as a metric of model fit.

3 A Hierarchical Reinforcement Learning Model Explains Individual Differences in Attention set shifting

3.1 Abstract

Attention set shifting refers to the ease with which the focus of attention is directed and switched. Cognitive tasks such as the widely used CANTAB IED reveal great variation in set shifting ability in the general population, with notable impairments in those with psychiatric diagnoses. The attentional and learning processes underlying this cognitive ability, and how they lead to the observed variation remain unknown. To directly test this, we used a generative computational modelling approach on two independent large-scale online general-population samples performing CANTAB IED (N > 700), with one including additional assessment of demographic variables and mental health problems. We found a hierarchical model that learnt both feature values and dimension attention best explained the data, and that compulsive symptoms were associated with slower learning and higher attentional bias to the first relevant stimulus dimension. Further, older people, those that spent less time in education and women showed more attentional bias to the first relevant dimension. These results establish a new model of cognitive processes underlying the CANTAB IED task, and suggest a possible mechanistic explanation for the variation in set shifting performance and its relationship to compulsive symptoms.

3.2 Introduction

It has long been suggested that the increased difficulty in performing extra-dimensional (ED) shifts compared to intra-dimensional (ID) shifts is due to attentional biases in learning (le Pelley et al., 2016; Mackintosh & Little, 2013; Trabasso et al., 1966), but these hypotheses have not been specified mathematically or formally tested. As there are a number of ways in which attention and learning might interact to produce the observed patterns of responses on a particular task, the mathematical specification of models and formal comparisons to test how well they fit the data are essential to provide a better understanding of these cognitive constructs. Many theoretical models of attention and learning have previously been formulated including Kruschke's connection models ALCOVE and EXIT (Kruschke, 1992, 2001). Whilst these models attempt to fully characterise the psychological processes behind human performance in a variety of tasks, they are also very complex and possibly over-parameterised (Paskewitz & Jones, 2020). Further, these models are theoretical accounts of typical human performance, but their parsimony in explaining variation in individual performance has not been examined; therefore, they are unable to provide mechanistic or normative explanations for the large individual differences observed in attention and learning tasks. On the other hand, building generative models that are fit to individuals' trial-by-trial choices can enhance the explanatory power, and provide a more rigorous test of the specified model. Furthermore, as these models estimate the

level of randomness in participants' choices, they can provide more precise participant-specific measures in the form of a small number of interpretable model parameters. Though increasingly popular in mental health research, few generative models of attention set shifting have been built.

The theory-driven field of computational psychiatry involves developing models by mathematically specifying our hypotheses of the cognitive processes involved in performing the task that best describe variation in symptoms between participants. Reinforcement learning models, in which agents use feedback to learn actions that maximise their total reward (Sutton & Barto, 1998), have been extensively applied to cognitive tasks where learning is involved, including those assessing cognitive flexibility (Daw et al., 2011). For example, Niv and colleagues developed the "dimensions task" to explore how selective attention aids learning about complex stimuli (Niv et al., 2015). It is similar to CANTAB IED in that participants are shown multidimensional stimuli and have to use feedback to infer the 'correct' feature on each block of the task. The authors showed that models of reinforcement learning on stimulus features fit participants' choices the best, as compared to reinforcement learning on full stimuli, or non-reinforcement learning models (Niv et al., 2015), and subsequently that attention biased both learning and decision making on this task (although attention was measured by eye-tracking so how attention itself might be learnt was not mathematically specified: (Leong et al., 2017)). Despite the similarities between the individual stages of these tasks, the different ordering of the stages means that the tasks are measuring distinct attentional process. In the dimensions task, the correct feature on each stage is randomly chosen, independent of other stages, whereas in IED, the 'correct' feature is from the same dimension for the first seven stages of the task, which strongly promotes the formation of an attentional set. Additionally, it has been suggested that attentional set shifts rely on transfer to novel exemplars, which do not truly exist in the dimensions task, meaning that intra- and extra-dimensional set shifts cannot be well-defined (Slamencka, 1968). The clearer separation of these processes in IED allows for a more straightforward interpretation of attentional set formation and cognitive flexibility (Downes et al., 1989). Furthermore, the dimensions task has not been used extensively in clinical populations, highlighting the utility of developing models for the CANTAB task.

Only two studies using CANTAB IED have implemented reinforcement learning models. The first of these only focused on the simple discrimination and reversal learning stages at the beginning of the task (Murray et al., 2008). The authors did not fit computational models to trial-by-trial data but used error scores from the relevant task stages to infer that patients with schizophrenia exhibit an impairment in basic reinforcement learning. As the analysis did not include the crucial set shifting stages, we are unable to draw conclusions about the underlying processes affecting their performance

from this study. A recent study fit a more sophisticated attention-weighted learning model to full IED data from controls and autism spectrum disorder participants (Yearsley et al., 2021). However, the authors did not report the parameter recovery data which is a crucial step in model validation, and they used pre-specified values for two parameters instead of estimating them from participants' data. This approach involves a non-trivial assumption and can result in biases to the other estimated parameters. Finally, both of these studies compared atypical groups screened with diagnostic criteria rather than exploring the relationships between specific symptoms and model parameters.

In this chapter we present an attempt to develop a full computational model of CANTAB IED that incorporates the set shifting stages and fully estimates participant-specific parameters by fitting to trial-by-trial data. We use a large dataset of unselected volunteers, tested online, and explore the mechanistic insights that the models provide. We validate the model comparison on a second large dataset of healthy volunteers and analyse how the model parameters relate to symptoms of common mental health disorders, focusing on compulsivity. Our modelling approach is inspired by previous models of the 'dimensions task' including feature-based reinforcement learning with attention mechanisms to explore whether these models are also able to capture participants' choices on the crucial set shifting stages.

3.3 Methods

3.3.1 Participants

Two independent datasets were collected online via Prolific Academic by Cambridge Cognition Ltd. Participants were recruited if they a) were over 18 years of age, b) were fluent in English, c) had not experienced a significant head injury (resulting in loss of consciousness), d) reported not having been diagnosed with an untreated mental health condition (by medication or psychological intervention) that had a significant impact on their daily life, e) had never been diagnosed with mild cognitive impairment or dementia. Participants' data was anonymous, and they provided their consent online before participating in the experiment. The first dataset includes 731 participants who completed the IED task. The second dataset includes 762 participants who completed the IED task and also several self-report mental health questionnaires. These sample sizes provide 95% power to detect associations of r = 0.13 at alpha = 0.05 (two-tailed).

3.3.2 CANTAB Intra-Extra Dimensional Set Shift Task

To assess attention set shifting, participants completed the CANTAB IED, originally designed as a computerised analogue of the Wisconsin Card Sort Task (Berg, 2010; Grant & Berg, 1948). The design

of the task is presented in Figure 3.1. On each trial, participants were presented with a choice between two stimuli, which for most of the task are compound stimuli comprising two dimensions – lines and shapes. The chosen stimulus was indicated with a mouse click which resulted in deterministic feedback indicating whether the choice was correct. Participants were provided the following instructions: "*This task will take around 7 minutes to complete. You can see 2 patterns. A rule exists telling you which one is correct. You need to try and discover this rule. At first, there is nothing to tell you which pattern will be correct. You have to guess and learn from the feedback. Select a pattern to start. We will tell you whether the pattern you selected was the correct or the incorrect one. Try and use the feedback to help you discover the rule. Once it is clear that you know the rule, it will be changed, but this will not happen very often. After it has changed, you will have to learn the new rule to continue being correct.*" On achieving six correct choices in a row, it is assumed the rule has been learnt, and they move on to the next stage where there is a new rule. If a participant completes 50 trials on a stage without achieving six correct choices in a row, the stage is failed and the task terminated. Throughout the task, the rule is that one of the features indicates the correct stimulus. Critically, this feature is from the line dimension for Stages 1-7, and the shape dimension in Stages 8-9 (in the second dataset, this was reversed such that the correct feature was from the shape dimension for Stages 1-7, and from the line dimension for Stages 8-9). During the task, participants must learn reversals (same stimuli but reversed such that the other feature of the same dimension becomes correct), intra-dimensional shifts (new stimuli and the correct feature is from the same dimension) and extra-dimensional shifts (new stimuli and the correct feature is from the other dimension). An increase in errors on Stage 8 (extra-dimensional shift) is typical as participants find it difficult to attend to the previously irrelevant dimension. No participants were excluded for the analysis of task data.

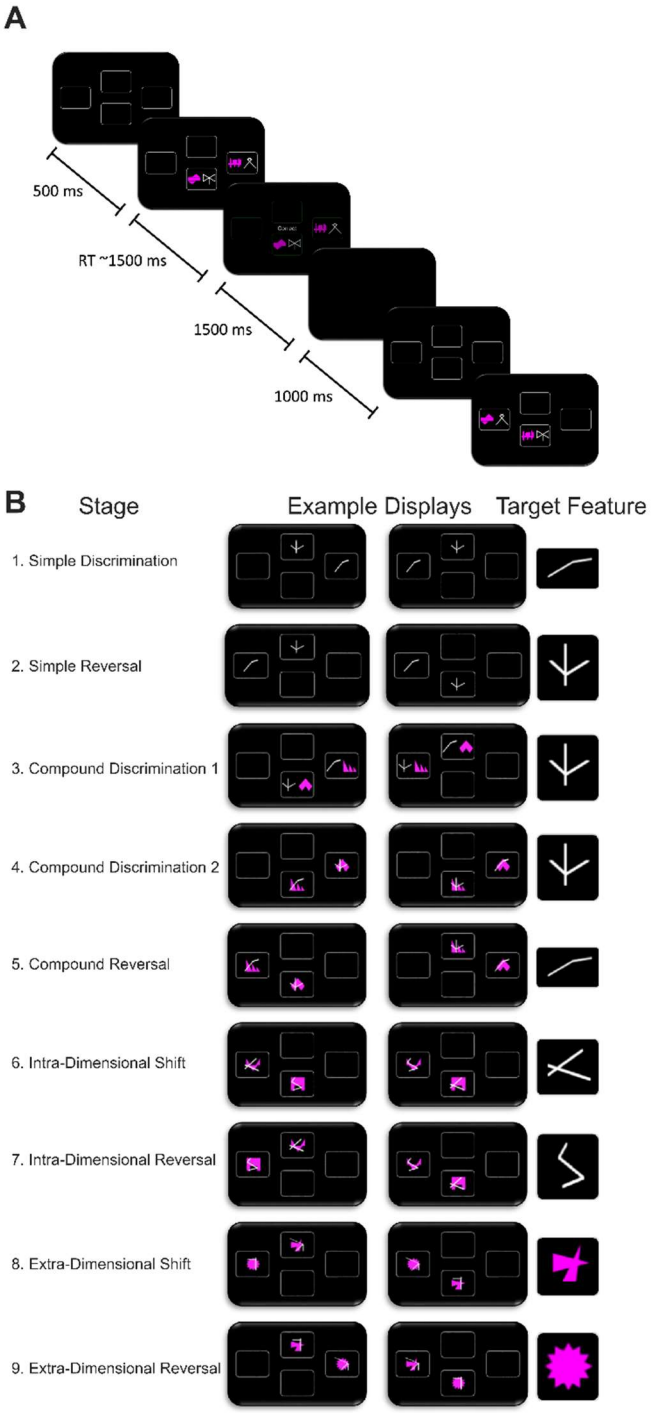

Figure 3.1 CANTAB IED Task Schematic.

A. Schematic of single example trial from Stage 3 of CANTAB IED. Participants are presented with two stimuli that are composed of one feature from each of two dimensions: pink shapes, and white lines. Participants select a stimulus and receive deterministic feedback that informs them of whether their choice was correct or incorrect. After the feedback, the screen briefly turns blank before the presence of two new stimuli indicates the start of a new trial. **B.** Illustration of all nine stages of CANTAB IED, displaying two example trials from each stage, as well as the target features for each stage. Participants need to learn that this feature indicates the correct stimulus for each trial of that stage.

3.3.3 Self-Report Questionnaires

In the second dataset, participants were additionally asked to provide their age, gender, level of education[1] and answer a series of questionnaires about mental health symptoms. Participants completed questionnaires assessing compulsive symptoms (Obsessive Compulsive Inventory Revised, OCI-R (Foa et al., 2002)), depressive symptoms (Self-Rating Depression Scale, SRDS (Zung, 1965)), anxious symptoms (State Trait Anxiety Inventory, STAI (Spielberger et al., 1970), and schizotypy symptoms (Short Scales for Measuring Schizotypy, SSMS (Mason et al., 2005)). Our key measure of interest was compulsive symptoms based on previous literature; however we collected additional symptom data and demographic variables for use as covariates in our analysis due to the known relationships between mental health symptoms and these variables.

3.3.4 Identifying clusters (K-means)

For each participant in the second dataset, an 'error trajectory' was determined as the number of errors at each of the stages of the task. K-means clustering using the *sklearn.cluster.KMeans* package in Python, version 3.7.1 was applied to these trajectories, treating the trajectory as a multidimensional point. This divides participants into a prespecified number (K) of clusters, based on the trajectory of their errors over the course of the task. Each participant was allocated to the cluster with the nearest mean trajectory. The algorithm was run 10 times with different initial mean values, and the best fitting output from these runs was used as the final clustering. This clustering was also used to predict cluster labels of model-simulated data for each participant, given their best-fitting parameters. Only participants with both real and model-simulated data from all nine stages of the IED task could be included in this analysis, leaving 611 participants. K was chosen to be three using the elbow method (Syakur et al., 2018; Thorndike, 1953) based on the screeplot of the sum of squared distances of samples to their closest cluster centre (Figure 3.3A). The participants excluded from the K-means analysis, due to failing the task early, formed an additional cluster, giving four behaviourally defined clusters.

3.3.5 Computational models

CANTAB IED data are traditionally analysed in terms of the number of errors per stage. However, these summary statistics do not make full use of the richness inherent in the dataset. Computational models, on the other hand, model the trial-by-trial choices of each participant, and therefore capture the underlying attention and learning dynamics that are necessary to complete the task. Thus, we

[1] 1: Left formal education before age 16, 2: Left formal education at age 16, 3: Left formal education at age 17-18, 4: Undergraduate degree or equivalent, 5: Master's degree or equivalent, 6: PhD or equivalent.

developed computational models to capture relevant attentional and learning processes to directly test whether they account for the variation in set shifting behaviour. Initial models were based on reinforcement learning (see below), whilst subsequent models included an additional layer where weights represent the allocation of attention to different stimulus dimensions. The three main models that we fit to participants' trial-by-trial choices are described below.

1. Feature Reinforcement Learning (fRL)

This model calculates the values (V) of stimuli (S) by summing the weights (W) of features (f) present in the stimuli:

$$V(S) = \sum_{f \in S} W(f) \tag{3.1}$$

All feature weights are initialised to 0. The stimulus values are then entered into a softmax probabilistic choice function:

$$p(S_i = S_{chosen}) = \frac{e^{\beta V(S_i)}}{e^{\beta V(S_i)} + e^{\beta V(S_{i'})}} \tag{3.2}$$

where $S_{i'}$ indicates the unchosen stimulus, and where β is the inverse temperature parameter, such that large β leads to more deterministic choices of the higher-valued stimulus, and small β leads to more random decisions that are less dependent on stimulus values. The model uses a reinforcement learning rule (Sutton and Barto, 1998) to update the weights of features in *both* stimuli on each trial, as follows:

$$W_t(f) = W_{t-1}(f) + \alpha(R_{s,t} - V_{t-1}(S)) \quad \forall f \in S \tag{3.3}$$

R represents the outcome of a particular stimulus, and t represents the trial number. This model does not treat each stimulus as independent, as it takes into account that stimuli share features. However, it does not consider that the features are part of two different dimensions, and thus cannot generalise across dimensions. Therefore, the next model incorporated an attentional component to estimate stimulus values, allowing it to capture within-dimension generalisation to novel stimuli.

The model's free parameters are: α (learning rate), β (choice determinism)

2. Combined Attention-Modulated Feature Reinforcement Learning (Ca-fRL)

This model incorporates dimensional attention weights to account for the attentional biases towards stimulus dimensions that might develop throughout the task, and capture the within-dimension generalisation to novel stimuli. These weights play a role in stimulus valuation, where they multiply feature weights from the corresponding dimension, thereby weighing the contribution of features from different dimensions according to how much attention is being paid to each one:

$$V(S) = \sum_{d \in [h,l]} W(a_d)W(f_d) \tag{3.4}$$

where a_d represents the attention to dimension d (of h: shape, or l: line), and f_d represents the stimulus feature from dimension d. More precisely, the attention weight for the initially relevant dimension, $W(a_d)$, is specified as $[\sigma(\theta)]$, and for the other dimension as $[1-\sigma(\theta)]$, where σ indicates a transformation by the logistic function. Thus, if one of these weights differs from 0, the other will also, but in the other direction. When both weights are close to 0, the dimensions are weighed equally. For instance, in the first dataset the line dimension was the initially relevant one, and stimulus values were calculated as follows:

$$V(S) = \sigma(\theta) \cdot W(f_l) + (1-\sigma(\theta)) \cdot W(f_h) \tag{3.5}$$

where f_l and f_h represent the line or shape feature within the current stimulus respectively. Stimulus values are then entered into a softmax probabilistic choice function as in the fRL model. All feature weights are initialised to 0. The initial value of θ (θ_0) is a free parameter inferred from the data and represents a dimension primacy effect - the extent to which the initially relevant dimension is attended to, and the second dimension is ignored when introduced on Stage 3. Standard backpropagation (Kelley, 2012; Rojas, 1996) updates the hidden weights and θ on each trial by differentiating the squared error loss (L) with respect to the weight being updated:

$$L = \frac{(V(S) - R(S))^2}{2} \tag{3.6}$$

$$W_t(f_d) = W_{t-1}(f_d) - \alpha \frac{\partial L}{\partial W_{t-1}(f_d)} = W_{t-1}(f_d) - \alpha \frac{\partial L}{\partial \theta_{t-1}} \frac{\partial \theta_{t-1}}{\partial W_{t-1}(f_d)} \tag{3.7}$$

$$\theta_t = \theta_{t-1} - \alpha \frac{\partial L}{\partial \theta_{t-1}} \tag{3.8}$$

Backpropagation of the hidden weights involves multiplication by θ (using the chain rule of differentiation – Equation 3.7). Therefore, when θ is high (more attention to initially relevant

dimension), the weights of features from the other dimension will be updated less on that trial, and vice versa; thus capturing the attentional processes missing from the fRL model. Notably, the same learning rate is used to update the feature and dimension weights, making the assumption that a *combined* process underlies the learning taking place.

The model's free parameters are: α (learning rate), β (choice determinism), θ_0 (dimension primacy)

3. Separate Attention-Modulated Feature Reinforcement Learning (Sa-fRL)

This model is identical to Ca-fRL except that while α is still the learning rate for updating feature weights, a different learning rate, ε, is used to update θ:

$$\theta_t = \theta_{t-1} - \varepsilon \frac{\partial L}{\partial \theta_{t-1}} \quad (3.9)$$

The hidden weights are updated using α, as in equation 3.6. This model allows us to test the possibility that identifiably *separate* learning rates underlie the learning of features values and dimensional attention allocation, which cannot be captured by the combined learning model.

The model's free parameters are: α (learning rate – features), ε (learning rate – dimensions), β (inverse temperature), θ_0 (dimension primacy)

3.3.6 Parameter Estimation and Recovery

Parameters were estimated as described in 2.3 Parameter Estimation and parameter recovery was assessed for each model (Table 3.1).

Model	Parameter	Range	Mean (±SD)	Recovery
fRL	Learning Rate	0 - 1	0.62 ± 0.04	0.49
	Choice Determinism	0 - ∞	0.91 ± 0.33	0.90
Ca-fRL	Learning Rate	0 - 1	0.92 ± 0.14	0.84
	Choice Determinism	0 - ∞	1.34 ± 0.43	0.87
	Dimension Primacy	-∞ - ∞	1.78 ± 0.97	0.91
Sa-fRL	Learning Rate - Features	0 - 1	0.91 ± 0.14	0.88
	Learning Rate - Dimensions	0 - 1	0.85 ± 0.24	0.67
	Choice Determinism	0 - ∞	1.34 ± 0.42	0.87
	Dimension Primacy	-∞ - ∞	1.73 ± 0.87	0.91

Table 3.1 IED Model Parameter Ranges and Recovery.

Free parameters of each model, along with summary statistics of their best-fitting values and recoverability.

3.4 Results

The first dataset includes 731 participants who completed the IED task, but for whom we have no other information. This data was used for model development. In the second dataset participants completed the IED task, several self-report mental health questionnaires and also provided demographic information. This dataset was used for model validation, clustering and subsequent analyses of model parameters. It includes 762 participants with an age range of 18 – 77 years (mean = 38.8, SD = 13.6), and of which 382 (50%) were women. The raw participant data in Figure 3.2 shows that the vast majority of participants make fewer than five errors on the first seven stages of the task. On Stage 8, when a feature from the previously irrelevant dimension is now completely predictive of the correct stimuli, there is a marked increase in variation in performance, with many participants failing this stage (indicated by the crosses).

3.4.1 Attention-Modulated Reinforcement Learning Accounts for Variation in ED Shift Performance

We first fit a feature reinforcement learning model (fRL) to participants' trial-by-trial choices. This model considers the dimensional composition of the task stimuli and has been previously used to model data from similar tasks with multidimensional stimuli (Niv et al., 2015). This model learns weights for individual stimulus features and calculates the overall stimulus value by summing the weights of its component features, thereby accounting for the tendency to select a stimulus with a particular white line feature (for example), regardless of the pink shape (for example) that it is paired with, if participants believe that this feature currently indicates the correct stimulus.

Figure 3.2A (top) shows posterior predictive checks comparing the errors per stage from fRL model-simulated data and from real data. The model-simulated error distributions match the real data fairly well for most stages, consistent with the notion that participants do not treat each stimulus as independent but take into account that stimuli share features. However, it is not able to capture the difficulty that many participants have on the ED Shift (Stage 8). Crucially, the model-simulated error distributions on the ID Shift (Stage 6) and ED Shift (Stage 8), in which participants are presented with novel stimulus features, are identical (and the same is true for their respective reversals: Stages 7 and 9). This is because the fRL model cannot take into account that the novel features are from the same dimensions, and thus does not generalise across dimensions, instead initiating new learning for unseen features. This pattern contrasts with the real data, where errors are generally lower when novel stimulus features are introduced but the relevant dimension remains the same (ID Shift), than when novel stimulus features are introduced and the relevant dimension changes (ED Shift). This is further evidenced by the relatively low correlation (r = 0.46) between real and model-simulated errors on the ED Shift stage (Figure 3.2A top right). In summary, the fRL model captures participants' ability

to weight stimulus features rather than whole stimuli, as shown by improved model fit on most stages, it does not capture their additional tendency to weight dimensions, as shown by poor model fit on stage 8.

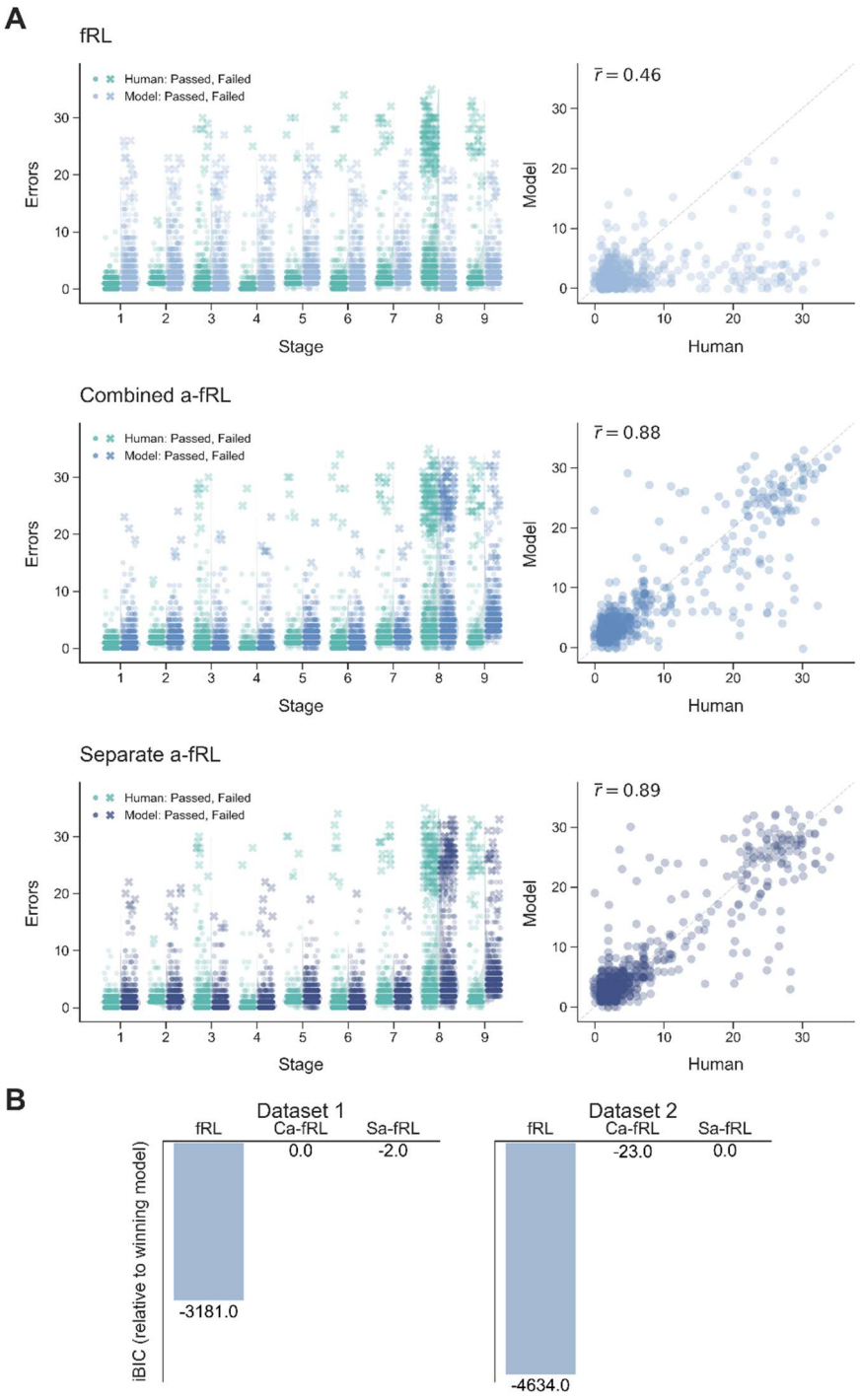

Figure 3.2 Qualitative and Quantitative Model Fits to IED data.

A. Qualitative comparison of real data with a model-simulated dataset for each participant given their best fitting parameters for models fRL (top), Ca-fRL (middle), and Sa-fRL (bottom). *Left:* Posterior predictive checks showing error distributions per stage of CANTAB IED. *Right:* ED Shift errors. Correlation coefficient indicates mean between real data and 10 model-simulated datasets. Jitter has been added to scatter points to aid visualisation. **B.** Quantitative model comparison with iBIC values for three models in two independent sets of data.

In order to account for dimension-based learning, we modified the above model to include a component that captures the dimension to which participants are attending. In line with previous work, this attention component biases both stimulus valuation, by overweighting the values of features from the more attended dimension, and learning, by updating the weights of features from the more attended dimension to a greater extent (Leong et al., 2017). The dimension attention weights themselves were updated using backpropagation. We tested two variations of this feature + attention reinforcement learning model, one with a single learning rate for both the feature weights and dimension weights (Ca-fRL), and one with a separate learning rate for each type of weight (Sa-fRL). Figure 3.2A (middle and bottom) shows that the addition of the attention layer markedly improves model performance, with qualitatively better matched distributions of errors between real and model-simulated data on the ID and ED Shift stages. This is confirmed by the greatly superior correlation between real and simulated datasets (Ca-fRL: r = 0.88, Sa-fRL: 0.89) on ED Shift errors. There is a clear vertical shift in the distribution of real and simulated errors on the ED reversal stages in Figure 3.2A, though the correlations between the two are still reasonably high (Ca-fRL: r = 0.67, Sa-fRL: r = 0.67).

Whilst the fits of Ca-fRL and Sa-fRL are hard to distinguish qualitatively using model simulations, a more formal model comparison considers both the model fit and the model complexity, providing an overall measure of model performance. The model comparison (Figure 3.2B) shows that Ca-fRL is the more parsimonious model in the first dataset, but that Sa-fRL is the more parsimonious model in the second dataset (indicated by the least negative iBIC score). As these iBIC values are very close, compared to the fRL model, it is not clear that the added dimension learning rate parameter in the Sa-fRL model sufficiently improves model fit to justify its added complexity. As we prefer simpler models for increased falsifiability and reduced overfitting, and due to the worse parameter recovery of the dimension learning rate parameter in the Sa-fRL model, we selected Ca-fRL as the most appropriate model for CANTAB IED data, thereby using it for our subsequent parameter analysis. Whilst it is somewhat surprising that such a simple learning model can capture almost all of the variation in set-shifting performance present in our sample, it is interesting that these kinds of models, which were initially created to describe how participants learn about multidimensional stimuli (Leong et al., 2017; Niv et al., 2015), are also able to account for set formation and shifting.

3.4.2 Slower Learning, Random Choices and Stronger Dimension Primacy Lead to Difficulties in Set Shifting

In order to assess the Ca-fRL model's predictions of overall task behaviour in more detail, we applied a K-means algorithm to cluster data based on participants' errors-per-stage trajectories. Using the elbow method (Syakur et al., 2018; Thorndike, 1953), the screeplot in Figure 3.3A shows that using

more than three clusters does not reduce the sum of squared errors substantially, so a three-cluster solution was chosen. The largest cluster (cluster 1) was made up of participants that score few errors on all stages of the task. The other two clusters were made up of participants that made more errors on the ED shift but were separated by their performance on the subsequent reversal. In this last reversal stage, participants should identify the correct feature from within the currently relevant dimension, however it is distinguished from the ID reversal on Stage 6, as it directly follows the ED shift, and therefore somewhat assesses the extent to which participants have shifted their attention to the newly relevant dimension. A final cluster (cluster 4) was added that was made up of the participants with incomplete IED task data (failed at Stage 8 or earlier) who could not be included in the K-means analysis. Model simulations of how the dimension attention weights change over trials for a participant from each cluster are shown in Figure 3.4.

We then tested whether simulated data from each participant using the Ca-fRL model would fall into the same clusters as participants' real data. Figure 3.3B shows that for the majority of participants, simulated data from their best fitting parameters using model Ca-fRL, was allocated to the same cluster as their original data. Some misalignment of real and model-simulated clustering is to be expected given that the K-means algorithm provides sharp cluster assignments, whereas real world behaviour cannot be so cleanly divided, and the assignment of some participants' data is more ambiguous. The largest misalignment is of participants in cluster 3 to cluster 4, which might be because the small number of participants in cluster 3 (score high errors on the extra-dimensional shift and the subsequent reversal) passed the extra-dimensional set shift by chance rather than having obtained an understanding of the key rule change. Nonetheless, this approach provides further evidence that the model is able to capture the main features of and differences between participants' task performance in our sample.

To highlight the mechanistic insights that a modelling approach provides, we examined how model parameters varied by these behavioural clusters. Figure 3.3C shows that compared to the participants in cluster 1 (who score few errors throughout), participants in clusters 2, 3 and 4 have lower learning rates, choose more randomly, and pay more attention to the initially relevant dimension in the early stages of the task. More specifically, learning rates for participants in cluster 3 seem to be particularly low, whilst dimension primacy seems to be particularly high for participants in cluster 4, explaining why many in the latter cluster fail the extra-dimensional shift stage altogether.

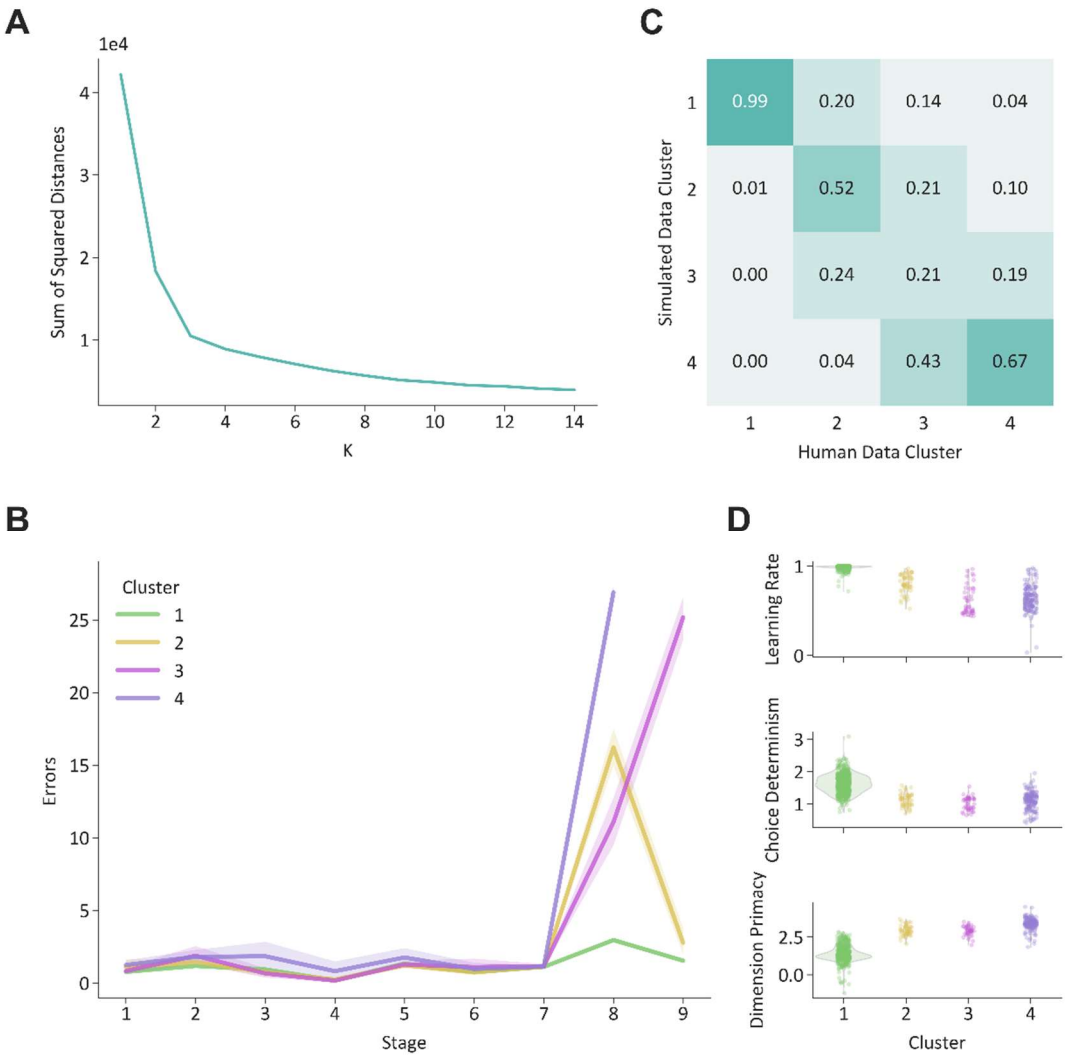

Figure 3.3 K-means Clustering of IED Data.

A. K-means clustering of error-per-stage trajectories. Screeplot of the sum of squared distances to cluster means with K means for different values of K. **B.** The clusters identified by a 3-cluster solution. The fourth cluster consists of participants who failed the task before reaching stage 9 (typically at stage 8, the ED shift) and could not be included in the analysis. Cluster sizes: Cluster 1: 532, Cluster 2: 50, Cluster 3: 42, Cluster 4: 138. **C.** Agreement between clustering of participants' real and simulated data. Each column shows the proportions of simulated data (from participants in a particular real data defined cluster) that were assigned to each cluster. **D.** Parameter distributions by clusters. Distributions of best fitting parameter values from Ca-fRL model, separated by cluster allocation.

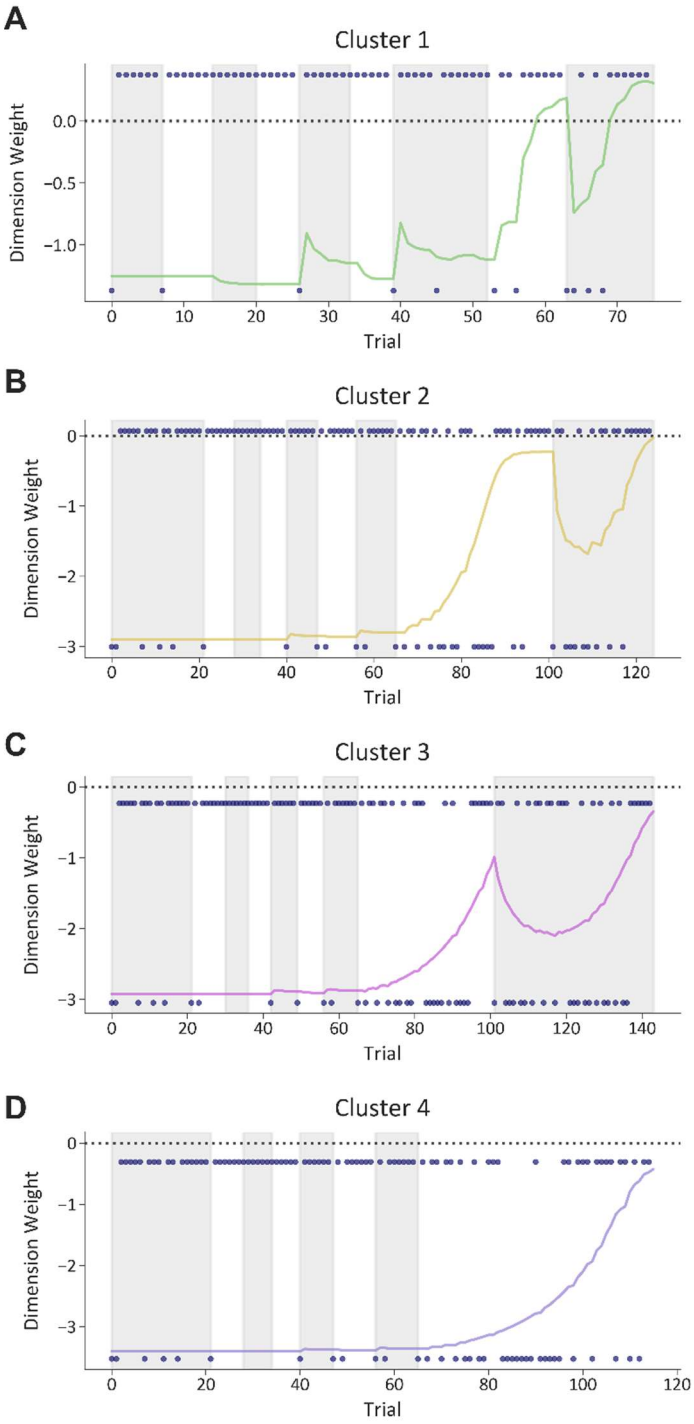

Figure 3.4 Simulations of Internal Model Values From Best-Fitting IED Model.

Model-simulated dimension weights and rewards for a selected participant from cluster 1 (**A**), cluster 2 (**B**), cluster 3 (**C**) and cluster 4 (**D**). Lines indicate the dimension weight: when less than 0, participants pay more attention to shapes (the first relevant dimension in these examples), but when more than 0, participants may more attention to lines. Shading indicates successive stages of IED. Blue dots indicate feedback on each trial: those at the top of the panel indicate correct choices, whilst those at the bottom of the panel indicate incorrect choices.

3.4.3 Slower Learning and Stronger Dimension Primacy Are Associated with Higher Compulsive Symptoms

To assess whether parameters from our best-fitting model, Ca-fRL, may be relevant to symptoms of mental illness, we examined their associations with a variety of symptom questionnaires: compulsivity, depression, state and trait anxiety, and schizotypy. Total score on the OCI-R questionnaire, which assesses compulsive symptoms, was significantly associated with the learning rate parameter ($r(760) = -0.13$, $p = 0.0003$) and the dimension primacy parameter ($r(760) = 0.12$, $p = 0.0008$) after applying a Bonferroni correction for the 15 comparisons performed ($\alpha = 0.0033$). Surprisingly, despite the well-established high associations between symptoms of mental health disorders in the literature, no other questionnaire scores were significantly correlated with model parameters (Table 3.2). The specificity of these relationships was confirmed with a Steiger's Z test (Steiger, 1980), which compares them to the next biggest correlation between parameters and symptoms scores, accounting for the association between the two symptom questionnaires of interest. For both parameters, this involved the relationship with state anxiety, and in both cases the association with compulsivity was significantly greater (learning rate: $r = -0.13$ v $r = -0.05$; $Z = 2.37$, $p = 0.009$; dimension primacy: $r = 0.12$ vs $r = 0.05$; $Z = 2.26$, $p = 0.01$).

To test whether these relationships were affected by the age, gender, or education level of participants, we conducted multiple regression analyses, predicting compulsivity symptoms from model parameters, including the demographic variables as covariates. Again, learning rate and dimension primacy were significant predictors of OCI-R score when evaluated in separate models, with both associations strengthening compared to the unadjusted associations (learning rate: $\beta = -0.82$, 95% CI = [-1.19, -0.46], $p = 0.00001$, overall model $F(4,757) = 16.32$, $R^2 = 0.079$; dimension primacy: $\beta = 1.66$, 95% CI = [0.86, 2.45], $p = .00005$, overall model $F(4,757) = 15.54$, $R^2 = 0.076$). This confirms that participants who exhibit more compulsive symptoms are slower to learn the task structure and show a greater attentional bias towards the initially correct stimulus dimension. The distributions of questionnaire scores are given in Table 3.3.

Table 3.2 Relationships Between IED Model Parameters and Symptom Questionnaires.

Spearman's rank correlations and p-values for all relationships between symptom questionnaire scores and untransformed parameters of best-fitting models. p-values were calculated using a bootstrapping procedure. OCI-R: Obsessive Compulsive Inventory-Revised; SDRS: Self-Rating Depression Scale; STAI-S: State Trait Anxiety Inventory-State; STAI-T: State Trait Anxiety Inventory – Trait; SSMS: Short Scales for Measuring Schizotypy. * indicates significant after correction for multiple comparisons.

		Parameters		
		Learning Rate	**Choice Determinism**	**Dimension Primacy**
Questionnaire	**OCI-R**	-0.13, 0.0003*	-0.09, 0.01	0.12, 0.0008*
	SDRS	-0.04, 0.29	-0.04, 0.24	0.03, 0.42
	STAI-S	-0.05, 0.17	-0.02, 0.55	0.05, 0.21
	STAI-T	-0.01, 0.88	0.02, 0.64	0.00, 0.97
	SSMS	0.01, 0.88	0.02, 0.50	-0.01, 0.76

Table 3.3 Descriptive Statistics of Questionnaire Measures.

Range of possible questionnaire scores, along with mean, standard deviation, and median of participant scores. OCI-R: Obsessive Compulsive Inventory-Revised, STAI-S: State Trait Anxiety Inventory – State, STAI-T: State Trait Anxiety Inventory – Trait, SRDS: Self-Rating Depression Scale, SSMS: Short Scales for Measuring Schizotypy.

	Measure	**Possible Range**	**Mean ± SD, Median**
Questionnaire	OCI-R	0 - 72	14.72 ± 10.93, 12
	STAI-S	20 - 80	36.68 ± 12.83, 35
	STAI-T	20 - 80	42.52 ± 14.31, 41
	SRDS	20 - 80	39.89 ± 10.54, 40
	SSMS	0 - 41	12.14 ± 7.07, 11

To assess which model parameters were related to age, gender and level of education, we calculated Spearman's correlations (with a bootstrapping procedure) and t-tests. After applying a Bonferroni correction for the nine comparisons made ($\alpha = 0.0055$), several remained significant. Age was positively associated with dimension primacy such that older people showed more primacy to the first relevant dimension ($r(760) = 0.11$, $p = 0.003$). Education level was significantly associated with choice determinism and dimension primacy such that those that spent longer in education were more deterministic, and showed less primacy to the first relevant dimension (choice determinism: $r(760) = 0.11$, $p = 0.002$; dimension primacy: $r(760) = -0.11$, $p = 0.003$). Finally gender was associated with all three model parameters such that men learnt faster, were more deterministic, showed less primacy to the first relevant dimension (learning rate: $t(760) = 3.54$, $p = 0.0004$, Cohen's $d = 0.26$; choice determinism: $t(760) = 3.44$, $p = 0.0006$, Cohen's $d = 0.25$; dimension primacy: $t(760) = -4.11$, $p = 4.44 \times 10^{-5}$, Cohen's $d = -0.30$).

3.5 Discussion

We implemented an algorithmic analysis of the CANTAB Intra-Extra Dimensional Set Shift Task. We showed that a hierarchical reinforcement learning model with two simple levels provides a parsimonious yet highly accurate account for participant's choices in two independent samples. In the model, lower-level weights represent the learnt values for stimulus features, and higher-level weights represent the learnt attention to stimulus dimensions. We also explored how model parameters were related to symptoms of common mental health disorders finding that lower learning rates and high dimension primacy were specifically associated with higher compulsive symptoms.

Our modelling analysis suggests a mechanistic explanation for how attention influences learning and decision making as well as for how the focus of attention is itself learnt and shifted. Our best-fitting model learns feature weights that represent current estimates of the associative value of features, and dimension attention weights that bias both the contribution of feature weights to stimulus valuation for action selection, but also the learning of the values of feature weights, as has been shown previously. Notably, our model extends previous algorithmic descriptions by suggesting that dimensional attention is itself learnt by simple prediction error-based update rules. More precisely, the dimension attention weights are updated based on tracking the predictive accuracy of the learnt values of their corresponding feature weights over time, in line with the idea that attention is directed to rewarding features (Mackintosh, 1975). Additionally, the rate of attentional learning is influenced by the strength of dimension attention itself, with more biased dimension attention slowing learning, in line with the idea that attention is updated faster when uncertainty is higher (Pearce & Hall, 1980). Thus, our results suggest that the attention is directed by both the expected reward and the

uncertainty of stimuli lending support to hybrid models of attention (le Pelley et al., 2012). More specifically, our best-fitting model uses backpropagation to achieve this - updating feature and dimension weights to reduce error on each trial. Despite the superior performance of backpropagation-based algorithms for predicting human performance in a range of tasks (Kell et al., 2018; Wenliang & Seitz, 2018), we acknowledge the historical scepticism around its implementation in the brain due to biological constraints (Crick, 1989). Recent research has offered biologically plausible approximations, such as using activity differences between sets of neurons in a local circuit to compute backprop-like weight updates using only locally available signals, providing new insights into possible implementations of backpropagation-like algorithms in the brain (Lillicrap et al., 2020).

The model suggests that a single type of underlying process, albeit in multiple instantiations, can explain performance across all IED task stages, including the extra-dimensional set shift. Reports that people with frontal lobe lesions, Parkinson's Disease or obsessive compulsive disorder demonstrate impaired performance on the extra- but not intra-dimensional set shift or simple reversal learning stages (Downes et al., 1989; Owen et al., 1991; Purcell et al., 1998) has led to the common assumption that the generalisation of pre-learnt rules to novel stimuli (intra-dimensional shift), and the shifting of attention or behaviour to new rules (extra-dimensional shift) are distinct cognitive processes. However, our model contains only a basic type of learning algorithm – albeit arranged hierarchically – and is able to account for choices on all task stages with multidimensional stimuli. There are two aspects to this finding. First, the need for a hierarchy does suggest the existence of separate processes. However, second, the fact that those processes are so similar suggests that the extra-dimensional set shifting process shares important attributes with simpler learning and reversal processes lower down in the hierarchy.

Some caveats to our study merit comment. First, the CANTAB IED has only one extra-dimensional set shift stage. This could result in noisier subject-specific parameter estimates and reduced sensitivity to identify more complex learning process, such as a separate learning rate parameter for the dimension weights. Typical computational modelling analysis of cognitive tasks involves repeated measurements of the cognitive construct of interest to obtain more reliable behavioural and algorithmic measurements. However, the introduction of novel stimuli and the specific ordering of task stages are thought to be crucial for the measurement of true attention set shifting and were intentionally considered in designing the task (Owen et al., 1991). Modelling analysis of tasks such as the 'dimensions task', akin to the Wisconsin Card Sort Task, which do not involve the introduction of truly novel stimuli, lends support to the idea of attention-weighted learning and decision making. The use of our specific algorithm to update attention weights on these task versions is yet to be tested and will be important to determine whether we could see evidence of a more complex process in a task

with multiple set shifts. Second, our algorithm predicts higher errors on stage 9, the extra-dimensional reversal, compared to real data. This indicates that our model does not fully capture the flexibility with which humans make the extra-dimensional set shift. Relative inflexibility is a well-known feature of model-free reinforcement learning algorithms, which can be overcome by using models that incorporate more task structure such as 'model-based' reinforcement learning or hypothesis testing algorithms, suggesting an avenue for future research (Daw et al., 2011; Song et al., 2020; Wilson & Niv, 2012). However, as the 'model-free' system described here was able to predict ED shift errors to a very high degree, any improvement in model performance is unlikely to justify the added complexity of these more structural model types (which are also only likely to come in to play in the last one or two stages of CANTAB IED). Furthermore, the learnt attention weights in our current model are akin to a biased hypothesis testing model where features and dimensions that were previously relevant are tested first.

Analysis of associations with mental health questionnaire data revealed a specific negative relationship between the learning rate parameter and compulsivity. This suggests that participants with higher compulsive symptoms require more information to update their estimates and adapt their behaviour in light of changes in the environment. The majority of our sample had very high learning rates (between 0.95 and 1), which indicates that updates of value estimates after a single trial with unexpected feedback are sufficient to change behaviour. Such one-shot learning is indeed optimal for maximising reward in the IED's deterministic environment. It is tempting to speculate that the ability to adapt behaviour swiftly relies on a better understanding of the structure of the task compared to participants with compulsive symptoms. These would be compatible with other modelling analyses of the probabilistic "two-step" task (Daw et al., 2011), which measures the extent to which participants arbitrate between 'model-free' and 'model-based' reinforcement learning systems. Whilst the former is relatively computationally cheap, it is slower due to reliance on learning from experienced feedback. The latter system benefits from fast learning and increased flexibility at the expense of increased computational complexity required to use a model or internal representation of the environment. A consistent finding in previous literature is an over-reliance on model-free behaviour in compulsive disorders (Gillan et al., 2016; Gillan et al., 2020; Voon et al., 2014).

Our sample consisted of unselected volunteers recruited from the general population. Despite the utility of online data collection, particularly for collecting a large number of participants, it is reliant on the pool of people who participate in online studies, which has the potential to introduce biases. Firstly, despite a very large sample relative to most studies in the field, the majority of participants performed very similarly, scoring few errors throughout the IED task. Similarly, the distributions of questionnaire scores were skewed towards the lower end of the scales, limiting the size of the

relationship that we could detect in this sample. This is reflected in the small effect size detected in the present study, which is fairly standard for online studies. However, it is worth noting that the correlation between model parameters and compulsive symptoms is greater than that between ED shift errors and compulsive symptoms, providing some evidence for the increased validity that a modelling approach is able to provide. Future research should focus on replicating this analysis in a sample with greater variation of symptom profiles.

In conclusion, we have shown that modelling analyses of CANTAB IED task performance are able to provide more precise explanations of behavioural differences. These explanations can then be leveraged to offer mechanistic insights into the symptoms of mental health disorders.

4 Individual Variation in Risky Decisions Is Related to Age and Gender but not to Mental Health Symptoms

4.1 Abstract

Risky decisions involve choosing between options where the outcomes are uncertain. Cognitive tasks such as the CANTAB CGT have revealed that patients with depression make more conservative decisions, but the mechanisms of choice evaluation underlying such decisions, and how they lead to the observed differences in depression, remain unknown. To test this, we used a computational modelling approach in a broad general-population sample (N = 753) who performed the CANTAB CGT and completed questionnaires assessing symptoms of mental illness, including depression. We fit five different computational models to the data, including two novel ones, and found that a novel model that uses an inverse power function in the loss domain (contrary to standard Prospect Theory accounts), and is influenced by the probabilities but not the magnitudes of different outcomes, captures the characteristics of our dataset very well. Surprisingly, model parameters were not significantly associated with any mental health questionnaire scores, including depression scales; but they were related to demographic variables, particularly age, with stronger associations than typical model-agnostic task measures. This study showcases a new methodology to analyse data from CANTAB CGT, describes a noteworthy null finding with respect to mental health symptoms, and demonstrates the added precision that a computational approach can offer.

4.2 Introduction

The most well-known example of a computational model applied to decision making is Kahneman and Tversky's Prospect Theory, which summarises performance in terms of parameters such as risk aversion (the degree to which participants avoid uncertainty) and loss aversion (the degree to which losses loom larger than gains) (Kahneman & Tversky, 1979). These types of models have been applied in mental health research where it has been shown that anxious patients are more risk averse, but not more loss averse, than healthy controls (Charpentier et al., 2017). Computational models based on Prospect Theory have also been applied to the CANTAB CGT. Romeu and colleagues modified traditional models to better fit CGT data from controls and groups of patients with various substance use disorders (Romeu et al., 2020). They found that whilst standard task measures such as "quality of decision making" (reflecting the tendency to make high-probability choices) and "risk adjustment" (the calibration between betting behaviour and probability), showed little difference between patients and controls, model parameters such as risk sensitivity and delay aversion did vary between groups, highlighting the added precision that the modelling approach can provide. However, in this study, the authors did not show some important aspects of model checking, such as parameter

recovery and correlations between individual-level summary measures (such as 'overall proportion bet') from model-simulated and real data. In particular, the latter is more specific than posterior predictive checks on the average of group behaviour and is important to highlight potential areas of model weakness (R. C. Wilson & Collins, 2019). Additionally, in this prior work the authors focused largely on capturing the impulsivity aspect of task performance due to its relevance in substance use disorders; however, mood and anxiety disorders are hypothesised to be more closely associated with changes in appetite for risk.

Here, we expand on this prior work by combining several approaches. First, we examine CGT behaviour in a large sample with a number of self-reported measures of psychopathology. This allows us to jointly assess the relationship between computationally defined decision making processes, psychopathology and demographic variables. Second, we build on a previous computational approach (Romeu et al., 2020) to develop a model which is fully validated. Third, we collect data online which enables fast scaling and replication in novel cohorts. The validated model can then be applied to existing datasets to understand the relationship between any group or psychopathology measures and well-defined computational processes.

4.3 Methods

4.3.1 Participants

Our dataset was collected online via Prolific Academic. Participants were recruited if they: a) were over 18 years of age; b) were fluent in English; c) had not experienced a significant head injury (resulting in loss of consciousness); d) had not been diagnosed with an untreated mental health condition (by medication or psychological intervention) that had a significant impact on their daily life; e) had never been diagnosed with mild cognitive impairment or dementia. Participants were anonymous, and they provided informed consent online before participating in the experiment. The dataset includes 762 participants who completed the CGT task and also several self-report mental health questionnaires. This sample size provides 95% power to detect associations of r = 0.13 at alpha = 0.05 (two-tailed).

4.3.2 CANTAB Cambridge Gamble Task

Participants completed the CANTAB CGT, as described previously (Rogers et al., 1999). The design of the task is presented in Figure 4.1A. Participants start with 100 points and are presented with ten boxes at the top of the screen on each trial - some of which are red, and the rest blue. They are told that one box has a token in it and that they must guess the colour of the box containing the token by selecting their choice at the bottom of the screen. They then have to bet a proportion of their points

that their guess is correct. The possible bets (0.05, 0.25, 0.50, 0.75 or 0.95 of their current points) are presented in a circle in the centre of the screen for two seconds each. Participants click on the circle when they see the amount they want to bet. The result of their choice is shown, and if they are correct, the points are added to their total. If not, the points are deducted. A new trial begins with different numbers of red and blue boxes.

Participants completed 8 practice trials in which they first completed the colour choice part of the task on its own, before the bet component was added; these trials were not included in task analysis. For the first 18 assessed trials, the stakes were shown in descending order, and for the subsequent 18 trials, in ascending order. Participants were excluded from analysis if they did not attempt all four blocks of the task (n = 2), selected either the earliest or the latest bets on all trials (n = 4), or selected the non-majority box colour on more than 50% of trials (n = 3) leaving 753 participants for modelling analysis.

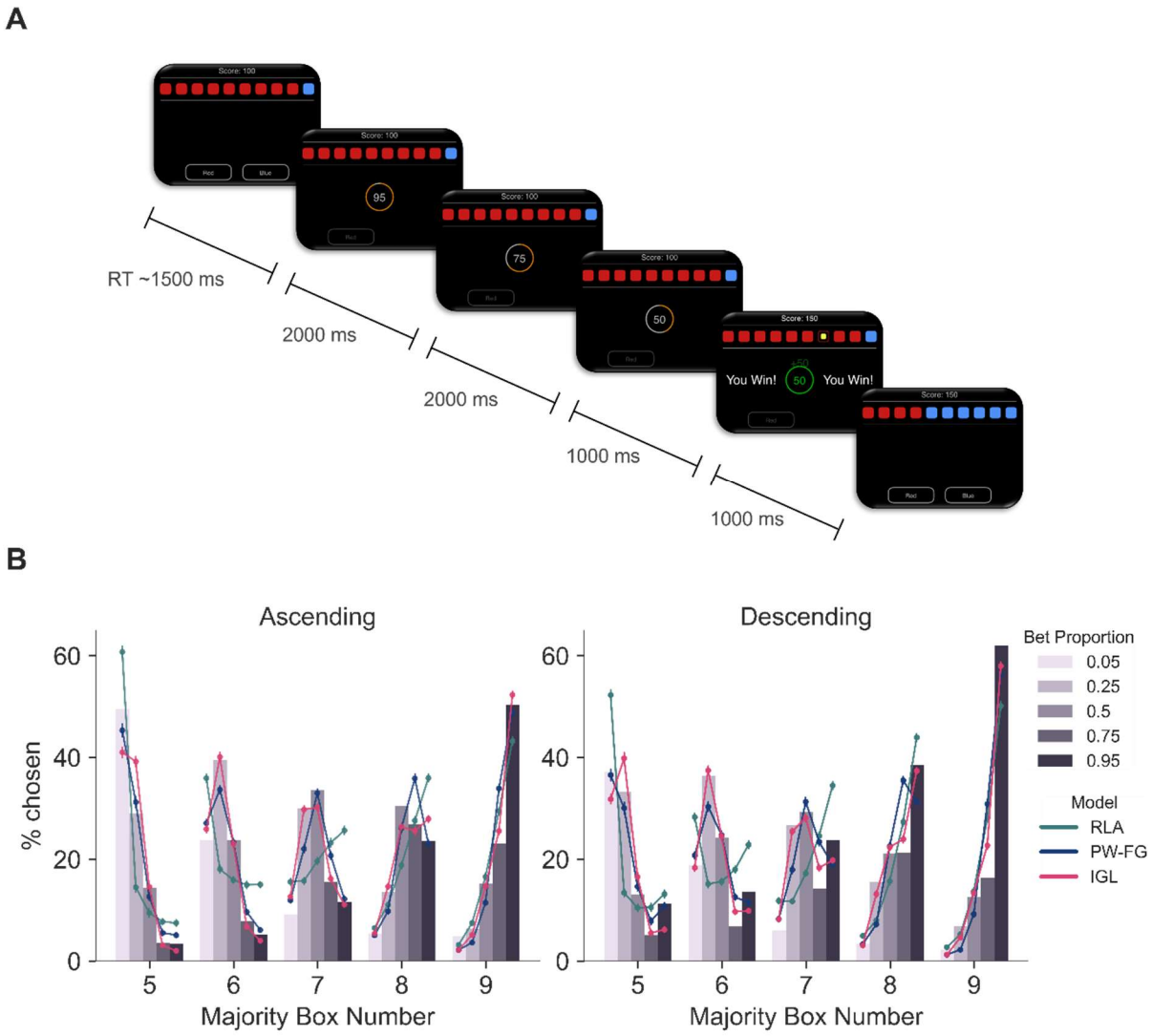

Figure 4.1 CGT Task and Betting Behaviour.

A. Schematic of a trial from the descending condition of the CANTAB CGT. Red and blue boxes are displayed at the top of the screen. After participants select which colour box they think the token is behind, bet options (0.95, 0.75, 0.50, 0.25, 0.05 of their current points – here 100) are presented sequentially in the centre of the screen. Participants click on the number they want to bet before the token is revealed. The points bet are added or subtracted from the total depending on whether the colour choice was correct or incorrect, and a new trial begins with a new box ratio and the new points total. **B.** Summary of real and model-simulated betting behaviour. Plots show the percentage of trials with each majority box number in which participants chose each bet proportion. Real data are shown with bars, while model-simulated data is shown by the coloured lines (points show the mean from 100 simulations, and error bars depict the standard deviation). Data are shown separately for the ascending (left) and descending (right) task conditions, and only trials on which participants chose the majority box colour (and all 5:5 trials) are included. RLA – Risk and Loss Aversion; PW-FG – Projected Wealth Fixed Gains, IGL – Inverse Gains and Losses.

4.3.3 Self-Report Questionnaires

Participants provided their age, gender and level of education[2]. They also completed questionnaires assessing depressive symptoms (Self-Rating Depression Scale, SRDS (Zung, 1965)), anxious symptoms (State Trait Anxiety Inventory, STAI (Spielberger, 1970)), impulsivity (Barratt Impulsiveness Scale, BIS-11 (Patton et al., 1995)), and anhedonia (Temporal Experience of Pleasure Scale, TEPS (Gard et al., 2006)).

4.3.4 Descriptive Measures of CGT Performance

CANTAB CGT data are typically analysed using a number of descriptive (or model-agnostic) measures (Deakin et al., 2004; Rogers et al., 1999), which we detail here as they are used to assess model performance throughout the paper:

1. Quality of Decision Making (QDM)

The proportion of trials on which the participant chose the majority box colour, calculated over all trials on which the number of boxes in each colour differed.

$$\mathrm{QDM} = \frac{\text{number of non 5:5 trials where participant chose the majority box colour}}{\text{number of non 5:5 trials}} \qquad (4.1)$$

2. Overall Proportion Bet (OPB)

The mean proportion of current points gambled by the subject on all gamble trials.

$$\mathrm{OPB} = \frac{\text{sum of the proportion bet on all trials}}{\text{number of trials}} \qquad (4.2)$$

3. Risk Adjustment (RA)

A measure of a participant's sensitivity to probability when betting. Higher values suggest that participants increase their bets with increasingly favourable odds (box ratio), while lower values suggest participants bet consistently irrespective of box ratio. This only includes trials on which participants chose the majority box colour and all 5:5 trials.

[2] 1: Left formal education before age 16, 2: Left formal education at age 16, 3: Left formal education at age 17-18, 4: Undergraduate degree or equivalent, 5: Master's degree or equivalent, 6: PhD or equivalent.

$$\text{RA} = \frac{\sum_n(\text{mean proportion bet when majority box number is n} \times \text{coeff}_\text{n})}{\text{mean propotion bet on trials where participant chose the majority box colour}} \quad (4.3)$$

$$\text{coeff}_\text{n} = \begin{cases} -2, & \text{n} = 5 \\ -1, & \text{n} = 6 \\ 0, & \text{n} = 7 \\ 1, & \text{n} = 8 \\ 2, & \text{n} = 9 \end{cases} \quad (4.4)$$

4. Delay Aversion (DA)

The difference between the mean proportion bet in the ascending and descending conditions. This only includes trials on which participants chose the majority box colour and the 5:5 trials.

$$\text{DA} = \text{mean proportion bet on ascending trials} - \text{mean proportion bet on descending trials} \quad (4.5)$$

4.3.5 Computational Models

While the above descriptive measures represent intuitive summary statistics of task performance, they do not provide insight into *why* participants are making the choices they are. Computational models, on the other hand, use a generative approach to specify how participants interpret the stimuli presented to them and use information to make decisions, and therefore capture the underlying cognitive mechanisms involved in completing the task. Thus, we developed computational models to capture trial-by-trial decision making and risk-taking processes to directly test which processes can account for individual differences in gambling task behaviour.

Each trial of the task requires two decisions - first to choose the box colour, and second to choose what proportion of points to gamble. The likelihood for each trial combines the model's predictions of these two choices in the following way: $p(chosen\ colour) \times p(bet|chosen\ colour)$, such that the bet choice is dependent on the prior colour choice.

All models make the same assumptions about the probability of choosing colours. Let c_t be the choice of colour on trial t, then:

$$p(c_t = red) = \frac{\tau(F_r)^\alpha}{\tau(F_r)^\alpha + (1-\tau)(10-F_r)^\alpha} \quad (4.6)$$

where F_r is the number of red boxes on the screen on the current trial (range 0 - 10), τ is a red-bias parameter such that higher values mean the participant is more biased to selecting red, and α is a sensitivity parameter that indicates how sensitive participants are to the box ratio on each trial,

equivalent to the slope of a logistic function. Higher values of α indicate that the participant chooses the majority box colour in a more deterministic manner.

The models then compute the probability of bets $p(b_t|c_t)$. The first step is the evaluation of the utility U of each bet option. This is where the different models differ, and we will turn to the different formulations of this below. Next, the utility of each potential outcome (a win or a loss of a certain magnitude, or the future wealth) is weighted by the probability of that outcome based on the choice of colour c_t. Finally, a linear delay factor is included which penalizes options that are presented later. This results in an overall value $V(b_t|c_t)$ for each bet b_t, conditioned on the first-stage colour choice c_t as follows:

$$V(b_t|c_t) = f(c_t)U(b_t, 1) + \big(1 - f(c_t)\big)U(b_t, -1) - \beta d(b_t) \quad (4.7)$$

where the function $d(b_t)$ takes on values {0, 0.25, 0.5, 0.75, 1} for increasing delays $d()$ of the bets, and the function $f(c_t)$ indicates the fraction (number of boxes/10) of the chosen colour on that particular trial. The form and meaning of U depends on the particular model as explained further below. Finally, the values V determine the probabilities of choosing each bet through a softmax function:

$$p(b_t|c_t) = \frac{e^{\gamma V(b_t|c_t)}}{\sum_{b'} e^{\gamma V(b'|c_t)}} \quad (4.8)$$

and we consider the joint likelihood over probability of colour and bet choices: $\mathcal{L}(\theta) = \prod_t p\,(c_t|\theta)p(b_t|c_t, \theta)$ for inference of parameters θ.

1. Risk and Loss Aversion (RLA)

The first model is based on the influential framework of Kahneman and Tversky (Kahneman & Tversky, 1979). Prospect Theory assumes that participants subjectively value the potential wins and losses on each trial. In the CGT, participants bet a proportion of their points which are either added or subtracted from their total, so the potential wins and losses are a product $b_t w_t$ of the bet b_t chosen on trial t, and the wealth w_t on that trial. The bet takes on a fixed set of proportions ($b_t \in .05, .25, .5, .75, .95$). In the RLA model, the subjective utility of the wins and losses is defined as follows:

$$U^R(b_t, r_t) = \begin{cases} \log(\rho b_t w_t), & if\ r_t = 1 \\ -\delta \log(\rho b_t w_t), & if\ r_t = -1 \end{cases} \quad (4.9)$$

where $r_t = 1$ indicates that the colour choice was successful, $r_t = -1$ that it was not. The nonlinear logarithm function alters the effective shape of the utility curve, resulting in risk-aversion in the domain of gains and risk-seeking behaviour in the domain of losses, with greater values of the risk-aversion parameter, ρ, exaggerating these effects. δ is a loss-aversion parameter, such that large δ means that losses are more aversive than the equivalent gain is rewarding. Note that we have used a log function (as opposed to the conventionally used power function) to aid numerical stability.

The parameters in this model were $\theta = \{\alpha, c, \rho, \delta, \beta, \gamma\}$ where α is the colour choice determinism, c the colour choice bias, ρ the risk aversion, δ the loss aversion, β the delay aversion and γ the bet choice determinism.

2. Projected Wealth (PW)

The RLA model as formulated above considers the potential wins and losses that would result on each trial. We next consider a related model (Bernoulli, 1738; von Neumann & Morgenstern, 1944) which assumes that the attractiveness of different bet options depends on the total projected wealth each option would result in, $w_t + b_t w_t r_t$. The utility U^W of each bet in the wealth model is defined as:

$$U^W(b_t, r_t) = \log(1 + \rho(w_t + b_t w_t r_t)) \quad (4.10)$$

Note that this model does not include a loss aversion parameter, and that increasing values of the risk aversion parameter, ρ, increases risk averse behaviour. The parameters in this model were $\theta = \{\alpha, c, \rho, \beta, \gamma\}$ where α is the colour choice determinism, c the colour choice bias, ρ the risk aversion, β the delay aversion and γ the bet choice determinism.

3. Projected Wealth Fixed Gains (PW-FG)

We next investigated the winning model from Romeu et al., which is similar to the PW model with the exception that the risk aversion parameter in the domain of potential gains is fixed to 1, and the parameter is only estimated in the potential loss domain with increasing ρ indicating increased risk aversion in the domain of potential losses. Again, we assume that participants subjectively value the potential total wealth, $w_t + b_t w_t r_t$.

$$U^F(b_t, r_t) = \begin{cases} \log(1 + (w_t + b_t w_t)), & if\ r_t = 1 \\ \log(1 + \rho(w_t - b_t w_t)), & if\ r_t = -1 \end{cases} \quad (4.11)$$

The parameters in this model were $\theta = \{\alpha, c, \rho, \beta, \gamma\}$ where α is the colour choice determinism, c the colour choice bias, ρ the risk aversion, β the delay aversion and γ the bet choice determinism.

4. Linear Loss Aversion (LLA)

We developed further models in order to improve the performance compared to the previously published PW-FG model described above. Our first novel model assumes that participants subjectively value the bet proportions b_t, independently from their current wealth (note that unlike in the above models, w_t does not enter the specification). Gains can be distorted through a power function, in which the risk aversion parameter, $\rho < 1$, represents risk aversion, but $\rho > 1$ represents risk seeking behaviour in the gains domain. Losses are scaled linearly by a loss aversion parameter.

$$U^L(b_t, r_t) = \begin{cases} {b_t}^{\rho}, & if\ r_t = 1 \\ -\delta b_t, & if\ r_t = -1 \end{cases} \quad (4.12)$$

The parameters in this model were $\theta = \{\alpha, c, \rho, \delta, \beta, \gamma\}$ where α is the colour choice determinism, c the colour choice bias, ρ the risk aversion, δ the loss aversion, β the delay aversion and γ the bet choice determinism.

5. Inverse Gains and Losses (IGL)

Our second novel model also assumes that participants subjectively value the bet proportions b_t, again independently from their current wealth.

$$U^I(b_t, r_t) = \begin{cases} {b_t}^{\rho}, & if\ r_t = 1 \\ -\delta {b_t}^{1/\rho}, & if\ r_t = -1 \end{cases} \quad (4.13)$$

In this model, the distortions of loss and gains are linked by an inverse power function, such that if the curve is concave ($\rho < 1$, risk averse) in one domain, the other is too, and if the curve is convex ($\rho > 1$, risk seeking) in one domain, the other is too. This breaks the assumption of Prospect Theory of opposite behaviour between domains - risk aversion in the gains domain, combined with risk seeking in the losses domain. Losses are again additionally linearly scaled by a loss aversion parameter δ.

The parameters in this model were $\theta = \{\alpha, c, \rho, \delta, \beta, \gamma\}$ where α is the colour choice determinism, c the colour choice bias, ρ the risk aversion, δ the loss aversion, β the delay aversion and γ the bet choice determinism.

4.3.6 Parameter Estimation and Recovery

Parameters were estimated as described in 2.3 Parameter Estimation and parameter recovery was assessed for each model (Table 4.1)

4.3.7 Sensitivity Analysis

We carried out a sensitivity analysis by removing participants whose bet choice data was not fit better by our winning model compared to chance performance. For this, we counted the number of times the model assigned the greatest probability to the bet that the participant actually chose and used a binomial test to compare this to the number of times we would expect it to happen by chance (where the probability of choosing each bet proportion was 1/5). We kept all participants where the outcome of this test was significant (suggesting that the winning model provided a significantly better fit to their data than chance) leaving a sample size of 727 (N = 26/753 (3.5%) excluded) for this analysis.

Model	Parameter	Possible Range	Mean ± SD, Median	Recovery (r)
1. Risk and Loss Aversion	Colour choice determinism	0 - ∞	8.30 ± 4.76, 8.48	0.71
	Colour choice bias	0 -1	0.50 ± 0.07, 0.50	0.50
	Risk aversion	0 - ∞	0.86 ± 0.01, 0.86	0.00
	Loss aversion	0 - ∞	1.90 ± 1.13, 1.67	0.87
	Delay Aversion	-∞ - ∞	0.22 ± 0.46, 0.15	0.85
	Bet choice determinism	0 - ∞	3.19 ± 3.00, 2.50	0.89
2. Projected Wealth	Colour choice determinism	0 - ∞	8.38 ± 4.81, 8.58	0.62
	Colour choice bias	0 - 1	0.51 ± 0.07, 0.50	0.57
	Risk aversion	0 - ∞	0.98 ± 8.06, 0.02	0.00
	Delay Aversion	-∞ - ∞	0.02 ± 0.04, 0.02	0.77
	Bet choice determinism	0 - ∞	33.27 ± 30.87, 24.15	0.86
3. Projected Wealth Fixed Gains	Colour choice determinism	0 - ∞	8.48 ± 5.11, 8.38	0.69
	Colour choice bias	0 - 1	0.51 ± 0.07, 0.51	0.61
	Risk aversion	0 - ∞	0.13 ± 0.21, 0.03	0.51
	Delay Aversion	-∞ - ∞	0.05 ± 0.13, 0.04	0.93
	Bet choice determinism	0 - ∞	18.47 ± 10.60, 16.23	0.89
4. Linear Loss Aversion	Colour choice determinism	0 - ∞	8.16 ± 4.53, 8.52	0.58
	Colour choice bias	0 - 1	0.50 ± 0.07, 0.50	0.53
	Risk aversion	0 - ∞	9.88 ± 20.06, 0.44	0.16
	Loss aversion	0 - ∞	1.29 ± 1.14, 0.99	0.43
	Delay Aversion	-∞ - ∞	0.08 ± 0.09, 0.04	0.71
	Bet choice determinism	0 - ∞	38.98 ± 35.68, 29.16	0.43
5. Inverse Gains and Losses	Colour choice determinism	0 - ∞	8.38 ± 5.01, 8.55	0.62
	Colour choice bias	0 - 1	0.50 ± 0.07, 0.50	0.53
	Risk aversion	0 - ∞	0.68 ± 0.35, 0.62	0.85
	Loss aversion	0 - ∞	2.65 ± 2.36, 1.99	0.95
	Delay Aversion	-∞ - ∞	0.07 ± 0.16, 0.05	0.91
	Bet choice determinism	0 - ∞	18.47 ± 12.62, 15.65	0.80

Table 4.1 CGT Model Parameter Ranges and Recovery.

Free parameters of each model, along with summary statistics of their best-fitting values and recoverability.

4.4 Results

Our final dataset consisted of 753 participants with an age range of 18 – 76 years (mean = 41.5, SD = 13.5), and of which 358 (50%) were women. In Figure 4.1B, the bars show the betting patterns in the raw data - people tend to pick smaller bets when the box ratio is more evenly matched (lower majority box number) but bet more points when there is greater discrepancy in the box colours (higher majority box number). Furthermore, there is a bias towards betting higher values in the descending condition compared to the ascending condition (descend vs ascend: $t(752) = 9.77$, $p = 2.63 \times 10^{-21}$), suggestive of delay-averse (impatient) behaviour.

Inspired by the modelling framework in Romeu et al., 2020, all our models had the same colour choice (Equation 4.6), probability weighting (Equation 4.7), delay aversion (Equation 4.7), and bet choice determinism (Equation 4.8) functions, while the key distinction between models came from the valuation functions (Equations 4.9 – 4.13).

4.4.1 Previous Models of Risky Decision Making Perform Poorly

We first assessed the performance of three previously published models for risky decision making. The first valuation function we used was inspired by the influential Prospect Theory (Kahneman & Tversky, 1979). The Risk and Loss aversion model (RLA) proposes that participants subjectively evaluate the magnitude that could be won or lost on each trial with risk aversion in the domain of gains, risk seeking behaviour in the domain of losses , and loss aversion (losses are more aversive than gains are attractive). In the behavioural economics literature, these models are typically fit to group behaviour and the average parameter estimates obtained. Here, we are more interested in fitting individual-level rather than group-level parameters to explore how differences in e.g., loss aversion and risk aversion lead to differences in task behaviour. When examining the predictions made by the RLA model, we found that it predicts that participants were most likely to choose either the largest or the smallest bet option on all trials. The green line in Figure 4.1B shows how the model predicts that the 0.05 or 0.95 bets are chosen the most often regardless of the box ratio, in contrast to the real data which displays a clear preference for intermediate bets when the majority box number is between 6 - 8. Therefore, the RLA model was not able to capture the key betting patterns observed in the real data (Figure 4.1B, Figure 4.2A).

The second established model we tested was inspired by Expected Utility Theory (Bernoulli, 1738; von Neumann & Morgenstern, 1944). The model assumes that participants subjectively evaluate their projected wealth following the outcome of each trial while making decisions, rather than the individual losses and gains, and therefore does not include a loss aversion parameter. Unlike the RLA model, this Projected Wealth model (PW) model was able to generate participants' choosing of

intermediate bet proportions (0.25, 0.50, 0.75). However, the model was not able to capture the behaviour of the most low-betting participants (Figure 4.2B).

The Projected Wealth Fixed Gains model (PW-FG; Romeu et al., 2020) is a previously published model that was specifically developed to describe behaviour on CANTAB CGT. It is similar to the PW model except for the adaptation that the risk aversion parameter was fixed to 1 for potential increases in cumulative points while a free parameter was estimated for the case of potential decreases to cumulative points. However, the PW-FG model exhibited the same weakness as the PW model, failing to account for the behaviour of participants who chose lower bets on average, even at high box ratios, as shown most clearly in Figure 4.1B. In particular it overestimates the percentage of participants that chose to bet 75% of their points when the majority box number was 7, 8 or 9. This weakness of the PW-FG model is further emphasised by the scatterplot of the overall proportion bet measure in real vs model-simulated data, where the failure to capture the behaviour of low-betting participants is clear (Figure 4.2C). Capturing this conservative betting behaviour is of particular importance as prior studies indicate that placing low bets on favourable gambles is related to anxious and depressive symptoms (Charpentier et al., 2017; Murphy et al., 2001; Rawal et al., 2013).

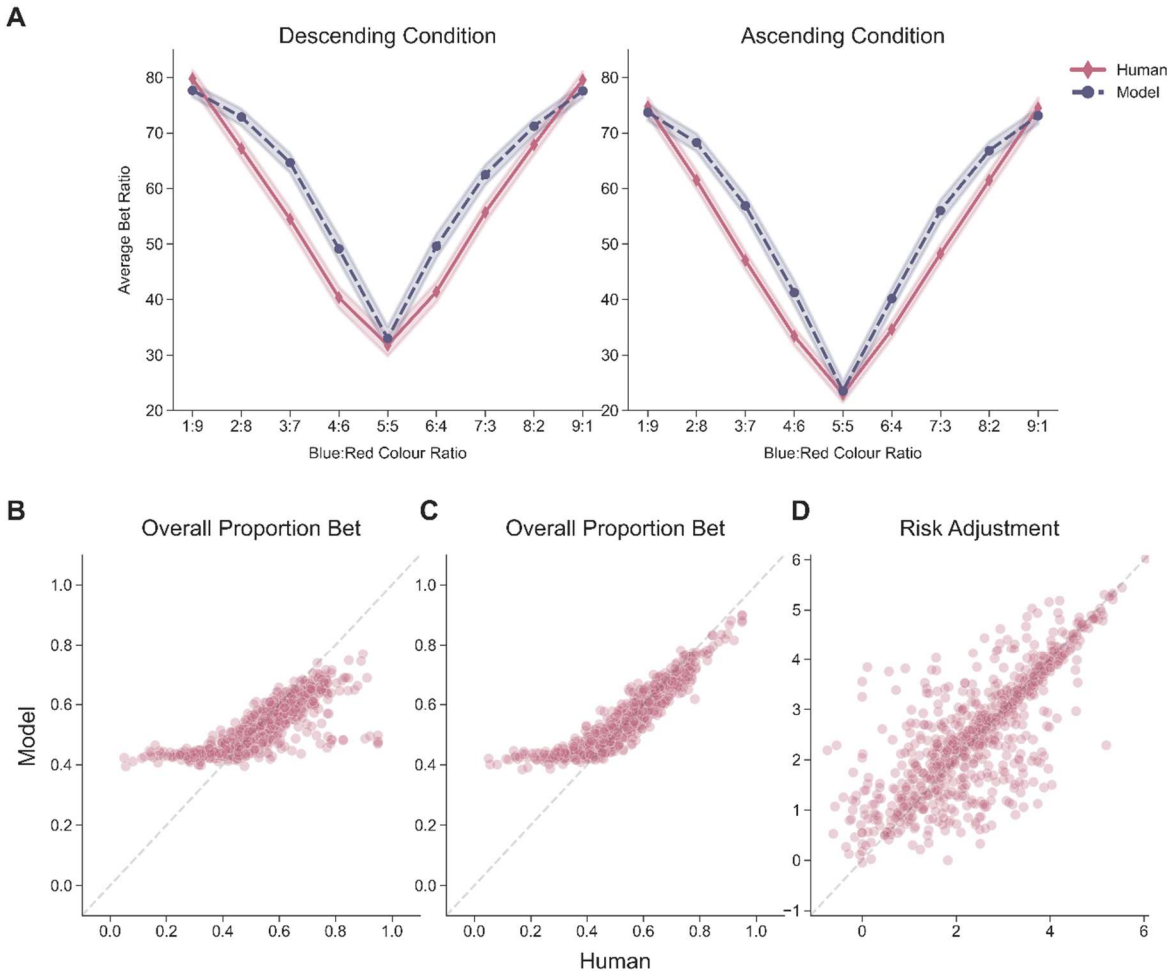

Figure 4.2 Posterior Predictive Checks Demonstrate Specific Model Weaknesses.

A. Average Group Fit of Risk and Loss Aversion Model. Real and model-simulated patterns of proportion bet averaged over subjects for each box ratio and task condition. Points indicate mean whilst bands represent 95% confidence intervals. **B.** Individual Differences in Overall Proportion Bet for Projected Wealth Model. Scatter plot of real vs model-simulated scores, r = 0.80. **C.** Individual Differences in Overall Proportion Bet for Projected Wealth Fixed Gains Model. Scatter plot of real vs model-simulated scores, r = 0.92. **D.** Individual Differences in Risk Adjustment for Linear Loss Aversion Model. Scatter plot of real vs model-simulated scores, r = 0.77. For the lower panels in this figure, the correlation between real and model data was calculated for 10 different simulations, and the average of these is reported here. The plots show the comparison with a single simulated dataset for visualisation purposes.

4.4.2 Participants are Risk Averse for both Gains and Losses, and Indifferent to Current Wealth

In order to overcome the limitations of the three models described above, we developed two novel models in which participants are assumed to subjectively evaluate the bet proportions themselves (0.05, 0.25, 0.50, 0.75, 0.95) thus rendering their current number of points irrelevant for decision making. The Linear Loss Aversion model (LLA) uses a power function for gains, and a linear function

was introduced for losses to prevent the convexity that led to the RLA model being unable to predict participants' choosing intermediate bets. Further, a loss aversion parameter was reincorporated to overcome the limitation of the PW and PW-FG models, and better describe the behaviour of consistently low-betting participants. This model performed much better than previous ones as shown by the promising correlation between real and model-simulated data, particularly for risk adjustment scores (Figure 4.2D). We then developed a final model to improve the predictions of participant-specific risk adjustment further, as this is a key feature of task performance. This model, Inverse Gains and Losses (IGL), also uses a power function for gains with the subjective valuation of losses now also determined by a power function and parameterised with the inverse risk aversion parameter ($1/\rho$). This has the effect that predicted behaviour is consistent between the domains of gains and losses, being either risk-seeking or risk-averse in both, in contradiction to the RLA model. This model was able to fully recapitulate the patterns of individual variability observed (Figure 4.1B).

Despite the utility of posterior predictive checks (Figure 4.1B, Figure 4.2) during our iterative model development procedure, a more formal model comparison approach is required. Figure 4.3A shows the average likelihood per trial, an indication of how well the model predicts participant choices on average, as well as the integrated Bayesian Information Criterion, which considers both the model's predictions and simultaneously penalises for added model complexity (a less negative score indicates a better model). These graphs both show that each successive model improves overall model performance, with the Inverse Gains and Losses model performing best on both measures. Figure 4.3B demonstrates how this winning model generates betting choices based on the box ratio presented to the participant on a particular trial, using the group median of estimated parameter values. Of note is the concavity of the valuation of the bet proportions in the far-left panel, which suggests that participants are indeed largely risk averse (with risk aversion parameters < 1) in the domain of both gains and losses (in contrast to Prospect Theory), and the quite minimal effect of delay aversion which is evident by comparing the middle two panels. Finally, Figure 4.4 shows that the model is able to capture both the group-level behaviour (Figure 4.4A), and the individual-level descriptive measures (Figure 4.4B) very well, as the correlations with the descriptive measures are all 0.89 and above.

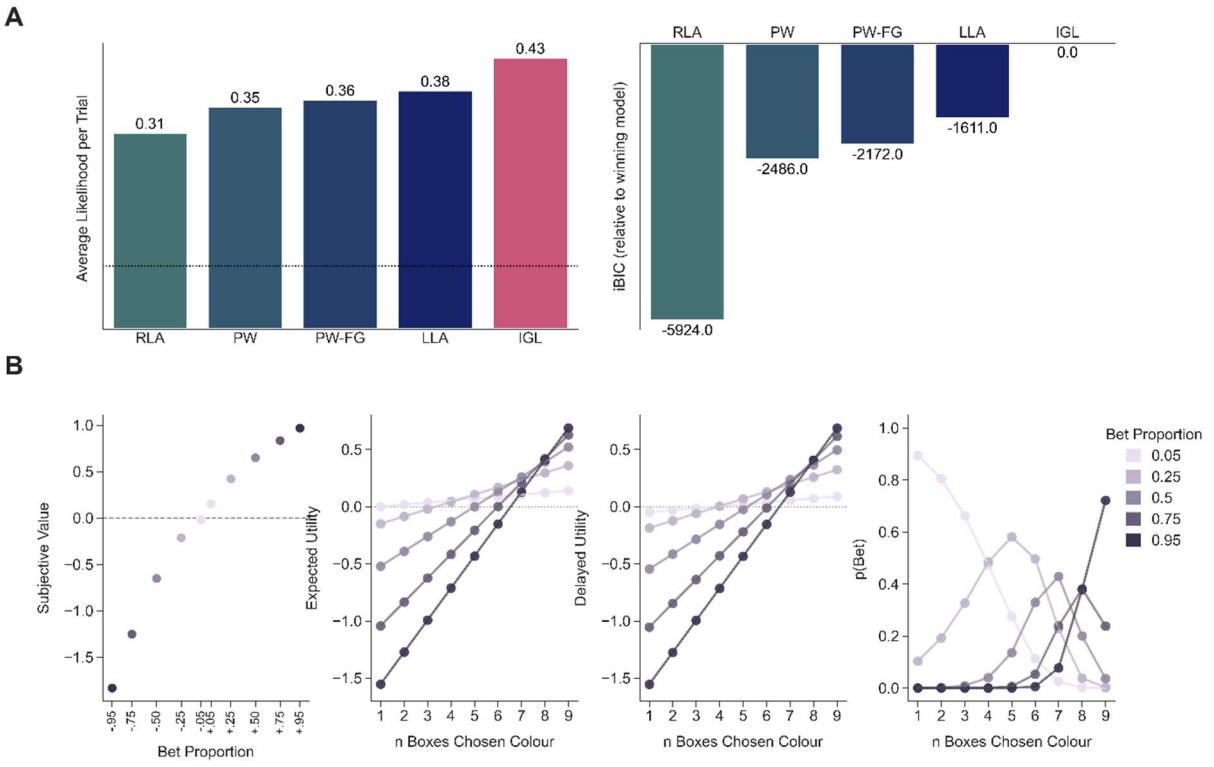

Figure 4.3 CGT Model Comparison and Model Simulations.

A. Model Comparison. Model performance assessed by average likelihood per trial (left) and iBIC (right). The dotted line indicates the average likelihood per trial for a model that makes choices at random – 0.1. **B.** Model Simulations. Internal model values simulated using the best fitting Inverse Gains and Losses (IGL) model. *Far Left:* Subjective valuation of potential wins and losses for each bet proportion. *Centre Right:* Win/loss values are weighted by their probabilities (n boxes of the chosen/unchosen colour) to give the expected utility for each bet proportion. *Centre Left:* Expected utilities are adjusted by the order in which they are displayed. This example is from the descending condition such that lower bet proportions are shown later, and thus penalised more. *Far Right:* The delayed utilities are passed through a softmax equation to give the probability of a participant choosing each bet. Importantly the IGL model predicts that participants are likely to make intermediate-level bets when the box ratio is between 6-8, consistent with real behaviour (cf Figure 1B). The medians of estimated parameters were used for simulations: risk aversion – 0.62, loss aversion – 1.99, delay aversion – 0.05, bet choice determinism – 15.65. The dotted lines at 0 aid visualisation. RLA – Risk and Loss Aversion; PW – Projected Wealth, PW-FG – Projected Wealth Fixed Gains, LLA – Linear Loss Aversion, IGL – Inverse Gains and Losses.

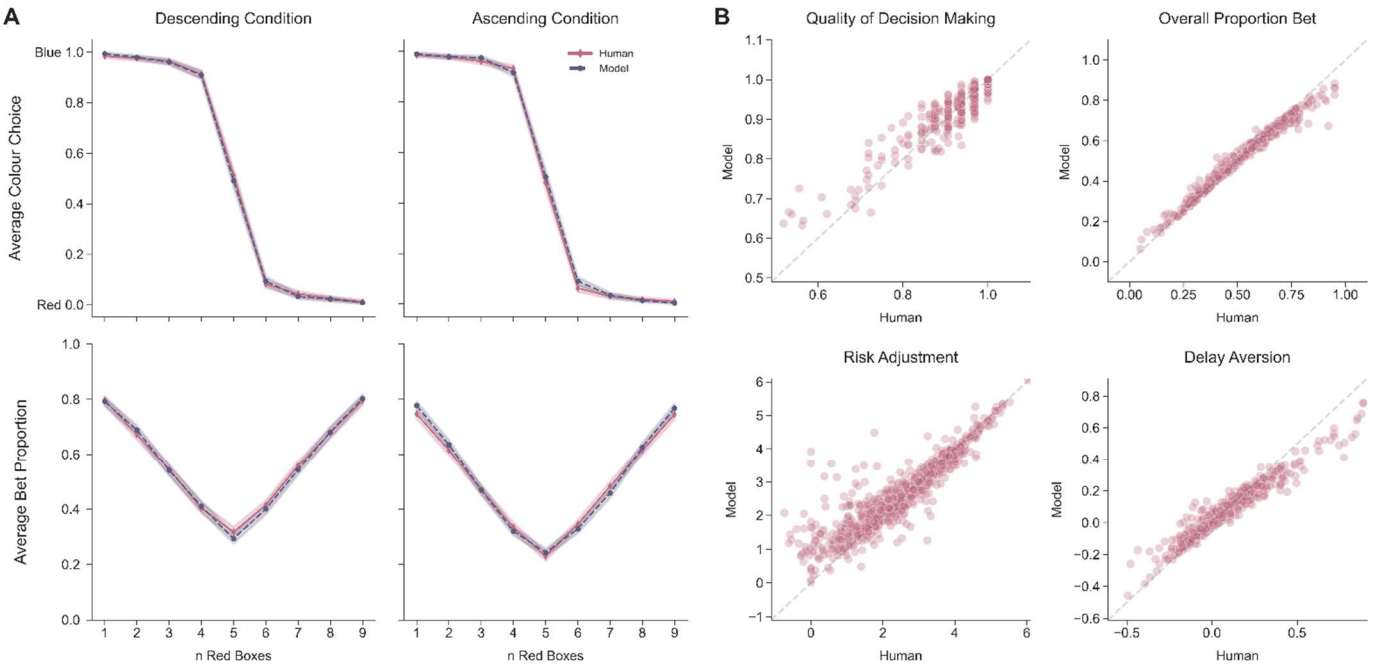

Figure 4.4 Qualitative Fits to CGT Data for Inverse Gains and Losses Model.

A. Average Group Fits. Average participant and model-simulated patterns of colour choice (top) and proportion bet (bottom) for each box ratio. Points indicate mean and bands represent 95% confidence intervals. **B.** Individual Differences of Descriptive Measures. Scatterplots of scores from real data against those from a model-simulated dataset for Quality of Decision Making, r = 0.94 (top left), Overall Proportion Bet, r = 0.99 (top right), Risk Adjustment, r = 0.89 (bottom left), and Delay Aversion, r = 0.96 (bottom right). The dotted line indicates y = x (perfect prediction). The correlation between real and model data was calculated for 10 different simulations, and the average of these is reported here.

4.4.3 Risk Aversion Increases with Age, and Women are More Risk Averse Than Men

To assess which model parameters were related to age, gender and level of education, we calculated Pearson's correlations and t-tests. Age was negatively associated with the risk aversion parameter (indicating older people are more risk averse) and delay aversion, but positively associated with loss aversion (risk aversion: $r(751) = -0.24$, $p = 1.49 \times 10^{-11}$; delay aversion: $r(751) = -0.09$, $p = 0.011$; loss aversion: $r(751) = 0.10$, $p = 0.0049$). Women were less deterministic in both the colour choice and betting choice part of the task and were also more risk-averse and loss-averse than men (colour choice determinism: $t(751) = 2.82$, $p = 0.0049$, Cohen's $d = 0.21$; bet choice determinism: $t(751) = 2.71$, $p = 0.0069$, Cohen's $d = 0.20$; risk aversion: $t(751) = 5.03$, $p = 6.23 \times 10^{-07}$, Cohen's $d = 0.37$; loss aversion: $t(751) = 2.45$, $p = 0.015$, Cohen's $d = -0.18$). Finally, higher education level was positively correlated with colour choice determinism ($r(751) = 0.08$, $p = 0.027$). After applying a Bonferroni correction to

account for the 18 comparisons performed (α = 0.0028), the relationships between risk aversion and age/gender remained significant (Figure 4.5).

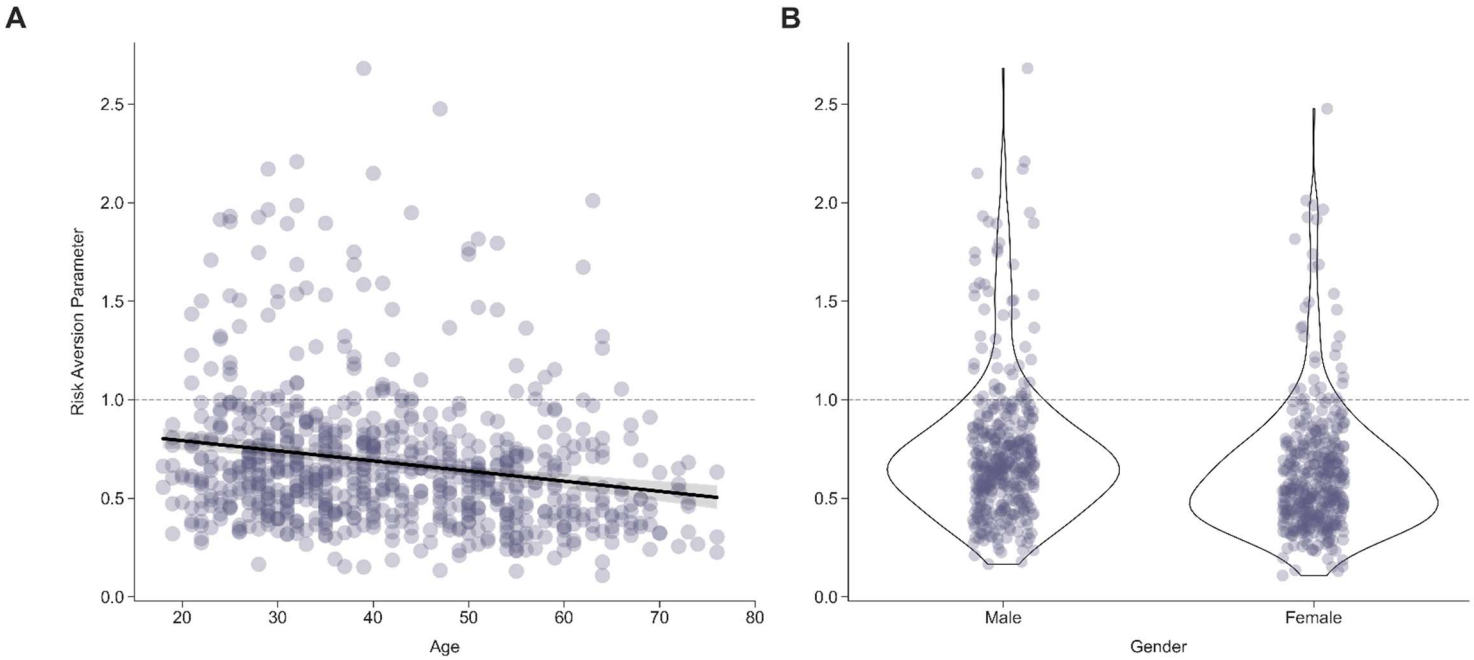

Figure 4.5 Associations Between CGT Model Parameters and Demographic Variables.

A. Older people are more risk averse. Lower values of the risk aversion parameter represent higher risk aversion, r(751) = -0.24. **B.** Women are more risk averse. Lower values of the risk aversion parameter represent higher risk aversion, t(751) = 5.03. Untransformed parameter values are used for statistical inference while transformed values ($\rho > 0$) are used for visualisation. Dotted line indicates an objective utility function, $\rho > 1$ indicates risk seeking behaviour whilst $\rho < 1$ indicates risk aversion.

Age and gender were also associated with the model-agnostic measures: age was negatively correlated with overall proportion bet, risk adjustment and delay aversion (overall proportion bet: r(751) = -0.14, p = 1.07×10^{-4}; risk adjustment: r(750) = -0.11, p = 3.80×10^{-3}; delay aversion: r(749) = -0.087, p = 0.017) whilst women bet less, and adjusted their bets less compared to men (overall proportion bet: t(751) = 4.13, p = 4.09×10^{-5}, Cohen's d = 0.30; risk adjustment: t(750) = 3.51, p = 4.72×10^{-4}, Cohen's d = 0.26)[3]. After applying a Bonferroni correction to account for the 12 comparisons performed (α = 0.0042), the relationships between overall proportion bet and risk adjustment with both age and gender remained significant. Crucially, the size of the relationship between age and the computational risk aversion parameter was significantly greater than that with the model-agnostic

[3] Note that the degrees of freedom vary slightly as some outcome measures are incalculable from certain data. For instance, delay aversion is incalculable if participants never chose the majority box colour in at least one condition. In our data, one participant did not obtain a risk adjustment score, and two participants did not obtain a delay aversion score.

overall proportion bet measure (r = -0.24 vs r = -0.14: Steiger's Z = 2.48, p = 0.0066), demonstrating the extra sensitivity conferred by the computational approach. The size of the effect of gender on risk aversion was also numerically greater than that with the model-agnostic overall proportion bet measure, though this difference was not statistically significant (r = 0.14, transformed from d = 0.3 vs r = 0.18, transformed from d = 0.37: Steiger's Z = 0.77, p = 0.2207). These results were unaffected by a sensitivity analysis in which participants whose data were not fit significantly better by our winning model than by chance were removed (Table 4.2).

Variable 1	Variable 2	Effect Size	p value
Risk Aversion Parameter	Age	r = -0.25	4.99×10^{-12}
Risk Aversion Parameter	Gender	Cohen's d = 0.38	4.02×10^{-7}
Overall Proportion Bet	Age	r = -0.15	3.84×10^{-5}
Overall Proportion Bet	Gender	Cohen's d = 0.32	2.29×10^{-5}
Risk Aversion Parameter	SDRS	r = 0.05	0.14
Risk Aversion Parameter	TEPS	r = 0.02	0.59
Loss Aversion Parameter	SDRS	r = 0.00	0.93
Loss Aversion Parameter	TEPS	r = -0.05	0.14
Overall Proportion Bet	SDRS	r = -0.01	0.88
Overall Proportion Bet	TEPS	r = 0.07	0.07

Table 4.2 Key CGT Relationships in a Sensitivity Analysis.

SRDS: Self-Rating Depression Scale, STAI-S: State Trait Anxiety Inventory – State, STAI-T: State Trait Anxiety Inventory – Trait, BIS-11: Barratt Impulsivity Scale, TEPS: Temporal Experience of Pleasure Scale.

4.4.4 Model Parameters are Not Associated with Mental Health Symptoms

To assess whether model parameters were related to mental health symptoms, we calculated Pearson's correlations between each pair of variables. However, we found no significant relationship between any of our model parameters and symptom scores, with all correlations $r \leq 0.07$ (Table 4.3). Model-agnostic measures of task performance were also not related to mental health symptoms, with all correlations $r \leq 0.07$ (Table 4.4). These results were unaffected by a sensitivity analysis in which participants whose data were not fit significantly better by our winning model than by chance were removed (Table 4.2). The distributions of questionnaire scores and model-agnostic measures are given in Table 4.5.

		Parameters					
		Colour choice determinism	Colour choice bias	Risk aversion	Loss aversion	Delay aversion	Bet choice determinism
Questionnaire	SRDS (746)	-0.03, 0.35	0.06, 0.12	0.06, 0.11	-0.00, 0.90	0.04, 0.33	-0.04, 0.29
	TEPS (745)	-0.05, 0.19	-0.03, 0.39	0.02, 0.53	-0.06, 0.12	-0.03, 0.49	-0.05, 0.18
	STAI-S (742)	-0.03, 0.40	0.06, 0.11	0.04, 0.23	-0.02, 0.68	0.03, 0.42	0.03, 0.47
	STAI-T (742)	0.02, 0.50	0.06, 0.09	0.02, 0.58	-0.01, 0.69	0.01, 0.88	0.02, 0.55
	BIS-11 (741)	-0.07, 0.04	0.02, 0.52	0.02, 0.54	-0.07, 0.06	0.06, 0.12	-0.05, 0.21

Table 4.3 Relationships Between CGT Model Parameters and Symptoms.

Pearson's correlations and p-values for relationship between symptom questionnaire scores and untransformed parameters of the best-fitting model (IGL). The bracketed value by each questionnaire gives the degrees of freedom for the corresponding analyses. SRDS: Self-Rating Depression Scale, STAI-S: State Trait Anxiety Inventory – State, STAI-T: State Trait Anxiety Inventory – Trait, BIS-11: Barratt Impulsivity Scale, TEPS: Temporal Experience of Pleasure Scale.

		Model-Agnostic Measure			
		Quality of Decision Making	Overall Proportion Bet	Risk Adjustment	Delay Aversion
Questionnaire	SRDS	-0.04, 0.31 (746)	-0.00, 0.96 (746)	0.01, 0.70 (745)	0.04, 0.24 (744)
	TEPS	-0.03, 0.42 (745)	0.07, 0.07 (745)	-0.05, 0.14 (744)	-0.05, 0.19 (743)
	STAI-S	-0.04, 0.32 (742)	0.01, 0.85 (742)	0.01, 0.70 (741)	0.04, 0.26 (741)
	STAI-T	0.01, 0.88 (742)	0.01, 0.78 (742)	0.05, 0.15 (741)	0.03, 0.43 (741)
	BIS-11	-0.05, 0.17 (741)	0.03, 0.39 (741)	-0.05, 0.21 (740)	0.06, 0.09 (740)

Table 4.4 Relationships Between CGT Model-Agnostic Measures and Symptoms.

Pearson's correlations and p-values for relationship between symptom questionnaire scores and model-agnostic measures of CGT task performance. Bracketed values give the degrees of freedom for the analysis. SRDS: Self-Rating Depression Scale, STAI-S: State Trait Anxiety Inventory – State, STAI-T: State Trait Anxiety Inventory – Trait, BIS-11: Barratt Impulsivity Scale, TEPS: Temporal Experience of Pleasure Scale.

Table 4.5 Descriptive Statistics of Questionnaire Scores and CGT Model-Agnostic Measures.

Range of possible questionnaire scores or model-agnostic measures, along with mean, standard deviation, and median of participant scores. SRDS: Self-Rating Depression Scale, STAI-S: State Trait Anxiety Inventory – State, STAI-T: State Trait Anxiety Inventory – Trait, BIS-11: Barratt Impulsivity Scale, TEPS: Temporal Experience of Pleasure Scale. QDM: Quality of Decision Making, OPB: Overall Proportion Bet, RA: Risk Adjustment, DA: Delay Aversion.

	Measure	Possible Range	Mean ± SD, Median
Questionnaire	SRDS	20 - 80	38.95 ± 10.25, 39
	STAI-S	20 - 80	36.97 ± 12.77, 35
	STAI-T	20 - 80	42.75 ± 13.26, 42
	TEPS	18 - 108	80.28 ± 11.44, 81
	BIS-11	30 - 120	57.99 ± 9.73, 58
Model-Agnostic Measure	QDM	0 – 1	0.96 ± 0.07, 1.00
	OPB	0 – 1	0.54 ± 0.15, 0.56
	RA	-2.4 – 7.2	2.43 ± 1.27, 2.44
	DA	-0.9 – 0.9	0.07 ± 0.19, 0.04

4.5 Discussion

In this study, we used a computational analysis of the CANTAB CGT to more precisely investigate the mechanistic relationships between task behaviour, symptoms of mental health disorders and demographic variables. We fit five different models, including two novel ones, and found that a novel model in which betting strategies are not influenced by the number of current points, and uses an *inverse* power function in the loss domain, captures the characteristics of a large online dataset very well. Somewhat surprisingly, we found no significant relationships between model parameters and symptoms of mental health problems, but we did find robust associations between the risk aversion parameter and age and gender, such that older people and women were more risk averse. It is noteworthy that these relationships with demographic variables were stronger than those with raw outcome measures such as overall proportion bet, highlighting the added precision and mechanistic insight gained from this modelling analysis.

Our best-fitting model suggests a mechanistic explanation for how participants approach risky decisions when provided with explicit information about the amounts and probabilities that may be won or lost. Our novel model (Inverse Gains and Losses) varies from traditional risky decision making models such as Prospect Theory in two main ways: 1) it assumes that participants' betting strategies are independent of their current number of total points; and 2) it assumes that participants are consistently either risk averse or risk seeking across the domains of gains and losses. Both of these

distinctions are likely to reflect the differences between tasks for which Prospect Theory models were developed and the CGT. The former were typically two-alternative forced choice tasks in which participants were required to choose between a risky gamble and a certain option (of winning or losing a number of points). As the current task asks participants to choose an amount to bet from five options, it involves a more fine-grained decision of how much one is willing to bet on a decision rather than just which option they choose. It is therefore likely to be more sensitive to characterising participants' risk preferences. Furthermore, it is possible that because of the additional complexity in the gambling component (due to there being more options), the current number of points becomes a less salient component than the overarching bet choice strategy. Finally, previous literature suggests that on average participants tend to be risk averse in the gain domain but risk seeking in the loss domain (Kahneman & Tversky, 1979). Due to the power function in the gain domain, the inverse power function in the loss domain, and the values of the estimated risk aversion parameter in our winning model generally being below one, our model suggests that when performing the CGT participants tend to be risk averse in *both* domains. Again, this is likely to be due to key differences in the CGT compared to traditional risky decision making tasks. Classical tasks are typically 'additive' in nature, such that the amount to be gained or lost is irrespective of the number of points the participant has. This contrasts with the CGT, where participants have to choose a bet from different proportions of their current number of points, which means that when following previous gains, participants have the potential to gain even more, often referred to as a 'multiplicative' feature. Multiplicative tasks can also often lead to bigger losses than additive tasks, and recent theories in ergodicity economics suggest that a multiplicative environment should therefore foster higher levels of risk aversion (Meder et al., 2021), which our results support.

Analysis of relationships between model parameters and symptoms of mental health problems revealed no significant findings. This result was somewhat surprising given the previous literature reporting more conservative risk attitudes in CGT in various groups with depressive symptoms (Mannie et al., 2015; Murphy et al., 2001; Rawal et al., 2013), as well as the previous algorithmic analysis of CGT in patients with substance use disorders (Romeu et al., 2020). However, our results are consistent with a recent study assessing CGT performance in an adolescent cohort (using standard model-agnostic measures) in which there was also no convincing evidence of a relationship between risk taking and depressive symptoms (Lewis et al., 2021), while clear relationships with age and gender were observed. It is possible that these inconsistencies are due to general population samples not capturing many participants at the more severe end of the symptom scales, and therefore limiting the size of any relationship. Alternatively, previous findings may have been due to chance or confounding variables. It is also possible that these models and parameters are not sufficiently sensitive to capture

more specific differences in reward seeking or risk-taking behaviour seen in depression. For example, Tavares et al found that unmedicated depressed patients had worse quality of decision making in the colour choice component of the task specifically on trials that followed a loss (Taylor Tavares et al., 2007), which could be further explored with learning models. This highlights an interesting possible follow-up computational study for analysis of this task and its relationship with depression.

We found relationships between many model parameters and demographic variables of which two survived correction for multiple comparisons: older people and women were found to show higher levels of risk aversion. This finding has been consistently reported in previous literature with large cohort samples and model-agnostic CGT outcome measures such as overall proportion bet (Deakin et al., 2004; Lewis et al., 2021). Here, we replicated these findings with the behavioural measures, but further showed that the relationships between these demographic variables and risk aversion, a parameter from a model-based analysis, were stronger. This demonstrates that a modelling analysis of task behaviour can lead to more precise measurements, and also that these parameters are more mechanistic in nature than traditional model-agnostic behavioural measures. Future research should implement this computational approach in existing or novel CGT datasets to better understand previous findings.

There are some caveats to our study that merit comment. Our data was collected online, which is important for accessing large samples for research, but there is less control over the participants that volunteer to take part in online studies. This might lead to biased samples and spurious correlations. However, as the associations with demographic variables have been reported before, including in large cohort studies, these are less likely to have led to the results reported here. Another caveat of online data collection specific to this task, is the difficulty of drawing a sharp distinction between true impulsivity and distraction. It is possible that some participants select the earliest presented bet every time to finish the task as quickly as possible, or that some let the timer run out and are not engaging with the task properly. We attempted to address this by removing a small percentage of individuals who always chose the first or last bets, and those whose data were not well fit by our winning model. This did not affect the results. However, it remains difficult to distinguish inattentive behaviour from poor decision making and impulsivity in online data collection.

In conclusion, we have presented a modelling analysis of CANTAB CGT in a novel, large dataset, and shown robust relationships with age and gender, but not mental health symptoms in an unselected, largely sub-clinical sample. This work highlights the added precision that computational models can provide to explore relationships with both demographic and mental health symptom variables.

5 Individual Variation In Subjective Probability Weighting Is Important But Unrelated to Catastrophising and Anxiety.

5.1 Abstract

Gambling tasks have long been used to study how people approach risky decisions. Previous research has suggested that elevated anxiety is associated with differences in risky decision making, but it remains unclear which specific facets or types of anxiety are most associated with these differences. Further, many existing studies of risky decisions have been unable to disentangle all of the components of these decisions, and in particular tend to neglect 'probability weighting' – how people's subjective weighting of probability differs from the true probability. We hypothesised that this component is highly relevant to catastrophising symptoms in particular, which are common in anxiety. To test this, we used a computational modelling approach in a broad general-population sample (N = 212) who performed a gambling task and completed questionnaires assessing symptoms of mental illness, including catastrophising. We found evidence that models incorporating a probability weighting function fit participants' data better than those without it, which replicated in a larger sample (N = 946). Despite this, we found no evidence for a relationship between probability weighting parameters and catastrophising; and whilst we found some evidence that people with higher general anxiety symptoms overweighted the probability of negative gamble outcomes relative to the positive ones in the initial sample, this result did not replicate when tested in the larger sample. This study showcases the importance of incorporating probability weighting parameters into studies of risky decision making and describes a noteworthy null finding with respect to mental health symptoms.

5.2 Introduction

Catastrophising refers to the tendency to predict the worst outcome. An example would be when scoring less highly than hoped on a maths test. Whilst some people might feel disappointed and make a mental note to do more preparation for the next test, an individual with high catastrophising might take this as a signal that they are incompetent at everything, and will never be able to get a job. These kinds of cognitions can be very distressing and have been implicated in common mental health disorders, particularly those associated with anxiety such as generalised anxiety disorder, obsessive compulsive disorder, and post-traumatic stress disorder (Gellatly & Beck, 2016). Whilst related to anxiety disorders, recent research has shown that catastrophising is a separate construct from anxiety and worry, suggesting it may be a trans-diagnostic symptom (Pike et al., 2021). Moreover, the importance of this symptom for mental health is exemplified by its targeting in cognitive behavioural therapy, an evidence-based talking therapy for common mental disorders (Cuijpers et al., 2016, 2019;

Hofmann et al., 2012): “decatastrophising” involves asking people to re-evaluate the likelihood and severity of the feared event. Despite its significance in mental health, the specification of what catastrophising entails remains imprecise. Beck's initial conceptualisation was that of cognitive distortion, which he termed magnification, defined as ‘inflation of the magnitude of [one's] problems and tasks'. He went on to refine this definition in 1979, as when an individual ‘always think[s] of the worst. It's most likely to happen to [them]’ (Beck AT et al., 1979). These definitions highlight two aspects of catastrophising: thinking of problems as more severe than they are, and thinking of problems as more probable than they are. It has been assumed that catastrophising cognitions comprise both of these, though this has not been investigated. The development of a catastrophising questionnaire, along with the recent advances in the field of computational psychiatry, offer a promising means by which to test these hypotheses.

Computational models are increasingly used to develop precise explanations of cognitive processes in the field of mental health (Adams et al., 2016). Kahneman and Tversky’s Prospect Theory (Kahneman & Tversky, 1979) is an influential model used to describe how people make risky decisions and has typically been applied to gambling tasks. In these tasks, participants are typically asked to choose between two options: a gamble (e.g. 50% chance of winning 50 points and 50% chance of losing 30 points) and a sure option (e.g. 100% chance of winning 15) in order to maximise their points. Crucially, compared to simply measuring the number of times the participant accepted the gamble, Prospect Theory offers a theoretical account of why participants make the choices they do in terms of how they subjectively value winning or losing different amounts of points. The subjective evaluation is specified mathematically and adjusted by setting the values of model parameters such as risk aversion and loss aversion. Risk aversion captures participants’ aversion to uncertainty, whilst loss aversion captures the extent to which participants find losses more aversive than gains are rewarding. The more recent Cumulative Prospect Theory (Tversky & Kahneman, 1992) includes an additional function that captures how people subjectively weight the probabilities that are presented to them. They suggested that people systematically overweight small probabilities and underweight large probabilities, and subsequently several one- or two-parameter functions have been suggested as ways to capture such variation. The advantage of using these models is that we gain more precise explanations for the observed behaviour, teasing apart the different behaviourally relevant components of interest. Here, we are most interested in the ability to measure and separate subjective probability weighting, due to its relevance to definitions of catastrophising – i.e., thinking of negative outcomes as more probable than they actually are.

Prospect Theory models have been used in previous research to disentangle components of risky decision making, and differences in behavioural parameters have been reported in anxiety-related disorders, particularly OCD. Fitting Prospect Theory models to gambling task data, Charpentier et al. found that risk aversion was higher in pathologically anxious patients compared to healthy controls, but that there was no difference in loss aversion (Charpentier et al., 2017). Sip et al. found increased loss aversion in unmedicated OCD patients as compared to medicated OCD patients and controls, though they did not estimate risk aversion (Sip et al., 2017). Neither of these studies incorporated the probability weighting aspect of Cumulative Prospect Theory models, and so they cannot distinguish whether apparently higher risk or loss aversion might instead be driven by differences in probability weighting (Charpentier et al., 2017). In another study, probability weighting parameters were estimated from a task in which participants had to choose between two gambles, each with three potential outcomes (Aranovich et al., 2017). Parameters were analysed between patients with OCD or hoarding disorder and healthy volunteers, and the authors reported that controls overweighted low probabilities and underweighted high probabilities (as noted by Kahneman and Tversky), whereas patients showed the opposite pattern. In a more recent study with OCD, GAD and social anxiety patients, there was no difference in probability weighting parameters between patients and controls (George et al., 2019). In summary, whilst there is some evidence of altered risky decision making in individuals with anxiety-related disorders, most of these studies did not estimate all parameters of the Cumulative Prospect Theory model, making it difficult to adjudicate between competing hypotheses as to what might be driving any observed differences. Further, these studies used adaptive task designs, which despite their popularity may bias parameter estimates. Until recently (Pike et al., 2021), it was not possible to readily measure catastrophising using a questionnaire, especially in a large-scale online context, so specific relationships with catastrophising have not been tested.

In this pre-registered study (Talwar et al., 2021), we use cognitive data from an unselected online sample on a novel gambling task, as well as mental health symptom data leveraging the behavioural variation inherent in larger samples (Gillan et al., 2016). We fit full Cumulative Prospect Theory models to task data to estimate subject-specific probability weighting parameters and assess their relationship with symptoms of catastrophising. Specifically, we aimed to test whether probability weighting parameters are related to symptoms of catastrophising.

5.3 Methods

5.3.1 Participants

Participants were recruited online via Prolific Academic to complete the Gambling Task and Abstract Reasoning Task (ART), and also several self-report mental health questionnaires. Participants were recruited if they confirmed they a) were between 18 and 60 years of age, b) were fluent in English, c) lived in the UK or US, d) did not have impaired, uncorrected vision or colour blindness, e) had never been diagnosed with mild cognitive impairment or dementia, f) had no history of head injury. Participants' data was anonymous, and they provided their consent online before participating in the experiment.

Original Sample

This sample was collected in July 2021. We recruited 212 participants, which provides over 90% power to detect effect sizes over r = 0.2 (in a one-tailed correlation test using a bivariate normal model) (Faul et al., 2009).

Replication Sample

This sample was collected in January 2022. We recruited 946 participants, which provides over 90% power to detect effect sizes over r = 0.1 (in a one-tailed correlation test using a bivariate normal model) (Faul et al., 2009).

5.3.2 Gambling Task

To assess how probability weighting affects risky decision making, participants completed a gambling task (Figure 5.1A). The task involves choosing between two options, which determine how many points participants win or lose on that trial. One option has 100% probability of a specified outcome, and the other option is a gamble with some probability (x%) of a positive outcome, and the complementary probability (100-x%) of a negative outcome. Participants are instructed to pick the outcome that they believe will maximise their points overall, but they do not get feedback on their decisions. There are three types of trials: gain-only, loss-only, and mixed (Figure 5.1A). In gain-only trials, the sure option is a gain of +10, +20, or +30, the positive outcome on the gamble is one of +20, +40, +60, whilst the negative outcome is always 0. The loss-only trials feature the equivalent negative outcomes, such that the positive outcome on the gamble is always 0, and the negative outcome is one of -20, -40, or -60 with the sure option being a loss of -10, -20, or - 30. For the mixed gamble trials, the sure option is always 0, whilst the positive outcome on the gamble is a gain of +40, +50, or +60, and the negative outcome is a loss of -10, -50, or -90. Furthermore, the probabilities of the good outcome in the gamble could be any of 0.3, 0.5, 0.75, 0.9, and 0.95. As there are nine distinct prospects, and

five different probabilities for each of three trial types, this gives a total of 135 trials. Participants were first shown the instruction page, which included example images of the options they might come across in the task. This was immediately followed by an instructions quiz that asked participants three questions to ensure that i) they understood that they will always choose between a risky gamble and a safe option, ii) they understood the meaning of the options and iii) they understood that they should aim to maximise points earned at the end of the task. If participants made errors on the quiz, they were redirected to the instructions page and asked to reread them before attempting the quiz again. Participants were randomly assigned to one of three pre-set pseudorandomised trial orders.

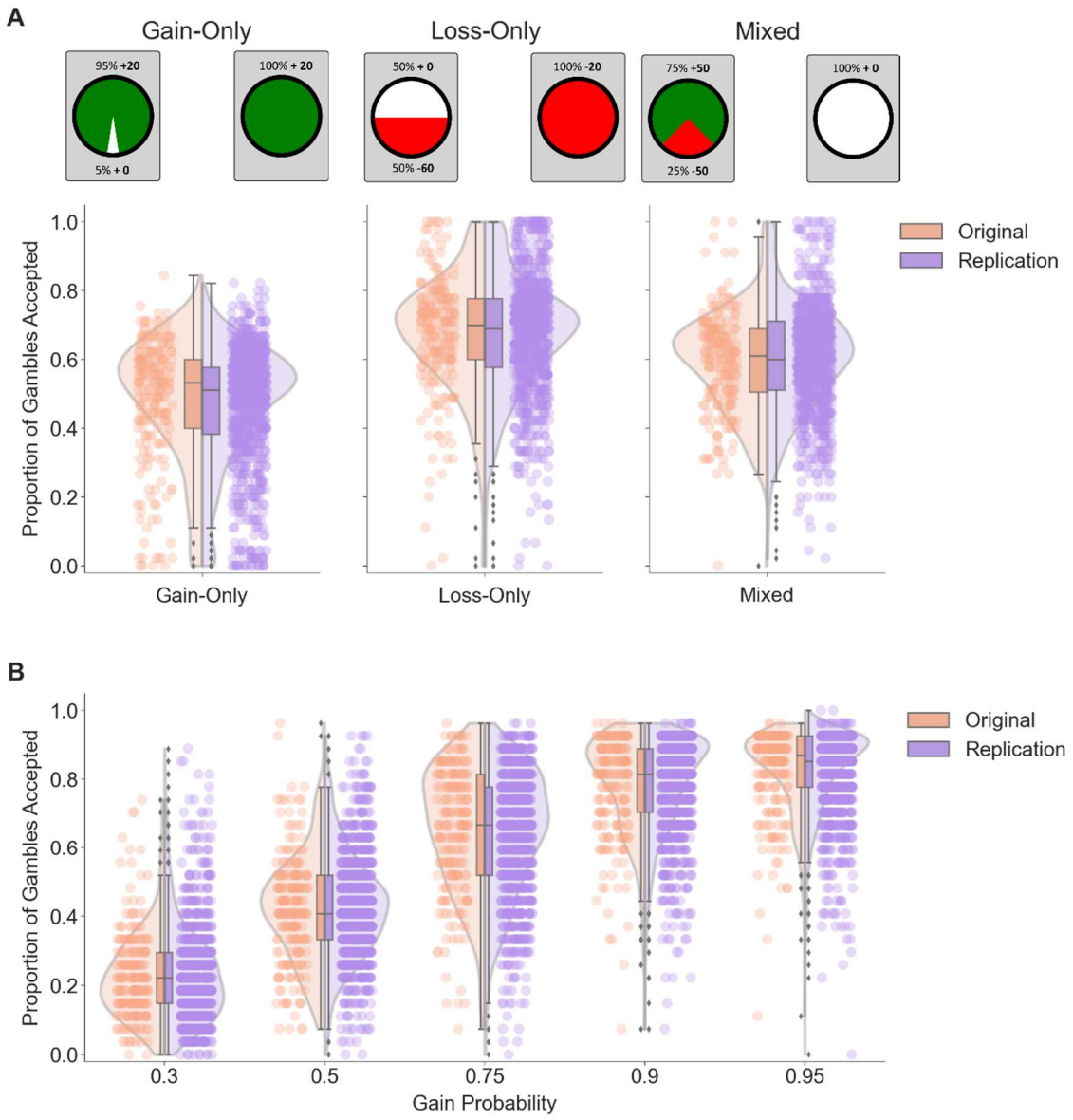

Figure 5.1 Gambling Task and Participants' Behaviour.

A. Example trials in which participants chose between a sure option and a gamble (top) and distribution of proportion of gambles accepted by each participant (bottom) for each trial type separated by study (original or replication). *Left*: Gain-only gamble. Participants chose between a sure option of winning a number of points (+10, +20, +30), or a gamble option in which there was an explicit probability of winning a number of points (+20, +40, +60) or of winning nothing. *Middle:* Loss-only gamble. Participants chose between a sure option of losing a number of points (-10, -20, -30), or a gamble option in which there was an explicit probability of losing nothing or of losing a number of points (-20, -40, -60). *Right:* Mixed gamble. Participants chose between a sure option of winning nothing, or a gamble option in which there was an explicit probability of winning a number of points (+40, +50, +60) or of losing a number of points (-10, -50, -90). The probabilities of the positive outcome in the gamble varied between 0.3, 0.5, 0.75, 0.9, and 0.95 and the negative outcome always had the complementary probability of occurring. Participants are most risk averse in the gain-only gamble condition, and most risk-seeking in the loss-only gamble condition, in line with classical Prospect Theory findings (Kahneman & Tversky, 1979). **B.** Distribution of proportion of gambles accepted by each participant for each gain probability separated by study (original or replication).

5.3.3 Abstract Reasoning Task

Participants completed a shortened version of the ART (Chierchia et al., 2019) as a proxy for general cognitive ability. On each trial, participants were asked to choose one out of four possible options that best completed a given pattern. Participants were given a maximum of 30 seconds to answer each problem, with a timer appearing for the last five seconds, after which the task progressed to the next problem. In this version of this task participants were asked to complete as many puzzles out of 40 as possible within four minutes, with the last trial being the one that began before the four-minute timer ends. A previous use of the shortened task confirmed that this version of the task preserves the psychometric properties of the original eight-minute task, such as a positive correlation between the ART task and digit span scores. Furthermore, in this previous analysis there were no floor or ceiling effects, a Cronbach's alpha of 0.6, and a split half reliability of 0.856, demonstrating acceptable internal validity (Harada-Laszlo et al., 2021).

5.3.4 Self-Report Questionnaires

Participants were additionally asked to provide their age and gender (which was coded as a binary variable with women: 0 and men: 1), and to complete questionnaires assessing catastrophising symptoms (Catastrophising Questionnaire, CAT (Pike et al., 2021)), anxiety symptoms (Generalised Anxiety Disorder 7-item Scale, GAD-7 (Spitzer et al., 2006)), depressive symptoms (Patient Health Questionnaire Depression Scale, PHQ (Spitzer et al., 1999)), trait anxiety symptoms (State Trait Anxiety Inventory – Trait Anxiety, STAI-T (Spielberger, 1970)), worry symptoms (Penn State Worry Questionnaire, PSWQ (Meyer et al., 1990)), and three questions to assess the impact of the COVID-19 pandemic (1: How confident are you of a return to a pre-pandemic life in the near future? 2: To what extent do you think your health would be affected if you catch the virus? 3: How much are pandemic-related lifestyle changes (e.g., mask wearing and social distancing) currently affecting your well-being?). As these questionnaires asked sensitive questions relating to mental health, participants were directed to local mental health charities in case they felt concerned about their answers.

5.3.5 Computational Models

We fit models (Table 5.1) to participants' trial-by-trial choices in the gambling task in order to capture their decision making and probability weighting patterns. We fit a number of models with different probability weighting functions and chose to use the best-fitting one for parameter inference to ensure that our model was as close to the true data-generating process as possible

All models use a prospect theory power function to evaluate gains and losses (Kahneman & Tversky, 2018):

$$V(gain) = gain^{\rho} \quad (5.1)$$

$$V(loss) = -\delta \times |loss|^{\rho} \quad (5.2)$$

Where V denotes the subjective value of a particular gain and loss, ρ is a risk aversion parameter, and δ is a loss aversion parameter (Figure 5.2A).

All models include the same function for calculating the expected value of each gamble:

$$EV(gamble) = \pi(p_{gain}) \times V(gain) + \pi(p_{loss}) \times V(loss) \quad (5.3)$$

Where π(p) is a probability weighting function that captures how participants subjectively weight objective probabilities seen on each trial. This weighting function is the key difference between the models.

All models use the same logit decision function to convert the expected values of the gamble option and sure option into probabilities of choosing each one on each trial:

$$p(choose\ gamble) = \frac{1}{1 + e^{-\gamma(EV(gamble) - EV(sure))}} \quad (5.4)$$

Where γ is an inverse temperature, or determinism, parameter that determines how noisy participants' choices are. Higher parameter values result in more deterministic choices.

As noted above, models differ in π(p) – the function that specifies how probabilities are weighted (Table 5.1). The first model is the Prospect Theory model suggested in Kahneman and Tversky's original prospect theory paper (Kahneman & Tversky, 1979), which does not include any probability weighting function. It therefore assumes that participants simply use the objective probabilities of the gamble options shown in the task when calculating expected values. For the other models with specified probability weighting functions, r is a curvature parameter that determines the non-linearity of the weighting function (Figure 5.2B). The Tversky and Kahneman (TK) weighting function was specified in their subsequent cumulative prospect theory paper (Tversky & Kahneman, 1992). At r = 1, the function is linear, and exhibits more rapidly diminishing sensitivity to probabilities at the boundaries (inverse S shape) with decreasing r (Figure 5.2B). Prelec I (PI) behaves similarly to the TK function, except that the fixed point is always 1 / e ≈ 0.36 (participants weight this probability objectively), and the function is also specified for r > 1, with a more marked S shape with increasing r (Prelec, 1998). The Goldstein-Einhorn (GE) model (Goldstein & Einhorn, 1987) includes an elevation parameter (s), where decreasing s < 1 indicates a general underweighting of probabilities (due to a high inflection point) and increasing s > 1 indicates an overweighting of probabilities (due to a low

inflection point). The 2-parameter function Prelec II (PII) behaves similarly to the GE function, though in this case, as s increases from 1, individuals underweight probabilities (Prelec, 1998). Further, the GE function produces more drastic changes in probability weighting for equivalent changes in parameter values, and the over or under-weighting is more pronounced for smaller probabilities.

We fit three different versions of each probability weighting model above, in which the weighting function for gains and losses were specified in different forms (Figure 5.2C):

a) probability weighting function fit to gains with the probability of losses defined as the complementary probability: p(loss) = 1 – p(gain).
b) probability functions and parameters (elevation parameter only for 2-parameter functions) are fit separately for gains and losses.
c) single probability weighting function and parameters for all probabilities - implying all probabilities are weighted similarly regardless of their valence.[4]

For the two-parameter probability weighting functions, two of these versions (b and c above) did not reach convergence during model fitting, and the parameter recovery was poor. This suggests that they were over parameterised, and the parameters were not identifiable (Table 5.2). Thus, the resulting set of models included one model with no probability weighting function (PT), six models with a one-parameter weighting function (TKa, TKb, TKc, PIa, PIb, PIc), and two models with a two-parameter probability weighting function (GEa, PIIa), giving a total of nine models (see Table 5.2 for a list of parameters in each model).

[4] This version was not specified in the pre-registration but was fit to all models for completeness.

Table 5.1 Probability Weighting Functions and Versions.

Top: Name, mathematical specification, and parameter constraints for the different types of probability weighting functions tested. Model 1 does not include a weighting function but uses objective probabilities instead. The weighting functions in Models 2 and 3 are specified with one parameter (r), whilst those in Models 4 and 5 are specified with two parameters (r, s), where r is a curvature parameter and s is an elevation parameter. The models have been named after the authors that first seemed to report them (Stott & Stott, 2006). **Bottom:** Name, and mathematical specifications for different versions of probability weighting models. In version a, we fit a probability weighting function for gains π with the probability of losses defined as the complementary probability. In version b, we fit separate probability weighting functions for gains and losses (π_{gain}, π_{loss}) with the probability of losses defined with its own parameters. In version c, the same probability weighting function π is used to weight gains and losses. p_{gain} and $1 - p_{gain}$ denote the objective probabilities of the positive and negative gamble outcomes respectively.

	Name	Probability Weighting Function	Parameter Constraints
1.	Prospect Theory (PT)	$\pi(p) = p$ (5.5)	
2.	Tversky-Kahneman (TK)	$\pi(p) = \frac{p^r}{(p^r + (1-p)^r)^{1/r}}$ (5.6)	$0 \leq r \leq 1$
3.	Prelec I (PI)	$\pi(p) = e^{-(-\ln p)^r}$ (5.7)	$r \geq 0$
4.	Goldstein-Einhorn (GE)	$\pi(p) = \frac{s\, p^r}{s\, p^r + (1-p)^r}$ (5.8)	$r, s \geq 0$
5.	Prelec II (PII)	$\pi(p) = e^{-s(-\ln p)^r}$ (5.9)	$r, s \geq 0$
	Version	**Gain Probability Weighting**	**Loss Probability Weighting**
1.	*a*	$\pi(p_{gain})$	$1 - \pi(p_{gain})$
2.	*b*	$\pi_{gain}(p_{gain})$	$\pi_{loss}(1 - p_{gain})$
3.	*c*	$\pi(p_{gain})$	$\pi(1 - p_{gain})$

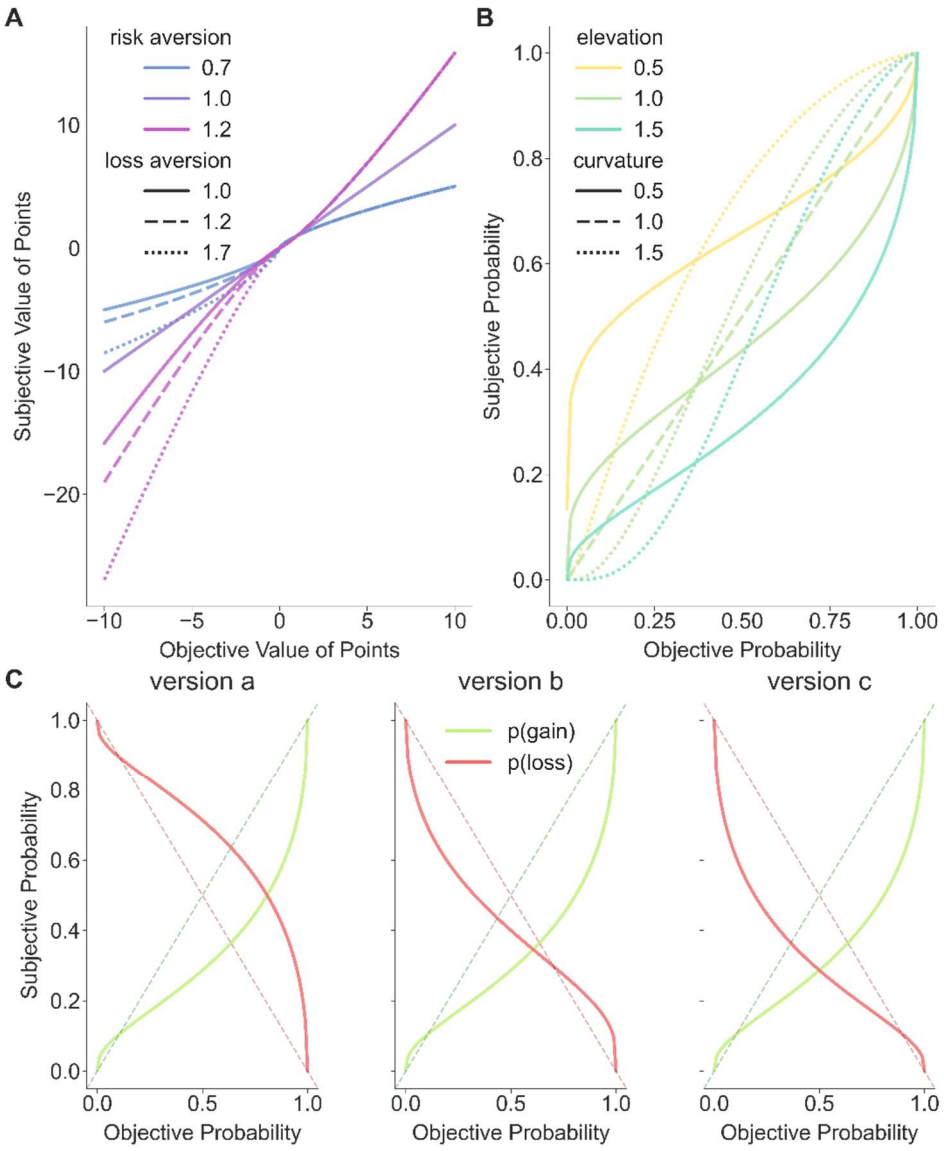

Figure 5.2 Cumulative Prospect Theory Model.

A. Value Functions. Line colour indicates how the subjective value of points changes for different values of the risk aversion parameter (ρ). Line pattern indicates how the subjective value of points changes for different values of the loss aversion parameter (δ). **B.** Probability Weighting Function. Line colour indicates how the subjective probability changes for different values of the elevation parameter (s) which affects the over or underweighting of probability. Line pattern indicates how the subjective probability changes for different values of the curvature parameter (r) which affects the curvature of the probability function. Models with one-parameter weighting functions (r only), have a fixed inflection point (as s = 1). **C.** Model Versions. Model versions all fit probability weighting parameters for gains but differ in how the probability of losses, p(loss), is specified. In version a (left), p(loss) is determined to be 1-p(gain). Subjective probabilities retain complementarity, shown by the functions being reflected in the y = 0.5 line. In version b (middle), p(gain) and p(loss) are specified by distinct parameters offering the flexibility of weighting gains and losses differently. In version c (right), the same parameters are used to specify the weighting function for gains and losses. p(loss) is equivalent to p(gain) for the same objective probability of loss/gain, shown by the functions being reflected in the x = 0.5 line. Dashed lines indicate objective probability weighting functions for gains (green) and losses (red) in a gamble. All panels in this figure use the Prelec IIa weighting function.

Table 5.2 Gambling Task Models Parameter Recovery.

Parameter recovery for each parameter in each of the 13 models tested. Recovery was assessed by simulating data for 300 participants with parameters drawn randomly from the estimated prior distribution (when fit to real data). The best-fitting (recovered) parameters for these data sets were found using the Expectation Maximisation procedure, and parameter recoverability is indicated by the Pearson's correlation between the simulated and recovered parameters. TK: Tversky-Kahneman, PI: Prelec I, GE: Goldstein-Einhorn, PII: Prelec II.

Model	Parameter	Recovery
Prospect Theory	Risk aversion (ρ)	0.75
	Loss aversion (∂)	0.88
	Choice determinism (β)	0.99
TKa	Risk aversion (ρ)	0.95
	Loss aversion (∂)	0.92
	Curvature (r)	0.93
	Choice determinism (β)	0.95
TKb	Risk aversion (ρ)	0.95
	Loss aversion (∂)	0.94
	Curvature gains (r_gain)	0.93
	Curvature losses (r_loss)	0.77
	Choice determinism (β)	0.96
TKc	Risk aversion (ρ)	0.96
	Loss aversion (∂)	0.93
	Curvature (r)	0.90
	Choice determinism (β)	0.93
PIa	Risk aversion (ρ)	0.96
	Loss aversion (∂)	0.89
	Curvature (r)	0.95
	Choice determinism (β)	0.93
PIb	Risk aversion (ρ)	0.96
	Loss aversion (∂)	0.90
	Curvature gains (r_gain)	0.88
	Curvature losses (r_loss)	0.86
	Choice determinism (β)	0.94
PIc	Risk aversion (ρ)	0.95
	Loss aversion (∂)	0.88
	Curvature (r)	0.91
	Choice determinism (β)	0.96
GEa	Risk aversion (ρ)	0.97
	Loss aversion (∂)	0.87
	Curvature (r)	0.85
	Elevation (s)	0.94
	Choice determinism (β)	0.93
GEb	Risk aversion (ρ)	0.03
	Loss aversion (∂)	-0.03
	Curvature (r)	0.76
	Elevation gains (s_gain)	0.08
	Elevation losses (s_loss)	0.76
	Choice determinism (β)	0.25
GEc	Risk aversion (ρ)	0.11
	Loss aversion (∂)	-0.05
	Curvature (r)	0.88
	Elevation (s)	0.66
	Choice determinism (β)	0.50
PIIa	Risk aversion (ρ)	0.96
	Loss aversion (∂)	0.89
	Curvature (r)	0.91
	Elevation (s)	0.87
	Choice determinism (β)	0.93
PIIb	Risk aversion (ρ)	0.30
	Loss aversion (∂)	0.59
	Curvature (r)	0.61
	Elevation gains (s_gain)	0.27
	Elevation losses (s_loss)	0.44
	Choice determinism (β)	0.19
PIIc	Risk aversion (ρ)	0.57
	Loss aversion (∂)	0.47
	Curvature (r)	0.85
	Elevation (s)	0.23
	Choice determinism (β)	0.46

5.3.6 Parameter Estimation and Recovery

Parameters were estimated as described in 2.3 Parameter Estimation and parameter recovery was assessed for each model (Table 5.2)

5.3.7 Statistical Inference

We used t-tests to assess whether the proportion of gambles accepted differed across conditions in the gambling task. We used Pearson's correlations to assess the effect size of the relationship between model parameters of interest and symptoms of interest and used linear regression to assess whether these relationships remained significant when accounting for relevant covariates. The scipy.stats package in Python was used to carry out t-tests, Pearson's correlations and linear regressions, and we set α to 0.05 for all inference, unless otherwise specified.

5.3.8 Sensitivity Analysis

Attention checks were determined from pilot data and were based on participants' performance on so-called 'easy trials', where the safe or gamble option was clearly an optimal choice (e.g., none of the gamble options give a better outcome than the safe option). There are two types of easy trials, and we used a cut-off of 50% incorrect on both trial types to identify inattentive participants. Additionally, participants who failed the attention quiz more than five times were classed as inattentive. Importantly, participants were not excluded from the primary analysis based on these attention checks, but the analysis was rerun without these participants to assess how this affected our conclusions.

5.4 Results

5.4.1 Descriptive Statistics

Our initial sample consisted of 212 participants of which 135 (64%) were women, with an age range of 18-58 years old. Figure 5.1A shows a summary of the proportion of gambles accepted in each condition of the gambling task data from our sample. Participants were most risk averse in the gain-only gamble condition, and most risk-seeking in the loss-only gamble condition (gain vs loss: $t(211) = -11.69$, $p = 1.19 \times 10^{-24}$; gain vs mixed: $t(211) = -7.83$, $p = 2.40 \times 10^{-13}$; loss vs mixed: $t(211) = 7.00$, $p = 3.54 \times 10^{11}$). This is consistent with results reported by Kahneman and Tversky: people are less willing to forego a sure gain for the prospect of a higher gain (risk averse in gain-only gambles), but more willing to gamble and potentially incur a greater loss for the possibility of losing nothing at all (risk-seeing in loss-only gambles) (Kahneman & Tversky, 1979).

5.4.2 Model Fitting

To test our prediction that the best-fitting model would be one with a two-parameter weighting function and distinct parameters for gains and losses, we compared model fit using two quantitative measures: average likelihood per trial and integrated Bayesian Information Criterion (Figure 5.3A). The models including probability weighting functions generally fit the participants' data better than the prospect theory model (PT, no probability weighting parameters; Table 5.1), suggesting that participants did not weight probabilities objectively. For the one-parameter models (TK and PI), version a consistently fit the data better than versions b and c (which also did not converge for the two-parameter models). This suggests that participants did not use distinct weighting functions for gains and losses (version b), nor the same one (version c) but rather that whilst weighting of probabilities is not objective, complementarity is retained when weighing up the possible outcomes of a gamble. This suggests that the probabilities of good outcomes are over or underweighted relative to the bad ones (version a; Figure 5.2C). Finally, based on both average likelihood per trial, and iBIC scores, the two parameter weighting functions (including an elevation parameter) fit the data more parsimoniously than the one parameter weighting functions, suggesting that a more precise specification of the weighting function is important for capturing the subtleties of risky decision making. In particular, the Prelec IIa (PIIa) model fits marginally better than the Goldstein-Einhorn model based on the iBIC score (Figure 5.3A). Qualitative model fits of the Prelec IIa model are shown in Figure 5.3B, where average correlations between the proportion of gambles accepted from real and model-simulated data are above 0.85 for all three gamble conditions.

Due to the strong overall fit to the real data, we selected the Prelec IIa model for our subsequent analysis and inference on the five model parameters: risk aversion, loss aversion, curvature (probability weighting), elevation (probability weighting) and determinism.

5.4.3 No relationship between model parameters and catastrophising

We were particularly interested in the relationship between catastrophising symptoms and probability weighting parameters. Our winning model, Prelec IIa, has two probability weighting parameters: one specifying the curvature of the weighting function and one specifying its elevation. We hypothesised that catastrophising questionnaire scores would be associated with higher values of the elevation parameter, s, as this suggests a general underweighting of the probability of gains, and a general overweighting of the probability of losses. However, there was no relationship between catastrophising scores and the elevation probability weighting parameter ($r = 0.07$, $p = 0.30$, Figure 5.4A), or indeed any other parameters in our winning model, including the curvature probability weighting parameter. This suggests that contrary to our predictions, there is no evidence that

catastrophising symptoms are related to how the probabilities of gambles outcomes are subjectively weighted.

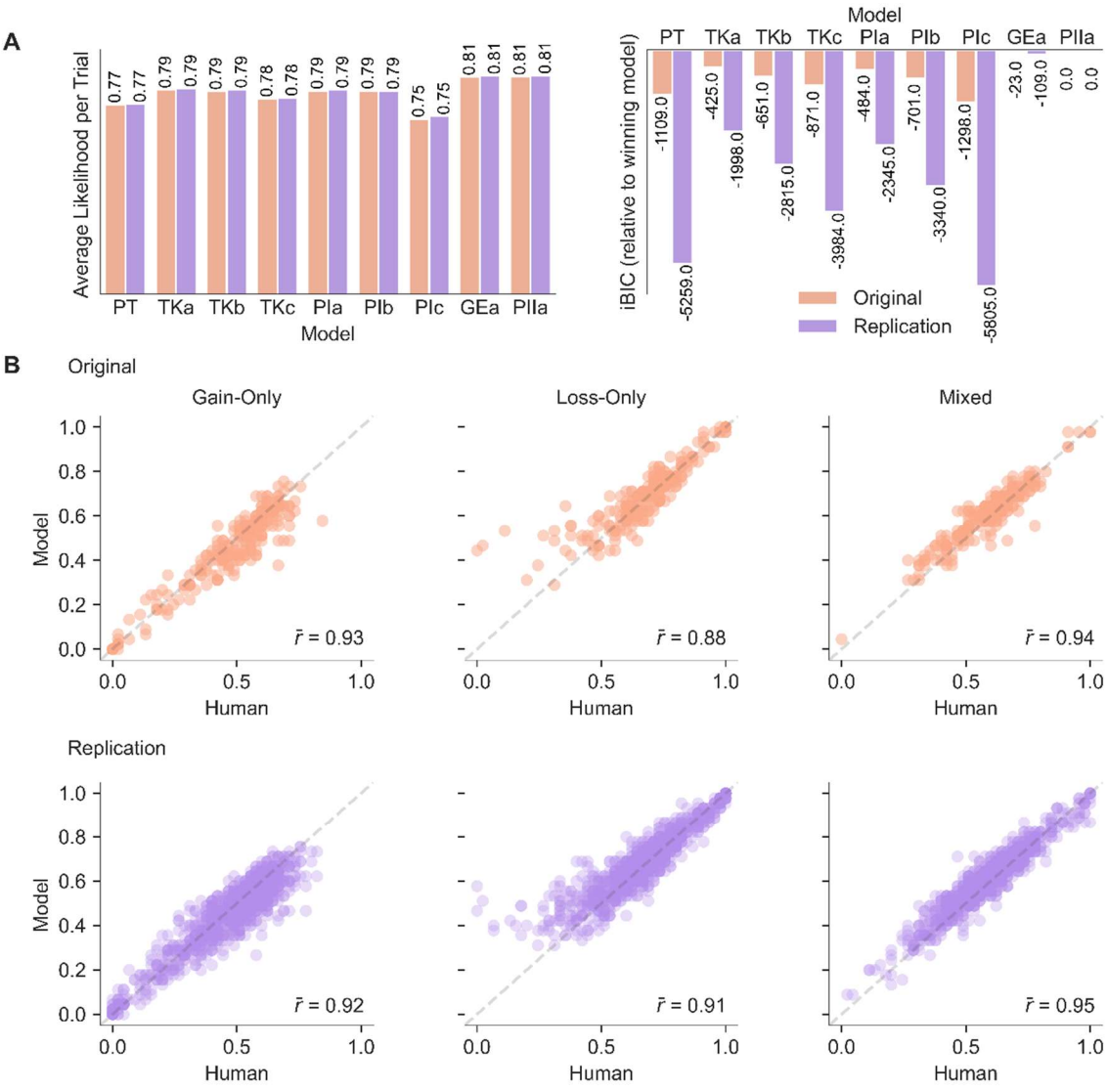

Figure 5.3 Gambling Task Model Comparison and Fit.

A. Qualitative Model Comparison. Average Likelihood per Trial (left) and iBIC (right) for each of the 9 valid models (for which the EM algorithm converged) that were tested for both the original and replication datasets. **B.** Qualitative Model Fit. Scatterplots of the proportion of gambles accepted by each subject from real and a model-simulated dataset in the gain-only (left), loss-only (middle) and mixed (right) conditions of the gambling task in the original (top) and replication (bottom) datasets. The correlation between real and model data was calculated for 10 different simulations, and the average of these ($\bar{r}$) is shown on each plot. PT: Prospect Theory, TK: Tversky-Kahneman, PI: Prelec I, GE: Goldstein-Einhorn, PII: Prelec II.

5.4.4 Evidence that anxiety is associated with overweighting the probabilities of negative outcomes

Due to our hypotheses involving the elevation probability weighting parameter, we conducted exploratory analyses to investigate the relationships between the elevation parameter and other mental health symptoms. We found a significant positive relationship between the elevation parameter from the winning model and GAD-7 scores ($r = 0.18$, $p = 0.01$, Figure 5.4A) suggesting that those who are more anxious overweight the probability of the negative gamble outcome and underweight the probability of the positive gamble outcome. There were no relationships between the elevation parameter of the winning model and any other symptom questionnaires. To test whether this finding was influenced by other symptom measures or demographic variables, we carried out a multiple regression predicting the elevation parameter from the GAD-7 questionnaire, all other symptom questionnaires, age, gender and performance on the ART task. GAD-7 scores were the only significant predictor of the elevation parameter (GAD-7: $\beta = 0.028$, 95% CI = [0.011, 0.045], $p = 0.001$; overall model $F(9,201) = 1.545$, $R^2 = 0.065$; Figure 5.4B). This provides preliminary evidence that general anxiety is related to overweighting the probabilities of negative gamble outcomes, particularly as the relationship between these variables is in the expected direction (overweighting the probability of negative outcomes could plausibly lead to anxiety or avoidance behaviours). In our sensitivity analysis, in which we excluded participants that did not pass the attention checks during the task, the results of the model comparison were unchanged. The relationship between GAD-7 and the elevation parameter remained significant ($r = 0.15$, $p = 0.03$), and GAD-7 remained the only significant predictor of the elevation parameter in the multiple regression covarying for age, gender, and other mental health symptoms (GAD-7: $\beta = 0.026$, 95% CI = [0.009, 0.043], $p = 0.003$; overall model $F(9,192) = 1.700$, $R^2 = 0.074$; Figure 5.5).

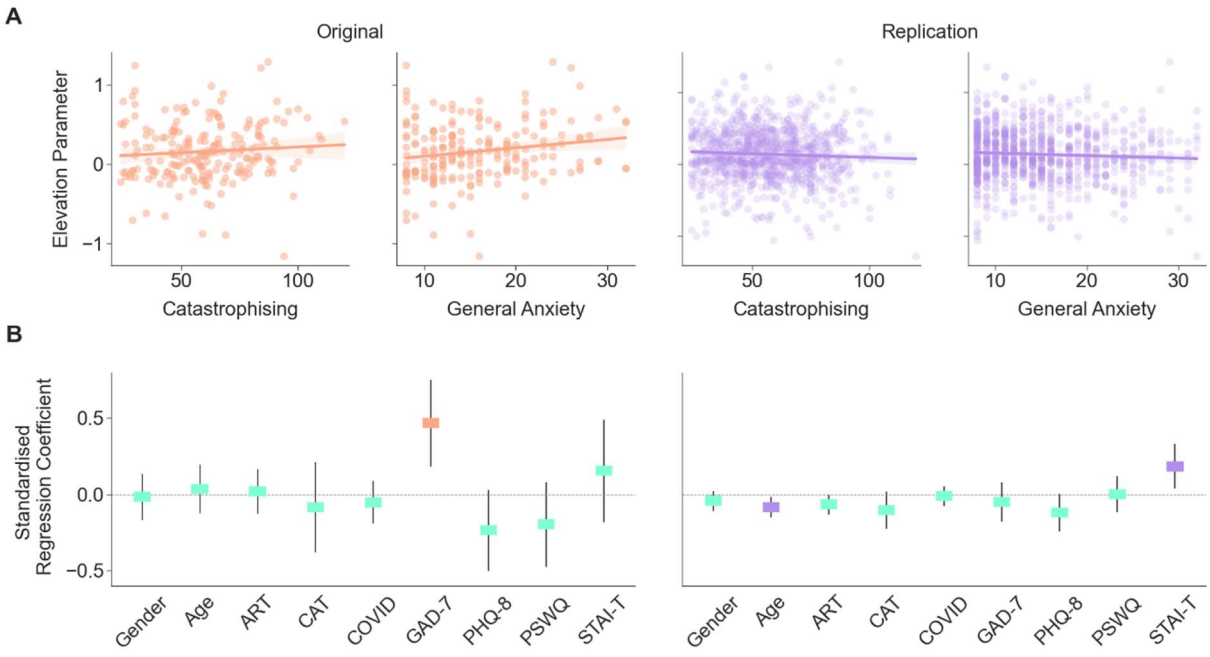

Figure 5.4 Relationship Between Symptoms and the Probability-Weighting Elevation Parameter.

A. Scatterplots showing the relationships of catastrophising symptoms and general anxiety symptoms with the elevation parameter (s) from the Prelec IIa model in the original (left) and replication (right) data sets. **B.** Regression coefficients and 95% confidence intervals from a multiple regression in which Gender, Age, ART, CAT, COVID, GAD-7, PHQ-8, PSWQ, and STAI-T predict the elevation parameter in the original (left) and replication (right) datasets. Standardised regression coefficients are used for plotting to aid visualisation. Orange/purple shade indicates significant predictors ($p < 0.05$). ART: Abstract Reasoning Task, CAT: Catastrophising Questionnaire, COVID: Covid Questionnaire, GAD-7: Generalised Anxiety Disorder 7-item scale, PHQ-8: Patient Health Questionnaire 8-item scale, PSWQ: Penn State Worry Questionnaire, STAI-T: State-Trait Anxiety Index Trait Anxiety.

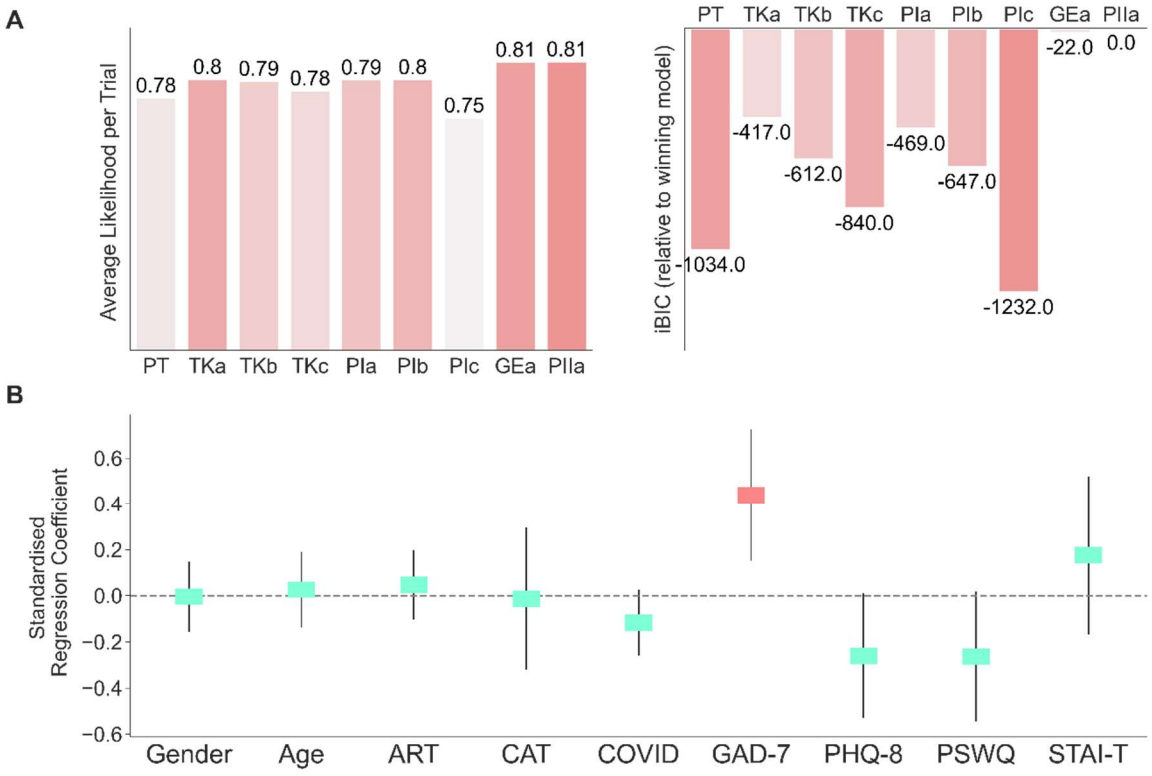

Figure 5.5 Results from Sensitivity Analysis of Gambling Task Data in the Original Dataset.

Results of analyses that exclude 10 participants who failed our pre-specified attention checks. **A.** Qualitative Model Comparison. Average Likelihood per Trial (left) and iBIC (right) for each of the nine valid models that were tested. **B.** Regression coefficients and 95% confidence intervals from a multiple regression in which Gender, Age, ART, CAT, COVID, GAD-7, PHQ-8, PSWQ, and STAI-T predict the elevation parameter. Standardised regression coefficients are used for plotting to aid visualisation. Pink shade indicates significant predictors ($p < 0.05$). ART: Abstract Reasoning Task, CAT: Catastrophising Questionnaire, COVID: Covid Questionnaire, GAD-7: Generalised Anxiety Disorder 7-item scale, PHQ-8: Patient Health Questionnaire 8-item scale, PSWQ: Penn State Worry Questionnaire, STAI-T: State-Trait Anxiety Index Trait Anxiety.

We also examined which model parameters were related to age, gender and ART scores. Age was negatively associated with loss aversion ($r(210) = -0.18$, $p = 0.0069$); whilst women had lower curvature probability parameters which indicates greater underweighting of probabilities associated with good outcomes ($t(209) = 2.36$, $p = 0.019$, Cohen's $d = 0.34$); and participants with higher ART scores also had higher curvature parameters (indicating more objective probability weighting), as well as higher risk aversion parameters (indicating lower risk aversion), higher loss aversion and less deterministic choices (curvature: $r(210) = 0.20$, $p = 0.0034$; risk aversion: $r(210) = 0.16$, $p = 0.018$; loss aversion: $r(210) = 0.31$, $p = 5.06 \times 10^{-6}$; determinism: $r(210) = -0.18$, $p = 0.0083$). After applying a

Bonferroni correction to account for the 15 comparisons performed (α = 0.0033), the relationship between ART scores and loss aversion remained significant.

5.4.5 Model fitting and relationships with demographic variables replicate, whilst relationships with mental health symptoms do not

In order to confirm whether general anxiety is related to the probability weighting parameter from our model, we carried out a replication study. The recruitment criteria, tasks and questionnaires were exactly the same as for the original sample (as specified in Methods). Our replication sample consisted of 946 participants of which 473 (50%) were women, with an age range of 18-61 years old. Figure 5.1 shows that the behaviour of participants in both samples on the gambling task is consistent. Figure 5.3A shows that based on the average likelihood per trial and the iBIC scores, the PIIa is the best fitting model for the replication data set, consistent with the results from the model comparison in the original sample. However, unlike in the original sample, when examining model parameters from the PIIa model, there was no significant association between general anxiety and the elevation probability weighting parameter ($r(944) = -0.06$, $p = 0.074$; Figure 5.4A), nor was general anxiety a significant predictor of the elevation parameter in a multiple regression that used other mental health symptoms as covariates (Figure 5.4B). In fact, all correlations between model parameters and symptoms of mental health questionnaires were below r = 0.1 (Figure 5.6), suggesting that even significant relationships found in this sample do not have a large enough effect size to be of interest, or meaningful clinical relevance.

We also examined which model parameters were related to age, gender and ART scores in our replication sample. Age was again negatively associated with loss aversion, as well as the curvature probability parameter, but positively associated with determinism (curvature: $r(944) = -0.15$, $p = 1.80 \times 10^{-6}$; loss aversion: $r(944) = -0.18$, $p = 1.85 \times 10^{-8}$; determinism: $r(944) = 0.12$, $p = 0.00030$); whilst women again had lower curvature probability parameters as well as lower risk aversion parameters (indicating higher risk aversion) (curvature: $t(944) = 7.79$, $p = 1.77 \times 10^{-14}$, Cohen's $d = 0.51$; risk aversion: $t(944) = 6.48$, $p = 1.51 \times 10^{-10}$, Cohen's $d = 0.42$); and participants with higher ART scores again had higher curvature parameters, higher risk aversion parameters (indicating lower risk aversion), higher loss aversion, and less deterministic choices (curvature: $r(944) = 0.13$, $p = 7.05 \times 10^{-5}$; risk aversion: $r(944) = 0.13$, $p = 6.21 \times 10^{-5}$; loss aversion: $r(944) = 0.14$, $p = 1.49 \times 10^{-5}$; determinism: $r(944) = -0.10$, $p = 0.0018$). After applying a Bonferroni correction to account for the 15 comparisons performed (α = 0.0033), all of these relationships remained significant.

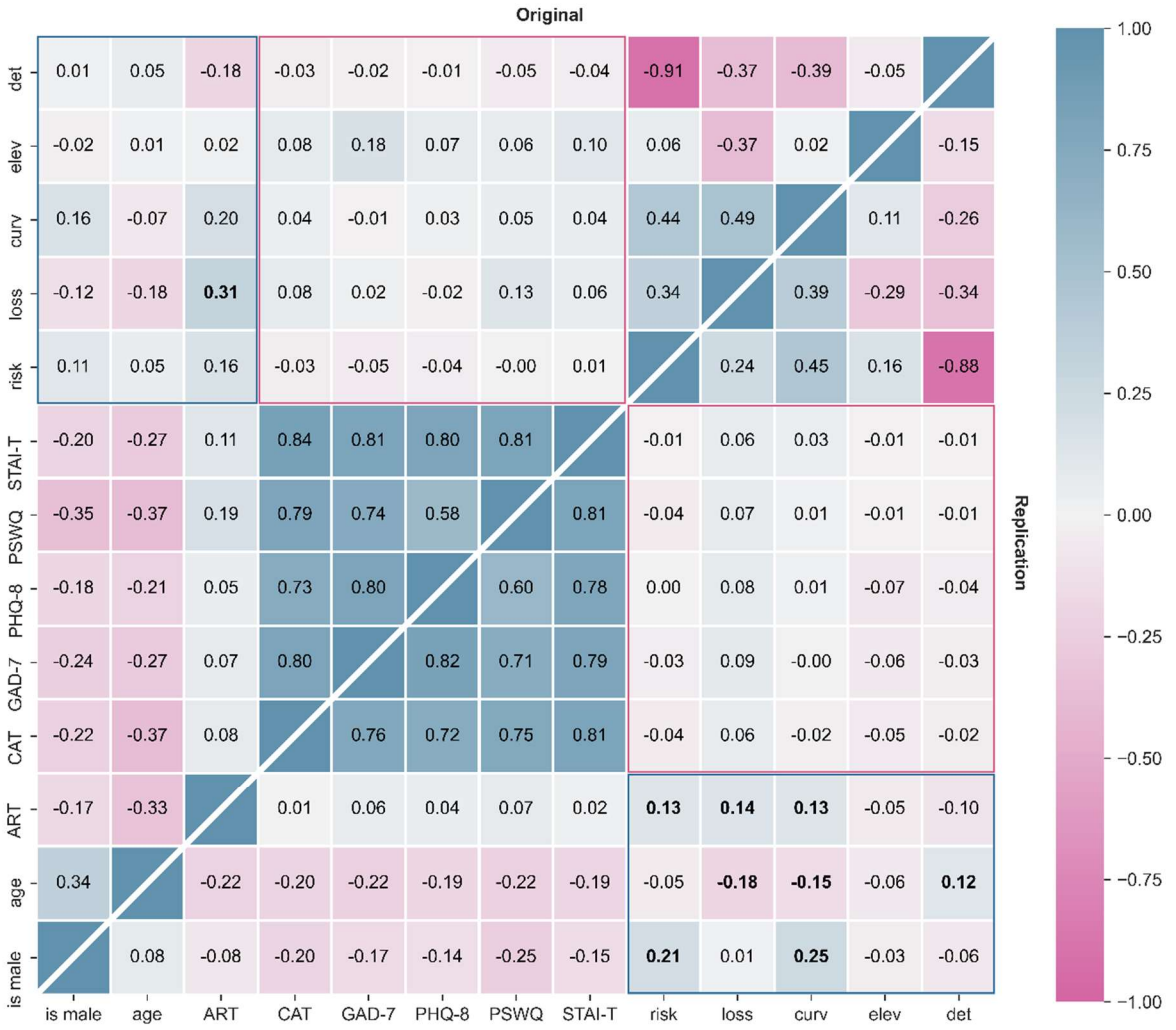

Figure 5.6 Relationships Between Key Variables in Original and Replication Datasets.

The relationships for the original dataset are shown in the upper triangle, whilst the relationships for the replication dataset are shown in the lower triangle. For comparability Cohen's d has been transformed to Pearson's r for the gender difference. Values in bold are significant after correcting for the multiple comparisons indicated by the coloured rectangles. ART: Abstract Reasoning Task, CAT: Catastrophising Questionnaire, COVID: Covid Questionnaire, GAD-7: Generalised Anxiety Disorder 7-item scale, PHQ-8: Patient Health Questionnaire 8-item scale, PSWQ: Penn State Worry Questionnaire, STAI-T: State-Trait Anxiety Index Trait Anxiety, risk: risk aversion parameter, loss: loss aversion parameter, curv: curvature parameter, elev: elevation parameter, det: determinism parameter.

To examine the differences between the two datasets, Figure 5.6 shows all correlations between key variables for both samples. Whilst the key relationship of interest, between general anxiety and the elevation parameter, did not replicate between samples, the overall structure of the relationships between variables seems consistent over both samples, as shown by the symmetry in shading

between the upper and lower triangles. More specifically, there are very high correlations between all symptom questionnaire measures, relatively high correlations between some model parameters, evidence of relationships between demographic variables and questionnaire measure or model parameters, but remarkably low correlations between model parameters and symptom questionnaires in general. This is worth noting, as it is these last relationships that cognitive neuroscience and mental health research typically focuses on, setting up hypotheses about behavioural signatures of mental health symptoms and carefully designing experiments to test these. Low correlations between questionnaire measures and mental health symptoms have been reported for other cognitive functions (Snyder et al., 2021), suggesting that format of assessment has a large influence on how an individual might respond. Due to the high correlation between age and gender in the first dataset, we carried out a multiple regression with the demographic variables as predictors for each model parameter in both datasets (Figure 5.7). This allows us to make comparisons about the relationships with parameters between the datasets more accurately. We further examined the differences between our datasets by visualising the distributions of key variables between the two samples (Figure 5.8 & Figure 5.9). It is noteworthy that four out of our six questionnaire measures are significantly different between the two samples: the catastrophising questionnaire, COVID questionnaire, PHQ-8 and GAD-7 (Figure 5.8). In all cases, the replication sample is skewed to lower scores, indicating less severe symptoms in this sample. Furthermore, the loss aversion and curvature parameters are significantly greater in the replication sample than the original sample.

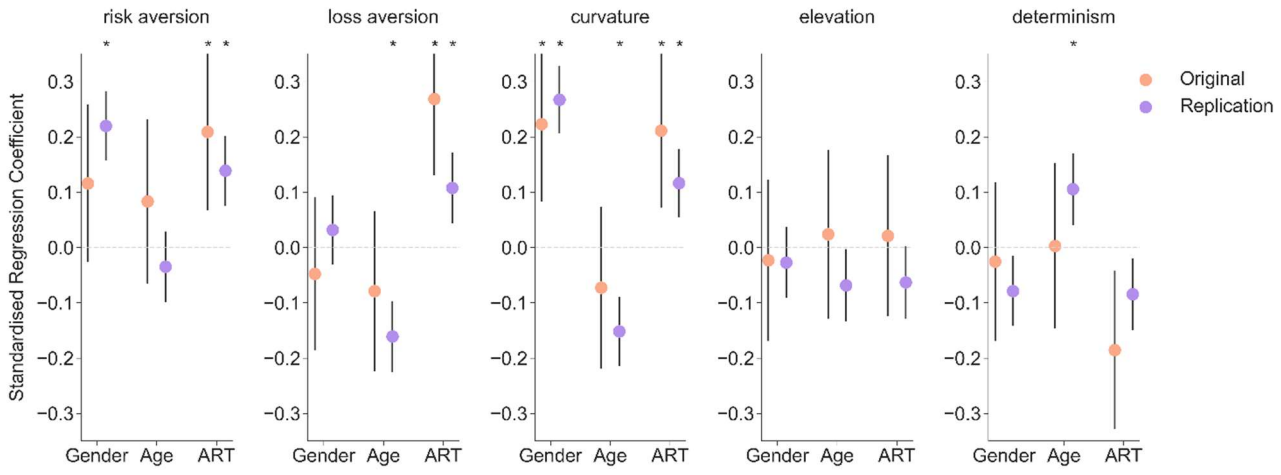

Figure 5.1 Multiple regression with Demographic Variables Predicting Each Model Parameter.
Standardised regression coefficients are plotted to aid visualisation. * indicates significant predictors after correcting for the 6 comparisons performed for each parameter ($p < 0.0083$). ART: Abstract Reasoning Task.

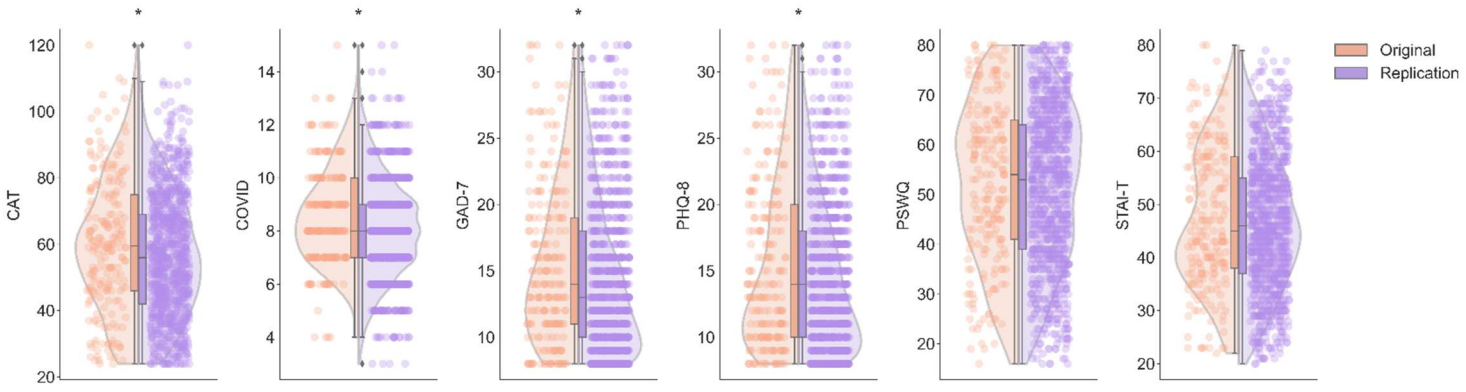

Figure 5.2 Distributions of Mental Health Symptom Questionnaires Between Datasets.

The y axes demonstrates the range of the data. * indicates significant t-test ($p < 0.05$). CAT: Catastrophising Questionnaire, COVID: Covid Questionnaire, GAD-7: Generalised Anxiety Disorder 7-item scale, PHQ-8: Patient Health Questionnaire 8-item scale, PSWQ: Penn State Worry Questionnaire, STAI-T: State-Trait Anxiety Index Trait Anxiety.

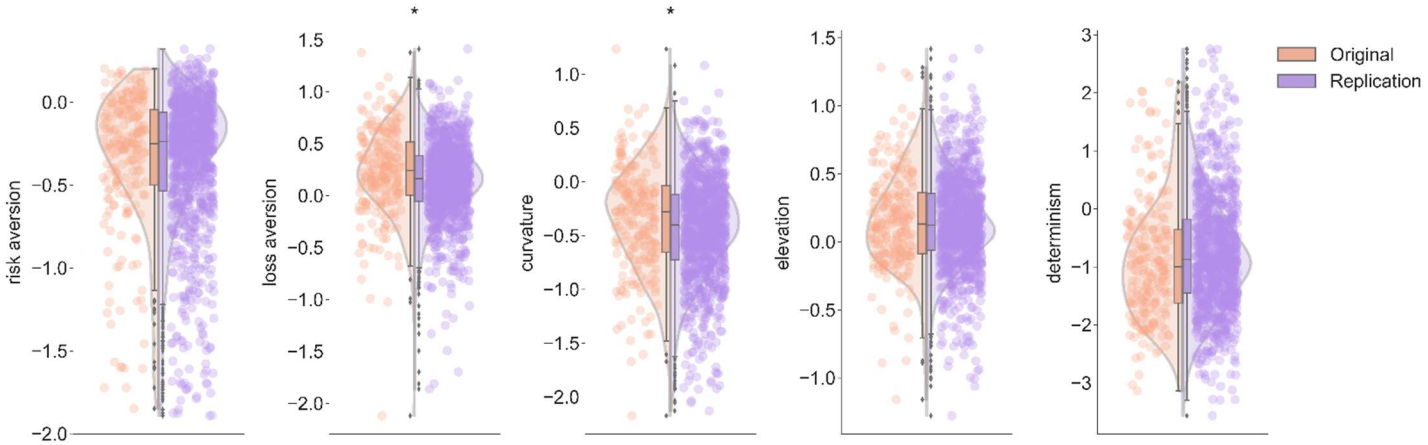

Figure 5.3 Distributions of Untransformed Model Parameters from the Winning Model Between Datasets.

The y axes demonstrates the range of the data. The untransformed parameters are distributed with a standard multivariate normal distribution. * indicates significant t-test ($p < 0.05$).

5.5 Discussion

In this study, we used a computational analysis of a gambling task to investigate the cognitive mechanisms underlying catastrophising, specifically whether it is related to difference in subjective probability weighting. To do this we fit a Prospect Theory model, and eight different Cumulative Prospect Theory models (which varied in their specification of the probability weighting function) to participants' data. The best fitting model included two probability weighting parameters (curvature and elevation). In this model the probabilities of gamble outcomes retain complementarity, such that

subjective probability weighting only affected whether good or bad gamble outcomes were overweighted relative to each other. Whilst we did not find a relationship with catastrophising, we found some preliminary evidence for a relationship with general anxiety in a discovery sample, though this was not evident in our larger replication sample. The results from the model comparison and the demographic variables (older people are more loss averse, women are less sensitive to differences in probability, and those with higher general cognitive ability are more sensitive to differences in probability, less risk averse, more loss averse, and make less deterministic choices) largely replicated across datasets, along with the overall structure of relationships between variables. The relationships between model parameters and mental health questionnaires were markedly low, compared to relationships between symptom questionnaires and also between symptom questionnaires and demographic variables.

Probability weighting models have not been extensively explored in the realm of mental health research, and as many different probability weighting functions exist in the literature, we tested how well a number of them fit to our data. The winning model (Prelec IIa) was the same in both of our datasets. This weighting function is specified by two parameters – curvature and elevation - which describe the probability of the positive gamble outcome, while the probability of the negative gamble is specified as the complementary probability. This suggests that the complementary of probabilities is generally retained during decision making, but that people may tend to pay more attention to either the good or the bad potential outcome, and therefore over- or under-weight this outcome relative to the other. The fact that most participants had elevation parameters between 0 and 1 suggests that the majority of participants overweight the probability of the positive gamble outcome relative to the negative one. Probability weighting has long been posited as an important part of risky decision making as specified in Cumulative Prospect Theory (Tversky & Kahneman, 1992), and our findings indicate that this model component is important in generating good fits to participant choices. Future studies investigating risky decision making should ensure they test these models, as they better explain participants' choices and also allow researchers to distinguish between different hypotheses.

Our primary hypothesis was that catastrophising would be related to probability weighting. However, in both samples, we found no significant relationship between catastrophising and probability weighting. We found a potential relationship between anxiety and probability weighting in our initial sample, but we failed to replicate this result in a larger sample. Moreover, all relationships between mental health symptoms and model parameters in our replication sample were weak (less than r = 0.1). Due to the large sample size in the second dataset, it is therefore tempting to speculate that there is little association between risky decision making and symptoms of anxiety and depression in

the general population, especially since the relationships with demographic variables did mostly replicate.

There are a few potential reasons for the different results between these samples that are worth mentioning. It is plausible that due to the relatively smaller sample size in the original dataset, the reported association between anxiety and probability weighting was due to noise - setting alpha at 0.05 means that if the null is true, 5% of the time one would expect to observe an effect of that magnitude or greater by chance. Further, it is always possible that there is sampling bias in online studies due to the little control we have over participants that choose to participate in research. One noteworthy difference between the two datasets is the unusually high correlation ($r = 0.34$) between age and gender in the original dataset, which close to disappears in the replication dataset. This is likely to be due to the timing of data collection between the two datasets, specifically as the first dataset was collected towards the end of July 2021, when a viral TikTok led to an unusual influx of young women entering into Prolific Academic studies around this time, skewing the demographics of the dataset. This skew is likely to have led to some of the results reported in the original dataset, such as the stronger relationships between age and gender with mental health questionnaires as compared to the replication dataset. Some modest differences are evident between the two datasets when examining the distributions of mental health symptoms, as the replication dataset is generally skewed to exhibiting less severe symptoms, which is significant in three of the five symptom questionnaires. We also found significantly lower scores in the replication sample for the COVID impact questionnaire. The timing of data collection between the two datasets (original: July 2021, replication: January 2022) coinciding with different stages in the COVID-19 pandemic may have led to this difference in questionnaire scores. While it is difficult to pinpoint the exact changes that might have led to this difference, due to participants being located in the US or UK and therefore governed by different national and international regulations around socialising and travel, there was a general relaxation of lockdowns and higher rates of vaccination by early 2022 that might have led to the observed differences. Despite these slight differences in symptom questionnaire scores between the datasets, the overall pattern of relationships between variables is largely similar, with a stark lack of relationships between model parameters and mental health symptoms.

This dearth of associations between parameters and symptom scores in the replication sample is somewhat surprising given the existing literature on risky decision making and mental health; however, there are some differences between our studies and previous studies that could explain these discrepancies. First of all, previous studies often included patients, whereas online studies typically aim to recruit samples that are representative of the general population. Whilst the latter

approach is beneficial for studying the variation of mental health symptoms that exist in the wider population at a subclinical level, the lower symptom severity in these samples limits the size of the relationship that are likely to exist. A recent study using another adaptation of a gambling task in a general population sample collected online also found no significant effect of anxious or depressive symptoms on economic decision making (Zbozinek et al., 2021). It is possible that whilst symptoms of mental health disorders exist in the general population to some extent, the greater severity observed in diagnosed patients is underpinned by distinct mechanisms.

Additionally, we adapted our task to vary the probabilities in the gambles, whereas often previous studies have used only 50:50 gambles. The incorporation of varying gambles increases the complexity of the task and therefore could have subtle effects on the interactions with mental health symptoms compared to previous studies. For instance, ART scores were related to model parameters in both samples, suggesting that general cognitive ability is an important factor in how people perform this task. In particular, higher ART scores were related to more objective probability weighting and making less deterministic choices. The former result suggests that those with higher general cognitive ability evaluate the information provided on a particular trial to a greater extent and base their decisions more closely on the information provided, rather than subjective evaluation of probabilities. The latter result is surprising, however, as more deterministic choices are optimal in this task. Is it possible that this result is due to the high correlation between the risk aversion and determinism parameters. It has also previously been shown that the precise task parameter settings used in gambling tasks can affect how participants behave (Peterson et al., 2021), which may also have knock on effects for the individual participant parameters that are estimated in different studies. Further, whilst we intentionally chose gamble probabilities to probe the behaviour we were interested in – not accepting the gambles even when the probability of the positive outcome was very high – it is possible that we did not vary the probabilities in the most appropriate range to reliably detect the subtle behavioural differences present in those with specific mental health symptoms. Finally, in the graphs showing the qualitative model fits, we can see that the model is not capturing the behaviour of the most risk-averse participants in the loss-only condition. Whilst this only includes a handful of participants in one task condition, it is worth noting, as this type of extreme risk aversion is what we hypothesised would be related to catastrophising.

In conclusion, we have presented a modelling analysis of a gambling task designed to detect probability weighting in two datasets. We highlight the importance of incorporating probability weighting parameters and find very low associations between model parameters and mental health symptoms in our data, as compared to associations with demographic variables.

6 General Discussion

6.1 Summary of Chapters

6.1.1 Chapter 3: A Hierarchical Reinforcement Learning Model Explains Individual Differences in Attention set shifting

We tested the attentional and learning processes underlying the individual variation in attention set shifting abilities using a generative computational modelling approach on two independent large-scale online general-population samples performing CANTAB IED. One sample included additional assessment of demographic variables and mental health problems. We found a hierarchical model that learnt both feature values and dimension attention best explained the data, and that compulsive symptoms were associated with slower learning and higher attentional bias to the first relevant stimulus dimension. Further, older people, those that spent less time in education and women showed more attentional bias to the first relevant dimension. These results establish a new model of cognitive processes underlying the CANTAB IED task, and suggest a possible mechanistic explanation for the variation in set shifting performance, and its relationship to compulsive symptoms.

6.1.2 Chapter 4: Individual Variation in Risky Decisions Is Related to Age and Gender but not to Mental Health Symptoms

We tested the mechanisms of choice evaluation in risky decisions, to understand the previously reported conservative behaviour in patients with depression, using a computational modelling approach in a broad general-population sample (N = 753). Participants performed the CANTAB CGT and completed questionnaires assessing symptoms of mental illness, including depression. We fit five different computational models to the data, including two novel ones, and found that a novel model that uses an inverse power function in the loss domain (contrary to standard Prospect Theory accounts), and is influenced by the probabilities but not the magnitudes of different outcomes, captures the characteristics of our dataset very well. Surprisingly, model parameters were not significantly associated with any mental health questionnaire scores, including depression scales; but they were related to demographic variables, particularly age, with stronger associations than typical model-agnostic task measures. This study showcases a new methodology to analyse data from CANTAB CGT, describes a noteworthy null finding with respect to mental health symptoms, and demonstrates the added precision that a computational approach can offer.

6.1.3 Chapter 5: Individual Variation in Subjective Probability Weighting is Important but Unrelated to Catastrophising and Anxiety

To test our hypothesis that catastrophising is driven by subjective probability weighting, we used a computational modelling approach in a broad general-population sample (N = 212) who performed a

gambling task and completed questionnaires assessing symptoms of mental illness, including catastrophising. We found evidence that models incorporating a probability weighting function fit participants' data better than those without it, which replicated in a larger sample (N = 946). Despite this, we found no evidence for a relationship between probability weighting parameters and catastrophising; and whilst we found some evidence that people with higher general anxiety symptoms overweighted the probability of negative gamble outcomes relative to the positive ones in the initial sample, this result did not replicate when tested in the larger sample. However, we did identify reliable associations between general cognitive ability, as measured by ART, and model parameters. This study showcases the importance of incorporating probability weighting parameters into studies of risky decision making and describes a noteworthy null finding with respect to mental health symptoms.

6.2 Computational Models Offer Precise Mechanistic Insights

6.2.1 Computational models are mechanistic descriptions

Computational psychiatry has been hailed as a method for formalising and testing more precise theories of the relationship between cognitive function and mental health problems (Adams et al., 2016; Montague et al., 2012). In this thesis, each experimental chapter develops a theory-driven computational model to describe performance on a specific task, offering a more mechanistic description of individual differences in behaviour. The mechanistic descriptions offered by these models are further supported by the thorough model checking that was carried out, such as good parameter recovery and excellent fits to participant data.

In Chapter 3, we developed a model of attention set shifting that describes the interaction between attention and learning processes on CANTAB IED and show that individual differences in attention set shifting ability can be explained by a relatively simple hierarchical reinforcement learning algorithm. Rather than using errors per stage to measure task performance, this model describes performance with three parameters: learning rate, dimension primacy, and choice determinism, and we show that lower learning rates, higher dimension primacy, and lower choice determinism lead to more errors on the extradimensional shift. Chapter 4 describes the development of models for CANTAB CGT based on classical economic decision making models. Our best-fitting model included six parameters: colour choice bias, colour choice determinism, risk aversion, loss aversion, delay aversion and bet choice determinism. Crucially, people's betting strategies do not depend on their current number of points, and that on this task, participants are generally risk averse in the domains of both gains and losses. These parameters provide more mechanistic explanations for participant's approach to decision making compared to typical outcome measures such as 'overall proportion bet' or 'risk adjustment'.

In Chapter 5, we explored subjective probability weighting and report evidence that people do not weigh the probabilities of gamble outcomes objectively. Our model also suggests that a two-parameter weighting function, including curvature and elevation parameters, fits the data best, and that participants retain complementarity when weighting probability. This level of description is considerably more fine-grained than that provided by coarse model-agnostic measures, such as proportion of gambles accepted, where the reasons underlying these choices remain ambiguous.

In Chapters 3 and 5, some of the mechanistic insights we gain are consistent with previous literature in relevant fields. For instance, it has long been suggested that attention and learning processes result in the added difficulty of extradimensional set shifts compared to intradimensional ones (though these had not necessarily been mathematically specified or formally tested by fitting to individual participant's data) (Kruschke, 2001; le Pelley et al., 2012, 2016; Trabasso et al., 1966). Furthermore, attention-biased learning algorithms have provided a good fit to other multidimensional stimulus learning tasks though they do not include extradimensional set shifts in the same manner as in CANTAB IED (Leong et al., 2017; Niv et al., 2015). It has also long been suggested that people do not weight probabilities objectively in gambling tasks (Tversky & Kahneman, 1992), which was further corroborated by our results reported here. Therefore, the insights gained from these models are consistent with previous literature, despite slight changes to the tasks used here. Speculatively, these attention and learning processes, or subjective probability weighting, might be somewhat more generally applicable mechanisms as they are not fully dependent on task context.

In Chapters 4 and 5, however, some of the insights from our modelling contradict previous research. Our best-fitting model for CANTAB CGT suggests that people's betting strategy does not vary with their number of points, and that they are generally risk averse in the domains of both gains and losses. These insights are in direct contrast to the well-established Prospect Theory which suggests that participants points or 'wealth' should impact their decision making behaviour, and that people are risk averse in the domain of gains but risk seeking in the domain of losses (Kahneman & Tversky, 1979). Furthermore, whilst the importance of probability weighting is consistent with the previously published Cumulative Prospect Theory model (Tversky & Kahneman, 1992), there does not seem to be consensus on the specification of the weighting function and our finding of the two-parameter Prelec function providing the best fit is inconsistent with earlier research (Stott & Stott, 2006). Therefore, the insights gained from these models somewhat contradict previous reports. Speculatively, this could be because types of economic risk related behaviours are more context dependent, and small changes in task paradigms between studies can have moderate effects on behaviour (Peterson et al., 2021). These insights offer an interesting area of future research to delve

deeper into cognitive mechanisms by evaluating how the context of an experiment can influence the behaviour and best-fitting models of participants.

6.2.2 Computationally derived parameters are associated with demographic variables

In all chapters, we found robust relationships between model parameters and demographic variables which allows us to make more specific inferences regarding the relevant aspects of model performance compared to when using traditional measures of task performance. For example, in Chapter 3, we reported that older people showed more primacy to the first relevant dimension, those that spent longer in education were more deterministic, and showed less primacy to the first relevant dimension and men learnt faster, were more deterministic, and showed less primacy to the first relevant dimension. In Chapter 4, we found that older people and women were more risk averse. Finally in Chapter 5, we reported that older people were less loss averse, weighted probabilities more objectively, but were more deterministic; that women weighted probabilities more subjectively, and were more risk averse; and that individuals with higher ART scores (an index of general cognitive ability) weighted probabilities more subjectively, were less risk averse, more loss averse, and made less deterministic choices. These findings provide a more precise understanding of why, for example, older people might in general score higher errors on the extradimensional set shift stage of CANTAB IED, or why women might choose to bet smaller amounts on the CANTAB CGT.

Some of the associations with demographic variables are comparable across chapters, and consistent with the general direction reported in the literature, such as women being more risk averse in Chapters 4 and 5 (Deakin et al., 2004; Lewis et al., 2021; van den Bos et al., 2013, 2014). However, some of these findings seem somewhat contradictory between chapters, such as spending longer in education being related to higher determinism in Chapter 3, but higher ART scores being related to lower determinism in Chapter 5. As both of these measures are used as a proxy for general cognitive ability, it is somewhat surprising that we find opposite relationships with the choice determinism parameter. Furthermore, we even find varying relationships with demographic variables between Chapters 4 and 5, despite these both incorporating a kind of gambling task, as well as similarly structured models whose mathematical specification of risk and loss aversion parameters is comparable. For example, in Chapter 4, we find that older people are more risk averse, but no relationship was found with loss aversion; whereas in Chapter 5, we find that older people are less loss averse, but no relationship was found with risk aversion. This could in part be due to the different participants with distinct behavioural characteristics that happened to make up these samples. However, it is also known that subtle variations to tasks can affect the observed behaviour, and that

similarly defined parameters are not comparable across different task variations, providing another potential explanation for these discrepancies (Eckstein et al., 2022; Peterson et al., 2021).

In summary, we showed that variation in key demographic variables affects performance on a number of cognitive tasks, and that these relationships can be further defined by exploring associations with computationally derived model parameters.

6.2.3 Computational models offer increased precision

One of the widely touted advantages of the computational approach is that it encourages an increase in precision when formulating and conveying theories about cognitive processes (see section 6.2.1). In Chapter 4, we showed that the extraction of computational parameters can lead to an increase in precision compared to the typical outcome measures. Here, we compared the size of relationship between age and gender using CANTAB CGT typical outcome measures (such as 'overall proportion bet' and model-derived parameters (such as 'risk aversion'). We found that the size of the relationships between the demographic variables and the computational risk aversion parameter were numerically greater than those with the model-agnostic task measure. Furthermore, the relationship between age and risk aversion was statistically greater than that with overall proportion bet when compared with Steiger's Z test. The increase in the size of these relationships demonstrates the extra sensitivity conferred by the computational approach. This is likely due to a number of factors, such as that computational parameters are derived using the full trial and task information and that the models utilised also account for the noise in human behaviour with choice determinism parameters rather than including random responding in the overall performance summary statistic. The increase in the size of the relationship also suggests that the model-derived risk aversion reflects a construct that more closely tracks age compared to the overall proportion bet measure, further strengthening our argument that model parameters provide additional mechanistic insight.

6.3 Minimal Associations Between Parameters and Symptoms of Mental Health Disorders

We initially set out to better define the relationship between cognitive processes and mental health disorders, thus, we were somewhat surprised to find that the associations between model parameters and symptoms of mental health disorders in this body of work were small.

6.3.1 Attention set shifting

In Chapter 3, we reported a significant (albeit modest) association between compulsivity scores, as measured by the OCI-R, and the learning rate and dimension primacy parameters derived from our best fitting model. This suggests a reason why people with more compulsive symptoms tend to score more errors on the extradimensional set shift stage – because they incorporate outcome feedback to

a lesser extent on each trial and demonstrate a stronger bias to the initially relevant stimulus dimension. While set shifting difficulties have been most consistently reported in OCD patients (Chamberlain et al., 2006, 2007; Purcell et al., 1998; Vaghi et al., 2017; Veale et al., 1996), it is still somewhat surprising that no other mental health symptoms were associated with model parameters given previous reports of difficulties performing this task in other patient groups (Elliott R et al., 1995; Gottwald et al., 2018; Kim KL et al., 2019; Liang S et al., 2018; Purcell et al., 1998; Purcell R et al., 1997). This could be because the association with compulsivity is stronger than with other symptoms of mental health and therefore is the only relationship big enough to be detected in a general population sample. It could also be because groups of people with high compulsivity are more mechanistically homogeneous than groups with other elevated mental health symptoms. Finally, the discrepancy in relationships reported here could also be because our best-fitting model partitions task performance in a manner that is most consistent with mechanisms of relevance to high compulsive symptoms.

6.3.2 Risky Decision Making

In Chapter 4, we did not find any significant relationships between model parameters and mental health symptoms. A preliminary data collection in Chapter 5 suggested a potential relationship between general anxiety and the elevation probability weighting parameter, but we failed to replicate this in a larger sample in which no associations of a noteworthy size were found. Again, the lack of associations in both studies is somewhat surprising given that previous literature has reported behavioural differences in CANTAB CGT in patients with depression (Mannie et al., 2015; Murphy et al., 2001; Rawal et al., 2013), and on gambling tasks similar to that used in Chapter 5 in patients with anxiety (Charpentier et al., 2017). Here, we consider the factors that might have led to these contradicting findings.

The most notable difference between the methodology applied to all chapters here, and to previous studies of mental health research is that we used a general population sample who were tested online, rather than patients with specific diagnoses who are tested in person. In fact, the aforementioned observations are not unique - other studies using general population samples have also reported a lack of associations with mental health symptoms, some using model agnostic measures on CANTAB CGT, and others using computationally derived parameters on modified gambling tasks (Deakin et al., 2004; Lewis et al., 2021; Zbozinek et al., 2021). One potential reason for these results is that symptoms of depression and anxiety actually have a minimal relationship with risky decisions overall, and that previous results are due to noise in small samples and the findings have been overstated by publication bias. Alternatively, it could be that cognitive processes are altered in patients with mental health disorders (as in the literature), but that sub-clinical symptoms of depression and anxiety are

unrelated to risky decision making behaviours. This would suggest a mechanistic distinction between patients with severe symptoms and those with milder symptoms and cast doubt on a key motivation for the 'symptomics' approach outlined in the introduction: that symptoms of mental health problems exist on a spectrum in the general population, which we would assume lead to detectable relationships with cognitive performance within large, unselected samples. Finally, it could be that there is a real relationship between these cognitive processes and mental health symptoms but that the size of this association is too small to be detected in general population samples due to the majority of participants not having very severe symptoms. This could have led to the dearth of associations observed here as the size of the relationships we can observe in these samples is limited.

Future studies will be required to untangle these possibilities, and further clarify the relationships between cognition and symptoms of common mental health disorders.

6.4 Limitations

There are a number of limitations in this body of work that must be considered when evaluating the results and conclusions drawn from them.

6.4.1 Online Data Collection

Whilst online data collection is particularly advantageous for psychological research due to the ability to rapidly and conveniently collect very large samples, there are many limitations associated with this form of data collection (Clifford & Jerit, 2014). For example, researchers have no control over the participants that enter their studies. This can result in samples which are heavily skewed in e.g., demographics as observed in the original dataset in Chapter 5, or in other variables that might not be detected as they are less frequently measured, but may nevertheless affect the overall conclusions (Bethlehem, 2010). One example is that people that self-select for online studies might be more likely to participate in a number of them, and therefore will be more familiar with cognitive tasks than the general public which could be reflected in the data collected (J. Chandler et al., 2014). In the worst-case scenario, many of the so-called study "participants" could even be bots designed to enter numerous studies, resulting in close-to-meaningless data (J. J. Chandler & Paolacci, 2017; Pozzar et al., 2020). However, this is unlikely as our data showed substantial variability in task performance consistent with that seen in human behaviour, and some of our key results (e.g., with compulsivity and demographic variables) are in line with what we would expect given previous research. Finally, there is no way to guarantee that even in a fully unbiased sample with human participants, that they are attending to the task, and not distracted or doing multiple things simultaneously. We tried to account for this in Chapters 4 and 5 by performing sensitivity analyses, removing participants who failed attention checks or whose data was not well fit by the model. In both cases, our main findings

were unaffected, though it is still possible that more specific and subtle effects could occur as a result of participant distraction and random responding in online data samples (Zorowitz et al., 2021).

6.4.2 Questionnaire Measures

Another potential limitation of this research is the questionnaire measures that were used to assess certain symptoms across all three chapters. These questionnaires are somewhat limited in their ability to assess the symptoms that they intend to measure, particularly as they rely on self-report which can be vulnerable to many cognitive biases and are thought to be less abstract and controlled compared to cognitive task measures. For instance, whilst they generally ask the participant about the preceding couple of weeks, it is known that questionnaires responses are subject to recall bias such that participants will typically respond based on how they are feeling in the moment, which is likely to add some noise to the results. This method of assessment was aimed at measuring relevant symptoms on a continuous scale; however it is possible that the constructs that we chose to assess in these studies do not assess the most appropriate aspect of mental health disorders to the cognitive tasks we used, which highlights another potential reason for the lack of associations with mental health symptoms observed, particularly in Chapters 4 and 5.

6.4.3 Computational Models

There are some general limitations of the computational approach that require discussion here. The process of selecting a best-fitting model with which to analyse participant data is apt for selecting the model that best explains individual variation (compared to competing models). However, it rests on the assumption that all participants use the same model or strategy to complete the task. This assumption does not always hold true, and the winning model will often provide a bad fit to at least some of the participants in the sample. We tried to account for this in Chapter 4 by removing participants who were not significantly better fit by the winning model than chance. This included 26 out of 753 participants, indicating that this is a potentially important consideration albeit for a very small percentage of participants. Another way to address this would be to use a model-averaging procedure, in which participants' performance can be described by weighting a couple of models that seem to describe different participant strategies well (Wasserman, 2000).

Another limitation of this approach is that some aspects of validity of the model-derived parameters have not been thoroughly assessed. For example, the test-retest statistics of model parameters are important to determine whether model parameters are more stable compared to model-agnostic task measures. Low reliability of measures would suggest that they are not providing particularly meaningful insights to participant behaviour. Whilst previous research suggests that model derived parameters tend to be more reliable than their associated model-agnostic outcome measures (Haines

et al., 2020; Mkrtchian et al., 2021), we cannot fully extrapolate these results to our tasks without testing them explicitly. This limitation is particularly relevant to Chapter 3, as the CANTAB IED task is known to suffer from poor test-retest reliability due to the extradimensional shift being substantially easier on subsequent attempts having passed it the first time.

Finally, the interpretability of mechanistic insights provided by the models is somewhat limited given the work done here in isolation. Whilst, all models were driven by existing cognitive neuroscience theories, and parameters were named descriptively according to their roles within these models, it is difficult to confirm whether the parameters actually represent something more neurally relevant without additional research. Further studies should focus on implementing these models whilst collecting neural data to assess whether signatures in the brain track internal model values, or whether differences between participants are consistent with their different parameter values. These findings would give more weight to the meaning of mechanistically relevant parameters and solidify the advantages of using computational models to analyse cognitive data.

6.5 Future Directions

The results of these studies have highlighted some promising avenues for future research. Here we discuss those that would be the most interesting, or impactful next steps. Both of these would further strengthen our understanding of cognitive processes, and the insights that can be gained from a computational analysis.

6..5.1 Exploring insights from computational models in patient studies

The use of diagnostic criteria in research is no doubt problematic, as outlined in the introduction. However, due to the differing results in previous case-control studies and the general population results reported here and elsewhere, applying a modelling analysis to case-control datasets could potentially lead to a number of additional insights. One suggestion for these differences that was previously mentioned is that distinct mechanisms lead to the mild symptoms seen in the general population, versus the more severe ones seen in patient populations, such that on a mechanistic level there is not a continuum between these groups. This could be investigated by performing the model fitting procedure in a patient group and exploring whether the winning model is consistent with those reported here. Observing the same model to be the best fit to participant data in both circumstances would give more weight to the theory that mental health lies on a continuous distribution in the general population, and that diagnosed patients are on the more severe ends of the distribution. In particular due to the discrepant results reported in previous studies using case-control approaches, it would be enlightening as to the null results reported here if the validated computational models could be analysed in patient groups.

With a computational analysis in these groups, we would be better placed to examine which parameters are relevant and therefore explore mechanistic insights. These studies might also be useful for guiding future studies in terms of which symptoms seem most strongly associated with model parameters, and therefore future research could be better defined to detect such relationships.

6.5.2 Examining the effect of pharmacological manipulation on model derived parameters

Gaining a deeper understanding of the mechanistic insights afforded by a computational analysis is essential if we are to fulfil the potential of computational methods in the field. One way to explore this would be to assess whether pharmacological manipulation had any effect on model parameters. For example, a future study to further explore our findings from Chapter 3 could use a pharmacological compound that modulates the cholinergic system (known to be important for cognitive flexibility) in healthy volunteers and assess whether model parameters are altered as a result of the drug challenge. The observation of a specific effect on model parameters, such as the dimension primacy, would have potential implications for targeted treatment of patients with compulsive symptoms in the clinic.

6.6 Conclusions

Our shallow understanding of mental health disorders contributes to the trial-and-error treatment approach in psychiatry clinical practice. The primary aim of this thesis was to better understand the relationships between mental health symptoms and certain cognitive processes. Specifically, two key methodologies were enlisted: measuring symptoms in large general population samples, and using a computational approach to analyse cognitive processes. We investigated whether utilising these could provide further insights into the mechanisms that underpin mental health disorders. Specifically, these methods were applied in the field of attention set shifting and risky decision making, where differences in task performance between those with mental health disorders and healthy controls have been previously reported. This thesis developed computational models that offer mechanistic descriptions of individual differences on CANTAB IED, CANTAB CGT, and a novel gambling task. It also highlights the added precision that a computational analysis can add. Finally, whilst we found several significant relationships between demographic variables and model parameters, limited associations were found between model parameters and mental health symptoms: that those with higher compulsive symptoms incorporate outcome feedback to a lesser extent after each trial and show a stronger bias to the first relevant stimulus dimension on a multidimensional task of attention set shifting.

Speculatively, these models demonstrate mechanistic insights into cognitive processes that vary with certain demographic variables, though this should be further investigated with neural measures or pharmacological studies. The surprising null results with respect to mental health symptoms are likely

due to the effect size between cognitive processes and mental health symptoms being too small to detect in these samples, or possibly that they do not even exist. Computational studies in patient populations are required to bridge the gap between case-control and general population studies and to uncover any major mechanistic differences between these groups.

Understanding the relationships between symptoms, diagnostic disorders, parameters and demographic variables will enhance our understanding of mental health. This thesis has reported preliminary findings by validating computational models for three specific tasks. We hope that future research will expand on this to better link these models to mental health symptoms, and ultimately improve our understanding and treatment of disorders in the clinic.

References

Ackerman, J. P., McBee-Strayer, S. M., Mendoza, K., Stevens, J., Sheftall, A. H., Campo, J. v., & Bridge, J. A. (2015). Risk-sensitive decision-making deficit in adolescent suicide attempters. *Journal of Child and Adolescent Psychopharmacology*, *25*(2), 109–113. https://doi.org/10.1089/CAP.2014.0041

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2016). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63. https://doi.org/10.1136/JNNP-2015-310737

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. https://doi.org/10.1176/APPI.BOOKS.9780890425596

Baek, K., Kwon, J., Chae, J. H., Chung, Y. A., Kralik, J. D., Min, J. A., Huh, H., Choi, K. M., Jang, K. I., Lee, N. bin, Kim, S., Peterson, B. S., & Jeong, J. (2017). Heightened aversion to risk and loss in depressed patients with a suicide attempt history. *Scientific Reports*, *7*(1). https://doi.org/10.1038/S41598-017-10541-5

Baldessarini, R. J., Tondo, L., Davis, P., Pompili, M., Goodwin, F. K., & Hennen, J. (2006). Decreased risk of suicides and attempts during long-term lithium treatment: a meta-analytic review. *Bipolar Disorders*, *8*(5 Pt 2), 625–639. https://doi.org/10.1111/J.1399-5618.2006.00344.X

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. https://doi.org/10.1016/0010-0277(94)90018-3

Bender, S., & Dittmann-Balcar, A. (2004). Review: in people with schizophrenia, lithium is ineffective as sole therapy, while evidence on augmenting antipsychotics with lithium is inconclusive. *Evidence-Based Mental Health*, *7*(4), 104–104. https://doi.org/10.1136/EBMH.7.4.104

Berg, E. A. (2010). A Simple Objective Technique for Measuring Flexibility in Thinking. *The Journal of General Psychology*, *39*(1), 15–22. https://doi.org/10.1080/00221309.1948.9918159

Berggren, N., & Derakshan, N. (2013). Attentional control deficits in trait anxiety: why you see them and why you don't. *Biological Psychology*, *92*(3), 440–446. https://doi.org/10.1016/J.BIOPSYCHO.2012.03.007

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, *5*, 175–192.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, *78*(2), 161–188. https://doi.org/10.1111/J.1751-5823.2010.00112.X

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PloS One*, *10*(9). https://doi.org/10.1371/JOURNAL.PONE.0137621

Cáceda, R., Nemeroff, C. B., & Harvey, P. D. (2014). Toward an understanding of decision making in severe mental illness. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *26*(3), 196–213. https://doi.org/10.1176/APPI.NEUROPSYCH.12110268

Chamberlain SR, Fineberg NA, Blackwell AD, Robbins TW, & Sahakian BJ. (2006). Motor inhibition and cognitive flexibility in obsessive-compulsive disorder and trichotillomania. *The American Journal of Psychiatry*, *163*(7), 1282. https://doi.org/10.1176/APPI.AJP.163.7.1282

Chamberlain SR, Fineberg NA, Menzies LA, Blackwell AD, Bullmore ET, Robbins TW, & Sahakian BJ. (2007). Impaired cognitive flexibility and motor inhibition in unaffected first-degree relatives of patients with obsessive-compulsive disorder. *The American Journal of Psychiatry*, *164*(2), 335–338. https://doi.org/10.1176/AJP.2007.164.2.335

Chandler, J. J., & Paolacci, G. (2017). Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Impostors. *Social Psychological and Personality Science*, *8*(5), 500–508. https://doi.org/10.1177/1948550617698203

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/S13428-013-0365-7/FIGURES/2

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*(12), 1014–1022. https://doi.org/10.1016/j.biopsych.2016.12.010

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., Ruhe, H. G., Turner, E. H., Higgins, J. P. T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J. P. A., & Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, *391*(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P., Churchill, R., Watanabe, N., Nakagawa, A., Omori, I. M., McGuire, H., Tansella, M., & Barbui, C. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet*, *373*(9665), 746–758. https://doi.org/10.1016/S0140-6736(09)60046-5

Clifford, S., & Jerit, J. (2014). Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. *Journal of Experimental Political Science*, *1*(2), 120–131. https://doi.org/10.1017/XPS.2014.5

Crick, F. (1989). The recent excitement about neural networks. *Nature 1989 337:6203*, *337*(6203), 129–132. https://doi.org/10.1038/337129a0

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-

analytic update of the evidence. *World Psychiatry*, *15*(3), 245–258. https://doi.org/10.1002/WPS.20346

Cuijpers, P., Noma, H., Karyotaki, E., Cipriani, A., & Furukawa, T. A. (2019). Effectiveness and Acceptability of Cognitive Behavior Therapy Delivery Formats in Adults With Depression: A Network Meta-analysis. *JAMA Psychiatry*, *76*(7), 700–707. https://doi.org/10.1001/JAMAPSYCHIATRY.2019.0268

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine 2013 11:1*, *11*(1), 1–8. https://doi.org/10.1186/1741-7015-11-126

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/J.NEURON.2011.02.027

Deakin, J., Aitken, M., Robbins, T., & Sahakian, B. J. (2004). Risk taking during decision-making in normal volunteers changes with age. *Journal of the International Neuropsychological Society*, *10*(4), 590–598. https://doi.org/10.1017/S1355617704104104

Deserno, L., Schlagenhauf, F., & Heinz, A. (2016). Striatal dopamine, reward, and decision making in schizophrenia. *Dialogues in Clinical Neuroscience*, *18*(1), 77–89. https://doi.org/10.31887/dcns.2016.18.1/ldeserno

Downes, J. J., Roberts, A. C., Sahakian, B. J., Evenden, J. L., Morris, R. G., & Robbins, T. W. (1989). Impaired extra-dimensional shift performance in medicated and unmedicated Parkinson's disease: Evidence for a specific attentional dysfunction. *Neuropsychologia*, *27*(11–12), 1329–1343. https://doi.org/10.1016/0028-3932(89)90128-0

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2022). The Interpretation of Computational Model Parameters Depends on the Context. *BioRxiv*, 2021.05.28.446162. https://doi.org/10.1101/2021.05.28.446162

Elliott R, McKenna PJ, Robbins TW, & Sahakian BJ. (1995). Neuropsychological evidence for frontostriatal dysfunction in schizophrenia. *Psychological Medicine*, *25*(3), 619–630. https://doi.org/10.1017/S0033291700033523

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6, 309*. https://doi.org/10.3389/FPSYG.2015.00309

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320. https://doi.org/10.1016/J.JAD.2015.09.005

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067–2076. https://doi.org/10.1017/S0033291713002900

Fried, E. I., von Stockert, S., Haslbeck, J. M. B., Lamers, F., Schoevers, R. A., & Penninx, B. W. J. H. (2020). Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. *Psychological Medicine*, *50*(16), 2682–2690. https://doi.org/10.1017/S0033291719002770

Fried EI. (2017). Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics*, *17*(5), 423–425. https://doi.org/10.1080/14737175.2017.1307737

Fried EI, & Nesse RM. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, *172*, 96–102. https://doi.org/10.1016/J.JAD.2014.10.010

Gard, D. E., Gard, M. G., Kring, A. M., & John, O. P. (2006). Anticipatory and consummatory components of the experience of pleasure: A scale development study. *Journal of Research in Personality*, *40*(6), 1086–1102. https://doi.org/10.1016/J.JRP.2005.11.001

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. *ELife*, *5:e11305*. https://doi.org/10.7554/ELIFE.11305

Gillan CM, Kalanthroff E, Evans M, Weingarden HM, Jacoby RJ, Gershkovich M, Snorrason I, Campeas R, Cervoni C, Crimarco NC, Sokol Y, Garnaat SL, McLaughlin NCR, Phelps EA, Pinto A, Boisseau CL, Wilhelm S, Daw ND, & Simpson HB. (2020). Comparison of the Association Between Goal-Directed Planning and Self-reported Compulsivity vs Obsessive-Compulsive Disorder Diagnosis. *JAMA Psychiatry*, *77*(1), 77–85. https://doi.org/10.1001/JAMAPSYCHIATRY.2019.2998

Ginn, S., & Horder, J. (2012). “One in four” with a mental health problem: the anatomy of a statistic. *BMJ*, *344*(7845), 31. https://doi.org/10.1136/BMJ.E1302

Gottwald, J., Wit, S. de, Apergis-Schoute, A. M., Morein-Zamir, S., Kaser, M., Cormack, F., Sule, A., Limmer, W., Morris, A. C., Robbins, T. W., & Sahakian, B. J. (2018). Impaired cognitive plasticity and goal-directed control in adolescent obsessive–compulsive disorder. *Psychological Medicine*, *48*(11), 1900. https://doi.org/10.1017/S0033291717003464

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404–411. https://doi.org/10.1037/H0059831

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox. *PsyArXiv.* https://doi.org/10.31234/OSF.IO/XR7Y3

Harada-Laszlo, M., Pike, A., Talwar, A., & Robinson, O. (2021). Exploring How Learning is Implicated in Catastrophizing by Modelling a Computerised Task. *MSc dissertation, University College London.*

Halahakoon, D. C., Kieslich, K., O'Driscoll, C., Nair, A., Lewis, G., & Roiser, J. P. (2020). Reward-Processing Behavior in Depressed Participants Relative to Healthy Volunteers: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *77*(12), 1286–1295. https://doi.org/10.1001/JAMAPSYCHIATRY.2020.2139

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/S10608-012-9476-1/TABLES/1

Hutton, S. B., Murphy, F. C., Joyce, E. M., Rogers, R. D., Cuthbert, I., Barnes, T. R. E., McKenna, P. J., Sahakian, B. J., & Robbins, T. W. (2002). Decision making deficits in patients with first-episode and chronic schizophrenia. *Schizophrenia Research*, *55*(3), 249–257. https://doi.org/10.1016/S0920-9964(01)00216-X

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, *7*(4), 1002028. https://doi.org/10.1371/journal.pcbi.1002028

Hyman, S. E. (2010). The Diagnosis of Mental Disorders: The Problem of Reification. *Annual Review of Clinical Psychology*, *6*, 155–179. https://doi.org/10.1146/ANNUREV.CLINPSY.3.022806.091532

Jablensky, A. (2016). Psychiatric classifications: validity and utility. *World Psychiatry*, *15*(1), 26–31. https://doi.org/10.1002/WPS.20284

Johns, L. C. (2005). Hallucinations in the general population. *Current Psychiatry Reports*, *7*(3), 162–167. https://doi.org/10.1007/S11920-005-0049-9

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/J.NEURON.2018.03.044

Keller, M. C., Neale, M. C., & Kendler, K. S. (2007). Association of different adverse life events with distinct patterns of depressive symptoms. *American Journal of Psychiatry*, *164*(10), 1521–1529. https://doi.org/10.1176/APPI.AJP.2007.06091564

Kelley, H. J. (2012). Gradient Theory of Optimal Flight Paths. *Https://Doi.Org/10.2514/8.5282*, *30*(10), 947–954. https://doi.org/10.2514/8.5282

Kendell, R., & Jablensky, A. (2003). Distinguishing Between the Validity and Utility of Psychiatric Diagnoses. *The American Journal of Psychiatry*, *160*(1), 4–12. https://doi.org/10.1176/APPI.AJP.160.1.4

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150. https://doi.org/10.1017/S0033291710001844

Kessler, R. C. (1994). The National Comorbidity Survey of the United States. *International Review of Psychiatry*, *6*(4), 365–376. https://doi.org/10.3109/09540269409023274

Kessler, R. C., Sampson, N. A., Berglund, P., Gruber, M. J., Al-Hamzawi, A., Andrade, L., Bunting, B., Demyttenaere, K., Florescu, S., de Girolamo, G., Gureje, O., He, Y., Hu, C., Huang, Y., Karam, E., Kovess-Masfety, V., Lee, S., Levinson, D., Medina Mora, M. E., … Wilcox, M. A. (2015). Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health Surveys. *Epidemiology and Psychiatric Sciences*, *24*(3), 210–226. https://doi.org/10.1017/S2045796015000189

Kim KL, Christensen RE, Ruggieri A, Schettini E, Freeman JB, Garcia AM, Flessner C, Stewart E, Conelea C, & Dickstein DP. (2019). Cognitive performance of youth with primary generalized anxiety disorder versus primary obsessive-compulsive disorder. *Depression and Anxiety*, *36*(2), 130–140. https://doi.org/10.1002/DA.22848

Korszun, A., Moskvina, V., Brewster, S., Craddock, N., Ferrero, F., Gill, M., Jones, I. R., Jones, L. A., Maier, W., Mors, O., Owen, M. J., Preisig, M., Reich, T., Rietschel, M., Farmer, A., & McGuffin, P. (2004). Familiality of symptom dimensions in depression. *Archives of General Psychiatry*, *61*(5), 468–474. https://doi.org/10.1001/ARCHPSYC.61.5.468

Kotov, R., Waszczuk, M. A., Krueger, R. F., Forbes, M. K., Watson, D., Clark, L. A., Achenbach, T. M., Althoff, R. R., Ivanova, M. Y., Michael Bagby, R., Brown, T. A., Carpenter, W. T., Caspi, A., Moffitt, T. E., Eaton, N. R., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., … Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/ABN0000258

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kruschke, J. K. (2001). Toward a Unified Model of Attention in Associative Learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. https://doi.org/10.1006/JMPS.2000.1354

Kyte, Z. A., Goodyer, I. M., & Sahakian, B. J. (2005). Selected executive skills in adolescents with recent first episode major depression. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *46*(9), 995–1005. https://doi.org/10.1111/J.1469-7610.2004.00400.X

le Pelley, M. E., Haselgrove, M., & Esber, G. R. (2012). Modeling attention in associative learning: Two processes or one? *Learning & Behavior, 40*(3), 292–304. https://doi.org/10.3758/S13420-012-0084-4

le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111–1140. https://doi.org/10.1037/BUL0000064

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/J.NEURON.2016.12.040

Leucht, S., Helfer, B., Dold, M., Kissling, W., & Mcgrath, J. J. (2015). Lithium for schizophrenia. *The Cochrane Database of Systematic Reviews*, *2015*(10). https://doi.org/10.1002/14651858.CD003834.PUB3

Levaux MN, Potvin S, Sepehry AA, Sablier J, Mendrek A, & Stip E. (2007). Computerized assessment of cognition in schizophrenia: promises and pitfalls of CANTAB. *European*

Psychiatry : The Journal of the Association of European Psychiatrists, *22*(2), 104–115. https://doi.org/10.1016/J.EURPSY.2006.11.004

Lewis, G., Srinivasan, R., Roiser, J., Blakemore, S. J., Flouri, E., & Lewis, G. (2021). Risk-taking to obtain reward: Sex differences and associations with emotional and depressive symptoms in a nationally representative cohort of UK adolescents. *Psychological Medicine*. *52(13),* 2805–2813. https://doi.org/10.1017/S0033291720005000

Liang S, Brown MRG, Deng W, Wang Q, Ma X, Li M, Hu X, Juhas M, Li X, Greiner R, Greenshaw AJ, & Li T. (2018). Convergence and divergence of neurocognitive patterns in schizophrenia and depression. *Schizophrenia Research*, *192*, 327–334. https://doi.org/10.1016/J.SCHRES.2017.06.004

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience 2020 21:6*, *21*(6), 335–346. https://doi.org/10.1038/s41583-020-0277-3

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J., & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review*, *42*, 72–82. https://doi.org/10.1016/J.CPR.2015.08.004

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/H0076778

Mackintosh, N. J., & Little, L. (2013). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science 1969 14:1*, *14*(1), 5–6. https://doi.org/10.3758/BF03336395

Mannie, Z. N., Williams, C., Browning, M., & Cowen, P. J. (2015). Decision making in young people at familial risk of depression. *Psychological Medicine*, *45*(2), 375–380. https://doi.org/10.1017/S0033291714001482

Mason O, Linney Y, & Claridge G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*(2–3), 293–296. https://doi.org/10.1016/J.SCHRES.2005.06.020

Meder, D., Rabe, F., Morville, T., Madsen, K. H., Koudahl, M. T., Dolan, R. J., Siebner, H. R., & Hulme, O. J. (2021). Ergodicity-breaking reveals time optimal decision making in humans. *PLOS Computational Biology*, *17*(9), e1009217. https://doi.org/10.1371/JOURNAL.PCBI.1009217

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495. https://doi.org/10.1016/0005-7967(90)90135-6

Miura, T., Noma, H., Furukawa, T. A., Mitsuyasu, H., Tanaka, S., Stockton, S., Salanti, G., Motomura, K., Shimano-Katsuki, S., Leucht, S., Cipriani, A., Geddes, J. R., & Kanba, S. (2014). Comparative efficacy and tolerability of pharmacological treatments in the maintenance treatment of bipolar disorder: a systematic review and network meta-analysis. *The Lancet. Psychiatry*, *1*(5), 351–359. https://doi.org/10.1016/S2215-0366(14)70314-1

Mkrtchian, A., Valton, V., & Roiser, J. P. (2021). Reliability of Decision-Making and Reinforcement Learning Computational Parameters. *BioRxiv*, 2021.06.30.450026. https://doi.org/10.1101/2021.06.30.450026

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

Murphy, F. C., Rubinsztein, J. S., Michael, A., Rogers, R. D., Robbins, T. W., Paykel, E. S., & Sahakian, B. J. (2001). Decision-making cognition in mania and depression. *Psychological Medicine*, *31(4)*, 679–693. https://doi.org/10.1017/S0033291701003804

Murray GK, Cheng F, Clark L, Barnett JH, Blackwell AD, Fletcher PC, Robbins TW, Bullmore ET, & Jones PB. (2008). Reinforcement and reversal learning in first-episode psychosis. *Schizophrenia Bulletin*, *34*(5), 848–855. https://doi.org/10.1093/SCHBUL/SBN078

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Owen, A. M., Roberts, A. C., Polkey, C. E., Sahakian, B. J., & Robbins, T. W. (1991). Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia*, *29*(10), 993–1006. https://doi.org/10.1016/0028-3932(91)90063-E

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*, 102371. https://doi.org/10.1016/J.JMP.2020.102371

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor Structure of the Barratt Impulsiveness Scale. *Journal of clinical psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679

Paulus, M. P., & Yu, A. J. (2012). Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, *16*(9), 476. https://doi.org/10.1016/J.TICS.2012.07.009

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*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Pedersen, A., Göder, R., Tomczyk, S., & Ohrmann, P. (2017). Risky decision-making under risk in schizophrenia: A deliberate choice? *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 57–64. https://doi.org/10.1016/J.JBTEP.2016.08.004

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/SCIENCE.ABE2629

Pozzar, R., Hammer, M. J., Underhill-Blazey, M., Wright, A. A., Tulsky, J. A., Hong, F., Gundersen, D. A., & Berry, D. L. (2020). Threats of Bots and Other Bad Actors to Data Quality Following Research Participant Recruitment Through Social Media: Cross-Sectional Questionnaire. *J Med Internet Res*, *22*(10), e23021. https://doi.org/10.2196/23021

Pratt, D. N., Barch, D. M., Carter, C. S., Gold, J. M., Ragland, J. D., Silverstein, S. M., & MacDonald, A. W. (2021). Reliability and Replicability of Implicit and Explicit Reinforcement Learning Paradigms in People With Psychotic Disorders. *Schizophrenia Bulletin*, *47*(3), 731–739. https://doi.org/10.1093/SCHBUL/SBAA165

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, *66*(3), 497. https://doi.org/10.2307/2998573

Purcell R, Maruff P, Kyrios M, & Pantelis C. (1997). Neuropsychological function in young patients with unipolar major depression. *Psychological Medicine*, *27*(6), 1277–1285. https://doi.org/10.1017/S0033291797005448

Purcell R, Maruff P, Kyrios M, & Pantelis C. (1998). Cognitive deficits in obsessive-compulsive disorder on tests of frontal-striatal function. *Biological Psychiatry*, *43*(5), 348–357. https://doi.org/10.1016/S0006-3223(97)00201-1

Purcell, R., Maruff, P., Kyrios, M., & Pantelis, C. (1998). Neuropsychological Deficits in Obsessive-compulsive Disorder: A Comparison With Unipolar Depression, Panic Disorder, and Normal Controls. *Archives of General Psychiatry*, *55*(5), 415–423. https://doi.org/10.1001/ARCHPSYC.55.5.415

Rawal, A., Collishaw, S., Thapar, A., & Rice, F. (2013). "The risks of playing it safe": a prospective longitudinal study of response to reward in the adolescent offspring of depressed parents. *Psychological Medicine*, *43*(1), 27–38. https://doi.org/10.1017/S0033291712001158

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, *170*(1), 59–70. https://doi.org/10.1176/APPI.AJP.2012.12070999

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*(1), 81–91. https://doi.org/10.1016/J.TICS.2011.11.009

Rogers, R. D., Everitt, B. J., Baldacchino, A., Blackshaw, A. J., Swainson, R., Wynne, K., Baker, N. B., Hunter, J., Carthy, T., Booker, E., London, M., Deakin, J. F. W., Sahakian, B. J., & Robbins, T. W. (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 1999 20:4*, *20*(4), 322–339. https://doi.org/10.1016/s0893-133x(98)00091-8

Rojas, R. (1996). The Backpropagation Algorithm. *Neural Networks*, 149–182. https://doi.org/10.1007/978-3-642-61068-4_7

Romeu, R. J., Haines, N., Ahn, W. Y., Busemeyer, J. R., & Vassileva, J. (2020). A computational model of the Cambridge gambling task with applications to substance use disorders. *Drug and Alcohol Dependence*, *206*, 107711. https://doi.org/10.1016/J.DRUGALCDEP.2019.107711

Rubinsztein, J. S., Fletcher, P. C., Rogers, R. D., Ho, L. W., Aigbirhio, F. I., Paykel, E. S., Robbins, T. W., & Sahakian, B. J. (2001). Decision-making in mania: a PET study. *Brain : A Journal of Neurology*, *124*(Pt 12), 2550–2563. https://doi.org/10.1093/BRAIN/124.12.2550

Rubinsztein, J. S., Michael, A., Underwood, B. R., Tempest, M., & Sahakain, B. J. (2006). Impaired cognition and decision-making in bipolar depression but no "affective bias" evident. *Psychological Medicine*, *36*(5), 629–639. https://doi.org/10.1017/S0033291705006689

Rush, A. J., Trivedi, M., & Fava, M. (2003). Depression, IV: STAR*D treatment trial for depression. *The American Journal of Psychiatry*, *160*(2), 237. https://doi.org/10.1176/APPI.AJP.160.2.237

Sachdev, P. S., & Malhi, G. S. (2005). Obsessive-compulsive behaviour: a disorder of decision-making. *The Australian and New Zealand Journal of Psychiatry*, *39*(9), 757–763. https://doi.org/10.1080/J.1440-1614.2005.01680.X

Seppälä, A., Pylvänäinen, J., Lehtiniemi, H., Hirvonen, N., Corripio, I., Koponen, H., Seppälä, J., Ahmed, A., Isohanni, M., Miettunen, J., & Jääskeläinen, E. (2021). Predictors of response to pharmacological treatments in treatment-resistant schizophrenia – A systematic review and meta-analysis. *Schizophrenia Research*, *236*, 123–134. https://doi.org/10.1016/J.SCHRES.2021.08.005

Severus, E., Taylor, M. J., Sauer, C., Pfennig, A., Ritter, P., Bauer, M., & Geddes, J. R. (2014). Lithium for prevention of mood episodes in bipolar disorders: systematic review and meta-analysis. *International Journal of Bipolar Disorders*, *2*(1), 1–17. https://doi.org/10.1186/S40345-014-0015-8/FIGURES/9

Slamencka NJ. (1968). A methodological analysis of shift paradigms in human discrimination learning. *Psychological Bulletin*, *69*(6), 423–438. https://doi.org/10.1037/H0025762

Smoski, M. J., Lynch, T. R., Rosenthal, M. Z., Cheavens, J. S., Chapman, A. L., & Krishnan, R. R. (2008). Decision-making and risk aversion among depressive adults. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(4), 567–576. https://doi.org/10.1016/J.JBTEP.2008.01.004

Snyder, H. R., Friedman, N. P., & Hankin, B. L. (2021). Associations between task performance and self-report measures of cognitive control: Shared vs. distinct abilities. *Assessment*, *28*(4), 1080. https://doi.org/10.1177/1073191120965694

Song, M., Cai, M. B., & Niv, Y. (2020). Learning what is relevant for rewards via serial hypothesis testing. *42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines*, Virtual, Online; 29 July 2020 through 1 August 2020; Code 182812. https://doi.org/10.32470/CCN.2019.1360-0

Spielberger C. D., Gorsuch R. L., Lushene R. E., Vagg, P. R., & Jacobs, G. A. (1983). Manual for the state-trait anxiety inventory. Palo Alto, CA: Consulting Psychologists Press.

Spitzer, R. L., Kroenke, K., & Williams, J. B. W. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA*, *282*(18), 1737–1744. https://doi.org/10.1001/JAMA.282.18.1737

Spitzer, R. L., Williams, J. B. W., & Skodol, A. E. (1980). DSM-III: The major achievements and an overview. *American Journal of Psychiatry*, *137*(2), 151–164. https://doi.org/10.1176/AJP.137.2.151

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stott, H. P., & Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *J Risk Uncertainty*, *32*, 101–130. https://doi.org/10.1007/s11166-006-8289-6

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1), 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Taylor Tavares, J. V., Clark, L., Cannon, D. M., Erickson, K., Drevets, W. C., & Sahakian, B. J. (2007). Distinct Profiles of Neurocognitive Function in Unmedicated Unipolar Depression and Bipolar II Depression. *Biological Psychiatry*, *62*(8), 917–924. https://doi.org/10.1016/J.BIOPSYCH.2007.05.034

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika 1953 18:4*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Trabasso, T., Anthony Deutsch, J., & Gelman, R. (1966). Attention in discrimination learning of young children. *Journal of Experimental Child Psychology*, *4*(1), 9–19. https://doi.org/10.1016/0022-0965(66)90048-8

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Tyagi, H., Apergis-Schoute, A. M., Akram, H., Foltynie, T., Limousin, P., Drummond, L. M., Fineberg, N. A., Matthews, K., Jahanshahi, M., Robbins, T. W., Sahakian, B. J., Zrinzo, L., Hariz, M., & Joyce, E. M. (2019). A Randomized Trial Directly Comparing Ventral Capsule and Anteromedial Subthalamic Nucleus Stimulation in Obsessive-Compulsive Disorder: Clinical and Imaging Evidence for Dissociable Effects. *Biological Psychiatry*, *85*(9), 726–734. https://doi.org/10.1016/J.BIOPSYCH.2019.01.017

Vaghi, M. M., Vértes, P. E., Kitzbichler, M. G., Apergis-Schoute, A. M., Flier, F. E. van der, Fineberg, N. A., Sule, A., Zaman, R., Voon, V., Kundu, P., Bullmore, E. T., & Robbins, T. W. (2017). Specific Frontostriatal Circuits for Impaired Cognitive Flexibility and Goal-Directed Planning in Obsessive-Compulsive Disorder: Evidence From Resting-State Functional Connectivity. *Biological Psychiatry*, *81*(8), 708. https://doi.org/10.1016/J.BIOPSYCH.2016.08.009

van den Bos, R., Homberg, J., & de Visser, L. (2013). A critical review of sex differences in decision-making tasks: Focus on the Iowa Gambling Task. *Behavioural Brain Research*, *238*(1), 95–108. https://doi.org/10.1016/J.BBR.2012.10.002

van den Bos, R., Taris, R., Scheppink, B., de Haan, L., & Verster, J. C. (2014). Salivary cortisol and alpha-amylase levels during an assessment procedure correlate differently with risk-taking measures in male and female police recruits. *Frontiers in Behavioral Neuroscience*, *7*(JAN), 219. https://doi.org/10.3389/FNBEH.2013.00219/BIBTEX

vanRossum, G. (1995). Python reference manual. *Department of Computer Science [CS]*, *R 9525*.

Veale DM, Sahakian BJ, Owen AM, & Marks IM. (1996). Specific cognitive deficits in tests sensitive to frontal lobe dysfunction in obsessive-compulsive disorder. *Psychological Medicine*, *26*(6), 1261–1269. https://doi.org/10.1017/S0033291700035984

von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. *Princeton, NJ: Princeton University Press*.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2014). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry 2015 20:3*, *20*(3), 345–352. https://doi.org/10.1038/mp.2014.44

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/JMPS.1999.1278

Wenliang LK, & Seitz AR. (2018). Deep Neural Networks for Modeling Visual Perceptual Learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *38*(27), 6028–6044. https://doi.org/10.1523/JNEUROSCI.1620-17.2018

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for the Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition. *Journal of Abnormal Psychology*, *114*(4), 494–504. https://doi.org/10.1037/0021-843X.114.4.494

Wilson, C. G., Nusbaum, A. T., Whitney, P., & Hinson, J. M. (2018). Trait anxiety impairs cognitive flexibility when overcoming a task acquired response and a preexisting bias. *PLoS ONE*, *13*(9). https://doi.org/10.1371/JOURNAL.PONE.0204694

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, *8,* e49547. https://doi.org/10.7554/ELIFE.49547

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience, 5,* 189. https://doi.org/10.3389/FNHUM.2011.00189

Woodrow, A., Sparks, S., Bobrovskaia, V., Paterson, C., Murphy, P., & Hutton, P. (2019). Decision-making ability in psychosis: a systematic review and meta-analysis of the magnitude, specificity and correlates of impaired performance on the Iowa and Cambridge Gambling Tasks. *Psychological Medicine*, *49*(1), 32–48. https://doi.org/10.1017/S0033291718002660

Yearsley, J. M., Gaigg, S. B., Bowler, D. M., Ring, M., & Haenschel, C. (2021). What Can Performance in the IEDS Task Tell Us About Attention Shifting in Clinical Groups? *Autism Research*, *14*(6), 1237–1251. https://doi.org/10.1002/AUR.2484

Zbozinek, T. D., Charpentier, C. J., Qi, S., & Mobbs, D. (2021). Economic Decisions with Ambiguous Outcome Magnitudes Vary with Low and High Stakes but Not Trait Anxiety or Depression. *Computational Psychiatry*, *5*(1), 119. https://doi.org/10.5334/CPSY.79

Zorowitz, S., Niv, Y., & Bennett, D. (2021). Inattentive responding can induce spurious associations between task behavior and symptom measures. *PsyArXiv.* https://doi.org/10.31234/OSF.IO/RYNHK

Zung WWK. (1965). A Self-Rating Depression Scale. *Archives of General Psychiatry*, *12*(1), 63–70. https://doi.org/10.1001/ARCHPSYC.1965.01720310065008