

The Relationship Between Extinction Learning Processes and Callous Unemotional Traits

Lindsay Kemp

School of Psychology, Faculty of Science

The University of Sydney

A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy

2022

Statement of Originality

I certify that the content of this thesis is my own work. This includes design and analysis of the studies contained within, as well as their present documentation. The content of thesis has not been submitted for any degree or other purposes. The assistance I received in preparing this thesis is acknowledged below.

Lindsay Kemp

Acknowledgements

This thesis was completed under the supervision of Dr Caroline Moul. I would like to acknowledge her assistance in the conception, design, and interpretation of the presently reported studies. I also recognise the input of Dr Evan Livesey on the development of these studies. The Sydney Informatics Hub (funded by the University of Sydney) provided direct assistance with a statistical analysis reported here (which is noted specifically where relevant). Their training programs were also informative to the data analyses in this thesis. Additionally, Dr Moul provided editorial feedback on the structure of the following thesis and assisted in proofreading. Further proofreading was completed by Ruby Brown.

I would like to thank Caroline for her help in securing employment opportunities and funding throughout my candidature, as this was vital in facilitating my candidature. Caroline's research inspired the current studies and I hope that the findings reported here can contribute to the ongoing development of this research. I also acknowledge that her encouragement was critical to the completion of this thesis. Finally, I would like to thank Ruby Brown for her personal commitment to supporting my research career. Without this support the following work would not have been possible.

0.0 Abstract

Callous Unemotional (CU) traits are an informative clinical construct, identifying individuals at risk of conduct disorder and treatment resistance. This thesis examines associative processes that may underlie this resistance. Specifically, extinction learning processes, which allow individuals to maintain expectations that match the outcomes of their behaviour. It is often assumed that there is an established link between CU traits and deficient associative learning of this kind. However, a systematic review of the current literature, described in this thesis, reveals that the evidence for this relationship is not consistent. This finding indicated the need for a study of this correlation using a well powered experimental methodology. As a result, six studies were carried out to develop a computer-based task able to measure extinction learning in large samples from the general population, with CU traits assessed using the Inventory of Callous Unemotional Traits (ICU). Findings from this series included repeated observations of CU trait associated extinction deficits. They also provided unique data on whether CU traits affect the relapse of extinguished behaviour. Furthermore, an unanticipated effect of cultural factors was observed. In consideration of these findings, observed parameters were utilised in the design of a large-scale confirmatory study with over 700 participants. This study also investigated the role of cultural variables in the structure and external validity of the ICU, as well as its association with extinction learning. This thesis therefore provides useful data on the relationship between CU traits and extinction learning, as well as on the cross- cultural validity of the ICU, representing a significant contribution to our understanding of the aetiology and phenotypic characteristics of CU traits.

0.1 Contents

1	Introduction.....	10
2	Reviewing Extinction Behaviour in Callous Unemotional Traits.....	31
3	Extinction Task Development.....	63
4	Modelling and Power Analysis	132
5	Cross Cultural Properties of the Inventory of Callous Unemotional Traits.....	151
6	Callous Unemotional Traits and Extinction Learning	185
7	General Discussion	211
8	References.....	228

0.1.1 Overview

In Chapter 1 of this thesis, I introduced a potential relationship between personality level deficits in social emotion (CU traits) and extinction learning processes. In Chapter 2 I systematically determined limitations to existing evidence for this relationship. I then addressed these limitations empirically in Chapter 3, which provided novel observations supporting the role of associative learning in CU traits. In Chapter 4 I utilised these observations to statistically inform the sampling strategy for later chapters. Employing this sampling strategy in Chapter 5, I examined cultural variation within this sample as a source of inconsistency in the measurement of CU traits. This analysis contributed to the implementation of a revised measure of CU traits in Chapter 6, in which I tested this measure as predictor of learning processes. I then discussed surprising outcomes of this test in Chapter 7, where I used the results of this thesis to suggest previously unconsidered hypotheses for the role of learning in CU traits.

0.2 List of Tables

Table 1.1 Comparable Contingency Reversals in the Experience of BFI and in Response Reversal Tasks.....	26
Table 2.1 Extracted Variables for Reviewed Response Reversal Studies.....	48
Table 2.2 Extracted Variables for Reviewed Extinction Studies	54
Table 3.1 Reinforcement Schedules for each Condition Throughout Pilot 1	71
Table 3.2 Reinforcement Schedules for Each Response.....	86
Table 3.3 Reinforcement Schedules for Each Response Lever and Contexts for Each Phase	95
Table 3.4 Descriptive Statistics for All Psychometric Scales	100
Table 3.5 ICU Scores, Grouped by English Language Background.....	101
Table 3.6 Reinforcement Schedules and Contexts for Each Phase and Condition in the Extinction Task	110
Table 3.7 Number of Trials for Each Pair of Stimuli in Each Block of the Reversal Task.....	111
Table 3.8 Descriptive Statistics for Psychometric Scales, Grouped by English as a First Language	119
Table 4.1 Reinforcement Schedules and Contexts for each Phase and Condition in the Extinction Task	136
Table 4.2 Extinction Model Statistics	140
Table 4.3 Renewal Model Statistics.....	144
Table 4.4 Power Analysis for the Renewal and ICU Score Relationship	147
Table 5.1 Descriptive Statistics for all Measured Scales	167
Table 5.2 Sample Demographics and Cultural Measures Split Across Ethnicity and Recruitment Source	169

Table 5.3 Demographic Descriptives in SRFES Clusters	171
Table 5.4 Confirmatory Factor Analysis for ICS Measure	172
Table 5.5 Demographic Descriptives in ICS Clusters	173
Table 5.6 Confirmatory Factor Analysis for ICU Measures	174
Table 5.7 ICU 11 Factor Loadings	175
Table 5.8 ICU 11 Measurement Invariance Over Grouping Variables	176
Table 5.9 ICU 11 Subscales' Correlations with External Measures	179
Table 6.1 Structure of the Extinction task.....	191
Table 6.2 Analysis of Total ICU Score as a Predictor of Extinction Responding in Western Participants.....	198
Table 6.3 Effect of ICU Subscales on Extinction Responding Controlling for Identified Covariates	200
Table 6.4 Effect of Total ICU on Renewal Responding for Western Participants	203
Table 6.5 Effect of ICU Subscales and Covariates on Renewal Responding	204

0.3 List of Figures

Figure 2.1 Impact of Reinforcement Probability on Extinction, Sheffield (1949)	37
Figure 2.2 Process for Identifying Eligible Studies	41
Figure 2.3 Tasks Evidencing Psychopathic Trait Related Behavioural Deficits.....	42
Figure 3.1 Quentin the Quokka Task Layout.....	73
Figure 3.2 Observed and Predicted Extinction Responses for Individuals in Pilot 1	74
Figure 3.3 Observed and Predicted Extinction Responses for Individuals in Pilot 2	78
Figure 3.4 Observed and Predicted Extinction Responses for Individuals in Pilot 3	82
Figure 3.5 Observed and Predicted Extinction Responses for Individuals in Experiment 1	89
Figure 3.6 Correlation Between ICU Scores and Slope Coefficients for Extinction Behaviour ..	90
Figure 3.7 Layout of The Task in Experiment 2	97
Figure 3.8 Observed and Predicted Extinction Responses for Individuals with English as a First Language in Experiment 2	102
Figure 3.9 Observed and Predicted Extinction Responses for Individuals Without English as a First Language in Experiment 2.....	103
Figure 3.10 Effect of English Language Background on Rapid Reacquisition	105
Figure 3.11 Layout of the Extinction Task in Experiment 3	114
Figure 3.12 Relationship Between ICU Scores and the Difference in Responding Between Training and Extinction Phases of the Control Condition in the Extinction Task	120
Figure 3.13 Observed and Predicted Extinction Responses for Individuals with English as a First Language in Experiment 3	122
Figure 3.14 Observed and Predicted Extinction Responses for Individuals without English as a First Language in Experiment 3.....	123

Figure 3.15 Number of Errors in Each Phase and Probability Condition of the Reversal Task, Split by ICU Scores	125
Figure 4.1 Individual Level Model Predictions for Extinction Behaviour	141
Figure 4.2 Aggregate Level Extinction Model Predictions Across ICU scores.....	142
Figure 4.3 Power Estimates for Different Sample Sizes Using Simulation of the Extinction Model	143
Figure 4.4 Renewal Model's Predictions of Individuals Responses	145
Figure 4.5 Renewal Model's Aggregate Prediction Across ICU Scores.....	146
Figure 4.6 Power Values for Simulations of the Effect of ICU scores on Renewal	147
Figure 6.1 Model of Extinction Responding for Different Levels of ICU 11 Callous and Uncaring Subscales	201
Figure 6.2 ICU Subscale Correlations (Original and Partialled)	202
Figure 6.3 Mean Predicted Values for Renewal Responses	205

0.4 Abbreviations

Antisocial Process Screening Device (APSD), behavioral family intervention (BFI), callous unemotional (CU), confirmatory factor analysis (CFA), Inventory of Callous Unemotional Traits (ICU), Individualism and Collectivism Scale (ICS), Levenson's Self-Report Psychopathy Scale (LSRP), intra/extra dimensional shift (ID/ED), orbitofrontal cortex (OFC), Psychopathy Checklist (PCL), Psychopathy Checklist – Revised (PCL-R), PCL screening version (PCL-SV), PCL-Youth Version (PCL-YV), Psychopathy Screening Device (PSD), psychopathic traits (PT), Self-Report of Aggression and Social Behaviour Measure (SRASBM), Self-Reported Fluency of English Scale (SRFES), State-Trait Anxiety Inventory (STAI), Triarchic Psychopathy Measure (TriPM), Urgency, Premeditation (lack of), Perseverance (lack of), Sensation seeking (UPPS), Wisconsin Card-Sorting Task (WCST).

0.5 Analyses

Analytical code developed for the thesis can be viewed at <https://osf.io/nz2sx/> and <https://osf.io/tfjqc/>. These repositories contain preregistrations relevant to research in Chapter 3, and Chapters 4-6, respectively. Data described in Chapters 5-6 are open and published in the relevant repository, however other data in this these were not ethically approved for public access. All analytical programming was conducted using R (R Core Team, 2020), and models were primarily fit using the lme4 (Bates et al., 2015) package. Graphs and data handling were completed with Tidyverse (Wickham et al., 2019) packages. Other utilised packages can be seen in the code linked above.

1 Introduction

Contents

1.0	What are Callous Unemotional traits?	11
1.0.1	Affective Features of CU traits	14
1.0.2	CU Traits and Conduct Interventions	15
1.1	Aetiology of CU Traits.....	16
1.1.1	Low Fear Hypothesis	18
1.1.2	Integrated Emotion Systems model	19
1.1.3	Differential Amygdala Activation Model	21
1.1.4	Impaired Integration Framework	23
1.2	The Significance of Learning Deficits in CU Traits	24
1.2.1	Learning Deficits and Conduct Problem Interventions	25
1.2.2	Extinction and Reversal Learning: Context Dependent Processes	26
1.2.3	Context Learning and Psychological Interventions	28
1.3	Thesis Aims and Hypotheses	30

Callous Unemotional traits (CU traits) are personality features characterised by poor empathy and muted emotional responsivity. The clinical interest in these traits is largely driven by the risk they impart for antisocial behaviour, and for resistance to interventions on this behaviour (Frick et al., 2014). Whilst accounts of this risk profile often cite CU traits' association with impaired fear recognition (e.g., Dadds et al., 2011), these traits are also linked to variation in emotion-independent learning processes (Hamilton et al., 2015). The particular focus of this thesis is variation in extinction learning, that is, learning about the absence of a previously expected outcome. Impaired extinction learning is a key process in theoretical accounts of the genetic aetiology underpinning CU traits (Blair et al., 2005; Moul et al., 2012). Furthermore, prominent interventions for antisocial behaviour act to address learned behavioural outcomes (Dadds & Hawes, 2006). In this introduction I will outline why extinction learning may be relevant to the reduced efficacy of these interventions in individuals with CU traits. I will also highlight that, despite the significance of this process, the pool of evidence linking impaired extinction learning to CU traits is limited. The absence of such evidence points to the need for further study of extinction learning as an explanatory process in the aetiology and clinical features of CU traits.

1.0 What are Callous Unemotional traits?

Throughout the history of psychological research into antisocial behaviour, deficits of social emotion have been noted as a key causal process. These deficits were formalised as “psychopathy” in the case studies of Cleckley (1976), which profiled the psychology of a group of antisocial individuals highly prone to recidivism. He identified 16 characteristic features of these individuals:

1. Superficial charm and good intelligence

2. Absence of delusions and other signs of irrational thinking
3. Absence of nervousness or psychoneurotic manifestations
4. Unreliability
5. Untruthfulness and insincerity
6. Lack of remorse or shame
7. Inadequately motivated antisocial behaviour
8. Poor judgment and failure to learn by experience
9. Pathologic egocentricity and incapacity for love
10. General poverty in major affective reactions
11. Specific loss of insight
12. Unresponsiveness in general interpersonal relation
13. Fantastic and uninviting behaviour with drink and sometimes without
14. Suicide rarely carried out
15. Sex life impersonal, trivial and poorly integrated
16. Failure to follow any life plan. (Cleckley, 1976, p. 337-338)

These features are widely cited throughout the modern literature. Indeed, most of the modern study of psychopathy derives from this initial formalization (Crego & Widiger, 2016), including the widely implemented psychometric tool; the Psychopathy Checklist – Revised (PCL-R) (Hare, 1991). This operationalization treats psychopathy as a dimensional personality trait, initially conceptualized with two factors, factor one being an interpersonal/affective factor (shallow affect, lack of remorse or guilt and callous/lack of empathy) and factor two a lifestyle factor (featuring, for example: impulsivity, irresponsibility and poor behavioural control) (Harpur et al., 1989). These traits do not map directly to antisocial behaviour itself, as psychopathic traits

occur in the non-offending general population (Gao & Raine, 2010). Further, a minority (15-25%) of US inmates meet PCL-R criteria for psychopathy (Hare, 1996). However, psychopathic traits are strongly correlated with antisocial behaviour. Meta-analysis has indicated that institutional infractions and recidivism are associated with greater Psychopathy Checklist scores, especially for Factor 2 (lifestyle) traits (Leistico et al., 2008). Contrastingly, Factor 1 (affective) scores are more predictive of impaired emotional recognition (Gillen et al., 2018). These affective features are addressed specifically by the callous unemotional (CU) trait construct. CU traits are less associated with disordered conduct than overall psychopathic traits (Andershed et al., 2018), and have particular importance in developmental settings (Frick et al., 2014).

A widely noted feature of adult psychopathic traits is a history of problematic conduct during childhood (Frick et al., 2014). Early onset conduct problems are predictive of both greater levels of aggression (Lahey et al., 1998) as well as antisocial behaviour that persists beyond adolescence (the life stage in which most antisocial behaviour is carried out) (Loeber et al., 1989). Furthermore, the prevalence of CU traits in children with conduct disorder has been estimated between 10 and 50%, with this subgroup exhibiting greater levels of aggression and cruelty (Kahn et al., 2012). This aggression is not accounted for by other factors such as age of onset, severity of conduct, and associated factors such as impulsivity (Frick et al., 2014). Furthermore, CU traits have been identified as a precursor for childhood aggression and emotional disturbance (Moran et al., 2008), as well as antisocial behaviour (McMahon et al., 2010), and psychopathic traits in adulthood (Burke et al., 2007). CU traits are therefore viewed as a risk factor for aggression and antisocial behaviour (Frick et al., 2014).

In light of these findings the Inventory of Callous Unemotional Traits (ICU), was developed to measure these personality features in children (Frick & White, 2008). Additionally,

the construct has been clinically formalized through the “limited prosocial emotions” specifier for conduct disorder in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). This specifier describes a “lack of remorse or guilt; callous-lack of empathy; unconcern about performance at school, work, or in other important activities; shallow or deficient affect.” (American Psychiatric Association, 2013). As this construct has gained clinical currency, its aetiology, prodromal features, and clinical implications have all become targets of research.

1.0.1 Affective Features of CU traits

Within the current literature there is a strong consensus that CU traits have predictive utility for a range of social and emotional deficits. As is characteristic of adult psychopathic traits, CU traits in youth predict lower levels of anxiety, as well as reduced responses to punishment cues (Frick et al., 2014). Further, psychopathic traits are also characterised by a disjunction between two forms of empathy (Blair, 2005). Blair (2005) synthesised neurological evidence suggesting that multiple discrete neurological systems are involved in empathic responding. Blair suggested that there was no evidence that mental representation of others’ internal states (cognitive empathy) was impaired in individuals with psychopathic traits. Rather, that such individuals show reduced emotional responses to others’ emotional expressions (affective empathy). This framework has been empirically applied in a study of children with CU traits, with measures of cognitive and affective empathy showing the dissociation proposed by Blair (2005) (Jones et al., 2010; Schwenck et al., 2012). Further studies have even demonstrated longitudinal continuity of specific affective empathy deficits in CU traits (Brouns et al., 2013).

Deficits in affective empathy continue to be a major component of the contemporary study of CU traits, and have been addressed empirically at a causal, bio-behavioural level (Moul et al., 2017). Affective empathy deficits appear to result from diminished endogenous attention

allocation to emotions expressed through the eyes, leading to particular deficits in the recognition of fear (which is predominantly expressed through widened eyes) (Dadds et al., 2008; Sebastian et al., 2014). This deficiency seems to cause some developmental delays in cognitive empathy (explicit comprehension of other's emotional state) which resolve during puberty (Dadds et al., 2009). However, impairments in affective empathy are also relevant to the intentional use of harmful actions to obtain goals. Known as pro-active aggression, this type of behaviour is thought to normally be inhibited by affective empathic responses, hence the greater likelihood of such behaviour in individuals with CU traits (Blair, 2017). These features raise challenges in the treatment of antisocial behaviour in children with CU traits (D. J. Hawes et al., 2014).

1.0.2 CU Traits and Conduct Interventions

Typically, psychological interventions for children exhibiting disordered conduct are designed to address the strategies and responses of parents to their child's behaviour (Dadds & Hawes, 2006). In the framework informing these interventions (the Patterson Coercion model (Patterson, 1982)), children's learning about the outcomes of goal-oriented behaviour (instrumental learning) is a key process. This model focusses on the emotional responses given to children's antisocial behaviour by parents - which can act to reinforce such behaviour. Resulting increases in a child's antisocial behaviour may then lead to parental avoidance of the child, and consequently prosocial behaviour does not receive parental attention (or reward). Thereby this model considers children's learning about the outcomes of both prosocial and antisocial behaviour. Accordingly, therapies such as behavioural family interventions (BFIs, Dadds and Hawes, (2006) Sanders, et al., (2000)) aim to change these outcomes. They achieve this by training parents to respond to antisocial behaviour with emotionally neutral strategies (e.g., time out), and to reward prosocial behaviour (though descriptive praise).

As well as their association with conduct problems, CU traits also appear to limit the efficacy of BFIs. This limiting impact was reliably observed through a systematic review of interventional research (D. J. Hawes et al., 2014). This review indicated that poor interventional responses were a result of the characteristics of high CU children, rather than alternative, correlated factors such as poor parenting skills. However, some authors have suggested that greater severity of disordered conduct accounts for the apparent treatment resistance seen in high CU children (Waller et al., 2013). Furthermore, a range of interventions have been applied to conduct problems, and CU traits do not limit efficacy in all cases (Wilkinson et al., 2015). For example (Dadds et al., 2012) found that therapy designed to improve emotion recognition had greater efficacy than a standard behavioural family intervention. The need to understand the psychological characteristics of individuals with these CU traits is clear, as these characteristics appear to precede antisocial behaviour in these individuals and may limit clinical interventions on such behaviour.

1.1 Aetiology of CU Traits

To understand why individuals with CU traits may have diminished responses to behavioural interventions, it is necessary to understand the aetiological theories which have addressed instrumental learning deficits concomitant with these traits. Much of the relevant theoretical body of work addresses the affective component of psychopathic traits, which can be considered as cognate to CU traits (Frick et al., 2014), and has strong genetic underpinnings (Blair et al., 2005). The heritability of CU traits in children, adolescents, and young adults has been estimated at 36-67% (Moore et al., 2019). Although this value is similar to estimates for other personality traits in children (Polderman et al., 2015), some studies document higher heritability for longitudinally stable CU traits (Blonigen et al., 2006; Henry et al., 2018).

Consistent with an elevated role of hereditary factors, Marshall & Cooke (1999) observed that negative experiences of childhood care and abuse were less predictive of adult psychopathic traits at higher levels of the trait scale. Similarly, Wootton et al. (1997) found that environmental factors such as poor discipline practices were strongly linked to conduct disorder in children without CU traits, but did not account for conduct problems where CU traits were present. However, more recent work has found that children with CU traits may indeed be sensitive to parenting quality, but that their response to poor parenting may be qualitatively different to that of non-CU children (Crum et al., 2015).

Due to the apparent heritability of CU traits, genetic processes are core to many mechanistic accounts of these traits. These theories have attempted to identify processes that bridge biological factors and social behaviour (Blair et al., 2005; Hamilton et al., 2015; Moul et al., 2012, 2017; Viding & McCrory, 2019). However, environmental factors remain significant determinants of CU trait expression. In a twin study, Takahashi et al. (2021) found that non-shared environmental factors had more influence on the long-term trajectory of CU traits (42%) than in their initial levels (21.7%). One such environmental factor of note is parental warmth. Waller et al. (2018) found that this factor played a more significant role in CU trait expression than genetic factors in sample of twin children. Further, Pasalich et al. (2011) found a correlation between CU traits and parenting characterised by reduced empathy and positive regard. Correspondingly, clinical interventions impacting the emotional harshness of parenting appear to have lasting impacts on the expression of CU traits (McDonald et al., 2011). It should also be noted that some expression of CU traits appears to stem directly from experiences of abuse leading to emotional withdrawal and coldness (Bennett & Kerig, 2014). This “Secondary” form of CU traits is often discriminated by the presence of internalizing symptomatology and anxiety

and is distinct from the genetically derived construct that is addressed in the theories described below (Craig et al., 2021).

1.1.1 Low Fear Hypothesis

Lykken (1995) attributed psychopathic traits to dispositionally muted fear responses. He compared the antisocial behaviour of psychopaths to the self-inflicted injuries that often accompany congenital analgesia: “Like the ability to experience pain, the fear mechanism is especially useful early in life before the individual’s judgment and reason are sufficiently dependable guides to behaviour.” (pg. 144). In support of this fear deficit Lykken cited three experimental paradigms in which individuals with psychopathic traits ostensibly display emotional deficits: evaluation of descriptions of frightening scenarios, skin conductance responses (autonomic arousal) to stimuli associated with painful electric shock, and passive avoidance learning. Lykken (1995) operationalised avoidance learning, or “passive avoidance”, using a task in which participants are presented with a series of decisions and must learn which of the options in this series lead to painful electric shock. Deficits in these tasks therefore illustrated an impaired experience of fear.

Lykken’s view has long been criticized, primarily for its failure to consider evidence showing links between fear independent deficits and psychopathic traits (Blair et al., 2005). For instance, Newman and Kosson (1986) showed such independence in a passive avoidance task. In their procedure subjects were presented with trials of eight different stimuli, four of which required responses, and four of which required responses to be withheld. In one condition subjects gained money for responding correctly, and lost money for responding incorrectly. Subjects classed as psychopathic showed deficits in learning in this condition. Significantly, in a further condition, which did not include monetary rewards, they showed no such deficit.

Therefore, in this study the variable regulating deficits associated with psychopathic traits was reward rather than fear.

Another contentious element of Lykken's view is its reliance on a classical concept of emotion wherein emotional experiences are categorized according to constructs present in our language (fear is an example). This view assumes that such categories have corresponding "essences" in our neurology, expressions, behaviour, and internal experience (Barrett, 2017). However, this assumption has been disputed in the contemporary literature. For example Barrett (2017) questions the existence of such essences, highlighting the dependence of theories such as Lykken's on inferred neurological phenomena that cannot be empirically demonstrated. More recent aetiological accounts of psychopathic traits have made a greater effort to account for a broader array of cognitive features of these traits and address the biological underpinnings of such features.

1.1.2 Integrated Emotion Systems model

Blair et al. (2005) attempted to address the lack of biological explanation in Lykken (1995) by considering the role of amygdala neurology in fear processing deficits. The Integrated Emotion Systems model referenced animal lesion studies, natural lesion case studies in humans, as well as neuroimaging and electroencephalogram data. This information was used to suggest that the functionality of the amygdala is essential to creating associations between stimuli and affective experiences. For example, in the passive avoidance task, responses to different stimuli lead to aversive or rewarding outcomes. An individual can therefore perform well in this task if they associate these stimuli with the affective experience of these outcomes. Contrastingly, if an individual is deficient in their capacity to make such associations, they will show impaired performance (as is the case in individuals with high psychopathic traits). The authors used this

logic to account for other behavioural deficits associated with psychopathic traits: autonomic responses to fear associated stimuli, and responses to emotional language. On this basis the integrated emotion systems model held that the aetiology of psychopathy involves abnormal amygdala function.

Blair et al. (2005) acknowledged that this mechanism is not explanatory of other instrumental learning deficits associated with psychopathic traits: reversal and extinction learning. An example of the response reversal paradigm is a study by Budhani and Blair (2005), which used a sample of children with conduct problems and varying levels of psychopathic traits. In this task subjects initially experienced that the selection of one stimulus leads to reward (cow picture → gaining points) whilst the selection of another leads to punishment (dog picture → losing points). In a second phase of the task, the contingencies are reversed, so that the participant must learn to choose the previously punishing stimulus and stop choosing the previously rewarding stimulus. Children with psychopathic traits showed poor reversal learning in this task.

Psychopathic traits are also associated with poor extinction learning. An example of an extinction learning task (as referred to by Blair et al. (2005)) is a study conducted by Newman et al. (1987). In their task, subjects could opt to play cards from a virtualized deck or quit the task. Initially, playing cards led to reward reliably, but throughout this task the probability of being punished for playing a card increased, requiring participants to withhold their behaviour in order to avoid punishment. The authors observed that high trait psychopathy was associated with response perseveration i.e., playing greater numbers of cards before withholding this behaviour. Blair et al. (2005) also cited Budhani et al. (2006); Budhani and Blair (2005), O'Brien and Frick

(1996), and Fisher and Blair (1998) as evidence for these behavioural deficits, which appeared to reflect poor inhibitory learning and could not be explained by amygdala based processes.

Whilst extinction and response reversal tasks are procedurally different, their common associative components are crucial to the account of Blair et al. (2005). The extinction paradigm is broadly treated as a measure of inhibitory learning about the cessation of an expected outcome (Bouton et al., 2021). It has therefore been applied to studies of exposure therapy for reducing fear in PTSD or drug cravings in addiction (Bouton et al., 2021). Contrastingly the reversal learning paradigm involves switching behaviour to a new response (rather than ceasing a behaviour), and has been considered a model of behavioural flexibility (Clark et al., 2004). Despite differences in their associative structure, both extinction and reversal paradigms have been used to understand increased rates of recidivism in individuals' psychopathic traits (Blair et al. 2005). This is because both tasks measure behavioural persistence in the face of ostensibly inhibitory outcomes. Furthermore, behavioural change in both these tasks is thought to depend on common neurological processes (Wilson et al., 2014). To explain psychopathic trait related deficits in reversal and extinction tasks Blair et al. (2005) pointed to the functionality of the Orbito-frontal Cortex (OFC). They cited evidence showing that lesions of this region impair responses to these tasks in animals (Butter, 1969). They also noted the established relationship between humans' neural activity in the OFC and their performance on response reversal and extinction tasks (Clark et al., 2004). Blair et al. (2005) therefore suggested that OFC function may mediate the role of genetics in the development of psychopathic traits.

1.1.3 Differential Amygdala Activation Model

Subsequent to the integrated emotional systems model, Moul et al. (2012) proposed an account of deficits in response reversal and extinction learning. This perspective relied solely on

the functionality of the amygdala. To achieve this, the Differential Amygdala Activation Model considered the functionality of substructures within the amygdala. This functional account is based on a lesion study of rat behaviour conducted by Balleine et al. (2003). These authors lesioned the basolateral amygdala (BLA) of subjects and trained them to respond instrumentally for food outcomes (the lesion did not impair their ability to learn these responses). After subsequently satiating these animals on a food outcome, lesioned animals continued to seek that food. Contrastingly, satiation led control animals to withhold responses for that food. The authors interpreted this result as evidence that the BLA is crucial for the brain's processing of the current incentive value of outcomes of behaviour. In the procedure of Balleine et al. (2003) this meant that lesioned rats could learn a response that lead to a rewarding outcome. However when the value of that outcome was been manipulated through satiation, these animals where neurologically incapable of adjusting their behaviour to reflect this diminished value.

The Differential Amygdala Activation Model applies this BLA impairment to response reversal deficits associated with psychopathic traits. In this account of such tasks (for example Budhani and Blair (2005)) an impairment of incentive representation would allow subjects to associate a response with reward (in this case points). However, after the response reversal manipulation these individuals would be impaired in their capacity to compare the new incentive value of the response outcome (points) to the value they had come to expect (loss of points), leading to apparent delays in their reversal learning. In addition to Budhani and Blair (2005), Moul et al. (2012) accounted for the results of a number of studies in this manner, specifically: Newman et al. (1987), Budhani et al. (2006) and Mitchell et al. (2002).

In addition to refining the concept of the amygdala's functionality in psychopathy deficits, the Differential Amygdala Activation Model also addressed evidence that attentional

processes contribute to the development of these traits. For example, Moul et al. (2012) note evidence from Dadds et al. (2006) showing that children with CU traits show normal levels of fear recognition when directed to look at the eyes of faces expressing fear. To account for this Moul et al. (2012) cited functional MRI studies showing that endogenous shifts in eye gaze to fear expressions are linked to amygdala activation. Thus, amygdala impairments could diminish such attentional shifts. The mechanisms underlying impairments in bottom-up attentional allocation were also the focus of the impaired integration framework (Hamilton et al., 2015).

1.1.4 Impaired Integration Framework

In contrast to the Differential Amygdala Activation Model, the impaired integration approach focussed on the network level of the brain, where combination of sensory information and existing encoded information occurs (Hamilton et al., 2015). Exemplifying this, Hamilton et al. (2015) point to the card playing task conducted by Newman et al. (1987) which I have described here previously. These authors found that psychopaths' tendency to perseverate on a response could be affected by attentional manipulations. Specifically, no perseveration was observed when participants were shown a record of the outcomes of their choices and given extra time to consider this feedback. The authors suggested that the observed perseveration effect resulted from a reduced capacity to compare the new experience of losing money to the memory of winning it throughout the game. Thereby they explain why procedures designed to facilitate this comparison reduced perseveration. Hamilton et al. (2015) contextualised this effect in a framework where high psychopathic traits are linked to limited attentional resources. They suggested that focusses of attention, such as reward seeking behaviour, may consume these limited resources, impairing an individual's ability to encode perceptual information.

1.2 The Significance of Learning Deficits in CU Traits

The description of the theories above does not constitute an exhaustive summary of each model, nor does it cover all the existing theories of psychopathic traits. However, what can be seen in these descriptions is the dependence of these four models on behavioural studies of extinction and response reversal learning. Despite this dependency, Blair et al. (2005), Moul et al. (2012) and Hamilton et al. (2015) collectively cite only six studies to support the existence of psychopathic trait based deficits in such learning (Budhani et al., 2006; Budhani & Blair, 2005; Fisher & Blair, 1998; Mitchell et al., 2002; Newman et al., 1987; O'Brien & Frick, 1996). However, a range of other research has examined instrumental learning processes in individuals with CU traits (e.g., Brazil et al., 2013; Dargis et al. 2017; De Brito et al., 2013; Gregory et al., 2015) and the consistency of results across this literature is unclear. Additionally, whilst instrumental learning deficits have been seen in adults and children (e.g., Budhani et al., 2006; Budhani & Blair, 2005) the effect of variables such as age do not appear to have been systematically examined. Furthermore, within this small pool of findings, the affective dimension (akin to CU traits) of psychopathic traits is not always distinguished from other factors of the psychopathic trait spectrum. For example, Budhani and Blair (2005) analysed the correlation between reversal learning performance and the APSD trait measure, which includes narcissism and impulsivity dimensions, in addition to a callous unemotional factors. As these authors only used the total score as a predictor, it is unclear which of these subdimensions contributed to the observed reversal learning. This creates ambiguity as to whether the construct measured in this study corresponds to that which is addressed in the theories described above. These weaknesses in the literature are significant not only because these effects are important to the accuracy of theoretical explanations of psychopathy, but also because of the implications that

extinction and reversal learning deficits have for interventions on CU trait related behaviours. There is therefore a substantial need for further research to resolve these issues.

1.2.1 Learning Deficits and Conduct Problem Interventions

Response reversal learning deficits are a significant process for understanding the clinical characteristics of CU traits. This is due to the relationship between the structure of reversal tasks and that of the Behavioural Family Interventions described above. The structure of these interventions can be simplified into a response-outcome model (Dadds & Hawes, 2006). This conceptualization divides the child's experience of BFI into two phases: pre-intervention and post-intervention. During pre-intervention experiences, the child is reinforced for antisocial behaviour, and receives no outcome for pro-social behaviour. If the goals of the intervention are met, antisocial behaviour is disincentivised or receives no outcome, whilst pro-social behaviour is reinforced (Dadds & Hawes, 2006). Learning about these contingencies is considered as a key process in the development and treatment of disordered conduct (D. J. Hawes & Dadds, 2005). As a result, the treatment resistance of individuals with CU traits has been attributed to punishment insensitivity, along with other factors including parental warmth, emotional awareness, and comorbidities (D. J. Hawes et al., 2014).

The associative structure of behavioural family interventions is also present in the response reversal task. Prior to reversal participants learn about a set of incentivized and disincentivised behaviours, then, after reversal (analogous to intervention), previously incentivised behaviours are disincentivised (see Table 1.1 for a summary of this comparison). This task may therefore serve as a behavioural model of behavioural family interventions. In the context of this model, deficits in such learning in children with CU traits (e.g. Budhani & Blair, 2005) can be seen as parallel to impaired responses to behavioural interventions.

Table 1.1 *Comparable Contingency Reversals in the Experience of BFI and in Response Reversal Tasks*

Model	Phase 1	Phase 2
BFI response-outcome model	Antisocial behaviour → reward	Antisocial behaviour → neutral responses or omission of reward
	Prosocial behaviour → no outcome	Prosocial behaviour → reward
Response reversal task	Response 1 → reward	Response 1 → punishment
	Response 2 → punishment	Response 2 → reward
Extinction learning task	Response → reward	Response → no reward

1.2.2 Extinction and Reversal Learning: Context Dependent Processes

The evidence described above signifies that studies of reversal learning may be used to model the clinical features of CU traits. This significance can be extended to studies of extinction learning, as extinction and response reversal tasks are assumed to measure the same learning mechanisms in the theories of both Blair et al. 2005, and Moul et al. 2012. Indeed, these paradigms share structural properties (see Table 1.1), which require participants to learn to withhold a previously reinforced response. This inhibitory learning process has been the subject of a long tradition of research, dating back to Pavlov (1927), which can offer novel predictions to for psychopathic and CU traits. Specifically, the relevance of context to this form of inhibitory learning.

Context is a key regulator of the expression of inhibitory learning, as reviewed by Bouton et al. (2021). These authors described how this principle of the associative learning literature was initially established through observation of the renewal effect. This effect can occur in extinction tasks where a behaviour is initially rewarded repeatedly (“trained”) before an

extinction phase of the experiment wherein behaviour is not rewarded (depicted in Table 1.1). If training occurs in a different context than extinction, a return to the training context often results in a ‘renewal’, or return, of the behaviour that had been previously extinguished. Bouton et al. (2021) articulated the widely accepted implication of this result: that the behaviour is not “forgotten” or “unlearned” during extinction, rather, a new memory of withholding that behaviour is formed. This memory’s influence on behaviour is determined by the presence of the extinction context. This accounts for why inhibition of behaviour ceases (or extinguished responding is “renewed”) upon leaving the extinction context. Whilst context is commonly manipulated through changing olfactory or visual features of the training environment, the concept has been extended to temporal features, and features linked to experience of the unconditioned stimulus itself (Bouton et al., 2021). Thus delays after extinction can lead to the return of trained behaviour (spontaneous recovery) as can exposure to the unconditioned stimulus (reinstatement) (Bouton et al., 2021).

Given existing cognitive and neurological observations in individuals with psychopathic traits, it is possible to hypothesise that these traits will impact contextual regulation of extinction learning. Whilst the impact of these traits on the role of context in extinction learning has not been previously examined, other factors have been observed to affect this process. Gershman et al. (2013) made such observations in a fear conditioning task with rat subjects. After establishing a contingency between a shock and a stimulus, these authors intermittently omitted the shock outcome during stimulus presentations, before fully extinguishing this contingency. This manipulation reduced spontaneous recovery and reinstatement of fear responding. In the authors’ terms, their manipulation obscured the discriminability of training and extinction experiences. Thereby, the contextual factors relevant to each of these experiences was less informative to the

animals' behaviour. Similarly, the behaviour of individuals with psychopathic traits has been seen to reflect diminished capacity to discriminate experiences of distinct outcome contingencies in learning tasks (e.g., Budhani & Blair, 2005). Given the theory substantiated by Gershman et al. (2013) it is possible that this lack of discrimination leads to diminished contextual control of extinction learning in those with psychopathic and CU traits.

This behavioural evidence is supported by neurological studies. Finger et al. (2011) observed that these individuals with psychopathic traits had diminished responsivity of the Orbitofrontal – cortex (OFC) during instrumental learning. The relationship between OFC activity and reversal learning has been formally modelled by Wilson et al. (2014). This model suggests that reduced activity in this region predicts slow extinction learning and diminished renewal of extinguished behaviour. Neurological data therefore supports the hypothesis that CU traits will be linked to lessened renewal of extinguished responding.

The attentional deficits noted in the impaired integration framework (Hamilton et al. 2005) also support this hypothesis. Specifically, this framework suggests that while individuals with psychopathic traits are engaged in incentive seeking behaviour, they will be less able to encode environmental features peripheral to that activity (i.e., contextual features). As established previously, encoding of such contextual features is necessary to experience relapse of extinguished responding. Despite the array of theoretical and behavioural evidence substantiating this hypothesis, I am not aware of studies addressing the contextual regulation of extinction learning in individuals with CU traits.

1.2.3 Context Learning and Psychological Interventions

The importance of this hypothesis is further supported by its clinical implications. As noted, children scoring highly on measures of psychopathic traits show deficient response

reversal (Budhani & Blair, 2005). This observation may be a result of diminished capacity to discriminate experiences of instrumental contingencies pre and post intervention (shown in Table 1.1) based on the learning model of Wilson et al. (2014). However, given the necessity of this discrimination in the formation of context specific inhibitory memories (Gershman et al., 2013), it may be possible that any behaviour learned during reversal may be less sensitive to context. These processes can be applied to behavioural interventions on antisocial behaviour (with the assumption that reversal tasks can validly model behavioural instrumental learning occurring in these interventions). Based on this model, it would be expected that such interventions would have less initial impact on problematic conduct in individuals with CU traits (reflecting delayed reversal learning). But any impact of these intervention may be better retained outside of the interventional context (reflecting the context insensitivity of this learning). Although this implication can be theoretically substantiated, I am not aware of empirical evidence regarding the contextual sensitivity of interventions on behaviour of individuals with CU traits.

Although unconsidered in the context of CU traits, contextual learning is commonly regarded as an important regulator of treatment efficacy in psychopathologies such as phobias and drug use. Treatment of phobia involves exposing individuals to the presence of feared stimuli in the absence of feared outcomes (for example a spider in the absence of a bite) in order to reduce expectation of those outcomes (Hofmann, 2008). Similarly, cue exposure therapies address addiction by extinguishing the association between drug paraphernalia and drug effects (Conklin & Tiffany, 2002). These interventions have short term efficacy, but are limited by relapse occurring when participants leave the therapy context (Conklin & Tiffany, 2002; Hofmann, 2008). Such relapse can be understood a context dependent renewal of the extinguished response (Podlesnik et al., 2017). Due to their ubiquity, context effects represent a

significant barrier to treatment efficacy (Conklin & Tiffany, 2002; Hofmann, 2008), and as a result efforts have been made to design pharmaceuticals (Singewald et al., 2015) and modify therapies in order to reduce the impact of context learning (Conklin & Tiffany, 2002; Podlesnik et al., 2017). These modifications include carrying out extinction procedures across a range of contexts in order to reduce the context specificity of learned behaviours (Conklin & Tiffany, 2002). The relevance of context effects in therapies dependent on extinction learning emphasises the need for investigations of these effects in the context of CU traits.

1.3 Thesis Aims and Hypotheses

The literature described above demonstrates that extinction learning processes are highly significant to CU traits. Currently, aetiological theories of these traits rely on a limited body of research to justify their implication of extinction learning. Further study of this learning also offers potential insight into unconsidered contextual factors relevant to the treatment of conduct problems in individuals with CU traits. This thesis therefore has three aims. Firstly, to assess the current extent of evidence supporting deficits in the extinction and reversal of instrumental behaviour amongst high CU individuals. Secondly, to generate data to support the reliability of this association. And thirdly, to demonstrate that CU traits predict limited renewal of extinguished behaviour. Such data would support the veracity of contemporary theories of CU trait aetiology. It would also implicate the contextual regulation of extinction learning in the instrumental behaviour of individuals with CU traits for the first time. .

2 Reviewing Extinction Behaviour in Callous Unemotional Traits

Contents

2.0.1	Predictions from Extinction Learning Literature	33
2.0.2	Aims and Hypotheses	38
2.1	Method	39
2.1.1	Eligibility Criteria	39
2.1.2	Search Methodology	40
2.1.3	Study screening and data extraction	41
2.2	Results	41
2.2.1	Psychometrics	42
2.2.2	Sample Sources	43
2.2.3	Response Reversal Tasks	44
2.2.4	Extinction Tasks	50
2.3	Discussion	55
2.3.1	Duration of Training and Extinction Phases	55
2.3.2	Reinforcement Probabilities.....	57
2.3.3	Context Changes	58

2.3.4 Traits and Populations..... 59

2.3.5 Future Studies 60

2.3.6 Implications and Conclusion..... 61

This review addresses my goal of assessing the existing evidence linking CU traits to extinction and reversal learning deficits. In the previous chapter it was shown that across major theories of CU trait aetiology, only a small pool of studies has been cited as evidence for these effects. As this literature as a whole has not been the subject of a review, it is unclear whether it is well represented by this research. The primary goal of this Chapter was therefore to systematically compile all relevant studies to determine the reliability of the association between psychopathic traits and extinction learning in experimental settings. I also sought to critique the validity of these experiments as measures of learning processes, as this validity is widely assumed in theoretical literature described in Chapter 1. To appraise this construct validity, I have presently outlined a range of behavioural effects that can be expected to occur in individuals with extinction learning deficits. I also documented operationalisations of psychopathic and CU traits utilised in this literature. This review therefore provides novel insight into the validity of methodologies used to measure extinction and reversal learning deficits in individual with CU traits, and the reliability of resulting observations.

2.0.1 Predictions from Extinction Learning Literature

Extinction learning is influenced by several factors, such as encoding of context (Bouton et al., 2021). It should be noted that the focus of this thesis is learning about goal directed behaviour (instrumental learning), rather than classical conditioning, although extinction and renewal are relevant to both processes. This focus is justified as instrumental learning procedures are particularly relevant as behavioural models of typical conduct interventions (see section 1.2.1). Therefore any evidence for moderators of this learning would be informative for the efficacy of such interventions. In the previous chapter, context learning was introduced as one such moderator. However, other known moderators are introduced below, and include the

duration of initial training, the duration of extinction (or reversal) training, the probabilistic contingency between outcomes and actions, and context changes across training and extinction. Each of these factors has an established impact on behavioural responses to tasks designed to measure extinction and reversal learning (Bouton et al., 2021). Therefore, the presence of these effects in the relevant CU trait research would support the validity of behaviour in these studies as a measure of these learning constructs.

2.0.1.1 Context

Context is one of the most studied regulators of the expression of extinction learning (Bouton et al., 2021) and should therefore impact the behaviour observed in studies of extinction and reversal learning in CU traits. In the prior chapter I developed the hypothesis that individuals with CU traits would exhibit less sensitivity of extinction learning to the context of that learning. Novel implications of this hypothesis included reduced relapse (or renewal) of pre-intervention behaviours for individuals high in CU traits. This review therefore provides an opportunity to systematically determine whether any evidence exists for this effect. Doing so will contribute to a primary aim of this thesis – demonstrating the relevance of renewal to CU traits.

2.0.1.2 Duration of Initial Training

Another significant determinant of the extinction of instrumental learning is the number of pairings between a behaviour and an outcome. This variable has been found to be an important determinant in animal studies of this learning. For example, Mackintosh (1963) trained rats to jump to one of two platforms which provided access to food, based on a black or white discriminative stimulus. It was found that rats who had received greater amounts of training were more resistant to extinction. That is, they took longer to stop exhibiting a preference for the trained platform after reinforcement had been omitted. Training appears to be a determinant of

behavioural flexibility in animal models. This observation emphasises the importance of early intervention on disorders of instrumental behaviour, as prolonged reinforcement of antisocial behaviour may impair inhibitory learning about that behaviour.

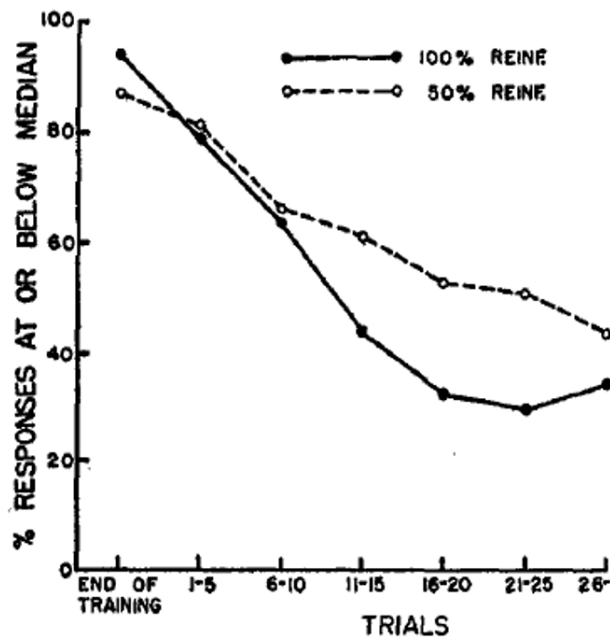
Despite the lack of direct evidence, theoretical work may provide grounds for a hypothesis about the interaction between CU traits and extended training. Moul et al. (2012) asserted that psychopathic traits are related to deficiencies in the value representation of outcomes. In animal models overtraining is also seen to be a cause of reduced outcome value representation. For instance, in a study by Holland (2004), a rewarding outcome was associated with nausea by pairing it with a lithium chloride injection. Rats typically ceased responses associated with this reinforcer. Such behaviour reflected the presence of an internal model wherein the trained response is causal to a food associated with nausea. Contrastingly, animals who received more initial training continued to respond for the nausea associated reinforcer. Thereby overtrained animals' behaviour did not reflect any internal representation of the outcome of that behaviour (Holland, 2004). As both psychopathic traits and overtraining have been viewed as causes of deficient outcome value representation, it can be expected that more extensive training will impair extinction or reversal learning in those with low PT or CU traits. These individuals' behaviour should therefore be commensurate with that resulting from the deficit present in those high on these traits.

2.0.1.3 Probabilities of Reinforcement

Another relevant factor for instrumental learning is the consistency of behavioural reinforcement. To examine the effects of reinforcement consistency, experimenters have manipulated the probability of reinforcement during the establishment of an instrumental response. This methodology is prevalent in the associative learning literature due to the

phenomenon of the partial reinforcement extinction effect, and is exemplified by Sheffield (1949). In this study rats were trained to run down an alley to receive food rewards. After experiencing food as a reward for this action on a 50% basis, animals were slower to extinguish their running behaviour than animals reinforced after 100% of runs (see Figure 2.1). Such findings are broadly construed to indicate that partial reinforcement reduces the discriminability of the extinction and training phases (Bouton et al., 2021).

A similar effect of reinforcement probability can be seen in the psychopathy literature. In their probabilistic choice task, Budhani and Blair (2005) established a reinforced choice and a punished choice before testing the effect of reversing these contingencies, a manipulation analogous to withholding reinforcement in an extinction paradigm. These authors also manipulated the probability of reinforcement in these contingencies, and the resulting behaviour was commensurate with the partial reinforcement extinction effect. That is, less consistent reinforcement led to slower behavioural adaptation after reversal of the behaviour's outcome. Notably, this effect was exaggerated for children with high levels of psychopathic traits. If this finding reflects a deficit in a process required for extinction learning, CU traits may exaggerate the impact of low reinforcement probabilities. Considering reversal learning as a model of behavioural intervention (see section 1.2.1), such results would suggest that inconsistent parenting should exaggerate the treatment resistance of dysfunctional behaviours in children with CU traits. The relevance of inconsistent parenting has also been noted empirically as factor causal to disordered conduct in such children (D. J. Hawes et al., 2014).

Figure 2.1 *Impact of Reinforcement Probability on Extinction, Sheffield (1949)*

Note. Reported on pg. 517 of Sheffield (1949). Lower values on the vertical axis indicates that animals are moving to the food location at a slower rate. The horizontal axis reflects the number of trials since reinforcement became unavailable.

2.0.1.4 Number of Extinction/Reversal Trials

Another important factor in studies of instrumental learning is the duration of extinction or reversal manipulations. It can be seen in the data from Sheffield (1949) (Figure 2.1) that whilst partial reinforcement slows extinction, the extinction learning curves for partially and continuously reinforced responding eventually converge. I would therefore expect that any extinction or reversal learning deficit will not be evident after sufficient extinction or reversal training has been delivered (although the duration of extinction required to mask these deficits is unclear). Extrapolating this hypothesis; an intervention on instrumental learning in children with CU traits (and deficient extinction learning) may show initial inefficacy, but this inefficacy may be surmountable with extended interventions.

2.0.2 Aims and Hypotheses

In the material introduced above it is evident that several variables are established determinants of extinction and reversal learning in both animals and humans. Therefore, observation of the effects of these factors is critical to demonstrating that the behavioural disparities evident in individuals with PT and CU traits are due to associative learning processes. As a result, the primary aim of this review was to apply the behavioural hypotheses introduced above to existing literature. These hypotheses were:

1. Context should regulate the expression of behaviour established during extinction, although to a lesser extent in individuals with high CU traits.
2. Behaviours undergoing prolonged training should extinguish or reverse at a slower rate. This may lessen apparent differences between individuals with high and low CU traits.
3. Intermittent or partial reinforcement should delay extinction or reversal of behaviour. This effect will impact high CU individuals particularly.
4. Extended extinction (or reversal) will normalise the behaviour of high CU individuals.

Evaluation of these hypotheses was intended to inform the validity of suggestions that learning impairments underlie CU traits (Blair et al., 2005; Moul et al., 2012). Further, observing the consistency of relevant studies would act to determine the reliability of the evidence for such learning impairments. Additionally, I have noted that studies referenced in major theories of CU traits do not directly measure CU traits specifically, rather they assess the subsuming construct of psychopathic traits (e.g., Budhani & Blair, 2006). As a result, I also aimed to document whether

the presently reviewed studies measured the affective features of psychopathic traits (or CU traits) specifically.

2.1 Method

The following methodology was designed to systematically identify any experimental studies measuring psychopathic traits (PT) or CU traits and learning processes. The systematic strategy involved an unbiased database search and text screening (summarized in Figure 2.2).

2.1.1 Eligibility Criteria

All studies considered in this review measured Psychopathic or CU traits using an established psychometric instrument. Additionally, they included some form of experimentally conducted task which I considered to be a measure of response reversal learning or extinction learning. Task inclusion was partially pre-specified, as I identified widely utilized learning tasks in the psychopathy literature for inclusion in my search terms. These tasks included:

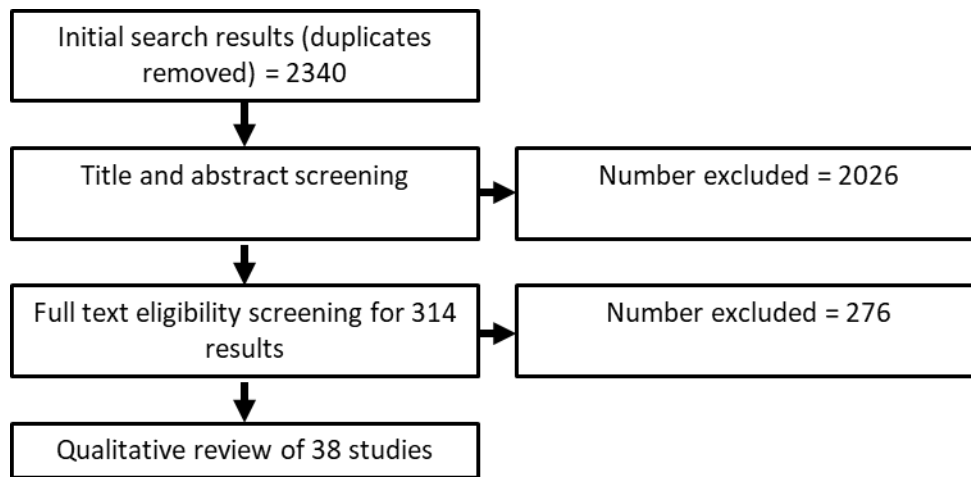
- The Probabilistic Response Reversal Task - e.g. Budhani and Blair (2005).
- The Wisconsin Card Sorting Task, which is frequently used to measure executive function. However it has been suggested that response reversal processes may be relevant to performance (Grant & Berg, 1948).
- Intra/Extradimensional Shift Task (includes response reversal phase) (Blair et al., 2001).
- Go/No Go paradigms which include reversals of correct responses to stimuli - e.g. Lapierre, Braun and Hodgins (1995).
- Response Perseveration Tasks, such as Newman et al. (1987) which have been referred to as a measure of extinction learning (see Blair et al. (2005)).

My eligibility criteria were that studies needed to report the effect of PT on the behavioural occurring in their utilised tasks. As the present hypotheses are behavioural, I

excluded studies that reported only neurological measures (this was the case in a number of studies of fear extinction). Additionally, all studies published prior to 1980 were excluded, as the first validated psychometric tool for PT was published in 1980 (Psychopathy Checklist (PCL) (Hare, 1980)). Studies measuring PT with scales developed prior to the PCL were also excluded. The final criterion was that only peer reviewed primary research was eligible; all reviews, theses and dissertations were excluded.

2.1.2 Search Methodology

On the 31st of May 2019 I carried out the following search strategy in four databases (PsychInfo, PubMed, Scopus and Web of Science). This search was then updated on the 4th of January 2021. The search included terms “psychopath”, “psychopathy”, “psychopathic” and “callous*” to detect references to either adult or child forms of the psychopathy construct. I used “No*go”, “perseveration”, “wisconsin card sorting”, “response reversal” and “*dimensional shift” terms to return the specific tasks I had determined to be relevant. The terms “conditioning”, “conditioned”, “associative”, “reinforcement” and “learn*” were designed to return any other studies measuring learning processes. Psychopathy related terms were combined with task related terms using the ‘and’ operator to narrow the scope of results. This search produced 2340 unique results.

Figure 2.2 *Process for Identifying Eligible Studies*

2.1.3 Study screening and data extraction

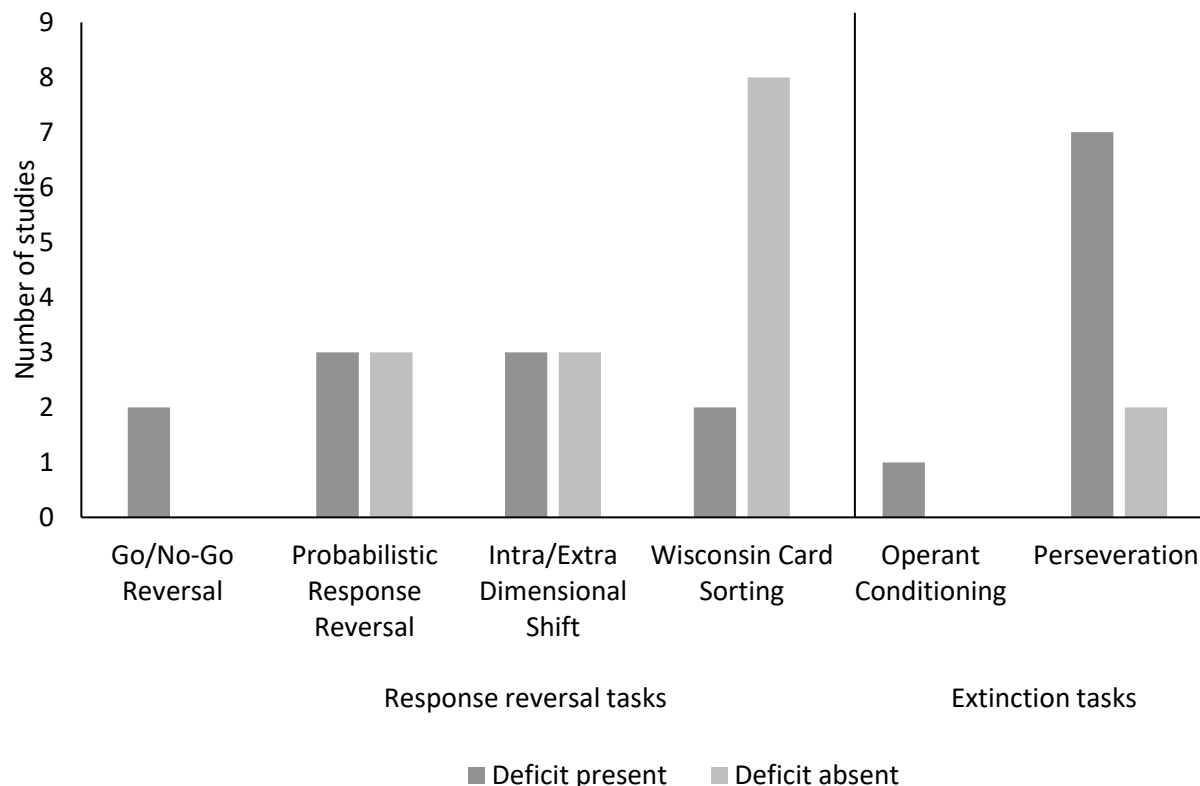
The eligibility criteria were applied to all results. If studies could not be excluded using their title and abstract, the full text was assessed. After screening, 38 papers remained. The papers were then coded using the following task related information: the type of task used, the number of trials in which task stimuli or responses were associated with outcomes (before and after changes in outcome contingencies), and the probability that each stimulus or response would lead to an outcome. For sample characteristics the age (child vs adult), size, and source of each sample group was collected, as well as the psychopathy measure utilized. Finally, I recorded whether there was a behavioural effect of PT or CU traits on responding, and if this effect was specific to a particular subscale of the measure. These data were considered in a qualitative manner, as substantial methodological variation precluded any quantitative aggregation of these data.

2.2 Results

The following synthesis describes the tasks that were utilised in the studies identified by the systematic search protocol. Variables that appeared to have some consistent impact on task outcomes were identified, along with other common features of the studies. Notably all studies

reviewed here utilised operationalisations of PT, rather than CU traits. Therefore, the results will primarily be described with reference to PT. This description includes an evaluation of whether there is consistent evidence of a PT related effect in each of the category of task: response reversal and extinction. A summary of the data extracted from the reviewed studies can be found in Table 2.1 for response reversal tasks, and Table 2.2 for extinction tasks. The overall distribution of findings consistent with PT associated learning deficits, compared to null findings is shown in Figure 2.3.

Figure 2.3 *Tasks Evidencing Psychopathic Trait Related Behavioural Deficits*



2.2.1 Psychometrics

Six measures of PT and CU traits were employed in the reviewed studies, most of which are derived from the Hare Psychopathy Checklist (PCL) (Hare, 1980). These scales included the PCL Revised (PCL-R, Hare (1991)), PCL screening version (PCL-SV, Hart et al. (1995)).

Additionally, two such scales were specifically developed for children; the PCL-Youth Version (PCL-YV, Forth et al. (2003)) and the Antisocial Process-Screening Device (APSD, Frick and Hare (2001)), which was originally termed the Psychopathy Screening Device (PSD). The original PCL scale comprised of two factors: an interpersonal/affective factor and an impulsive/antisocial factor (Harpur et al., 1989). This model has been applied to later scales, which have also been considered under a three-factor model. In this model one factor reflects a callous, deficient affect, another measures narcissism and grandiosity and a final factor measures impulsivity (Cooke & Michie, 2001; Frick et al., 2000).

Two studies utilized the Triarchic Psychopathy Measure. This device measures three factors, defined as follows (Patrick, 2010). Meanness, designed to measure callousness and affective characteristics, and which correlates with the affective PCL-R factor. Boldness, measuring grandiosity and features addressed by the interpersonal subscale of the PCL-R. And disinhibition, which reflects impulsivity and features of the PCL-R lifestyle factor. Of the studies that recorded effects of psychopathic traits only six included any analysis based on the underlying factors – or subscales – of these measures. I have noted these where they occur, otherwise any effects documented are those of the overall PT measure.

2.2.2 Sample Sources

Out of 32 studies using adult samples, four recruited exclusively from the general community whilst the remainder sampled, at least partially, from a forensic setting. Forensic samples were from detained or criminally convicted populations, and were sometimes selected for diagnosed anti-social personality disorder (ASPD). As the presence of this diagnosis was not relevant to the current hypotheses, I did not record this variable.

All six studies using samples of children recruited either from a population with conduct problems (defined by clinical assessment) or from a school for children with behavioural difficulties. The presence of attention deficit disorder (ADHD) was also assessed in some studies but was not subject to any analysis here. Control samples were often obtained from the general community.

2.2.3 Response Reversal Tasks

This systematic review identified four methodological paradigms wherein respondents were required to reverse learned responses to given stimuli. Primary features of these tasks are described below and tabulated in Table 2.1.

2.2.3.1 Probabilistic Response Reversal Task

Our search yielded six papers that used a probabilistic response reversal task (see Table 2.1). This task typically consisted of a number of pairs of stimuli (2 – 10) presented in discrete trials. Each option from a pair was probabilistically associated with reward (all studies used points), with one pair being rewarded more often than the other. If a choice did not incur reward, it led to loss of points. After a certain number of ‘acquisition’ trials the likelihood of reward for each stimulus in a pair reversed such that the option previously rewarded less often became the most frequently rewarded option and vice versa.

Of the six studies, three demonstrated deficits in task performance amongst high PT groups, whilst the other studies showed no significant differences. There were no clear task parameters or sample characteristics that were determinants of PT’s effects. All these study’s high PT samples came from antisocial populations. Additionally, all studies used male samples, except (Finger et al., 2008), which had 64% males in the PT group. Overall group sample sizes were often quite small (< 30), especially in neuro-imaging studies by Finger et al. (2008) and

Gregory et al. (2015). These studies may therefore have had adequate power for their main neurological effects, but not for the behavioural effects of interest in this review. Overall, there was not consistent evidence of response reversal learning deficits within this small group of studies.

2.2.3.2 Go/No-Go Tasks with Reversal

Two studies used go/no-go procedures to measure reversal learning (see Table 2.1). In these tasks participants were required to respond to one stimulus (go signal) and withhold responding to another (stop signal) in sequential trials. Lapierre et al. (1995) found that if the required response of the signals was reversed, high PT individuals continued in the initially trained behaviour for longer than low PT individuals. This result reflected reduced ability to reverse established behaviour as seen in some probabilistic response reversal results.

A variation on this procedure was conducted by Brazil et al. (2013). Each of the signals were probabilistically related to another set of preceding stimuli. One of these stimuli was predictive of the go signal, preceding 80% of trials. The other stimulus was not predictive, preceding the go signal on the other 20% of trials. All participants could learn to react faster to the go signal when it was preceded by the predictive stimulus. After 500 trials, these predictive contingencies were reversed. As an outcome of this manipulation, high PT individuals showed a deficit in reversal of their response to the predictive stimulus. However, this only occurred in an iteration of the experiment where there were instructions indicating that there would be a predictive relationship between the cues and the stop and go signals.

2.2.3.3 Intra/Extra Dimensional Shift Task

The intra/extra dimensional shift (ID/ED) task described in the reviewed literature also included response reversal manipulations. As with the probabilistic response reversal paradigm,

this task was designed to test behavioural flexibility in the face of changing contingencies between stimuli and outcomes. However, in this task, stimuli varied in two dimensions, allowing contingency changes to occur within a predictive dimension (i.e., a reversal manipulation). Outcomes could also become contingent on the other stimulus dimension (Extra-Dimensional shift, e.g., shifting from the shape of stimulus to the colour), or shift to different stimuli but remain contingent on the same stimulus dimension (Intra-Dimensional shift, shifting from square to triangle). This task presented a fixed order of these shifts, based on a protocol that was adhered to across the literature.

In animal studies (Dias et al., 1996) there has been attribution of the reversal component of this task to OFC function. Furthermore, the psychopathy literature presents evidence that deficits on this task, where present, occur solely in the response reversal manipulation. This has allowed researchers to claim that psychopathy is linked to reversal processes rather than dimensional shift processes. This effect is not reliable, as it is only observed in three of six studies identified here (see Table 2.1). Factors such as sample characteristics may therefore moderate this effect. Additionally, these results are difficult to link to the hypotheses of the present study as none of the relevant variables are manipulated in this paradigm.

2.2.3.4 Wisconsin Card-Sorting Task

The Wisconsin Card-Sorting Task (WCST) required responses based on changing dimensions of stimuli (equivalent to the extra-dimensional shift in the ID/ED shift task) and has been frequently used in the psychopathy literature as a measure of executive function (Pera-Guardiola et al., 2016). Historically, there has been some consideration of the associative structure of the task. Grant and Berg (1948) examined the effects of greater numbers of rewarded responses on dimension shift behaviour. The results showed that shifts occurred more quickly

when more reinforcements had been given, suggesting training effects. In the reviewed literature Yang et al. (2011) observed increases in ‘perseverative errors’ (responding according to the previously rewarded dimension) which were associated with reduced cortical thickness within participants with PT. This finding was potentially suggestive of a deficit akin to that observed in the response reversal literature. However, the rest of the literature does not support the dependence of the WCST on a process involved in response-reversal performance. The only other study to find a PT dependent increase in perseverative errors also found an increase in errors in general (Kim & Jung, 2014). None of the other six studies found such perseverative effects of PT. WCST results therefore seemed to reflect the absence of dimensional shift deficits in high PT individuals observed in the ID/ED task

Table 2.1 Extracted Variables for Reviewed Response Reversal Studies

Task	Paper	PT deficit (traits analyzed)	PT measure	Number of trials per phase	Reinforcement probabilities tested (%)	Age	PT group (sample source) – N
Go/No-Go Reversal	Brazil et al. (2013)	Yes (categorical difference – overall score)	PCL-R	500	80	Adult	Experiment 1: high (forensic source) – 18, control (community source) – 18, Experiment 2: high (forensic source) – 21, control (community source) – 19
	Lapierre et al. (1995)	Yes (categorical difference – overall score)	PCL	50 acquisition, 150 reversal	100	Adult	High – 30, low – 30 (forensic source)
Intra/Extra Dimensional Shift	Baliouis et al. (2019)	No	PCL-R	Unspecified – cited Mitchell et al. (Mitchell et al., 2002)		Adult	High (forensic source) – 27, low (forensic source) – 55, control (community source) – 20
	Blair et al. (2001)	No	PSD	8-50		Child	High – 19, low – 23 (conduct problem source)
	Dolan (2012)	No	PCL-SV	6-50	100	Adult	High (forensic source) – 33, medium (forensic source) – 28, low (forensic source) – 35, control (community source) – 49
	Dargis, Wolf, & Koenigs (2017)	Yes (categorical difference and continuous effect of overall score)	PCL-R	6-50		Adult	High – 51, medium – 24, low – 39 (forensic source)
	Mitchell, Avny, & Blair (2006)	Yes (categorical difference – overall score)	PCL-R	8-unspecified		Adult	High (forensic source) – 5, low (forensic source) – 5, control (community source) – 5
	Mitchell, et al. (Mitchell et al., 2002)	Yes (categorical difference – overall score)	PCL-R	8-unspecified		Adult	High – 15, low – 17 (forensic source)
Probabilistic Response Reversal	Budhani & Blair (2005)	Yes (categorical difference – overall score)	APSD	20	100, 90, 80, 70	Child	High – 19, low – 22 (conduct problem source)
	Budhani, Richell, & Blair (Budhani et al., 2006)	Yes (categorical difference – overall score)	PCL-R	40	100, 80	Adult	High – 20, low – 17 (forensic source)
	De Brito et al. (2013)	No	APSD	40	100, 80	Adult	High (forensic source) – 17, low (forensic source) – 28, control (community source) – 21
	Finger et al. (2008)	No	APSD & PCL-YV	20 (probability 100), 25 (probability 80)	100, 80	Child	High (Conduct problem source) – 14, low (ADHD source) – 14, control (community source) – 14
	Gregory et al. (2015)	No	PCL-R	20	100, 80	Adult	High (forensic source) – 12, low (forensic source) – 20, control (community source) – 18
	Mitchell, Fine, et al. (2006)	Yes (Factor 1 (Interpersonal/Affective) and Factor 2 scores (Impulsive/Antisocial) – continuous predictors)	PCL-R	30-20	100	Adult	High – 18, medium – 30, low – 40 (forensic source)
Wisconsin Card Sorting	Bętkowska-Korpała et al. (2013)	No	PCL-R			Adult	62 (forensic source, traits analyzed continuously)
	Hare (1984)	No	PCL	10 – 128	100	Adult	High – 14, medium – 16, low – 16 (forensic)
	Ishikawa, Raine, Lencz, Bihle, & Lacasse (2001)	No	PCL-R			Adult	High (forensic source) 16, high (community source) 13, low (community source) 26

Task	Paper	PT deficit (traits analyzed)	PT measure	Number of trials per phase	Reinforcement probabilities tested (%)	Age	PT group (sample source) – N
	Kim & Jung (2014)	Yes (categorical difference and continuous effect of overall score)	PPI-R			Adult	High- 15, low -15 (community source)
	Krakowski & Czobor (2017)	No	PCL-SV			Adult	102 (community source traits analyzed continuously)
	Lapierre et al. (1995)	No	PCL			Adult	High – 30, low – 30 (forensic source)
	Mol, Van Den Bos, Derks, & Egger (2009)	No	PCL-R			Adult	High – 17, low – 36 (forensic source)
	Pera-Guardiola et al. (2016)	No	PCL-R			Adult	High (forensic source) – 31, medium (forensic source) 47, low (forensic source) -13, control (community source) – 24
	Pham, Vanderstukken, Philippot, & Vanderlinden (2003)	No	PCL-R	6 – 48		Adult	High – 18, low – 18 (forensic source)
	Yang et al. (2011)	Yes (overall score – continuous predictor)	PCL-R	10 - 128		Adult	High – 27, low – 32 (community source)

2.2.4 Extinction Tasks

Three methodologies were identified that included extinction of a response. These are described below, where their reliability as measures of learning deficits is also considered. The primary features of these studies are tabulated in Table 2.2.

2.2.4.1 Operant Conditioning Task

In this task reported by X. Chen et al. (2020), participants could receive points for pressing the spacebar of a keyboard. After training this response, points were withheld in an extinction phase. It was seen that individuals with higher meanness scores on the TriPM responded to extinction with an initial increase in their rate of pressing that exceeded lower scoring participants. They also went on to display more responding at the end of the extinction period. This result is consistent with a deficit in extinction learning.

2.2.4.2 Response Perseveration Task

In this procedure, developed by Newman et al. (1987), participants were given a virtual deck of cards and told to choose whether to play a card from the deck or quit the task over multiple trials. Initially each trial led to reward (money) at 100% probability. However, after every block of 10 trials the probability that a trial resulted in the loss of money increased by 10%. Participants could play a maximum of 100 cards and the number of cards played before choosing to stop was measured. This metric indexed the speed at which participants learned that reward was being omitted. In seven of the nine studies I reviewed, the high PT group played more cards from the deck (Table 2.2). The results of this task have been interpreted as evidence that psychopaths exhibit an extinction deficit (Blair et al., 2005; Mitchell, Avny, & Blair, 2006; Nigg, 2000). However, Newman et al. (1987) made no attribution of this result to an extinction learning process. Furthermore, the present hypotheses cannot be applied to the results seen in this

task, as no manipulation of the relevant variables has ever occurred in this experimental paradigm.

There is also some indication within the reviewed literature that the source of PT associated deficits in this task are not due to associative learning processes. Newman et al. (1987) found that forcing participants to wait for 5 seconds after feedback before moving to the next trial reduced the PT deficit in this task. Congruently, Molto et al. (2007) found that the number of cards played in this task was associated with faster response times, indicating that high PT individuals were taking less time to make decisions in this task. Furthermore, the lifestyle factor of the PCL-R was predictive of behaviour in this task, rather than the affective factor. This may indicate that there was increased impulsivity in responding in this task, related to trait impulsivity measured by the lifestyle factor of the PCL-R. Such a result would therefore not reflect any variation in learning processes.

Further support for the role of impulsivity in producing deficits on this task can be seen in the findings of Ribes-Guardiola et al. (2020). These authors reported that the response perseveration deficit was related to the boldness subscale of the TriPM. This result contrasted to those of the operant conditioning task described above, where-in the meanness subscale, which is correlated with the affective features of other PT measures, was solely predictive of a psychopathy deficit (X. Chen et al., 2020). The operant conditioning task measures extinction learning as a reduction in reward seeking behaviour; a measure with a long history in the study of learning (e.g. Skinner (1938)). Contrastingly, the response perseveration task required participants to opt out of the game before they were recorded as having learned that rewards were absent. This element of the task may have made it a better measure of participant's boldness (risk taking tendency) than of their extinction learning capacity.

Other studies showed a clearer link between the perseveration deficit and the emotional features of psychopathy. Fisher and Blair (1998) claimed that both the CU dimension and the impulsivity/conduct dimension of the PSD predicted more cards being played. Similarly, Barry et al. (2000) reported that within conduct problem children, higher levels of CU traits lead to poorer performance. However, this effect was isolated to this group; non CD children's performance was not affected by the presence of CU traits. Also in a child population, O'Brien and Frick (1996) found that there was an overall effect of the CU subscale of the PSD, again over a clinical population only. Whilst there seems to be evidence that the emotional component of PT was predictive of performance on this task, this evidence is restricted to clinical populations. Notably this task was never performed on a purely non-clinical or non-forensic population, aside from in Mitchell et al. (2006) where differences within the community control group were not examined. Therefore, although there were some results linking increased response perseveration to the affective features of psychopathy, these features have never been fully experimentally dissociated from disordered conduct.

2.2.4.3 Aversive Pavlovian Conditioning

The aversive Pavlovian-learning paradigm was considered in this review. However, it does not concern the establishment of goal directed behaviour (instrumental learning) that is basis of the hypotheses introduced in this chapter, and the focus of the primary hypotheses of this thesis. To generate learning, these studies pair a stimulus with an aversive outcome over repeated trials. This procedure results in automatic fear responses to the stimulus (such as elevated physiological arousal and startle responses). Deficits in these responses have been linked to psychopathic traits (Lykken, 1957). As a result, this paradigm has been the subject of extensive research; the present search produced 24 aversive conditioning studies. However, only ten

included an extinction phase where the aversively conditioned stimulus was presented alone. Of these, six measured only physiological responses to stimuli. The remaining four studies were not set up to test extinction effects; only Rothmund et al. (2012) analysed differences in responding between training and extinction. Birbaumer et al. (2005) acknowledged that their study was not designed to analyse extinction processes. Indeed, none of these studies mentioned extinction in the rationale for their experimental design, rather it seemed to be treated as a test period where no unconditioned stimuli were present.

The diminished impact of aversive conditioning in psychopathy (Lykken, 1957) also caused problems for the interpretation of extinction responding in the conditioning studies reviewed here. Only Rothmund et al. (2012) observed that conditioning of stimulus discrimination in the training phase was consistent between high and low PT groups. This was significant for the analysis of extinction learning, as comparisons between groups could be made from an equivalent initial response rate or intensity. Thereby, Rothmund et al. (2012) appears to be the only study that has ever generated the conditions required to make behavioural observations of aversive inhibitory learning in PT. This study observed that higher PT individuals produced more complete extinction of self-reported arousal responses to stimuli. Ratings of contingencies between stimuli and the aversive outcome also appeared to decrease faster for high PT individuals. However, these data were not analysed statistically by the authors. An extinction study designed to address these issues would be valuable, as the existing data suggests that extinction is faster in high PT individuals, which is inconsistent with widely espoused accounts of learning deficits in these populations (e.g., Blair et al. 2005).

Table 2.2 *Extracted Variables for Reviewed Extinction Studies*

Task	Paper	PT deficit	PT measure	Number of trials per phase	Reinforcement probabilities tested (%)	Age	PT group (sample source) – N
Aversive Pavlovian conditioning	Birbaumer et al. (2005)	Unclear	PCL-R	16 conditioning/ 8 extinction	100	Adult	High (forensic source)- 10, low (community source) – 10
	Flor et al. (2002)	Unclear	PCL-R	60 conditioning/ 50 extinction		Adult	High – 9, low – 12 (community source)
	Rothmund et al. (2012)	Unclear	PCL-R	48 conditioning/24 extinction		Adult	High (forensic source) – 11, low (community source)11
	Veit et al. (2002)	Unclear	PCL-R	16 conditioning/ 16 extinction		Adult	High (forensic source)- 4, low (community source) – 7
Operant conditioning	X. Chen et al. (2020)	Yes (Meanness Subscale – continuous predictor)	TriPM	8 conditioning trials (4-minute duration), 12 extinction trials (2-minute duration)	3.33, yoked random interval schedule.	Adult	75 (community source, traits analysed continuously)
Perseveration	Barry et al. (2000)	Yes (categorical difference – overall score)	PSD	10 trials*10 blocks	100%, decreasing by 10% after every trial	Child	High (ADHD + conduct problem source) – 16, low (ADHD + conduct problem source) – 12, control (ADHD source) – 58, control (clinical source) – 68
	Blair & Cipolotti (2000)	Yes (categorical difference – overall score)	PCL-R			Adult	High – 5, low – 5 (forensic source)
	Fisher & Blair (1998)	Yes (Factor 1 (Callous/Unemotional) and Factor 2 (Impulsiveness/Conduct Problems) – continuous predictors)	PSD			Child	39 (conduct problem source, traits analysed continuously)
	Mitchell et al. (2006)	No	PCL-R			Adult	High (forensic source) – 5, low (forensic source) – 5, control (community source) – 5
	Molto et al. (2007)	Yes (Factor 2 (Impulsive/Antisocial) continuous predictor)	PCL-R			Adult	High – 9, medium – 19, low – 11 (forensic source)
	Newman et al. (1987)	Yes (categorical difference – overall measure)	PCL-R			Adult	High 36 low 36 (forensic source)
	Obrien & Frick (1996)	Yes (categorical difference – CU subscale)	PSD – callous unemotional subscale			Child	High (conduct problem source) – 8, high (clinical source) – 22, low (community source) – 45
	Ribes-Guardiola et al. (2020)	Yes (Boldness subscale – continuous predictor)	TriPM			Adult	222 (community source, traits analysed continuously)
	Vitale & Newman (2001)	No	PCL-R			Adult	High – 11, low – 73 (forensic source)

2.3 Discussion

Through the systematic review described above, I was able to assess the reliability of existing evidence for PT based deficits in response reversal and extinction learning. It was plain that evidence for this association is largely inconsistent (see Figure 2.3). In the following discussion I have applied the hypotheses of this study (see section 2.0.2) to the results of the reviewed studies. These hypotheses regarded the roles of training and extinction durations, context changes, and reinforcement probabilities in moderating learning deficits. The following discussion thereby assesses validity of reviewed measures of instrumental learning. Accordingly, I have also considered other factors that impacted the interpretability of this pool of evidence. It should be noted that whilst hypotheses in this chapter concerned CU traits, the reviewed studies only utilised operationalisations of the subsuming PT construct. As a result, considers where PT measures may have been influenced by factors that are not relevant to the CU trait construct.

2.3.1 Duration of Training and Extinction Phases

An important factor in the literature I have reviewed here is the number of trials that participants experience in each phase of reversal or extinction paradigms. My hypotheses for these variables were that increased training, and extended exposure to reversed or extinction contingencies, would both mask learning deficits. Unfortunately, no study deliberately manipulated each of these variables independently, so there is no direct evidence for such effects.

Whilst the literature reviewed here cannot inform these hypotheses, variation in training and extinction durations may explain inconsistencies in the literature. For instance, there was a greater reversal deficit evident in the adult version of the response reversal task by Budhani et al. (2006) compared to these authors' study of child participants (Budhani & Blair, 2005), which may have been due to effects of participants' age. However, as well as using a different subject

population, Budhani et al. (2006) also implemented 40 trials in training and reversal phases of their study, whereas in Budhani and Blair (2005) 20 trials were used. Despite this, there was no evidence of deficits in another study using 40 trials per phase (De Brito et al., 2013), whereas another using 24 or fewer trials did show effects (Mitchell, Fine, et al., 2006). Clearly a host of controlled and uncontrolled variables may determine variation in results across studies. However, these inconsistencies point to the need for systematic variation of training and extinction duration.

Mitchell, Fine, et al., (2006) was the only design in which the amount of training of contingencies varied. Poor reversal performance in individuals with PT was seen after a reversal of contingencies trained for 24 trials, but not for a reversal of contingencies trained for 12 trials. Whilst this may reflect a heightening of reversal deficits after greater training this study was not designed to measure the effect of this variable, so it does not provide direct evidence of this effect. However, it further reinforces that inconsistencies in experimental design could cause the poor reliability of the presently reviewed literature.

More observations relevant to the effect of training duration were made by Brazil et al., (2013). In one condition of this study no PT associated reversal deficits were seen after 500 trials of training and 500 of reversal, which was the highest trial number utilized over all the studies I reviewed. However, in a condition where instructions were given that highlighted the presence of probabilistic contingencies in the task, high PT individuals showed a reversal deficit. Whilst no direct statistical comparisons were made across these instruction conditions, the results appear to be driven by better reversal performance in the instruction condition amongst subjects without PT. One account of these results, informed by studies of animal models (Holland, 2004), may be that after 500 trials participants behaviour may have been less informed by the goal of

responding. Thereby, poor outcome representation could have impaired all participants' ability to respond quickly to changes in the contingencies of the task, normalising PT based learning deficits. However, the presence of instructions may have promoted attention to outcomes, amending the impediment to learning in individuals with low PT, but not those with high PT. This result could reflect the hypothesis that excessive training slows learning in a manner similar to the deficits underlying PT. However, Brazil et al. (2013) cannot be meaningfully interpreted in this way as statistical comparison across instruction conditions was not a component of their study. Further, no other study in this literature has systematically manipulated this variable.

Notably, across all the reviewed studies, behaviour was either measured at the end of reversal or extinction phases or averaged over their duration. This means that the literature cannot indicate whether differences in the rate of learning across reversal/extinction trials are associated with PT. Overall, there are findings showing that the number of such trials may moderate response reversal deficits in the reviewed literature. However, no results directly apply to such an effect.

2.3.2 Reinforcement Probabilities

I hypothesised that lower contingencies between responses and outcomes would exaggerate the response reversal deficit seen for high PT individuals. This pattern could be observed in this literature; two studies showed that the size of the PT associated response reversal deficit increased as the probability of reinforcement after responding decreased. This effect was found by Budhani and Blair (2005) in a sample from a school for children with emotional and behavioural difficulties, and by Budhani, et al. (Budhani et al., 2006) in an adult forensic sample. Additionally Budhani and Blair (2005) suggested that the absence of response reversal deficits in an ID/ED study by Blair et al. (2001) was due to the 100% probability of

reinforcement for behaviour in that task. They argue that at this level of reinforcement the change in contingency is too salient for high PT individuals not to notice. However, whilst deficits at 100% reinforcement rates were absent for children in the probabilistic response reversal task (Budhani & Blair, 2005) and the ID/ED task (Blair et al., 2001) they have been seen in adult forensic populations (Budhani et al., 2006; Dargis et al., 2017; Dolan, 2012; Mitchell et al., 2002; Mitchell, Avny, et al., 2006; Mitchell, Fine, et al., 2006). This indicates that other factors, such as age, may mediate the relationship between reinforcement probabilities and reversal deficits. In conclusion, only two studies directly indicate that variation in reinforcement probabilities affects response reversal deficits in psychopathy, so evidence of this effect is present but limited.

2.3.3 Context Changes

None of the reviewed studies directly addressed my hypothesis for context effects. Specifically, that context changes would exaggerate PT associated learning deficits, due to decreased encoding of contextual information in high PT individuals. However, some results were applicable to this hypothesis. In Budhani et al. (2006) participants learnt about two sets of stimuli at once, with trials of each pair interspersed. Exposure to each pair was staggered such that when the reversal phase of one pair started, exposure to another pair finished and a new pair of stimuli began. The initiation of trials of these novel stimuli may have therefore acted as a contextual shift in the task. My hypothesis suggested that high PT individuals may have reduced capacity to use such a change to prompt new learning about changed outcome contingencies. Such an impairment could therefore explain slower reversal of responding in high PT individuals (compared to those low in PT) seen in Budhani et al. (2006).

Consistent with this interpretation, the methodology of Finger et al. (2008) contained no such structural indication of the differences between phases (as the order of contingency presentations was randomized) and produced no reversal deficit. Contextual information may have also led to the observation of an extinction learning deficit in high PT individuals by X. Chen et al. (2020) in an operant conditioning task. In this task there were changes in the duration of trials between training and extinction phases. Further, the order of these trials changed after the initiation of extinction. Whilst none of these studies actively manipulated contextual information, this literature indicates that context changes should be considered in future studies. Furthermore, future researchers should avoid inadvertently including contextual information about the task structure, as this factor may confound the effects of controlled variables.

2.3.4 Traits and Populations

Further variables that were found to be likely determinants of results in this literature were the characteristics of utilised samples (including age and clinical status). The only evidence of a reversal deficit in children with PT was seen by Budhani and Blair (2005) in a conduct disordered population. Response reversal performance in this population has therefore not been directly compared to the general population. A direct comparison of reversal performance across levels of PT in each of these populations would determine whether factors associated with PT in conduct disordered populations are responsible for the effect observed by Budhani and Blair (2005). Such a study is necessary as it is currently undetermined whether reversal deficits in children are predicted by PT or CU traits themselves, or by their interaction with conduct disorder. This would contribute to understanding the aetiology of CU traits independently from that of the clinical correlates of these traits.

The handling of trait measures in this literature also limited its applicability to CU traits. All the psychometrics used in the reviewed studies had a factor that was relevant to the affective dimension of PT (or therefore to CU traits). However, only Fisher and Blair (1998), O'Brien et al. (1994), Mitchell et al. (2006) and X. Chen et al. (2020) showed specifically that the affective dimension of PT predicted behavioural deficits. Such data distinguishes the effects of the affective features of PT from those of lifestyle and behavioural correlates of these features. This is significant as such behavioural correlates may be confounding to learning effects. Gray et al. (2019) found that the Lifestyle/Antisocial dimension of PCL-R and PCL-SV predicted increased negative urgency (reactivity to distress) on the Urgency, Premeditation (lack of), Perseverance (lack of) and Sensation seeking scale. Interestingly this measure of trait impulsivity was negatively correlated with increased scores on the Affective/Antisocial dimension of PCL scales. This is significant as negative urgency also predicts poor response reversal learning amongst frequent cocaine users, who have higher scores on this measure than control subjects (Fernández-Serrano et al., 2012). Therefore, in studies that have not determined which factor of the PCL predicts reversal deficits, such deficits could be explained by impulsivity traits.

2.3.5 Future Studies

The evidence reviewed here showed limited applicability to the predictions I derived from the associative learning literature. However, this evidence does point to a variety of factors that should be targeted by future studies. Most basically, a study that examines both extinction learning and response reversal learning would be both novel and informative. Deficits on both these paradigms have been attributed to the same psychological process (e.g., Blair et al. 2005). However, no study has sought to measure a corresponding performance deficit in reversal and extinction tasks.

In this review, I have highlighted that the manipulation of basic task parameters can be a useful tool in making inferences about the aetiology of psychopathic traits. Crucially, none of the reviewed studies have ever set out to measure the effect of trial numbers and context changes on the response reversal deficit. Further, reinforcement probabilities were manipulated in a limited number of studies. This is surprising given that these variables are the common subjects of experimentation in the extinction learning literature. Studies examining extinction learning should consider testing for renewal, as no study of CU traits has done so. This would provide novel insights into the context sensitivity of learning in high CU individuals and may inform theoretical predictions about patients' retention of conduct interventions.

2.3.6 Implications and Conclusion

This review determined that there was mixed evidence supporting response reversal and extinction learning deficits as correlates of PT. These deficits are commonly assumed to be reliable by various aetiological theories of CU traits and PT (Blair et al., 2005; Moul et al., 2012). Therefore, this review indicated the need to resolve inconsistencies in this literature in order to support the validity of these theories. This can be achieved by experimental examination of moderating factors identified in this review, including sample characteristics, trait measures and task parameters. This work should be completed before these deficits are accepted as established features of PT and CU traits.

Overall, there was little evidence to support the hypotheses developed in this study. Two studies indicated that lower reinforcement probabilities exaggerated these learning deficits, as hypothesised. Considering this learning as a model of behavioural intervention (see section 1.2.1), this result is consistent with observations that inconsistent parenting contributes to the reduced efficacy of such interventions in high CU children (D. J. Hawes et al., 2014). However,

these findings should be replicated to support their reliability. I was unable to find clear support for hypotheses regarding the role of training duration, extinction/reversal duration and context in determining extinction/reversal deficits in high PT individuals. This shortfall in the present literature needs to be addressed in order to validate assertions of the role of associative processes in the aetiology of PT and CU traits. Studies systematically varying these task parameters would also be a useful means of modelling the associative processes involved in behavioural interventions. Overall, this study fulfilled a major aim of this thesis: to assess existing evidence for extinction learning deficits in individuals with CU traits.

3 Extinction Task Development

Contents

3.0.1	The Importance of Testing Renewal	65
3.0.2	Validating Extinction Learning	66
3.0.3	Excluding Behavioural Confounds	66
3.0.4	Aims and Hypotheses	68
3.1	Pilots: Establishing an Extinguishable Response	69
3.1.1	Participants and Procedure.....	69
3.2	Pilot 1	70
3.2.1	Methods.....	70
3.2.2	Results.....	73
3.2.3	Discussion	75
3.3	Pilot 2: Response Costs.....	76
3.3.1	Methods.....	76
3.3.2	Results.....	77
3.3.3	Discussion	79
3.4	Pilot 3: Response Instructions.....	80
3.4.1	Methods.....	80

3.4.2	Results.....	81
3.4.3	Discussion.....	82
3.5	Experiment 1: An Effortful Response.....	84
3.5.1	Methods.....	85
3.5.2	Results.....	88
3.5.3	Discussion.....	90
3.6	Experiment 2: An Alternative Response.....	92
3.6.1	Methods.....	93
3.6.2	Results.....	100
3.6.3	Discussion.....	105
3.7	Experiment 3: Renewal.....	108
3.7.1	Methods.....	109
3.7.2	Results.....	118
3.7.3	Discussion.....	125
3.8	General Discussion.....	128

This chapter addressed my aim to provide empirical evidence supporting suggestions of a link between CU traits and extinction learning. In the previous chapter the need for such evidence was reinforced through systematic review of relevant literature. This review highlighted that despite being implicated in the modern aetiology of CU traits (Blair et al., 2005; Moul et al., 2012), evidence for the role of extinction learning processes in CU traits is limited. The review additionally identified several factors that affected the validity and reliability of existing research, and these were therefore considered in the present study.

An additional goal of this study was to generate novel observations of the relationship between CU traits and the renewal of extinguished responding. In Chapter 1 the importance of evidencing this relationship was established on the basis of its novel implications for conduct interventions. In Chapter 2 I discussed the utility of contextual learning in validating the attribution of task behaviour to learning processes. The study conducted in that chapter also highlighted the complete dearth of evidence regarding this effect. These factors substantiate the need for the present investigation.

3.0.1 The Importance of Testing Renewal

As discussed in Chapter 1, extinction learning is a context specific inhibition of established responding (Bouton et al., 2021). The context dependency of extinction learning is widely measured using a renewal paradigm: after initial conditioning, extinction occurs in a new context. Subsequent return to the initial training context causes responding to renew, indicating context dependent expression of inhibition learned in extinction (Bouton et al., 2021). An extinguished behaviour is highly liable to return given a range of triggers. For example, in addition to the renewal effect, reinitiating rewards for lever presses after extinction can lead to rapid reacquisition of behaviour (Bouton et al., 2021). These effects are collectively referred to

as relapse effects. Despite research showing the relevance of such relapse in behavioural interventions on other psychopathologies (Conklin & Tiffany, 2002; Hofmann, 2008), this process has not been considered as a determinant of interventional success in the clinical context of CU traits (as outlined in section 1.2). Before considering this process in clinical contexts, it is important to establish that relapse (or renewal) is related to CU traits in experimental settings.

3.0.2 Validating Extinction Learning

In addition to the clinical significance of renewal effects, renewal has utility in the design of a valid extinction experiment. As extinction learning is a context specific effect, deficits in extinction learning should be accompanied by deficits in renewal, and relapse effects in general. This view is also supported by recent learning models discussed in Chapter 1 (Wilson et al., 2014). Therefore, a novel means of testing the theory that CU and PT traits are associated with extinction learning deficits is to test whether high CU individuals exhibit co-occurring deficits in extinction and renewal. In this study, I therefore aimed to develop an experimental paradigm capable of detecting such an effect.

3.0.3 Excluding Behavioural Confounds

In addition to being a validating element of extinction learning, renewal or relapse effects improve the interpretability of behaviour in extinction tasks. This is because relapse tests are not confounded by variables often associated with PT and CU traits such as impulsivity (Cardinale & Marsh, 2020). Impulsivity can act as a confound due to its association with impaired response reversal (Fernández-Serrano et al., 2012). In the previous chapter I identified that some demonstrations of extinction deficits were potentially explained by this confound (see section 2.3.4). In contrast, the hypothesis under examination here would require high CU individuals to produce greater rates of responding in extinction and lower rates of responding in a relapse test.

Therefore, any increased tendencies to respond would not explain these effects. Examination of relapse effects also rules out unintended impacts of contextual learning on results. Such effects were noted as potential determinants of historical results in the prior chapter (see section 2.3.3).

Other aspects of the experimental design utilised in this study were also intended to exclude confounds of extinction effects. This design was informed by the basic associative structure used in relapse studies by Bouton et al. (2012). This meant training a simple instrumental response with intermittent reinforcement. Following training, rewards were omitted for the extinction phase. The use of this instrumental reward learning design methodologically avoided aversive outcomes, which are historically thought of as being appraised differently by people with CU traits (Lykken, 1995). To further dissociate any responding in this study from the social features of CU traits I additionally aimed to design a task in which the instrumental behaviour had a prosocial component. This meant that in order to display any extinction deficit, high CU individuals would need to act in a manner ostensibly more prosocial than low CU individuals. These design choices were made so as to minimise the influence of antisocial attitudes or behaviours on the findings, allowing for improved validity in observations of extinction learning processes.

This discrimination of the behavioural effects of associative learning from those of antisocial behaviour was also achieved through the population sampled for this study. Frequently in this area of research, samples are derived from incarcerated or offender populations, where psychopathic traits are more prevalent. However, these populations show a range of executive function deficits (De Brito et al., 2013), which may confound evidence of specific deficits in extinction learning. I therefore used a general population sample to minimize the role of such

confounding factors, as have prior studies of PT and associative learning (e.g. X. Chen et al., 2020)

No prior study has utilised a scale designed to measure CU traits specifically, rather, they have assessed the affective component of psychopathic trait measures (See 2.3.4). This study therefore utilised a specific CU trait measure to allow for improved attribution of results to this construct. This measure also excluded effects of factors associated with the lifestyle component of PT measure, which I suggested as potential sources of effects documented in the review of Chapter 2.

3.0.4 Aims and Hypotheses

The aims of this study were to provide evidence of a relationship between CU traits and delays in extinction of a trained response. This aim was motivated by a need identified in Chapter 2 to add to existing evidence of such effects, which is presently unreliable. Additionally, I aimed to demonstrate reductions in the renewal of extinguished behaviour at higher levels of CU traits. This would act to validate assertions that the extinction behaviour associated with CU traits results from associative processes. I hypothesised that continuously higher levels of CU traits would predict less extinction, as has been previously documented for the affective component of PT (X. Chen et al., 2020). My novel prediction was that higher levels of these traits would also predict less renewal of responding following a return to a training context after extinction.

3.1 Pilots: Establishing an Extinguishable Response

In this series of pilot studies, I aimed to develop a task which could measure the extinction of an instrumental response in large samples. These pilots were therefore focussed on establishing the experimental design and analysis for the task. As a result, they did not involve measurement of CU traits, or participant demographics.

3.1.1 Participants and Procedure

Participants in these pilots (and the following experiments) were undergraduate students from first and second-year courses, who received course credit as compensation for their participation. Recruitment and procedures were approved by the University of Sydney Human Research Ethics Committee protocol 2018/573 – “Reward learning and Personality”. Participants attended the study in person, and all experimental materials were delivered via Windows PCs. After completing the participant consent form, participants completed all experimental procedures using a mouse and viewed all materials on a dell monitor (resolution = 1920x1080 pixels (width x height)). Auditory stimuli were delivered through over-ear headphones. This procedure applies at all subsequent pilots and experiments. Code for the analyses reported below are available at <https://osf.io/nz2sx/> (see Chapter3Analysis.html).

3.2 Pilot 1

In order to establish the ideal parameters for the task, the effects of different reinforcement schedules were examined. Rather than testing hypotheses for these effects, an exploratory approach was taken (due to the novel nature of the experimental design).

3.2.1 Methods

3.2.1.1 *Participants*

Thirty-one first year psychology students were recruited to complete this pilot.

3.2.1.2 *Design*

During this experiment participants experienced three sequentially presented conditions. Responses during each condition were reinforced on a predetermined schedule, which varied between a Training Phase and an Extinction Phase (summarized in Table 3.1). During the Training Phase, participants were required to achieve a number (randomised between 20 and 30) of rewards, after which the experiment moved to the extinction phase, during which responding was measured for 100 seconds.

Two of the conditions allowed for the comparison of the rate of reinforcement on responding. During the “Fast Condition” reinforcement occurred on a variable interval (VI)3 schedule during the Training Phase. This meant that after a reinforced response another reinforcement could only be obtained by responding after a varying interval of 3 seconds (on average). The “Slow Condition” was reinforced on a VI6 schedule. During the Extinction Phase responses were not reinforced. A third condition (Control Condition) controlled for changes in motivation throughout the study and was reinforced on a VI3 schedule during both phases. The order in which conditions were presented was randomised across participants.

Table 3.1 *Reinforcement Schedules for each Condition Throughout Pilot 1*

Condition	Training	Extinction phase
Fast	3	Ø
Slow	6	Ø
Control	3	3

Note. All numbers refer to VI schedules, and Ø indicates no reinforcement.

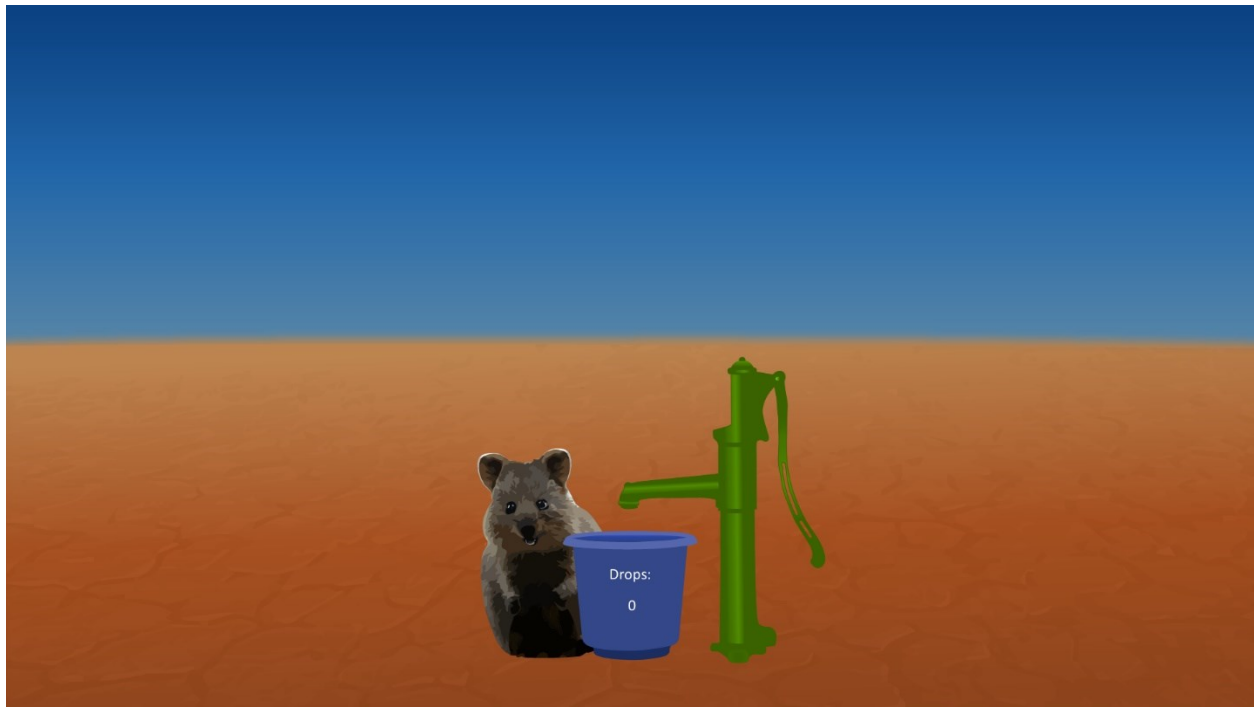
3.2.1.3 *Materials*

To minimize the cost of delivery (and therefore the applicability of the task to large samples) I limited the task duration to 20 minutes. Additionally, a narrative context was designed to create a concept of reward, so as to remove the costs involved in using material rewards. I expected that this feature would be a valid way of measuring learning about this outcome, as prior studies have shown meaningful variation in tasks using represented reward outcomes such as points (Budhani & Blair, 2005). The use of a narrative context to motivate instrumental behaviour, rather than points, was designed to be engaging for both adults and children. In so doing I hoped to allow for the potential of future implementations of the task in child populations.

The task was programmed digitally using the Psychtoolbox (Brainard, 1997) Matlab package and utilized visual stimuli designed in Adobe illustrator (shown in Figure 3.1). By clicking on the handle of a water pump, participants were able to move the handle, sometimes causing a drop of water that fell into a bucket, causing a drop sound to play and contributing 1 point to a count of ‘drops’. It was made explicit that the goal of the task was to collect as much water as possible to help an animal – “Quentin the Quokka” (a small Australian marsupial) through the following instructions:

Thank you for participating in this experiment. In this task we would like you to help Quentin the Quokka who is lost in the desert. Quentin has found water pumps in the desert, but he needs your help to pump them. You will be able to click on the handle of a pump in order to operate it. Operating the pump may produce a drop of water that Quentin will collect in a bucket. However, some pumps take time to refill, so you will not get a water drop every time. Additionally, sometimes Quentin will decide that it is time move on to find a new pump. You will need to collect as much water as possible to help Quentin live in the desert. Once you have understood these instructions, please press the spacebar to begin the task.

After completing the requirements to progress to another pump, participants were explicitly told that they were moving to a new pump by the following instructions, which appeared on a black screen: “Quentin has decided to search for a new pump. He will drink some water and store the rest, so you will need to fill up the bucket again. Please press the spacebar to continue the task.” After moving to a new pump, the score was reset. Additionally, the colour of the pump changed (to a new random colour), aiding discrimination of each pump.

Figure 3.1 *Quentin the Quokka Task Layout*

3.2.1.4 Data Analysis

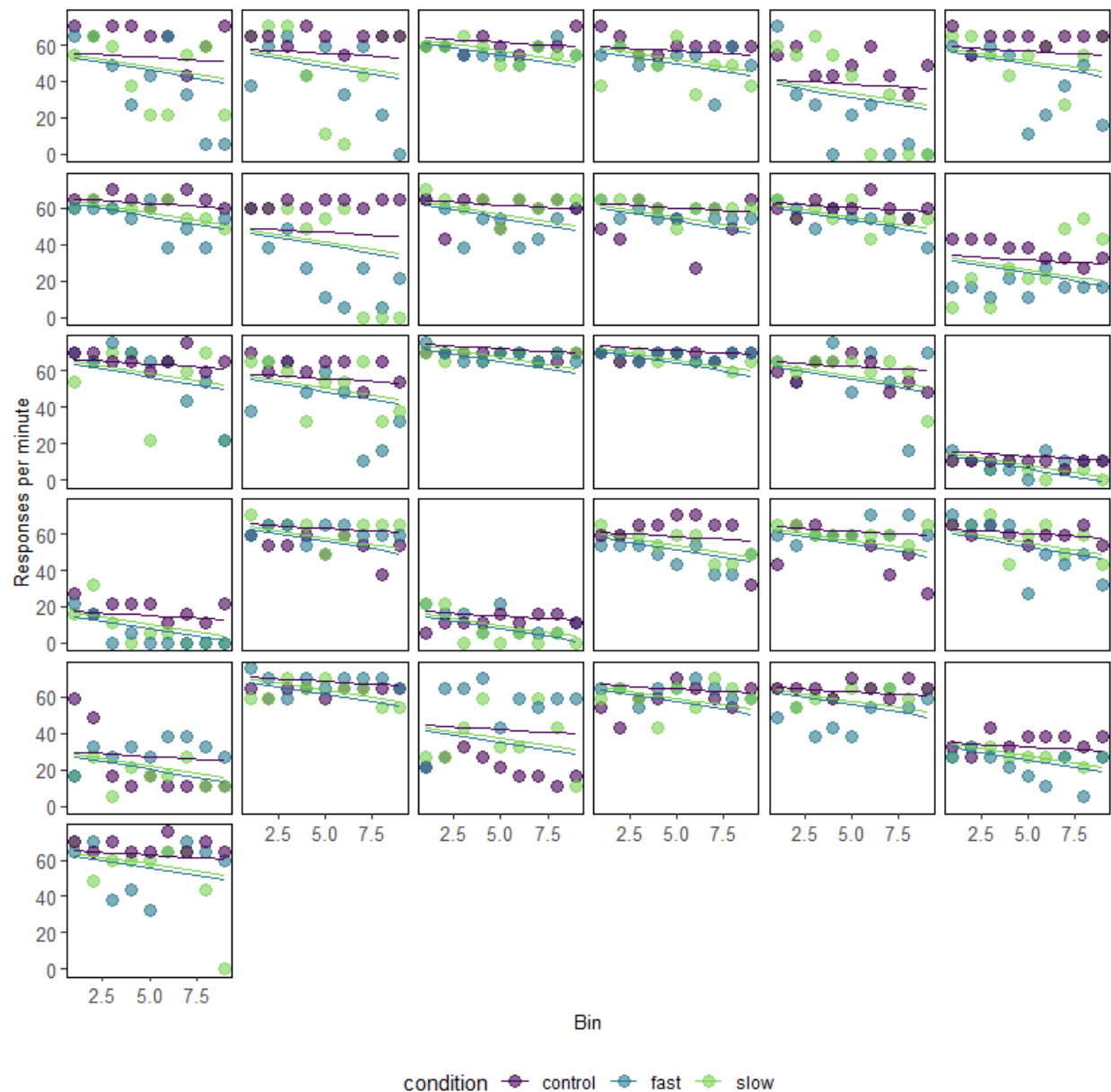
Responding during the Extinction Phase of all conditions was binned into nine 11 second bins and converted to responses per minute. I then fit a mixed effects model designed to determine which conditions lead to greater reductions in behaviour throughout the extinction phase (when compared to the Control Condition). Specifically, the analysis modelled the interaction between the bin variable and two effect coded variables (see DeBruine & Barr (2021)) comparing the Control Condition to both the Fast Condition and the Slow Condition. It included terms for random intercepts across participants to account for within-subjects variation, thereby accounting for any overall differences in response rate between participants.

3.2.2 Results

The model was fit successfully. It demonstrated that participants' responding during extinction (shown in Figure 3.2) reduced at a greater rate across bins in the Slow Condition

compared to the Control Condition ($\beta = -1.00$, 95% CI = $[-1.76, -0.24]$, $t = -2.58$, $p = .010$). A similar effect was seen for the Fast Condition compared to the Control Condition ($\beta = -1.10$, 95% CI = $[-1.86, -0.34]$, $t = -2.84$, $p = .005$).

Figure 3.2 *Observed and Predicted Extinction Responses for Individuals in Pilot 1*



Note. Each plot shows data for separate individuals. Observations of response rate in each bin of the Extinction Phase of this pilot are shown as points, coloured by condition. Values predicted by the mixed effect model are shown as lines for each condition. In most individuals, predicted

values can be seen to decrease across bins faster for the Fast and Slow conditions compared to the Control condition.

3.2.3 Discussion

I was able to observe extinction of responding in this pilot study. Both extinction conditions (trained under “fast” and “slow” variable intervals) lead to greater reductions in behaviour throughout the Extinction Phase of the study than did the Control Condition (where rewards were not omitted). This indicated that behaviour consistent with extinction learning could be produced in this novel task.

Although this result was consistent with the aim of this pilot, complete extinction of responding (i.e., response rates close to zero) did not occur in many instances. This may introduce complexities to the interpretation of extinction effects in future experiments using this procedure. This complexity arises from the uncontrolled processes that may have driven responding in the absence of expectation of reward (such as exploratory behaviour or demand characteristics of the experimental context). To best measure extinction effects in this study, it is important to minimize these uncontrolled motivators of responding.

3.3 Pilot 2: Response Costs

In this pilot I aimed to replicate the demonstration of extinction behaviour achieved in the first pilot. Additionally, to produce a more complete extinction of responding than that which was observed in Experiment 1, I introduced a cost for responding. Past studies have observed that participants continuously respond if the action required is easy (e.g. keystrokes or mouse clicks), regardless of whether or not rewards are being provided (Bradshaw et al., 2015). This problem has been avoided by introducing a cost for instrumental responses. For example, Weiner (1962) found that introducing a cost (in the form of point deductions) to responding for a reward delivered on a fixed interval schedule led to peaks of responding at that interval. Contrastingly, in the absence of cost, responding did not reflect temporal variations in reward expectancy. I therefore expected that point costs for instrumental responses would create a stronger contingency between reward rates and responding in this task, improving its capacity to measure extinction learning.

3.3.1 Methods

3.3.1.1 Participants

Twenty-one undergraduate psychology students participated in this experiment.

3.3.1.2 Procedure and design

Procedures and overall task design for this experiment was equivalent to Experiment 1, except that after every five responses one reinforcement was subtracted from the running total of reinforcement.

3.3.1.3 Materials

I utilised the task described in Pilot 1, with the inclusion of the following instructions displayed after the initial task instructions:

Additionally, operating the pump will make you thirsty, so Quentin will need to share the collected water with you. Therefore, for every five clicks on the pump handle, one drop of water will be subtracted from the bucket. Remember that pumps take time to refill.

Therefore, leaving several seconds between each click on the pump will work best.

Additionally, instructions between each pump were also amended:

Quentin has decided to search for a new pump. He will drink some water and store the rest, so you will need to fill up the bucket again. Remember that operating the pump will make you thirsty, so Quentin will need to share the collected water with you. Therefore, for every five clicks on the pump handle, one drop of water will be subtracted from the bucket. Pumps take time to refill. Therefore, leaving several seconds between each click on the pump will work best. Please press the spacebar to continue the task.

The statement “leaving several seconds between each click on the pump will work best” was meant to suggest that participants should respond with respect to an interval schedule of reward. This was included because if participants assumed that reinforcements were delivered on a ratio schedule there would be no means of avoiding the response cost. To represent the response cost, one point from the score displayed on the bucket was subtracted from the total score after every 5 clicks.

3.3.1.4 Data analysis

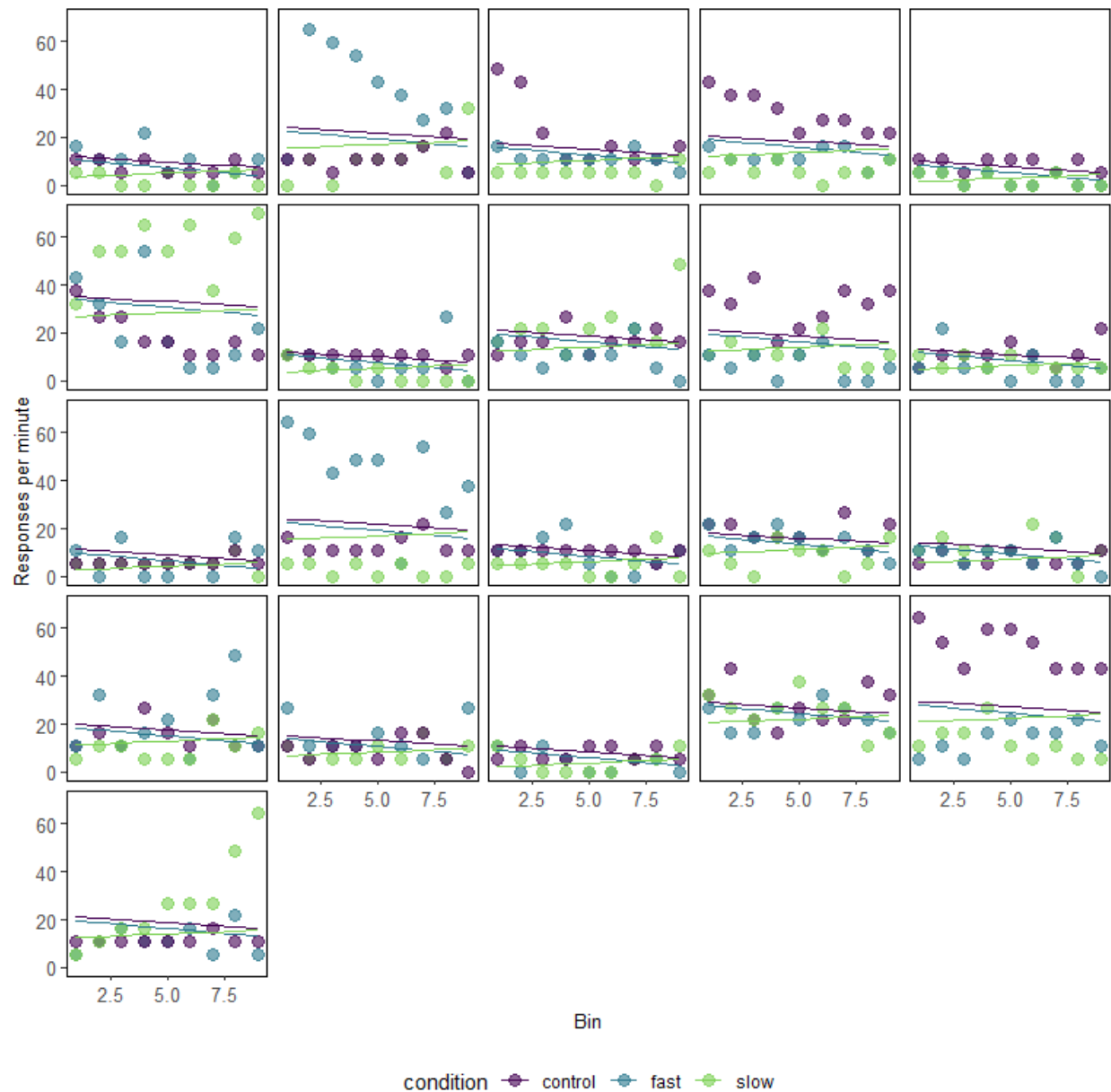
As there were no changes to the overall design of this experiment, analyses were carried out as per Pilot 1.

3.3.2 Results

The model was successfully fit to extinction phase responding. The model indicated that individuals responding during the Extinction Phase (shown in Figure 3.3) decreased more across

bins in the Control Condition compared to the Slow Condition ($\beta = 0.93$, 95% CI = [0.08, 1.80], $t = -2.13$, $p = .034$). The model did not detect any difference in the change in behaviour across bins when comparing the Fast and Control conditions ($\beta = -0.26$, 95% CI = [-1.12, 0.60], $t = -0.59$, $p = .553$).

Figure 3.3 *Observed and Predicted Extinction Responses for Individuals in Pilot 2*



Note. See annotations for Figure 3.2. The effect described in the present data is observable in by comparing fitted lines for slow and control conditions.

3.3.3 Discussion

It appears that floor effects may have limited observations of extinction in this version of the task. Responding at the beginning of extinction was very close to the minimum response rate for many individuals (as indicated by inspection of Figure 2). This suggests that responding could not be extinguished from the rate that was established in training, which may have been limited by the presence of response costs. These results therefore demonstrate that this version of the task would not be suitable for testing individual difference in extinction learning.

The inclusion of response costs may have also introduced a confound relevant to the nature of CU traits. This is because some theories treat the disregard for costs or punishments seen in high CU trait PT individuals as the primary explanatory mechanism for the development of antisocial behaviour in these people (Patterson & Newman, 1993). The presence of costs in this experiment means that punishment insensitivity could be a possible explanation for any low rates of extinction learning. Therefore, to better link variation in extinction behaviour in this task to variation in associative processes (rather than punishment insensitivity) responses costs should not be included in the design of the present task.

3.4 Pilot 3: Response Instructions

As Pilot 2 indicated that response costs lead to unwanted floor effects, I removed these costs for Pilot 3. As an alternative means of addressing the issue of responses occurring in the absence of reinforcement, I altered the narrative framing of the task to create a reason for participants to cease responding when they no longer expected rewards. This exploratory manipulation was intended to create a stronger relationship between rates of responding and participants' expectation of reward, and thereby increase the validity of my task as a measure of extinction learning.

3.4.1 Methods

3.4.1.1 *Participants, Procedure and Design*

Thirty-two undergraduate students underwent the same procedures and task design as described in Pilot 1.

3.4.1.2 *Materials*

This task was again consistent with pilot 1, with additional instructions following the initial task instructions:

Remember that pumps take time to refill. Therefore, leaving several seconds between each click on the pump will work best.

This instruction was intended to create the concept that participants should match their rate of responding to the VI schedule that they experienced. The text continued:

Sometimes, if the water runs out, Quentin will decide that it is time move on and find a new pump. However, if you are still operating the pump he might stay.

And a variation of the instructions that appeared between each pump:

Quentin has decided to search for a new pump. He will drink some water and store the rest, so you will need to fill up the bucket again. Remember that pumps take time to refill. Therefore, leaving several seconds between each click on the pump will work best. Sometimes, if the water runs out, Quentin will decide that it is time move on and find a new pump. However, if you are still operating the pump he might stay.

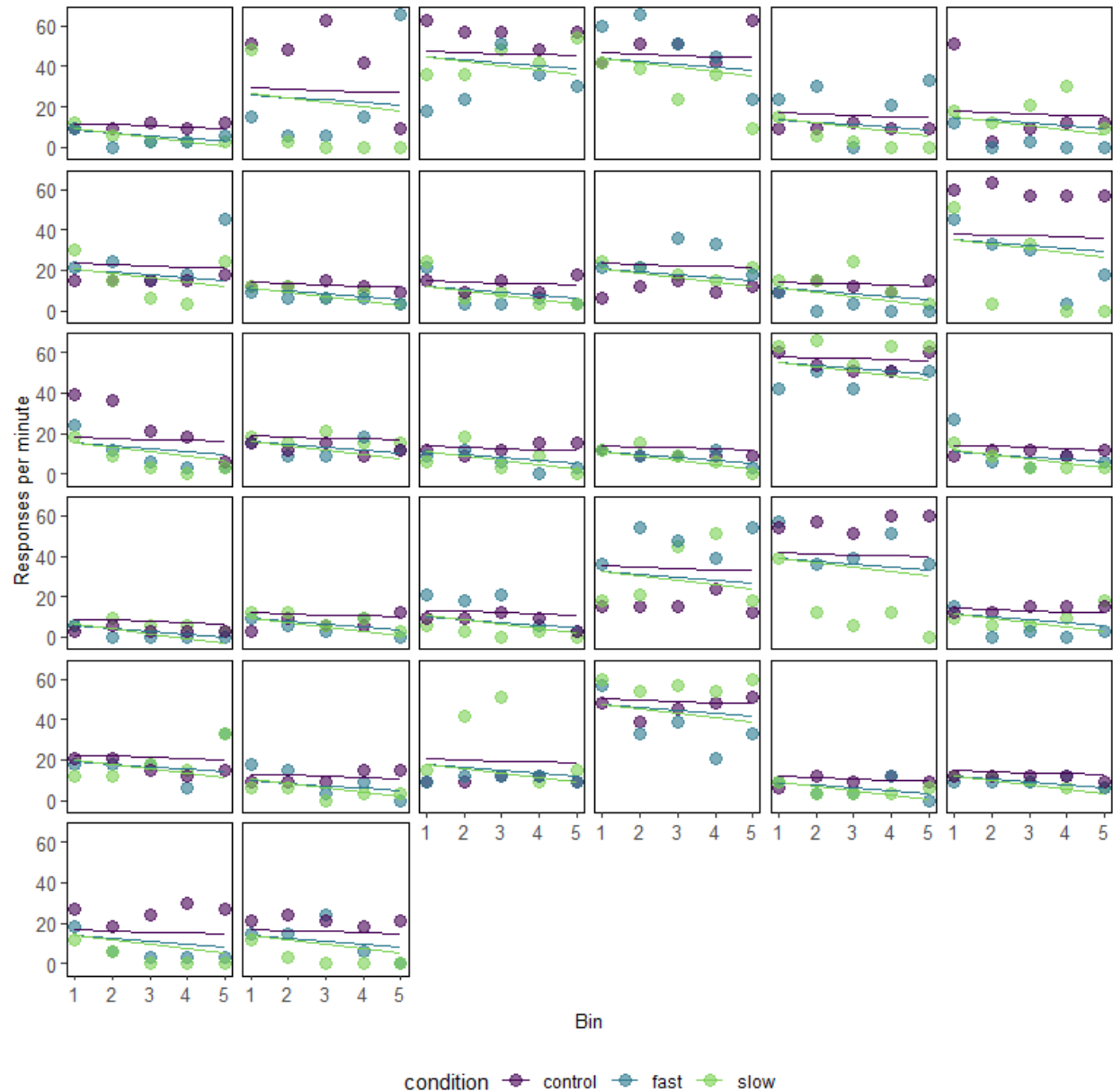
The final sentence of these sets of instructions was designed to provide participants with a reason, consistent with the task narrative, to cease responding when they no longer expected a reward.

3.4.1.3 Data analysis

Data for this pilot were binned into five bins throughout the extinction phase, aside from this variation, the analysis matched that utilised in Pilot 1.

3.4.2 Results

The model was successfully fit to responding from the extinction phase of the task (see Figure 3). As seen in Pilot 1, there was a greater reduction in responding throughout the extinction phase of the Slow condition than that which occurred in the Control Condition ($\beta = -1.65$, 95% CI = [-3.28, -0.02], $t = -1.97$, $p = .049$). This comparison did not reach significance for the Fast Condition ($\beta = -0.85$, 95% CI = [-2.49, 0.78], $t = -1.02$, $p = .308$).

Figure 3.4 *Observed and Predicted Extinction Responses for Individuals in Pilot 3*

Note. See Figure 3.2 for annotations. The effect described by the model can be seen in the larger negative gradient of the Slow Condition in comparison to the Control Condition.

3.4.3 Discussion

As observed in Pilot 1 this pilot indicated that reductions in behaviour consistent with extinction learning can be observed in this novel task. Unlike the first pilot, this was only evident for the condition trained under the VI3 schedule (Slow Condition). The Fast Condition did not

display the same effect size as Pilot 1, and greater variability was seen in the effect of the Slow condition. This increased behavioural variability may have been introduced by the additional instructions in this version of the task. Indeed, one possible interpretation of these instructions was that behaviour could be withheld in order to complete the task faster. Therefore, omitting these instructions could help to avoid this unwanted motivation for extinction behaviour, and to minimize the complexity of the task.

3.5 Experiment 1: An Effortful Response

The three pilot studies produced multiple observations of extinction behaviour in the novel extinction task. However, I observed that high rates of responding at the end of the extinction phase often occurred; indicating that not all responding may have not been contingent on reward expectation. The current experiment therefore aimed to test an alternative task feature designed to increase the contingency between participants' reward expectation and their response rates. The new design required participants to click and drag the lever, rather than simply clicking on it to respond. This manipulation necessitated more attention to the task. Additionally, the added complexity of the response was intended to increase the effort required to respond. In animal research, responses requiring greater effort have been seen to be more tightly correlated with interval schedules (Skinner & Morse, 1958). I therefore anticipated that this design change would improve the contingency between reward and response rates in this task.

As I had observed repeated evidence of extinction behaviour in the pilot studies, I aimed to use this task to address one of the main aims introduced in this study. Specifically, to provide evidence to support reports of extinction deficits in high CU individuals. To operationalise CU traits in this experiment I utilised the Inventory of Callous Unemotional traits (ICU). This measure has not previously been used in a learning study of CU traits or PT, therefore I did not utilise any categorisation of scores on this measure. This approach avoided any assumptions regarding the structure of the relationship between this measure and extinction learning (apart from linearity). Additionally, due to complexity in the ICU's associations with anxiety (Cardinale & Marsh, 2020), anxiety was included as a covariate in this experiment. This inclusion determined whether variation in anxiety accounted for behaviour in the task, which has been documented in some reinforcement learning tasks (Mkrtchian et al., 2017). The hypothesis tested

in this experiment was that responding would diminish to a lesser extent throughout the extinction phase for individuals with higher CU traits.

3.5.1 Methods

3.5.1.1 Participants

This experiment involved 48 undergraduate students. Thirteen were male, 35 were female, 13 were male, and the average age was 19.38 (SD = 1.59).

3.5.1.2 Procedure

The procedure matched that of the pilot studies, with the addition of the questionnaires presented after other materials, so as to avoid any contextual influence of the questionnaire materials on task responding.

3.5.1.3 Design

The structure of this task is outlined in Table 3.2. Participants were sequentially presented with two conditions, each of which underwent a training and extinction phase. The Experimental Condition was trained on a VI 1.5 schedule and was not reinforced during the extinction phase. The Control Condition was reinforced at VI 1.5 through both phases, so as to control for time-based effects on responding, allowing for statistical isolation of the effect of non-reinforcement in the Experimental Condition. The VI 1.5 schedule was half that used in the Fast Condition of the pilot studies and was chosen despite results showing that slower VI schedules were more effective in pilot studies. This choice was made to encourage high response rates despite the additional time and effort required to make a response in this experiment. After training there was an Extinction Phase of 140 seconds, with the last 40 seconds of the control condition going unreinforced. This 40 second period was intended to allow participants to learn that all pumps would run out, as consistent with the narrative context of the task.

Table 3.2 Reinforcement Schedules for Each Response

Pump	Training	Extinction phase
Experimental	1.5	Ø
Control	1.5	1.5

Note. Numbers refer to VI schedules, Ø indicates no reinforcement.

3.5.1.4 Materials

Extinction Task. The materials for this task matched those described in Pilot 1. However, in this experiment, participants had to click and drag the handle of a pump up to a set point and then release it in order to make a response. The initial instructions were therefore amended to read:

Thank you for participating in this experiment. In this task we would like you to help Quentin the Quokka who is lost in the desert. Quentin has found water pumps in the desert, but he needs your help to pump them. In order to operate a pump, click on the handle of a pump and drag it up, when you have lifted it all the way it will lock into place and you will need to release the mouse. Operating the pump may produce a drop of water that Quentin will collect in a bucket. However, some pumps take time to refill, so you will not get a water drop every time. Quentin will sometimes decide that a pump does not make enough water, and he will move on to a new pump. Your task is to collect as much water as possible to help Quentin live in the desert. Once you have understood these instructions, please press the space key to continue.

Psychometrics. Callous unemotional traits were measured using the youth version of the ICU traits, which was developed in a population of children and youths up to 20 years of age. It has also been shown to assess consistent traits amongst older adults (age range: 17-62) (Byrd et

al., 2013; Kimonis et al., 2013) and has been used in experimental work with an adult sample (Szabó et al., 2017). I therefore assumed that the ICU would be a valid self-report measure of CU traits in a population of undergraduates. Additionally, anxious personality traits were controlled for using the trait subscale of the State-Trait Anxiety Inventory (STAI) (Spielberger, 1983) as a potential covariate.

3.5.1.5 Data Analysis

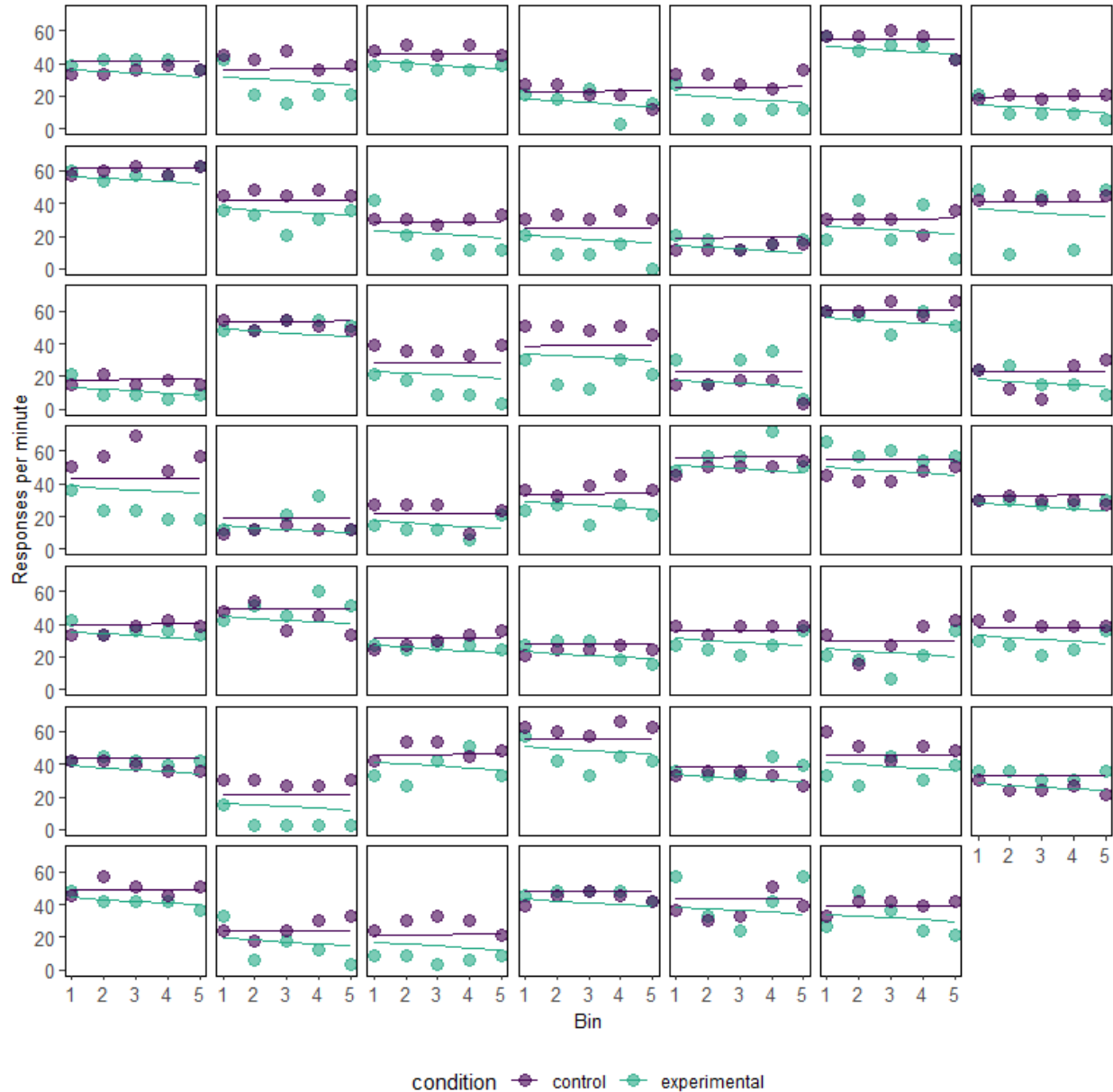
Responding during the extinction phase of the study was binned into 20 second bins and converted to responses per minute for both response levers. The initial analysis determined if the Experimental Condition led to a greater reduction in behaviour throughout extinction than the Control condition. Only the first 5 bins (100 seconds) of the Extinction Phase were examined for this comparison, as this period was reinforced in the Control Condition (unlike the subsequent 40 seconds). To test the relevant effects, a mixed effects model was fit to the responding occurring during the each of the conditions, as described in Pilot 1.

An additional model was defined to examine whether extinction behaviour was related to individual differences scores. This model was developed with the assistance of the Sydney Informatics Hub, which is funded by the University of Sydney. It was fit to the data from all seven bins from the Extinction Phase of the Experimental Condition. To estimate individual differences in rate of extinction, it included a term for random slope effects over bins (time) at the participant level of analysis. This model also produced intercept terms which were treated as a covariate to control for individual differences in initial rate of responding. To test for covariate effects on the rate of extinction variable, it was regressed on total STAI scores (z-transformed) and response rate intercepts. Significant predictors from this regression were then included in a

regression with total ICU scores (z transformed), to determine whether ICU scores predicted change in responding during extinction.

3.5.2 Results

The average ICU score was 19.75 (SD = 7.38), and that of the STAI was 44.98 (SD = 11.18). The model comparing the effects of condition on responding during extinction was fit successfully. It showed that responding decreased significantly more over time (bins) in the Experiment Condition than in the Control Condition ($\beta = -1.21$, 95% CI = [-2.25, -0.17], $t = -2.28$, $p = .023$) (Figure 3.5).

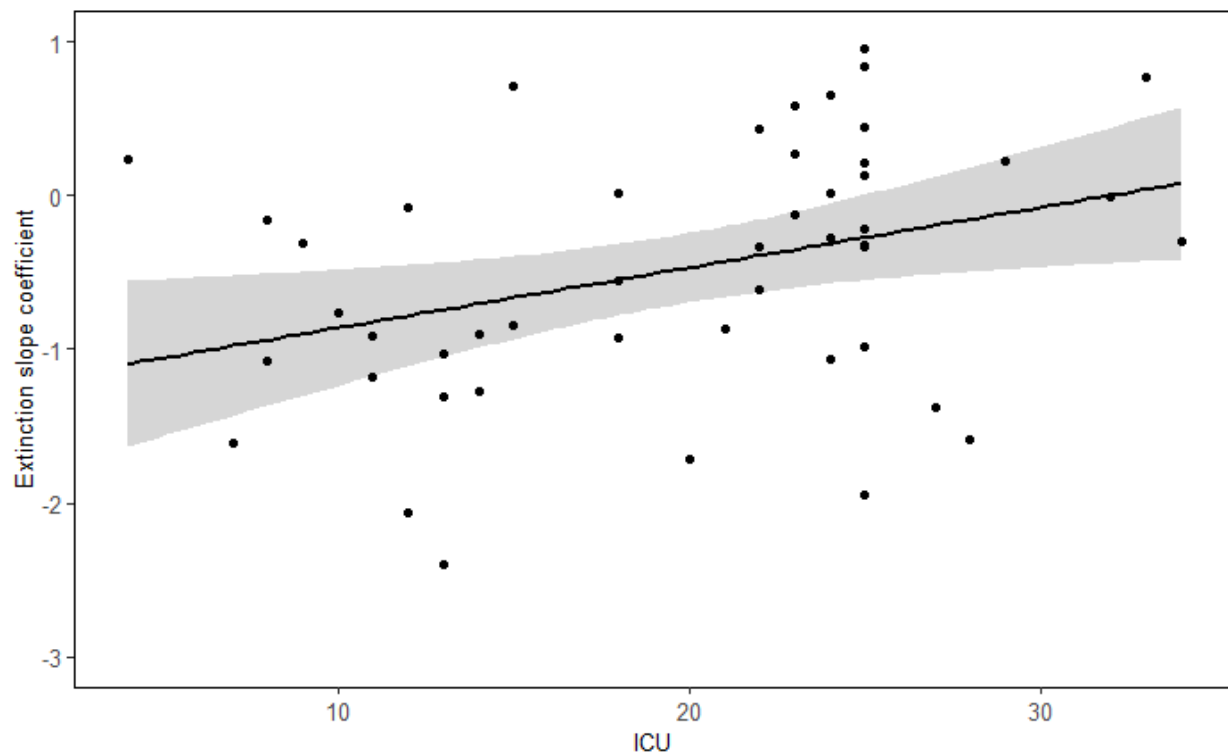
Figure 3.5 *Observed and Predicted Extinction Responses for Individuals in Experiment 1*

Note. Each plot shows an individuals' responses. The response rate in each bin of the extinction phase of this experiment is shown as a point, coloured by condition. Values predicted by the mixed effect model are shown as lines for each condition. The effect of condition can be seen in the more negative slope of predictions for the Experimental Condition.

The covariate block of the model testing change in rate of responding during extinction showed that this variable was not related to STAI scores ($\beta = -0.07$, 95% CI = $[-0.31, 0.16]$, $t = -0.63$, $p = .529$). However, response rate intercepts were a significant predictor ($\beta = 0.02$, 95% CI

= [0.00, 0.04], $t = -2.19$, $p = .034$), and were therefore included in the regression testing the effect of ICU scores. This second regression showed a positive effect of ICU scores ($\beta = 0.27$, 95% CI = [0.04, 0.49], $t = 2.42$, $p = .019$) (Figure 3.6).

Figure 3.6 *Correlation Between ICU Scores and Slope Coefficients for Extinction Behaviour*



Note. Negative coefficients reflect reductions in responding during the extinction phase of the Experimental condition.

3.5.3 Discussion

In this experiment I observed that CU traits predicted less reduction of behaviour during extinction. This result is consistent with the hypothesis that CU traits predict lower rates of extinction learning. Although this effect was small, the capacity to detect it in the general population was promising for the aims of this thesis. A large proportion of research in this field has been conducted in forensic or clinical populations (see chapter 2) which may be expected to have greater severity of CU traits than the populations recruited here, and therefore have greater

power to detect associates of those traits. This result was also novel in that it was the first such finding for the ICU measure of CU traits. This meant that the present extinction effect can be specifically linked to CU traits, rather than broader components of psychopathic traits which have been addressed in much of the prior literature (reviewed in Chapter 2).

Despite this encouraging result, several aspects of the methodology warranted improvement. As there was only one behaviour possible in the task, it may have been the case that high CU individuals enacted this behaviour impulsively – as impulsive traits are known to correlate with CU traits (Poythress & Hall, 2011; Salekin, 2017). Indeed, as most participants maintained some level of responding (even after 100 seconds of reward omission) there may be some behaviour in this task that was driven by factors such as exploration, impulsivity, or demand characteristics, which do not reflect reward expectation. Such responding therefore limits the ability of this task to detect extinction learning. Controlling for impulsivity may therefore improve the validity of the present result.

A further means of improving the analysis above would have been to consider the change in responding between training and extinction. In the analysis described for this experiment, any adjustment in rates of responding occurring within the first 20 second bin of extinction would not have been detected, as the rate of responding at the end of training was not considered. Additionally, increasing VI schedules may have been a means of slowing adjustments in behaviour after the onset of extinction (as participants should have anticipated longer periods of reward omission). Such a methodological change would maximise the capacity of this task to measure significant behavioural processes occurring within the first bin of extinction. Therefore, despite the promising nature of the present result reported here, a number of methodological improvements were warranted.

3.6 Experiment 2: An Alternative Response

As I had continued to observe effects consistent with extinction in both the pilots and Experiment 1, I aimed to replicate and extend these observations in Experiment 2 by including tests for renewal effects. This inclusion was consistent with my aim to demonstrate deficits in renewal amongst individuals with CU traits. Although Experiment 1 showed an association between CU traits and Extinction, I aimed to improve the sensitivity of the methodology to extinction learning by introducing another variation to the task designed to improve the contingency between response rates and reward expectation. This design comprised two response “pumps”, one of which was subject to experimental manipulations, and the other of which provided an alternative response for participants to make when they did not expect the extinguished pump to produce rewards.

In this experiment I also intended to address a possible confound identified in Experiment 1, where reduced extinction behaviour could be attributable to impulsive traits, a known correlate of CU traits (Poythress & Hall, 2011; Salekin, 2017). This was significant as the Urgency, Premeditation (lack of), Perseverance (lack of), Sensation seeking (UPPS) impulsivity measure has been implicated in deficits in reversal learning (Fernández-Serrano et al., 2012). I therefore aimed to measure any such effects by including this measure of impulsive traits in the analysis. I also intended to measure immediate responses to extinction in this Experiment, in addition to the overall change in responding occurring throughout extinction. This was identified as a potentially informative measure in the discussion of the previous experiment.

This experiment allowed me to test the hypothesis that reductions in responding in extinction would be slower amongst individuals with higher levels of CU traits, resulting in slower rates of change. I also tested the hypotheses that these traits would predict less renewal of

responding following a return to a training context after extinction. Additionally, this study considered rapid reacquisition as an alternative procedure implicated in contextual learning (Bouton et al. 2021). I predicted that, as for renewal, less rapid reacquisition would occur in individuals high on CU traits.

3.6.1 Methods

3.6.1.1 Participants

I recruited 37 participants for this study, 19 were paid participants from the university community, and a further 18 were first year psychology students. Two participants were excluded, one due to failure to engage in the experimental task, another due to technical failure during the experiment. This left a final sample of 35 in which there were 21 females and 14 males. Fourteen of the participants indicated that English was not their first language. The average age was 24.09.

3.6.1.2 Design

This task was designed to test two relapse effects: renewal and rapid reacquisition, as well as extinction learning (for a summary of the design see Table 3.3). Subjects experienced three pairs of response levers that were presented sequentially in a random order. One of these pairs allowed for the measurement of rapid reacquisition ('Rapid Reacquisition Condition'), a further pair which acted as a baseline for measurement of the renewal effect ('Renewal Control Condition'), and the final pair measured renewal itself ('Renewal Condition'). During presentations of each of the pairs of levers, training and extinction phases occurred and were followed by a test of learning. As the renewal occurs due to changes in context after extinction learning, I varied the context for the Renewal Condition. This variation was such that the training context was different from the extinction context and returned to the training context during the

test phase. For the other conditions context remained constant (Table 3.3). The order of conditions was randomized across participants. To keep the duration of the task consistent with prior versions I did not include a control condition (for prior the prior Experiment and pilots the control condition was not subject to reinforcement omission, so as to control for changes in responding over time not caused by extinction). As prior iterations of the task had consistently produced differences in responding between control and extinction conditions, I was confident that the task could measure changes in responding due to reward omission.

For all conditions, one lever was reinforced on a VI 2 schedule (the ‘Experimental Pump’) and the other on a VI 6 schedule (‘Alternative Pump’) (see Table 3.3). The ‘Training Phase’ continued until participants had earned a certain number of rewards from the Experimental Pump (randomly selected from a uniform distribution ranging from 15 to 25). This phase was immediately followed (with no indication to participants) by the ‘Extinction Phase’ in which the Experimental Pump was no longer reinforced. This phase lasted for 80 seconds. Finally, the task moved (again without any notification to participants) to the ‘Relapse Test Phase’, which lasted 60 seconds for each pair of pumps but involved different reinforcement schedules across conditions. Responses in this ‘Relapse Test Phase’ were not reinforced for the Renewal Condition and the Renewal Control Condition, whilst the Rapid Reacquisition Condition responding was reinforced as for the Training Phase.

Table 3.3 *Reinforcement Schedules for Each Response Lever and Contexts for Each Phase*

Condition	Feature	Training	Extinction Phase	Relapse Test Phase
Renewal Control	Experimental	2	Ø	Ø
	Alternative	6	6	Ø
	Context	A	A	A
Renewal	Experimental	2	Ø	Ø
	Alternative	6	6	Ø
	Context	A	B	A
Rapid Reacquisition	Experimental	2	Ø	2
	Alternative	6	6	6
	Context	A	A	A

Note. All numbers refer to VI schedules, and Ø indicates no reinforcement.

3.6.1.3 Procedure

The procedure matched that of the pilot studies, with questionnaires completed after the extinction task.

3.6.1.4 Materials

The layout of the task is shown in Figure 3.7. Participants were able to click, drag up the handle of a depicted water pump, and release the mouse click to produce water droplets. Two pumps were available, of which one at a time could be lifted at any time throughout each condition. Other elements of the task were the same as has been described in Pilot 1 (Note that alterations to the layout of the task are shown in Figure 7 and that the instructions made reference to “pumps” rather than a “pump”).

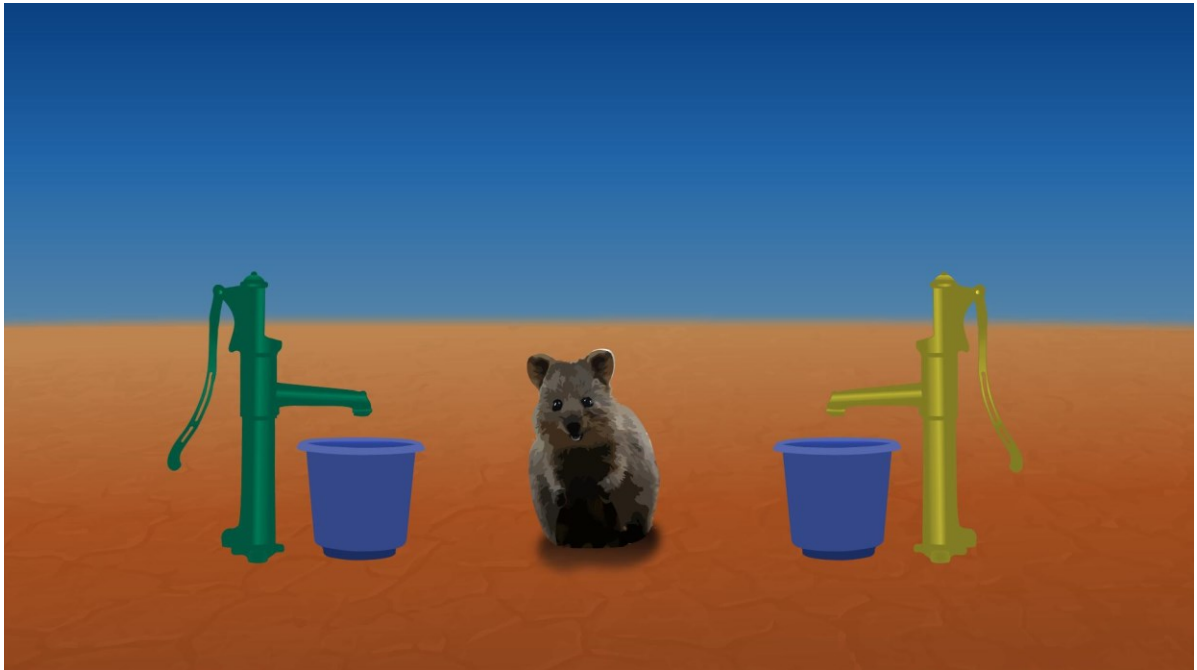
At the beginning of each condition a day-time or night-time scene was randomly allocated to act as context A or B (Table 3.3), this prevented any confounding effects of a particular scene across all participants. Although contexts could be consistent across conditions, discrimination of conditions was aided by changes in the colour of displayed pumps and presentation of the following instructions (as described in Pilot 1):

Quentin has decided to search for a new pump. He will drink some water and store the rest, so you will need to fill up the bucket again. Please press the spacebar to continue the task.

Between context changes during the Renewal Condition (which occurred as described in the design section), participants were shown the following instructions for 5 seconds, depending on whether the context was shifting from daytime to night-time or vice versa: “Quentin asks you to take a break, Night falls...” or “Quentin asks you to take a break, the sun rises...”.

Figure 3.7 *Layout of The Task in Experiment 2*

A.



B.



Note. Moveable pump handles are shown in various colours that were randomly allocated in the task. Panel A Shows the daytime context and Panel B the night-time context.

3.6.1.5 Psychometrics

As for the previous experiment, I used the ICU and STAI questionnaires to assess CU traits and trait anxiety. I additionally implemented the UPPS scale to measure impulsivity. This scale was developed as a measure of impulsivity in undergraduates by Whiteside & Lynam (2001) whose factor analysis indicated five components: urgency, lack of premeditation, lack of perseverance and sensation seeking. This scale was measured on a four-point Likert scale ranging from “agree strongly” to “disagree strongly”. Here, total scores on the UPPS were used to measure the impact of impulsive traits on task behaviour.

3.6.1.6 Data Analysis

The number of lever presses on the experimental pump in each condition were recorded throughout this experiment. These data were then collated into bins as a response rate/minute metric. For Training Phase data (in each condition), the length of the phase was recorded and divided into 22 bins within which responding rates were calculated. The other phases of each condition, which were a set length, were divided into 5 second bins (16 for the Extinction Phase and 12 for the Relapse Test Phase). I controlled for covariates (gender, STAI, and UPPS scores (totalled and z-transformed)) in the analyses described below by testing whether they predicted the behavioural effect in the initial block of each regression. Only significant covariates were reported, and were included in the second block, in which the effect of CU traits was tested. Statistics for the effect of task manipulations were reported from this second block. Additionally, as I had noted variation in English language background in recruited participants, I tested whether this variable (measured by the question: “Is English your first language?”) influenced the effects of interest. To achieve this, I included this variable as a moderator term for effects of

ICU on behaviour. This acted to account for any effects of task or psychometric comprehension resulting from language fluency.

Extinction Learning. Two separate analyses were used to model extinction responding. One took an approach consistent with the prior experiment – examining change in responding over bins of extinction. The other model was designed to include any immediate changes occurring at the start of extinction by comparing total change in behaviour between training and extinction phases. This was carried out on data from the renewal-control and rapid reacquisition conditions, which did not include context changes. The model tested for the interaction between phase, ICU score and English background to determine firstly, if CU traits moderated the change in behaviour after omission of reward in the Extinction Phase, and secondly if such an effect depended on English language background. This model included a term for random intercepts at the participant level of the analysis to account for within-subjects effects.

To investigate differences in the rate of extinction learning, a second mixed effects model was utilised, following the approach developed Experiment 1. This analysis was applied to binned data across the Extinction Phase of the Renewal Control and Rapid-Reacquisition Conditions. It included a term for random slopes across bins of extinction at the participant level of analysis. This term which generated linear estimates of behavioural change across the data. These estimates were then regressed on an interaction term including ICU scores and English language background, with the intercept of each slope included as a covariate to control for the effect of initial response rate on responding. This allowed me to test for the effect of CU traits on learning rate.

Renewal Effect. To determine the presence of a renewal effect, the difference in responding between the last four bins of the Extinction Phase and the first four bins of the

Relapse Phase during the Renewal Condition was compared to the same difference within the Renewal Control Test. These bins were chosen to concentrate my analysis on the immediate response to the change in context. To make the necessary comparison I fit a mixed effects model with a term for random intercepts at the participant level and fixed effects terms for the interaction between four factors: condition (Renewal vs Renewal Control), phase (Extinction vs Relapse Test), ICU score and English language background. This determined whether CU traits moderated any renewal of responding after the onset of the training context in the relapse phase, and if this depended on English Language background.

Rapid Reacquisition. The difference between responding in the last four bins of the Extinction Phase and all bins of the Relapse Phase for the Rapid Reacquisition Condition was analysed. As for prior analyses, I included the interaction between English language background, ICU score and this phase variable to test for a relapse effect in this condition. The random intercept of participants' response rate was also modelled.

3.6.2 Results

Descriptive statistics for psychometric measures in this study are shown in Table 3.4. Additionally, ICU descriptive statistics, grouped by English language background, are shown in Table 3.5.

Table 3.4 *Descriptive Statistics for All Psychometric Scales*

scale	mean	sd	min	max	alpha	MIC
UPPS	108.00	14.33	81	147	0.87	0.13
STAI	46.34	10.00	28	69	0.88	0.28
ICU	24.71	9.21	8	40	0.85	0.18

Note. MIC = mean inter-item correlation. alpha = Chronbach's alpha.

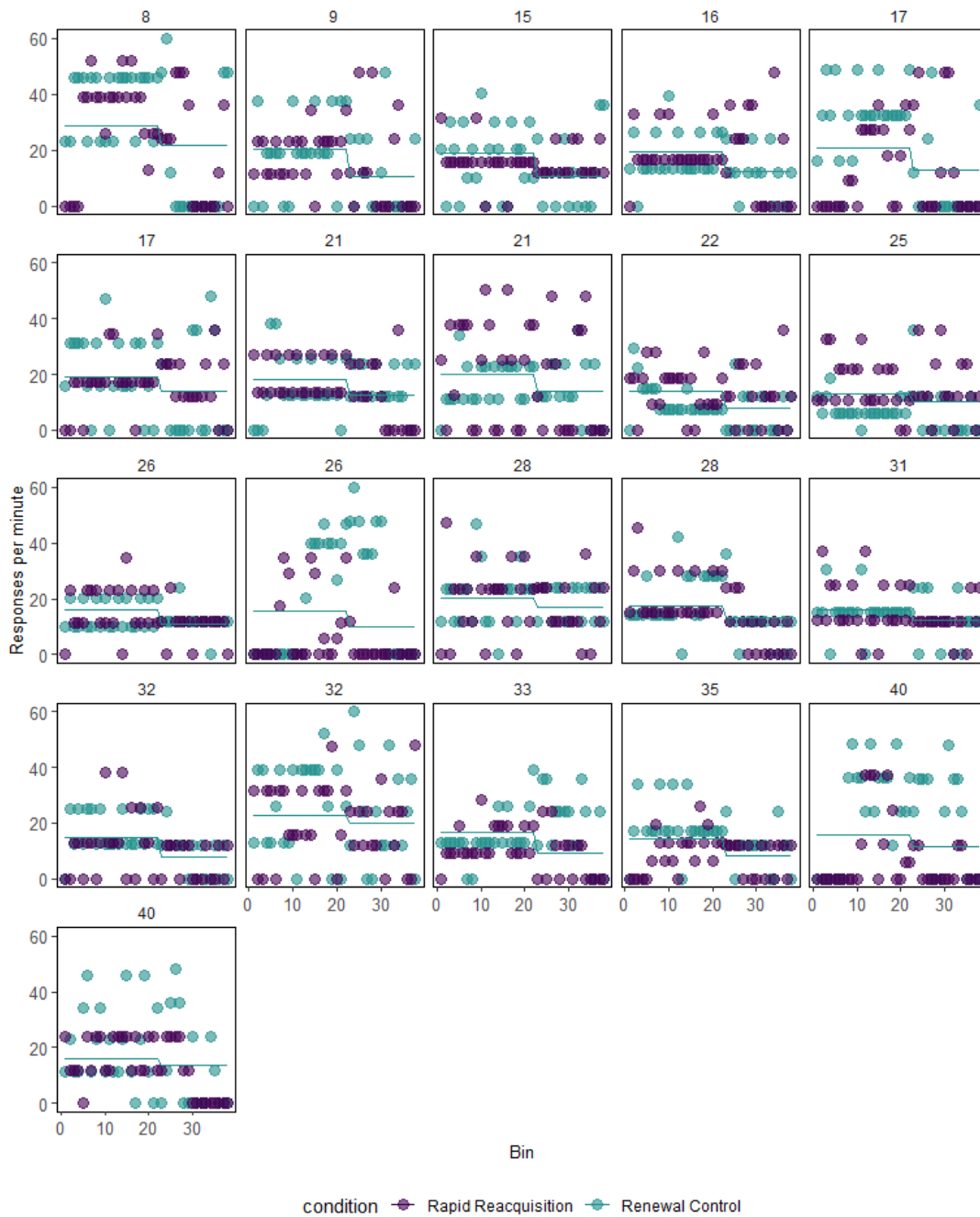
Table 3.5 *ICU Scores, Grouped by English Language Background*

English first language?	Mean	SD	Min	max	Cronbach's alpha
Yes	24.86	9.12	8	40	.85
No	24.50	9.70	9	40	.85

3.6.2.1 Total Extinction Effect

Covariate analysis of responding during the Training and Extinction phases of the Renewal Control and Rapid Reacquisition conditions revealed a significant reduction in responding between these phases ($\beta = -6.66$, 95% CI = $[-7.88, -5.44]$, $t = -10.69$, $p < .001$). I found that STAI scores were a significant covariate such that as scores increased, responding decreased more between phases ($\beta = -1.69$, 95% CI = $[-2.74, -0.65]$, $p = .002$). UPPS was not found to be a significant covariate ($\beta = 0.96$, 95% CI = $[-0.10, 2.03]$, $p = .075$), nor was gender ($\beta = 0.04$, 95% CI = $[-1.92, 2.00]$, $p = .964$). Whilst accounting for the direct effect of STAI scores, I found that ICU scores significantly determined the change in responding between phases, such that higher levels of traits were associated with more reduction in responding ($\beta = -2.56$, CI = $[-4.07, -1.13]$, $t = -3.46$, $p = .001$). However, this effect depended on subjects' English language background ($\beta = 4.65$, 95% CI = $[2.72, 6.59]$, $t = 4.72$, $p < .001$). Further (post-hoc) analyses of each English language group separately (continuing to account for variation in STAI) showed that the relationship between ICU scores and change in responding between phases for participants without English as a first language was negative ($\beta = -1.94$, 95% CI = $[-3.39, -0.49]$, $t = -2.62$, $p = .009$) (Figure 3.8). Whereas participants with English as a first language showed a positive relationship, where higher traits predicted less reduction in responding ($\beta = 1.44$, 95% CI = $[0.04, 2.84]$, $t = 2.02$, $p = .044$) (Figure 3.9).

Figure 3.8 *Observed and Predicted Extinction Responses for Individuals with English as a First Language in Experiment 2*



Note. Each observation of response rate in each bin of the Training Phase (bins 1-22) and the Extinction phase (bins 23 to 38) of this experiment is shown as a point, coloured by condition. Values predicted by the mixed effect model (specifically, the model fit to the entire sample, with English language background as a moderator variable) for each phase are shown as lines. Each plot is labelled by the individuals ICU score. The difference in predicted values between phases can be seen to reduce as ICU scores increase.

Figure 3.9 *Observed and Predicted Extinction Responses for Individuals Without English as a First Language in Experiment 2*



Note. Please see the notes for the prior figure for a description of the plotted values. Here the difference in predicted values between phases can be seen to increase as ICU scores increase.

3.6.2.2 *Rate of Extinction*

When analysing the rate of reduction of responding in the Extinction Phase of the Renewal-Control and Rapid Reacquisition conditions I found that the individual's intercept (or initial response rate in their mixed effects model) was a significant covariate ($\beta = -0.09$, $CI = [-0.12, -0.06]$, $t = -6.10$, $p < .001$). Other individual difference variables were not significant predictors of the extinction slope term (UPPS: $\beta = 0.06$, $CI = [-0.07, 0.19]$, $t = 0.93$, $p < .358$, STAI: $\beta = -0.11$, $CI = [-0.24, 0.02]$, $t = -1.73$, $p < .094$, gender: $\beta = 0.09$, $CI = [-0.16, 0.33]$, $t = 0.74$, $p < .465$). Accounting for the direct effect of intercept, rate of behavioural change was not predicted by ICU score ($\beta = -0.02$, $CI = [-0.21, 0.17]$, $t = -0.21$, $p = .836$), English language background ($\beta = -0.05$, $CI = [-0.20, 0.31]$, $t = 0.67$, $p = .672$), or their interaction ($\beta = -0.04$, $CI = [-0.29, 0.21]$, $t = -0.32$, $p = 0.75$).

3.6.2.3 *Renewal*

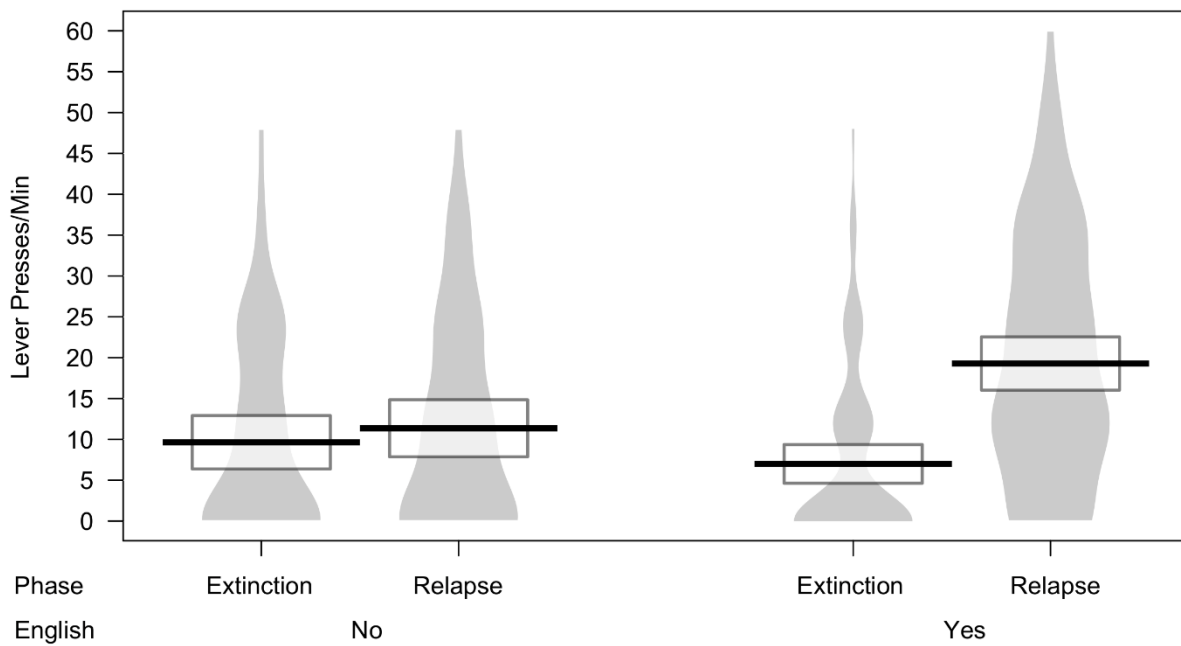
The covariate model for the renewal effect indicated no significant covariates. The model testing the effect of ICU scores and their interaction with language background showed that there was significantly more responding after the transition between the end of Extinction and the beginning of Relapse Test phases in the Renewal Condition than in the Renewal Control condition ($\beta = 6.27$, $CI = [0.77, 11.78]$, $t = 2.24$, $p = .026$). However, this effect was not significantly moderated by ICU score ($\beta = 2.52$, $CI = [-2.90, 7.94]$, $t = .91$, $p = .362$), English language background ($\beta = -0.94$, $CI = [-8.08, 2.02]$, $t = -0.26$, $p = 0.795$) or their interaction ($\beta = -5.49$, $CI = [-12.64, 1.65]$, $t = -1.51$, $p = 0.131$).

3.6.2.4 *Rapid Reacquisition*

The covariate model for rapid reacquisition showed no significant covariates. In the model testing for effects of ICU scores, there was no significant difference in responding

between the end of the Extinction and the beginning of the Relapse Test phases in the rapid reacquisition condition ($\beta = 1.69$, $CI = [-2.45, 5.85]$, $t = 0.79$, $p = .428$). However, English language background interacted significantly with this phase variable ($\beta = 10.58$, $CI = [-2.45, 5.85]$, $t = 3.83$, $p < .001$) (Figure 3.10). This interaction reflected a greater increase between extinction and reacquisition for participants with English as a first language. ICU scores did not significantly moderate the effect of the reacquisition phase on responding ($\beta = -0.68$, $CI = [-4.77, 3.41]$, $t = -0.32$, $p = .75$).

Figure 3.10 *Effect of English Language Background on Rapid Reacquisition*



Note. Dark lines show the mean (surrounding boxes show the 95% CI) of responding in the last four bins of Extinction and the first four of Relapse phases in the Rapid Reacquisition condition, split by English language background (Yes = English was first language, No = not first language). Rotated probability density plots are shown in grey.

3.6.3 Discussion

These results were consistent with the hypothesis that the amount of reduction in overall responding from training to extinction was lessened at higher levels of CU traits. However, this was only the case for individuals with English as a first language. This groups' behaviour was

consistent with that observed in Experiment 1, where the differences in responding between phases was related to CU traits. This effect is analogous to other results seen in the psychopathy literature, such as reversal learning deficits (Budhani & Blair, 2006), where high PT individuals take longer to adjust their responding after a change in contingencies between phases of the task.

Despite observing an extinction effect consistent with my hypothesis, I did not see a corresponding effect of CU traits on the change in responding occurring throughout the extinction phase. This finding was consistent with Experiment 1, where I speculated that unexplained maintenance of behaviour during extinction may limit the capacity of the methodology to detect such an effect. In light of those findings, I introduced an alternative response to this experiment. This was designed to remove some of the unexplained responding occurring during extinction in Experiment 1, which limited possible rate of change in behaviour. Inspection of Figure 3.8 and 3.9 indicated that I did not achieve this aim: behaviour during extinction was highly variable, even in the absence of reward. It was possible that this was due to exploratory behaviour of participants swapping between responses.

Variability in responding with the extinction phase may have meant that any extinction rate effect may have required greater power to detect than the overall extinction effect. This overall effect was larger and more consistent, possibly because the onset of the extinction phase was a salient event where the experimental response lever instantly became an ineffective source of reward. These data therefore suggested that the inclusion of an alternative response introduces unnecessary variability to the data. It further reinforced previous indications that the difference between training and extinction rates of responding may have been a better index of extinction learning than the rate of behavioural change during extinction.

Another, unexpected, feature of the sample in this study was the discrepant findings for non-native English speakers. I saw that the effect of CU traits on extinction was reversed for these participants; higher CU traits led to greater reductions in responding. This finding did not coincide with any differences in levels or reliability of the ICU measure between the English language groups. However, there was evidence that participants who did not have English as their first language were less likely to resume responding when the response outcome was reintroduced (during the Relapse Phase of the Rapid Reacquisition Condition). This set of observations suggests that these participants may have had a different interpretation of the task. Any difference in participants' understanding of the extinction task may have contributed to the different relationship between extinction and CU traits in the English as a second language subsample. However, this difference would need to be examined systematically to provide a more conclusive explanation.

A notable result for this study was the presence of a renewal effect. Such demonstrations of renewal in human subjects are limited (Crossley et al., 2014), and this task's potential capacity to measure this process should be noted (especially given its potential applicability in child populations). The absence of any relationship between CU traits and relapse effects in this experiment was difficult to interpret as this is the first test of these effects using this paradigm. Given the large variation in estimated effect magnitude this methodology may not have had sufficient statistical power to detect any of these effects where present. This suggests that any future study aiming to confirm the presence or absence of such relationships should be sure to achieve adequate power to observe the effect sizes documented here.

3.7 Experiment 3: Renewal

In this Experiment I aimed to simplify the methodology used in Experiment 2 and replicate my observation of a renewal effect. These outcomes would permit a finalised methodology applicable to a larger confirmatory sample. Over the course of this series of studies I utilised many variations on the basic task structure tested in Pilot 1, with the aim of increasing the contingency between participants' response rate in the task and their expectation of reward. The best such variation was that made in Experiment 1, which increased the attention and effort required to make a response by requiring participants to click and drag the pump lever. All other variants introduced potential confounds or unwanted variability in behaviour. In this Experiment, I therefore utilized the response mechanism developed in Experiment 1 whilst slowing the VI reward schedule, as the pilot studies had shown extinction effects more reliably under longer VI schedules. These features were chosen so as to provide the best circumstances for observing the hypothesized association between CU traits and reduced extinction learning.

In addition to these changes, I also implemented the test of renewal developed in Experiment 2. This test was included to provide a means of testing my hypothesis that relapse effects would be smaller for individuals with greater levels of CU traits. The renewal test was chosen over the rapid reacquisition test because it produced a larger and more consistent effect in Experiment 2. Additionally, it would provide direct evidence that the processing of context should be considered in studies of cognitive differences associated with CU traits. As the presence of a renewal effect was a novel observation in my last experiment, I aimed to replicate the effect in a larger sample. This procedure served the aim of this study to demonstrate reductions in the renewal of extinguished responding in individuals with high CU traits.

Additionally, in order to show that the extinction effects observed using the novel extinction task were related to those documented in the literature reviewed in Chapter 2, I included a replication of the reversal task used in Budhani and Blair (2006). In doing so I aimed to demonstrate that the level of variation in CU traits amongst non-clinical adults can be linked to reversal deficits. Despite extinction and reversal deficits being interpreted as evidence for the same aetiologies of PT and CU traits (Blair et al., 2005; Moul et al., 2012), such a result would be the first evidence of co-incident reversal and extinction learning deficits. I therefore hypothesized that CU traits would predict greater inaccuracy in responding after a reversal of response contingencies in this task.

3.7.1 Methods

3.7.1.1 *Participants*

Recruitment occurred in two stages for this study. In the first, 80 first year psychology students were recruited. Of these, one was excluded due to technical failure. A further 11 were excluded because they did not meet the task performance criteria outlined in the following sections. Thirty-Two of these participants indicated that English was not their first language. After analysis of Experiment 2 indicated that the effect of CU traits on extinction behaviour was significantly mediated by this language measure, I decided to recruit a full sample of 80 participants with English as a first language. This alteration to the sampling strategy provided a full dataset whilst avoiding unexplained effects of English language history. This additional sample was recruited by advertising the study only to subjects with English as a first language. This additional recruitment included 41 participants, one of whom was excluded for failing to meet English language criteria, and a further three were excluded due to failing to meet task performance criteria. This left 73 participants with English as a first language, 52 of whom

identified as female (21 as male) and whose average age was 19.36 (SD = 1.95). Of participants without English as a first language 29 were female and 2 were male, the average age was 20.23 (SD = 3.48).

3.7.1.2 Design

Extinction Task. Participants were sequentially presented with two conditions, with the order varied randomly across participants. In each, a single response lever underwent a training phase in which lifting the lever produced rewards on a VI 4 schedule with an SD of 1.5s. This phase lasted until participants had received a number of rewards selected randomly from a uniform distribution ranging from 17 to 23 (20 on average). Following training, a 100 second extinction phase occurred followed by a further 20 second ‘Renewal’ phase in which responses continued to go unreinforced. One of the two conditions acted as a test and the other as a control. For the experimental condition, participants were exposed to different contexts during training and extinction, and the training context was present during renewal (ABA context schedule). For the Control Condition the context was consistent throughout, acting as a baseline to compare to the effect of context changes in the experimental condition (AAA context schedule, see Table 3.6).

Table 3.6 *Reinforcement Schedules and Contexts for Each Phase and Condition in the Extinction Task*

Condition		Training	Extinction phase	Renewal phase
Experimental	Reinforcement	4	Ø	Ø
	Context	A	B	A
Control	Reinforcement	4	Ø	Ø
	Context	A	A	A

Note. All numbers refer to the variable interval (VI) schedules (as number of seconds), and Ø indicates no reinforcement.

Reversal Task. On each trial of this task (based on that described by Budhani and Blair (2006)) participants chose one of two stimuli, and subsequently either received or lost 100 points. Each of the stimuli was probabilistically related to each of these outcomes, with one stimulus always being more likely to lead to the rewarding outcome. For three of the pairs, termed ‘reversing pairs’, the most reinforced stimulus was switched after 40 trials (acquisition phase), and presented for a further 40 trials (reversal phase). Three further ‘non-reversing pairs’ were presented for 40 trials in total and were included to prevent participants from expecting that all pairs of stimuli would reverse. Each of the three pairs of stimuli in the reversing and non-reversing sets of pairs were reinforced according to the following probabilities: $[1 - 0]$, $[.8 - .2]$, $[.6 - .4]$. Throughout the experiment trials of a reversing pair were interspersed with trials of a non-reversing pair. The onset of new pairs was staggered such that new non-reversed pairs initiated when the reversal phase of the reversing pair began (see Table 3.7). The remaining two ‘dummy pairs’ were used to intersperse with reversing pairs in portions of the design in which reversed and non-reversed trials did not overlap, the two stimuli in each of these pairs was reinforced at probabilities of 1 and 0. The order of each probability of reversed and non-reversed pairs was randomized between subjects.

Table 3.7 *Number of Trials for Each Pair of Stimuli in Each Block of the Reversal Task*

Stimulus Pair	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7
Dummy 1	20						
Reversing 1	40	40					
Non-reversing 1		20	20				
Reversing 2			40	40			
Non-reversing 2				20	20		
Reversing 3					40	40	
Non-reversing 3						20	20
Dummy 2							10

3.7.1.3 Procedure

Participants were asked to complete two computer-based tasks to examine reward learning. Participants completed the self-report questionnaires on the computer after the behavioural tasks had concluded. All other aspects of the procedure are as described for Pilot 1.

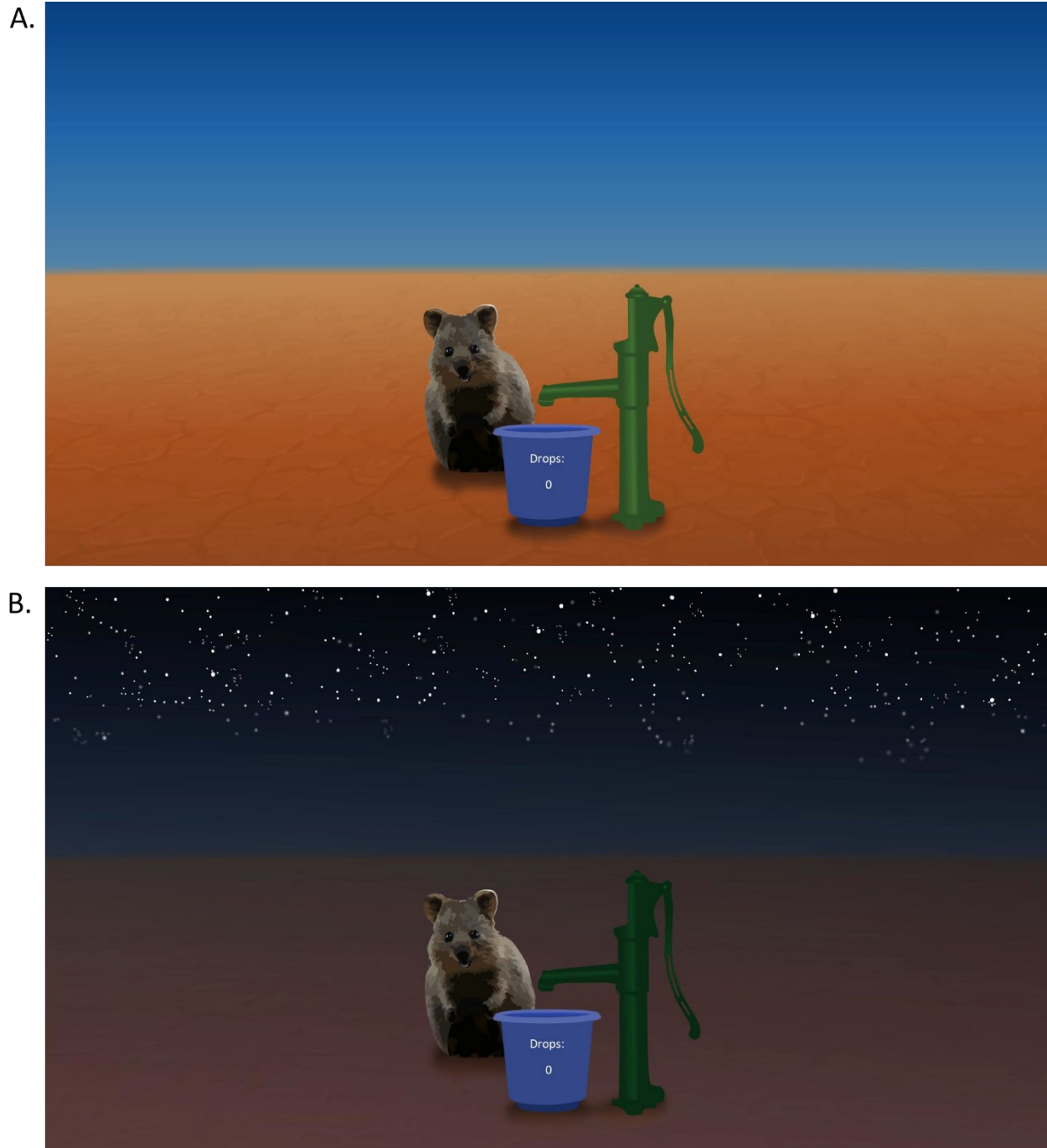
3.7.1.4 Materials

Extinction Task. This task ran as described in Experiment 1. Additionally, the two contexts developed in Experiment 2 (a daytime scene and a night-time scene) were utilised here (Figure 14). The task material therefore ran as follows. The initial instructions read:

Thank you for participating in this experiment. In this task we would like you to help Quentin the Quokka who is lost in the desert. Quentin has found water pumps in the desert, but he needs your help to pump them. In order to operate a pump, click on the handle of a pump and drag it up, when you have lifted it all the way it will lock into place and you will need to release the mouse. Operating the pump may produce a drop of water that Quentin will collect in a bucket. However, some pumps take time to refill, so you will not get a water drop every time. Quentin will sometimes decide that a pump does not make enough water, and he will move on to a new pump. Your task is to collect as much water as possible to help Quentin live in the desert. Once you have understood these instructions, please press the space key to continue.

After reading these instruction participants were randomly allocated to complete either condition first, and a night-time or day-time scene was randomly selected as the initial context for that condition, as was a colour for the displayed pump (see Figure 3.11). Participants could click and drag the handle of the lever, causing water droplets to fall into the bucket (adding to the displayed total) on the basis of the reinforcement schedules outlined in Table 3.6. After

completing all schedules relevant to a given condition participants received the following instructions: “Quentin has decided to search for a new pump. He will drink some water and store the rest, so you will need to fill up the bucket again. Please press the spacebar to continue the task.” When moving to a new condition, the pump colour and initial context were randomly reallocated. For the Experimental condition, context changes occurred according to the sequence described in Table 3.6. When changing context, the following instructions were displayed for five seconds: “Quentin asks you to take a break, Night falls...” or “Quentin asks you to take a break, the sun rises...”, as appropriate for whether the context changed from night to day or vice versa.

Figure 3.11 *Layout of the Extinction Task in Experiment 3*

Note. The movable water pump lever is shown in green, with the record of received reinforcers shown on the bucket. Panel A shows the daytime context and Panel B the night-time context.

Reversal Task. For this task participants were shown the following instructions, taken unchanged from Budhani and Blair (2006, p. 554):

Pairs of animals will appear on the screen. On each go you have to choose one of these animals and the computer will tell you if your choice was correct or wrong. If it is correct you will win 100 points. If it is wrong you will lose 100 points. Each animal will sometimes be correct and sometimes be wrong, but one of the animals will tend to be correct more often than the other one. Find out which animal is usually correct and choose that animal every time. Stick with it even if it is occasionally wrong. At some point it may change so that the other animal is usually correct, in which case you should choose that one every time.

Stimuli in this task consisted of Snodgrass images (Snodgrass & Vanderwart, 1980) over a coloured square sized 146 x 146 pixels. There were 16 possible colours and images were randomly assigned to one of these colours between subjects. Images were also randomly allocated to each of the probability and pair type conditions. On each trial, two stimuli were shown, each randomly allocated to the left or right side of the screen. After stimulus onset participants could press the left or right arrow key to select a stimulus on the corresponding side of the screen, and trial progression was self-paced. After a choice was made feedback appeared (either ‘you win 100 points’ or ‘you lose 100 points’) for 900 ms and a blank screen was shown for an inter-trial interval of 500 ms.

Psychometrics. As described for prior experiments I utilised the ICU, STAI and UPPS to measure constructs of CU, anxiety, and impulsivity traits respectively.

3.7.1.5 *Data Analysis*

To test for effects of CU traits on task performance, I used ICU score as a continuous predictor in mixed effects models. I controlled for covariates (gender, STAI and UPPS scores) as described in Experiment 2, by testing their effects in an initial model and including significant predictors in regressions testing the effect of ICU scores. Participants were excluded if they provided only one response throughout the questionnaires (to avoid participants who did not make a serious attempt to respond to these material), and if they met the exclusion criteria for the reversal task described below. One participant was also excluded as they were identified as an unexplainable outlier in their ICU measure responses. Pre-registration of the planned analyses, as well as the final script utilized for analyses in this study can be viewed at <https://osf.io/nz2sx/>. The code initially registered for this task (FINALANALYSISextinctionDeficitsInCUTraits.R) contains all primary analyses described below. Post-hoc analyses and additional visualisations are contained within the same file as prior analyses (Chapter3Analysis.html).

Extinction Task. Twelve bins were formed for responding in the Training Phase. The duration of these bins varied based on the amount of time it took for participants to complete this phase (mean = 9.19 s, SD = 1.83). Ten bins of 10 seconds duration were created for the Extinction Phase and two 10 second bins for the Renewal Phase.

Extinction Learning. To detect effects of CU traits on extinction I analysed two measures of extinction learning. This procedure followed the approach developed in Experiment 2, where one model examined change in responding across extinction, and the other modelled change between training and extinction. These models were fit to data from the control condition, as this condition did not include context changes, and primary analyses were conducted on data from native English speakers.

To examine differences in responding across phases of the task, a model was constructed for the interaction between ICU scores and phase. This model compared training responding to that in the other, non-reinforced phases: extinction and renewal. This model accounted for within subjects effects with a term for random intercepts of individuals' responding. A further analysis included the full sample and examined whether language background moderated the relationship between ICU scores and extinction, as this effect was observed in Experiment 2. This moderation analysis was deemed as post-hoc as this effect had not been considered during the design of this study (and was therefore not included in the preregistered analysis).

I also analysed change of responding occurring in the extinction phase with mixed effects modelling of responses. This model allowed participant identity to generate random slopes of response rate over the course of extinction. These linear slopes were taken as learning rates, which were then used in a multiple regression to test for an effect of CU traits, whilst treating the intercepts of these models as a covariate.

Renewal Effect. To test for differences in renewal of responding I compared the last two bins of responding in the extinction phase to the bins of the renewal phase across the test and control conditions. This test was conducted for participants with English as a first language. To test for the role of CU traits in this effect I included ICU scores in this interaction, creating a three-way term with phase, condition, and ICU scores.

Reversal Task. Participants were required to have completed eight consecutive correct trials in the acquisition phase of the [.8 – .2] and [1 – 0] reinforcement probability conditions of the reversing pairs to be included in the analysis. This criterion was derived from the methodology of Budhani and Blair (2006). Correct trials were those in which the stimulus with the greatest probability of reward was chosen. Only data from the [.8 – .2] and [.6 – .4]

conditions were analysed as performance in the [1 – 0] condition was too close to ceiling. The relationship between response reversal and CU traits was tested by deriving two measures from responding to the reversing stimulus pairs. Firstly, I analysed the total errors in each phase of each condition, including these two factors in an interaction term with ICU scores. This determined whether the change in behaviour across phases depended on CU traits and whether the probability of reinforcement in each condition moderated such a relationship. Second, I examined the rate at which individuals learned to reverse their behaviour. To achieve this, I analysed the trial-by-trial error data throughout the reversal phase of conditions. I applied a logistic mixed effects model with random slopes over trials at the level of the participant to these binary data. This produced a term for the change in error rate in the reversal phase that I analysed in a regression to determine the effect of CU traits on learning rate.

I also made an exploratory analysis of whether language group influenced the total errors seen in each phase of the task. This was achieved by fitting the same model as described above, with the inclusion of all participants data and a term for the interaction between Phase, language group and ICU scores. As Budhani et al. (2006) treated psychopathic traits categorically when testing for associations with response reversal, an additional analysis tested whether there was a different result for the model of total errors in each phase of the task using median split ICU scores as a predictor (rather than the continuous total ICU score).

3.7.2 Results

Table 3.8 shows the descriptive statistics for all psychometric scales grouped by English language background.

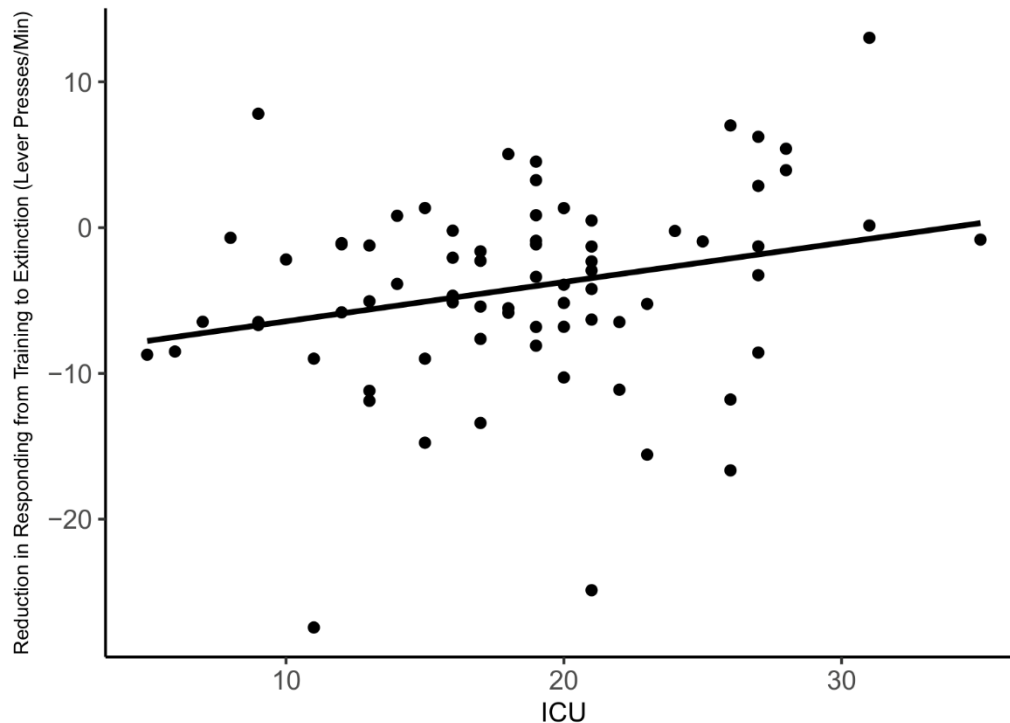
Table 3.8 *Descriptive Statistics for Psychometric Scales, Grouped by English as a First Language*

English	Scale	Mean	SD	Min	Max
Yes	ICU	18.53	6.33	5	35
	STAI	45.60	8.47	26	65
	UPPS	106.56	14.58	61	149
No	ICU	20.35	5.23	11	34
	STAI	47.58	7.93	27	63
	UPPS	102.90	14.26	67	127

3.7.2.1 Extinction Task

Overall Change in Responding. For responding occurring during the control condition (in native English speakers) I saw a significant effect of phase such that less behaviour was seen in the extinction phase ($\beta = -4.13$, 95% CI = $[-5.05, -3.20]$, $p < .001$). This reduction was predicted by STAI scores, with higher scores predicting less reduction in responding ($\beta = 1.34$, 95% CI = $[0.35, 2.33]$, $p = .008$). Controlling for the effect of STAI scores, ICU scores showed a significant effect, such that higher scores predicted less reduction in behaviour between training and extinction ($\beta = 1.69$, 95% CI = $[0.77, 2.61]$, $p < .001$) (Figure 3.12). A post-hoc test showed that responding was significantly higher in the training phase (across both conditions) at higher levels of ICU scores ($\beta = 0.56$, 95% CI = $[0.04, 1.08]$, $p = .034$).

Figure 3.12 *Relationship Between ICU Scores and the Difference in Responding Between Training and Extinction Phases of the Control Condition in the Extinction Task*

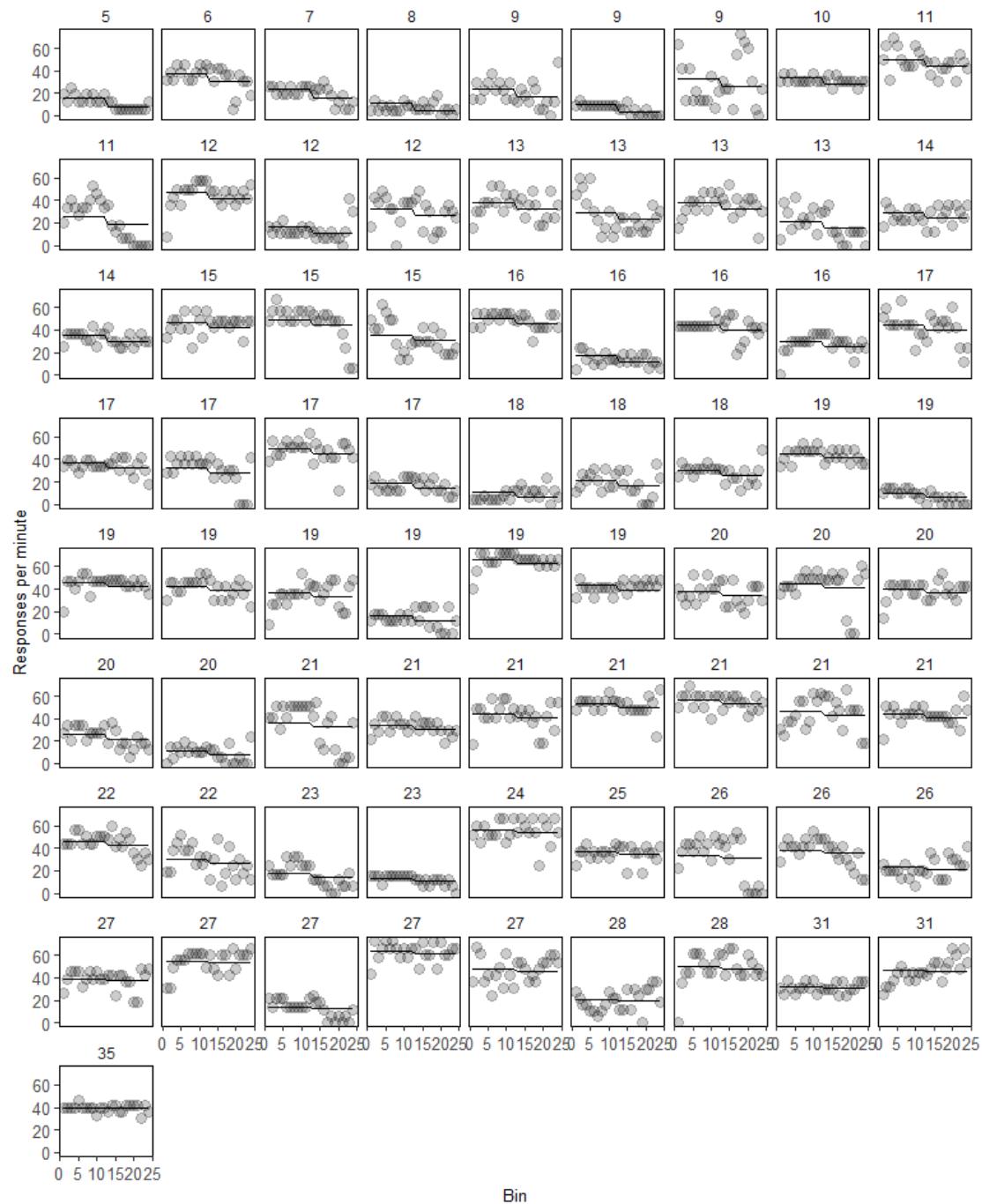


Note. These values are not adjusted for differences in STAI scores and the displayed regression is not directly derived from the statistical model reported. It is intended to illustrate the observed direction of relationship between ICU scores and reduction in responding between.

A further exploratory analysis tested whether language background moderated the effect described in the paragraph above. The analysis indicated no significant covariates and, as in the original analysis, showed a significant reduction in responding between training and extinction ($\beta = -5.37$, 95% CI = $[-6.87, -3.87]$, $p < .001$). However, unlike the original analysis, this reduction was significantly larger as ICU scores increased ($\beta = -2.59$, 95% CI = $[-4.30, -0.88]$, $p = .003$). This discrepancy was due to the role of English language background which moderated the effect of ICU scores ($\beta = 4.22$, 95% CI = $[2.28, 6.15]$, $p < .001$). This effect was such that for participants with English as a first language, greater ICU scores predicted less change in

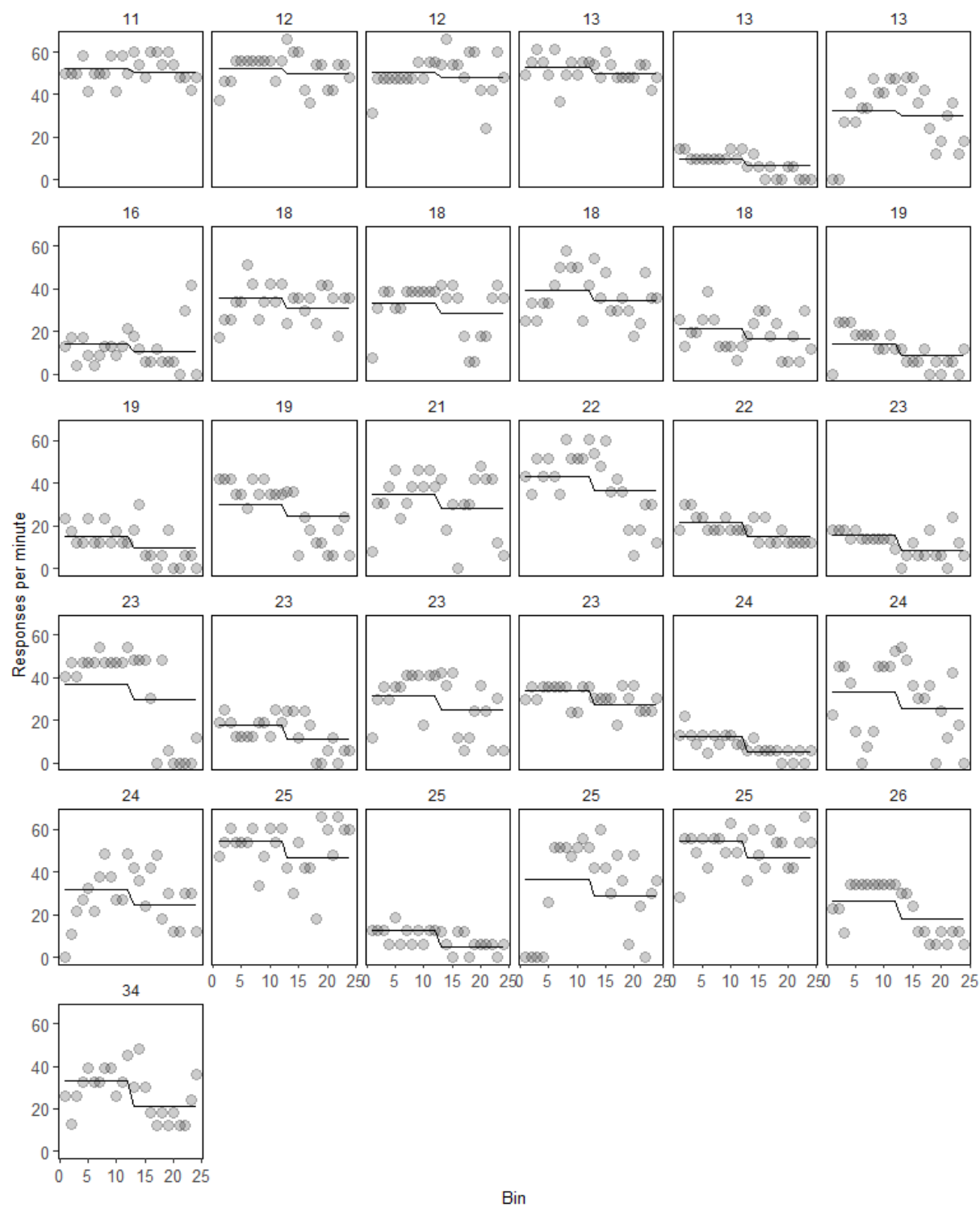
responding between phases, comparatively non-native English speakers exhibited more change in responding between phases across levels of ICU scores (Figures 3.14 and 3.15).

Figure 3.13 *Observed and Predicted Extinction Responses for Individuals with English as a First Language in Experiment 3*



Note. Each observation of response rate in each bin of the Training Phase (bins 1-12) and the Extinction phase (bins 13 to 24) of this experiment is shown as a point. Values predicted by the mixed effect model for each phase are shown as lines. Each plot is labelled by the relevant individual's ICU score. The difference in predicted values between phases can be seen to reduce as ICU scores increase.

Figure 3.14 *Observed and Predicted Extinction Responses for Individuals without English as a First Language in Experiment 3*



Note. See the caption for Figure 3.13 for a description of this plot. The difference in predicted values between phases can be seen to increase as ICU scores increase.

Rate of Change in Responding. I did not find that the rate of reduction of behaviour in the extinction phase was related to covariate measures or ICU scores ($\beta = 0.12$, $CI = -0.09, 0.32$, $p = .268$).

Renewal. In my analysis of the renewal effect, I saw that a greater increase in behaviour occurred after the renewal phase in the experimental condition than in the control condition ($\beta = 13.6$, $CI = 1.15, 26.06$, $p = .032$). However, ICU scores did not interact significantly with this effect ($\beta = -0.50$, $CI = -1.13, 0.19$, $p = .126$), or covariate measures.

3.7.2.2 *Reversal Task*

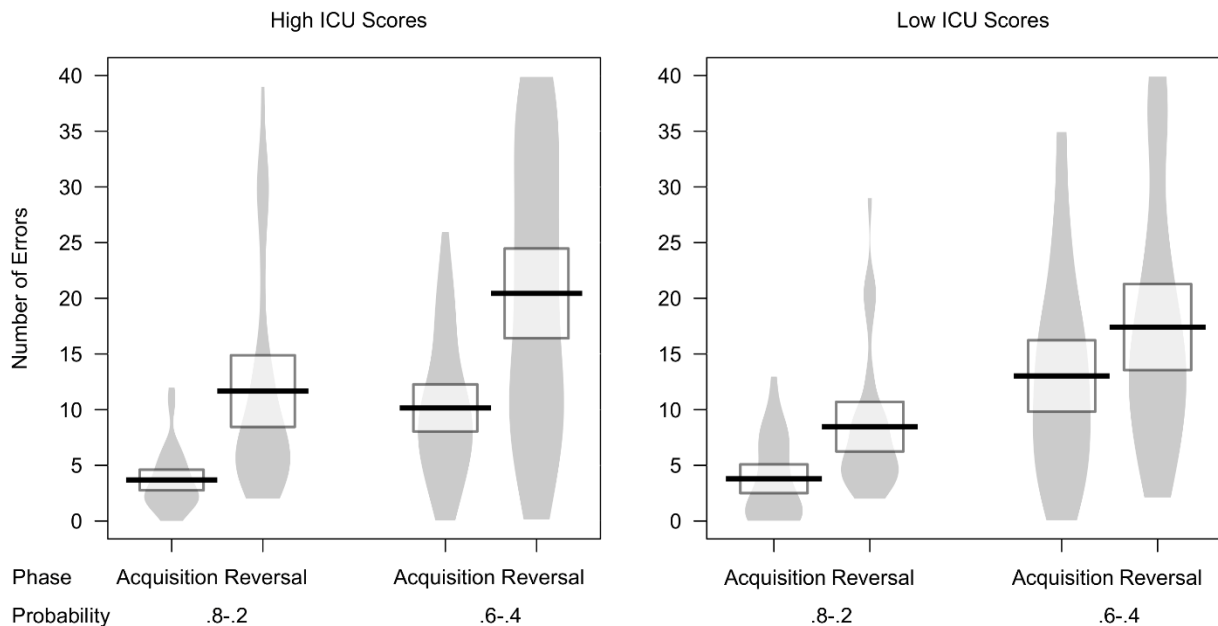
There were significantly more errors in the reversal phase of the [.8 – .2] and [.6 – .4] conditions than in the acquisition phase ($\beta = 6.99$, $CI = 5.04, 8.94$, $p < .001$). There were also more errors overall in the [.6 – .4] condition than in the [.8 – .2] condition ($\beta = 5.48$, $CI = 3.53, 7.43$, $p < .001$). However, this probability variable was not seen to interact with phase ($\beta = 0.77$, $CI = -1.98, 3.53$, $p = .580$). None of the tested covariates were found to be significant. ICU score did not moderate the increase in errors in the reversal phase, although this effect approached significance (for a greater increase at higher levels of ICU score) ($\beta = 1.90$, $CI = -0.05, 3.86$, $p = .056$) (see Figure 3.15). Linear change in error rate throughout reversal was also not seen to depend on ICU scores ($\beta = -0.01$, $CI = -0.02, 0.00$, $p = .057$), although again the effect approached significance.

Further analysis of the reversal task was carried out, which included non-native English speakers. This analysis showed a significant increase in errors between phases ($\beta = 6.35$, $CI = 3.32, 9.38$, $p < .001$) which was not predicted by any covariates, or ICU scores ($\beta = 0.16$, $CI = -3.29, 3.60$, $p = .929$). The interaction between ICU scores and English language background was

also not predictive of any variance in the increase in errors between phases ($\beta = 1.67$, $CI = -2.24$, 5.58 , $p = .402$), nor was English language background itself ($\beta = 0.80$, $CI = -2.79$, 4.39 , $p = .66$).

I also carried out a post-hoc analysis of reversal task behaviour using a median split of ICU scores (creating High ($ICU \geq 19$) and Low ($ICU < 19$) groups). There was no evidence that changes in error rates between acquisition and reversal phases were moderated by membership to High or Low ICU groups ($\beta = -1.93$, $CI = -5.90$, 2.04 , $p = .339$).

Figure 3.15 *Number of Errors in Each Phase and Probability Condition of the Reversal Task, Split by ICU Scores*



Note. Bars and surrounding boxes show the mean $\pm$ 95% CI, superimposed on rotated probability density plots in grey. Note that ICU scores were not treated categorically in the analyses and are displayed as such here for illustrative purposes (using a median split to define high and low ICU scores).

3.7.3 Discussion

These results fulfilled the aim of this study to replicate extinction and renewal effects in a simplified version of Experiment 2. The extinction task revealed less overall reduction in

responding during the extinction phase at higher levels of CU traits, as hypothesised. This was a similar result to that seen in my prior studies. Furthermore, I again did not see that CU traits predicted the rate of behavioural change during extinction, adding to my observations that this measure was less sensitive to individual differences in extinction learning. I was also able to replicate a renewal effect in this task, indicating that my novel design may be an effective tool for measuring relapse processes. Although I did not observe a statistical relationship between renewal and CU traits, this experiment provided an estimate of the magnitude of this effect. This effect was in the direction hypothesized where higher levels of CU traits predict lower levels of renewal. This effect size can therefore be used to estimate appropriate sample sizes for testing the reliability of this effect.

An additional aim of this experiment was to demonstrate that I was measuring the same learning deficit as Budhani et al. (Budhani et al., 2006) by including a replication of their response reversal task. The results of this task were approaching significance in the direction of a CU associated reversal deficit. This marginal effect may be accounted for by the distribution of CU traits in this sample. Only a small proportion of subjects (7%) reached a clinically relevant cut-off of 28 defined by recent research (Docherty et al., 2017). Comparatively, approximately half of the sample used by Budhani et al. (2006) met the equivalent cut-off. The more moderate range of traits in this study may mean that more statistical power is required to achieve a reliable observation of the reversal learning deficit effect. Alternatively, a host of other unmeasured characteristics may have varied between the current sample of primarily female undergraduates and the sample described in Budhani et al. (2006) who were incarcerated males (for example: executive function and motivation for participation). Another difference between the current study and Budhani et al. (2006) was that these authors used a categorical analysis of CU traits.

However, in a post-hoc categorical analysis I did not see any evidence of a reversal deficit, therefore this methodological difference is unlikely to account for the absence of the effect in this study. These findings indicate that the examination of this task could not provide useful information for the nature of learning in this sample.

Another, unexpected, feature of the sample in this study was the discrepant findings for non-native English speakers. Here I observed a similar result to that seen in Experiment 2 where the effect of CU traits on extinction was reversed for these participants; with higher CU traits leading to greater reductions in responding. This was not due to any differences in levels of CU traits between the English language groups. However, in Experiment 2 these participants were less likely to resume responding when the response outcome was reintroduced. This finding may suggest that these participants had a different interpretation of the task. Supporting this interpretation, I saw no systematic effects of English language history on the reversal task, which was a much simpler procedure (it had no narrative content and gave direct feedback for each response). This study was not designed to differentiate between effects occurring due to comprehension of the task materials, and those caused by variations in the validity of the ICU between these groups. Future studies would need to include measures designed to tease apart these effects in order to explain the causes of my observations here.

3.8 General Discussion

In this series of experiments, I was able to reliably observe responding consistent with extinction learning. In the three pilot studies this was observed as greater rates of reduction in responding during periods in which responding was not reinforced, compared to periods in which rewards were available. These results suggested that my novel experimental design could be used as a measure of extinction learning. In the subsequent three experiments I observed that the change in responding between training and extinction was a more useful measure of individual differences in extinction learning than rates of change in responding occurring during extinction. As hypothesized this overall reduction in responding was found to be less pronounced in individuals with higher levels of CU traits. This result therefore fulfills my aim of supporting historical findings which have linked PT and CU traits to deficits in extinction learning (reviewed in Chapter 2). However, this result was not corroborated by renewal and reversal effects (it should be noted that due to analyses utilised here, the absence of these effects does not constitute evidence against the extinction hypothesis examined in this study (Masson, 2011)).

Although the extinction behaviour in this study was consistent with my extinction learning hypothesis over multiple samples and variations of the task, there remained questions regarding alternative explanations for differences in responding in the absence of reward. I was evident that participants did not cease responding altogether under extinction conditions in any variation of the task. This observation may indicate that participants' persistence on the task could account for the observed effect, rather than differences in extinction learning. In mitigation of this potential confound, I was able to control for impulsive traits using the UPPS measure. Furthermore, the methodology of this study accounted for individual differences in overall

response rate. Considering these features of the study, differences in overall motivation to respond to the task was not a satisfactory explanation for the present extinction results.

Other measures were also implemented to minimise the impacts of uncontrolled effects on responding. Specifically, I designed several methodologies intended to reduce responding in the absence of reward expectation. These efforts were warranted as the interpretation of this task as a measure of extinction learning assumes that responses are contingent on reward expectation. However, none of these methodological changes were able to improve the contingency between reward and response. These factors may have particularly affected the statistical models I used to estimate change in responding over the course of extinction, which did not yield significant associations with CU traits. In contrast, overall differences between training and extinction appeared to capture more meaningful variation. This analysis was introduced in Experiment 2 in order to capture variation occurring between the end of training and the start of extinction, which was not considered in the models initially developed for behaviour in this task. However, it may have been possible to improve the efficiency and error rate of these analyses by creating a single model for both these effects.

Tests for renewal in this study were intended as a means of discriminating behaviour attributable to extinction learning from that caused by other processes. I predicted that CU traits would be associated with reductions in renewal. Therefore, any increases in overall responding on the task will not be able to account for both higher responding in extinction and lower responding in renewal. This study demonstrated a methodology capable of detecting a renewal effect (as seen in two separate samples). Additionally, the association between renewal and CU traits was in the predicted direction. However, it was not statistically reliable, despite the presence of hypothesised extinction effects. This may have been because changes in context

elicited less consistent behavioural responses than omission of the reward, which could be viewed as a more salient cue for responding. Therefore, the observations made in this series of experiments can be used to design a sample with the statistical power to make a meaningful conclusion about this hypothesis.

A further procedure designed to validate the current methodology as a measure of extinction learning was the implementation of a reversal task. This task, described by Budhani and Blair (2006), has been interpreted as evidence for learning processes in major aetiologies of CU traits (Moul et al., 2012). Utilisation of this task was therefore a means of tying the features of CU traits measured in the current study to those addressed in CU trait aetiologies. However, the results of this task in this study were very unclear. Whilst CU traits did not predict deficits in reversal performance, this relationship approached significance. As discussed previously, it is not clear whether sample characteristics or a lack of statistical power accounted for this result. An alternative means of linking extinction and reversal behaviour in this study would have been to directly correlate them. However, such an analysis would not have clarified whether any variance shared by these tasks was relevant to CU traits.

An unexpected outcome of these studies were the repeated findings that language background determined the relationship between CU traits and extinction behaviour. This study did not measure sufficient information about the individuals with and without English as a first language to be able to confirm causes for this difference. It is likely that non-native English speakers recruited to this study were predominantly East Asian, due to the demographic make-up of the international student population in Australia (Department of Education Skills and Employment - Australian Government, n.d.). Therefore, language comprehension may have caused differences observed for this group, as poor confidence in English has been identified in

this population (Sawir, 2005). Alternatively, the structure or validity of the ICU may have varied between these groups, potentially due to cultural factors. Allen et al. (2020) have observed that the structure of the ICU varies over English and Chinese samples. Therefore, factor analyses of the ICU measure and more detailed description of participants' English language capacity and cultural background is necessary to understand cultural influences on the measures used in this study.

In addition to between subject variation in responses to the task developed here, there may also be variation within individuals across time. To estimate this intra-individual variation (or task reliability) future studies could employ a test-retest methodology. Resulting reliability estimates would determine whether this task is measuring a stable psychological characteristic (or "trait") that can validly be correlated with other traits (in this case CU traits). Such validation would improve the utility of the task for future research.

Overall, this series of experiments has consistently demonstrated variation in extinction behaviour occurring across levels of CU traits in a non-clinical, non-forensic population. This finding supports a growing body of evidence showing that the processes involved in psychopathic traits and CU traits can be investigated in readily accessible populations (X. Chen et al. 2020). Furthermore, it is the first such finding using the ICU operationalisation of CU traits, and therefore strengthens suggestions that learning deficits are specifically relevant to the affective features of psychopathic traits (Blair, 2017). Although this finding was not supported by corresponding deficits in renewal, the present experimental design showed promise as a means of measuring this effect. This study therefore demonstrates a useful paradigm for testing the relationship between CU traits and extinguished learning processes.

4 Modelling and Power Analysis

Contents

4.0	Methods.....	135
4.0.1	Data.....	135
4.0.2	Modelling Extinction Task Behaviour	136
4.0.3	Power Analysis.....	138
4.1	Results.....	139
4.1.1	Extinction.....	139
4.1.2	Renewal.....	143
4.2	Discussion.....	148

The discussion of Chapter 3 indicated the need for further analytical investigation of several results. Primarily, as the sample collected in Experiment 3 did not show any relationship between ICU scores and renewal of extinguished behaviour, it was unclear whether this was due a population level absence of such an effect, or insufficient statistical power to detect it. This chapter therefore investigated the statistical power of that analysis and estimated a sample size likely to allow observation of that renewal effect. A secondary issue observed in Chapter 3 was that two models were used to describe extinction behaviour, and they did not have consistent results. This issue was addressed by creating a single model to capture the differences previously described in separate models.

In Chapter 3, two analyses were used to model extinction learning effects. One of these considered changes in responding between training and extinction phases of the task across levels of ICU scores. The other fit linear functions to individuals' responding over time during extinction and correlated these slope coefficients with ICU scores. This analysis of rate of change in responding was designed to capture individual variation in extinction learning. However, it was evident that there were meaningful changes in behaviour between the last bin of training and the first bin of extinction, which were not captured in this model. These immediate effects of reward omission appeared to make the overall reduction in responding between phases a more useful measure of extinction.

This modelling approach was unsatisfactory as it portioned information from one set of responses into two models. This portioning limited the information analysed by each model, as well as unnecessarily inflating the probability of observing type 1 errors (by adding hypothesis tests to the analysis), therefore limiting the statistical power of these models. Accordingly, in this chapter I aimed to develop a model which was sensitive to both the initial change in responding

between training and extinction phases, as well as reductions in behaviour throughout the extinction phase.

In addition to this consideration of extinction modelling, the present study addressed renewal effects observed in Experiment 3 of Chapter 3. These effects were not found to vary with individuals' ICU scores. However, estimated model parameters were in the hypothesised direction, where high ICU scores predicted less renewal of extinguished responding. In the prior Chapter I suggested that the lack of statistical significance in this effect may have been the result of inadequate statistical power. This was explicable as the renewal effect was a response to changes in task context, which may have been less salient than the omission of reward which generated extinction effects.

The current study therefore aimed to statistically determine the sample size that would provide a strong probability of detecting the renewal and extinction effects observed in my prior research. This power analysis used model parameters estimated from the data reported in Experiment 3 of Chapter 3. Specifically, I utilised data from native English speakers as only this group displayed the hypothesised extinction effects in that study. As the ICU has only been validated for undergraduates in a western (English speaking) context (Kimonis et al., 2013), variation in language background may have caused uncontrolled differences in interpretation of this scale. Utilisation of this data from native English speakers therefore offered the most valid means available for modelling the relationship between ICU scores and renewal, and thereby informing appropriate sampling strategies. It was crucial to justify this methodological parameter in order for subsequent research to successfully address the aims of this thesis.

4.0 Methods

Scripts developed for the following analysis can be viewed on the preregistration for the study utilising this power analysis (described in Chapter 5 and 6) <https://osf.io/tfjqc/> (see PowerAnalysis_editsForThesis.html). All relevant analyses and output are described in detail in the relevant markdown document¹ associated with that registration.

4.0.1 Data

The present study utilised data from the task and questionnaires described in Chapter 3, Experiment 3. The structure of that task is shown in Table 4.1. This data set also included responses to ICU, STAI, and UPPS measures. Only data from participants with English as a first language was utilised. Exclusion criteria based on the reversal task in Chapter 3 were not employed, to allow for generalisation of findings to studies unable to measure and apply such criteria. However, any participants who gave a single response to all questionnaire items were excluded here as in prior analyses. This led to a sample of 80 participants, 59 of these were female, 21 were Male, and their median age was 19 (18 – 20 interquartile range). Mean total ICU score was 18 (SD = 7), mean UPPS score was 107 (SD = 15), and mean STAI score was 46 (SD = 9).

¹ PowerAnalysis_editsForThesis.html

Table 4.1 *Reinforcement Schedules and Contexts for each Phase and Condition in the Extinction Task*

Condition		Training	Extinction phase	Renewal phase
Experimental	Reinforcement	4	Ø	Ø
	Context	A	B	A
Control	Reinforcement	4	Ø	Ø
	Context	A	A	A

Note. This table is taken from Chapter 3, Experiment 3. All numbers refer to the variable interval (VI) schedules (as number of seconds), and Ø indicates no reinforcement. Contexts A and B were each one of either night-time or daytime settings (visualized in Figure 3.11)

4.0.2 Modelling Extinction Task Behaviour

The following models were specified as mixed effects models (fit using lme4 (Bates et al., 2015)). As in prior analyses, within-subjects effects were accounted for through a random effects term for each individual participants' intercept of responding. The utility of these models was considered by visual inspection of their predictions for individual's behaviour as well as for the aggregate. As described in Chapter 3, the dependence of effects of interests on potential covariates (gender, STAI scores and UPPS scores) was tested using by testing the interaction between these variables and model terms describing extinction or renewal effects. If any covariates were found to moderate these terms, they were included in the test of ICU scores in order to account for their effects. Assumptions of residual normality were checked by examining qqplots for large departure from expected values. Similarly, homoscedasticity was checked using scatter plots of residuals against fitted values. The effect of outliers on the model was examined by excluding individuals whose response sets resulted in large cook's distance statistics. Any notable impact of this exclusion on model results was reported here.

4.0.2.1 Extinction

As for analyses in Chapter 3, extinction effects were modelled using response rate data from the training phase (12 bins), extinction phase (10 bins), and renewal phase (two bins) of the

Control Condition. The renewal phase was treated as an additional two bins of the extinction phase, as no context changes occurred in this condition. The novel model considered in the present analysis differed in structure to these previous analyses in that it was designed to detect changes in response rates between training and extinction as well as across extinction. It included a term that represented the average change in responding across bins of this data, dependent on whether responding occurred in the training or extinction phase. This term represented the average response across the structure of the extinction test. The effect of individual difference variables on extinction was tested through the interaction of these variables with a term comparing mean responding during training to the change in behaviour occurring during extinction. This term did not consider any changes in behaviour within the training phase because such changes were not relevant to the extinction learning effect of interest.

4.0.2.2 Renewal

Modelling of the renewal effect was consistent with Chapter 3. The model was fit to data from the final two bins of extinction as well as the two bins of renewal in both control and experimental conditions. It tested the difference between conditions, to control for overall differences in responding resulting from the occurrence of context changes in the Experimental Condition. The model also tested for the interaction between individual difference factors and changes in responding between extinction and renewal, as dependent on the task condition. This approach isolated the effects of changes in context occurring during the Experimental Conditional from the general change in behaviour occurring over bins of the Control Condition (due to the absence of reinforcement).

4.0.3 Power Analysis

In order to estimate the appropriate sample sizes required to have a justifiable chance of detecting the parameters of the models described above, I applied the methodology described by DeBruine & Barr (2021). This analysis involved simulating individual sets of responses relevant to extinction and renewal effects. Concordantly, for the power estimation of the extinction effect, a full set of responses was generated for the 12 training bins and 12 extinction bins. For the renewal effect, responses were generated for the final two bins of extinction and the two bins of renewal phases in both experimental and control conditions. To achieve this, each simulated participant received a score on the ICU (and any relevant covariate measures) generated from a normal distribution (or multivariate normal distribution if covariates were relevant) estimated from observed data. Each individual also received a random intercept value for their responding generated from a normal distribution with mean zero and variance value derived from the relevant model. Responses were then generated by entering these values into the specified model equation, which included fixed effects terms for extinction or renewal effects as well as their interaction with ICU traits and covariates where relevant. Any predicted values less than zero were set to zero in order to match the observable data for the extinction task.

Using this process, samples of N simulated respondents were generated and the relevant extinction or renewal model was then fit to these simulated data sets. For a given N this process was repeated one thousand times and the proportion of observed p values less than an alpha of .05 was calculated. This proportion represented the estimated power for that sample size. The criteria utilised in order to determine adequate statistical power of a sample size was .8 (Cohen, 1977). For the effect of ICU scores on renewal, N s of 200 to 500 were tested in increments of 50. Upon the sample which produced a power approximating .8, five additional sample sizes were

tested in increments of five from that approximate value. This provided a more specific sample size estimate. I expected the extinction effect to have higher observed power (given that a significant relationship between extinction and ICU scores was seen in section 3.7.2.1). Therefore, values of N ranging between 25 and 175 (in increments of 50) were tested for the extinction effect.

In addition to testing these sample sizes, differing sizes of the interaction between ICU and renewal were tested. This provided an indication of sensitivity of power estimates to changes in the size of this effect. The tested values included the lower limit of the 95% confidence interval for this effect. As the coefficient for this term was negative this value represented the largest expected effect. Additionally, as the term was not statistically significant the upper confidence interval value was greater than zero, which would be inconsistent with the hypothesised effect. Therefore, instead of testing the upper limit of the confidence interval, a value of -1 was chosen, which was between the observed effect and zero.

4.1 Results

4.1.1 Extinction

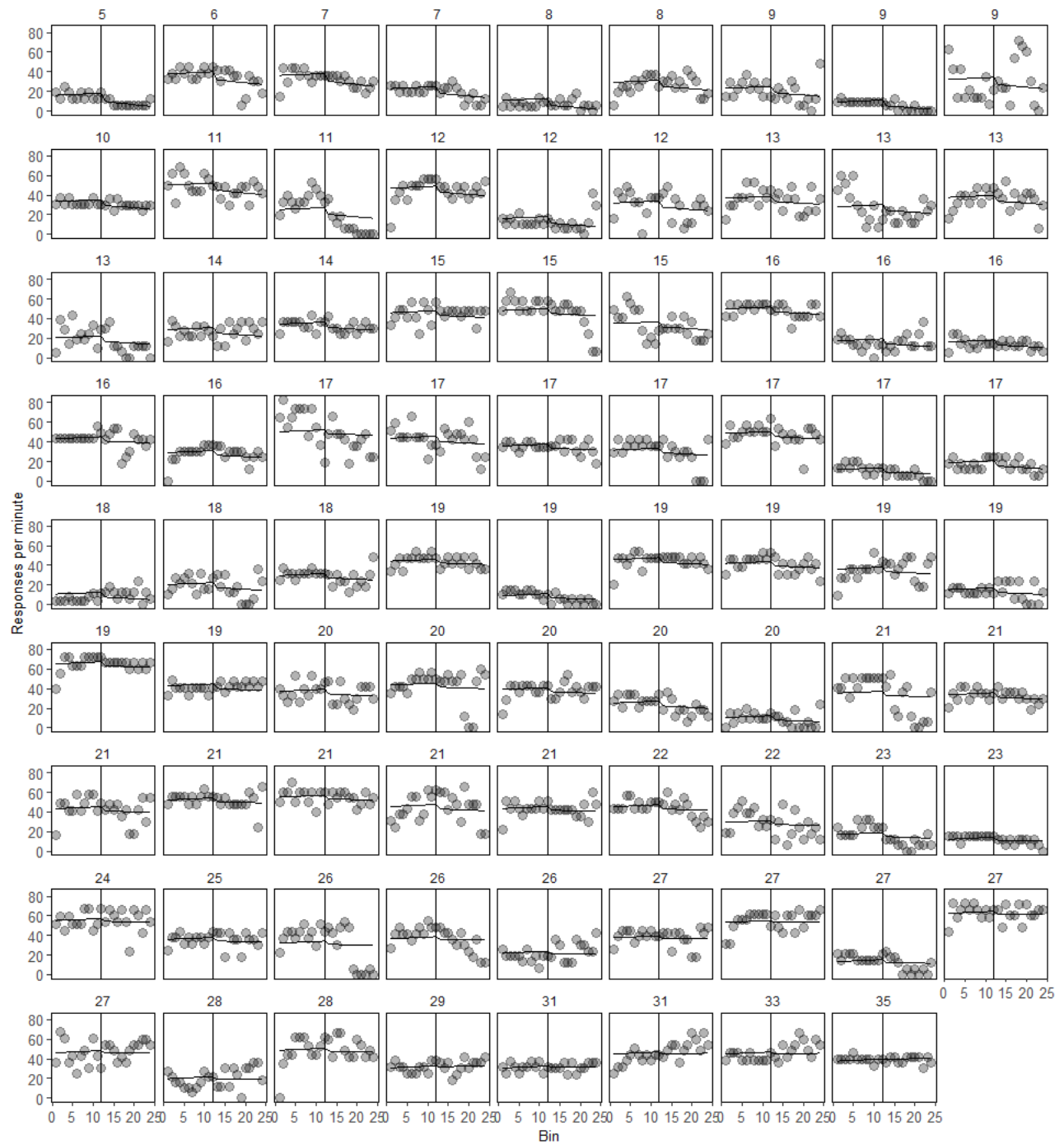
The extinction model was successfully fit to the data, showing no violations of model assumptions. Statistics produced by the model are shown in Table 4.2. The model showed that change in responding over bins was significantly more negative during the extinction period than during training. In the covariate analysis higher STAI scores were significantly predictive of less reduction in behaviour during the extinction phase. However, this effect was not maintained when modelled with the extinction effect of ICU scores. Higher ICU scores were predictive of significantly less reduction in behaviour during extinction. At the individual level the model predicted responding as depicted in Figure 4.1. At the sample aggregate level, the model

predicted responses shown in Figure 4.2 for minimum and maximum ICU scores. Simulation of responses using this model indicated that a sample size of between 25 and 75 would have a power greater than .8 (Figure 4.3).

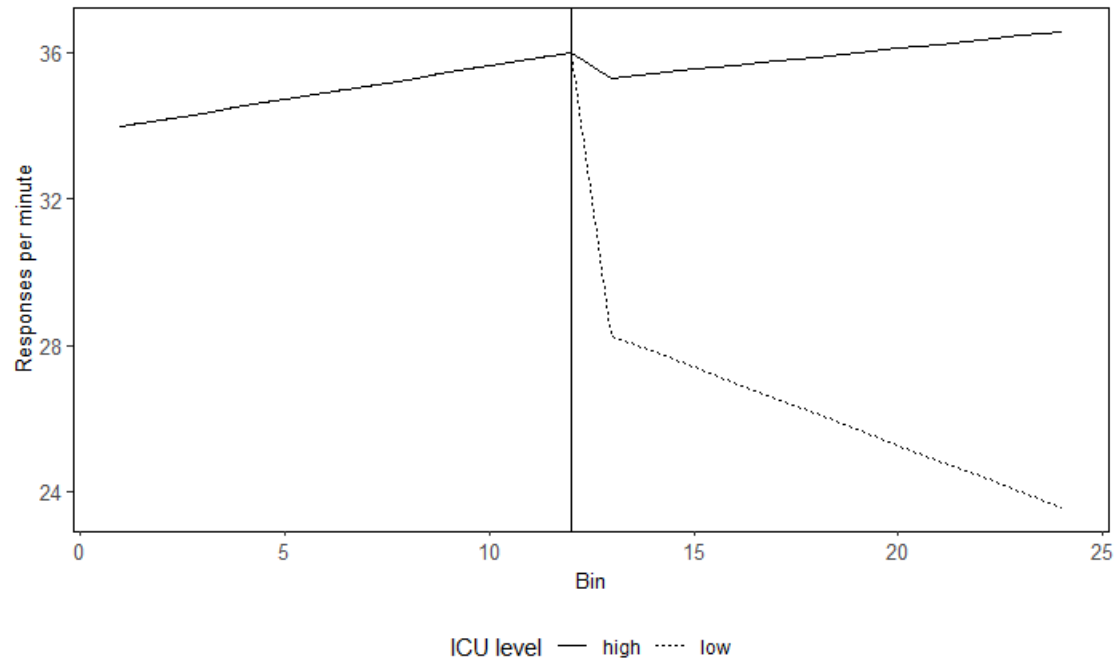
Table 4.2 *Extinction Model Statistics*

Predictors	covariate test			ICU test		
	Estimates	CI	p	Estimates	CI	p
(Intercept)	33.89	30.79 – 37.00	<0.001	33.81	30.77 – 36.85	<0.001
bin * phase	-0.34	-0.42 – -0.26	<0.001	-0.37	-0.43 – -0.30	<0.001
bin * extinction	0.05	0.00 – 0.10	0.047	0.04	-0.01 – 0.08	0.127
phase * STAI						
bin * extinction	-0.07	-0.18 – 0.04	0.187			
phase * gender						
bin * extinction	-0.01	-0.06 – 0.03	0.582			
phase* UPPS						
bin * extinction				0.12	0.07 – 0.17	<0.001
phase * ICU						
Random Effects						
σ^2		96.29			95.15	
τ_{00}		194.78 _{participant}			186.96 _{participant}	
ICC		0.67			0.66	
N		80 _{participant}			80 _{participant}	
Observations		1920			1920	
Marginal R^2 /		0.020 / 0.676			0.029 / 0.672	
Conditional R^2						

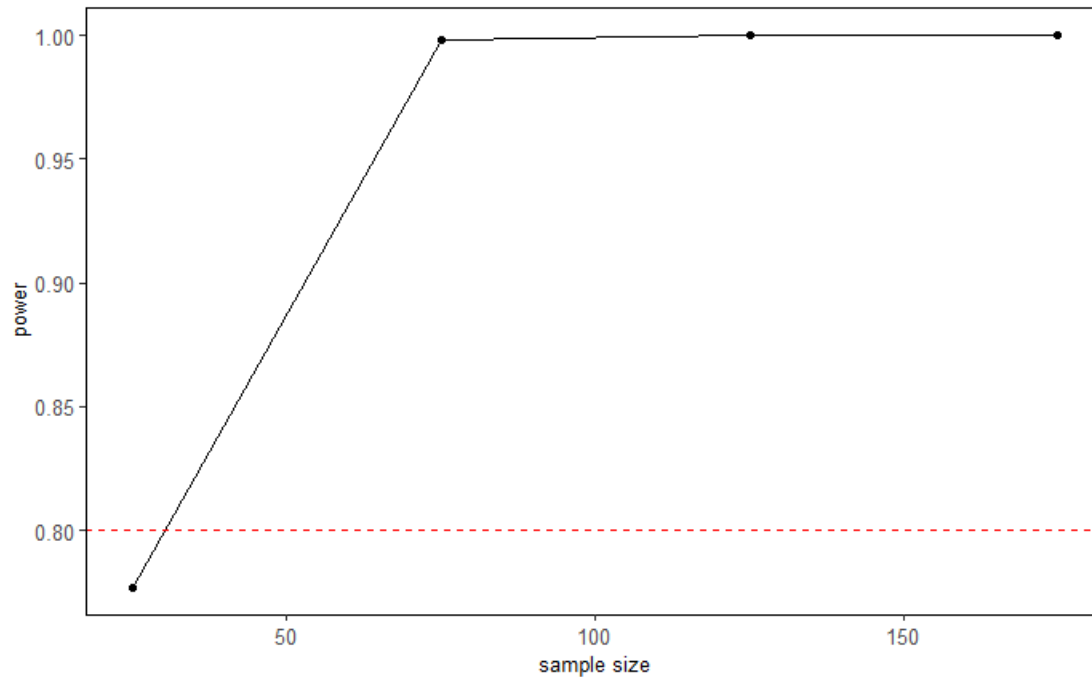
Note. CI = 95% confidence interval. Marginal R^2 = variance accounted for by the fixed effects in the model, conditional R^2 = variance accounted for by entire model, σ^2 = model error term, τ_{00} = variance of random intercepts, ICC = intraclass correlation. The random effects term hierarchized data at the participant level of analysis.

Figure 4.1 *Individual Level Model Predictions for Extinction Behaviour*

Note. Observed response rates of each individual shown as shaded dots, predicted response rates shown as lines. Vertical line delineates training and extinction phases. Plots are labelled with individuals' ICU scores.

Figure 4.2 *Aggregate Level Extinction Model Predictions Across ICU scores*

Note. High and low ICU values are minimum and maximum observed values. Predicted values are generated from the “ICU test” model reported in Table 2, with the value of STAI set to 0 (mean). Vertical line delineates training and extinction phases.

Figure 4.3 *Power Estimates for Different Sample Sizes Using Simulation of the Extinction Model*

Note. Power estimates derived from simulation of interaction effect between bin, extinction phase and ICU scores in the “ICU test” model shown in Table 4.2.

4.1.2 Renewal

The renewal model was fit to the data as specified in the methodology and all statistical assumptions were met. Output from this model is shown in Table 4.3. When testing for covariates the renewal term of the model was not significant, nor it was not statistically dependent on levels of STAI, gender or UPPS. In the ICU test the renewal term was significant indicating a greater change in responding between extinction and renewal phases during the Experimental Condition compared to the Control Condition. This effect did not vary significantly over levels of the ICU. Individual level fit of this model is visualised in Figure 4.4 and the aggregate predictions of the model are visualised in Figure 4.5.

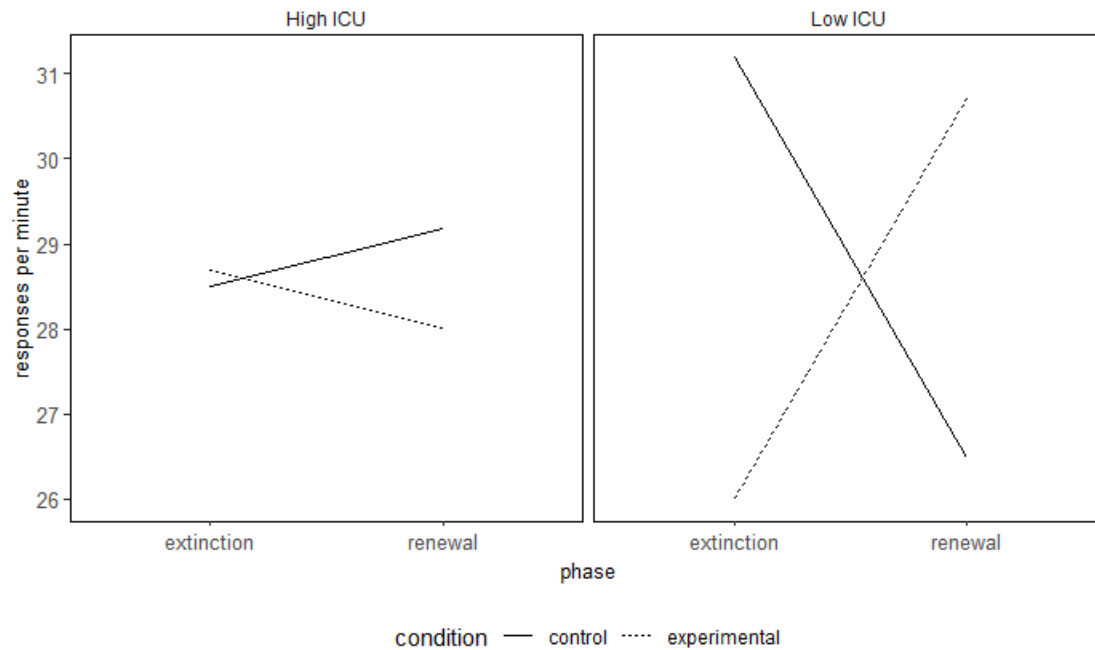
Table 4.3 *Renewal Model Statistics*

Predictors	Covariate Test			ICU Test		
	Estimates	CI	p	Estimates	CI	p
(Intercept)	28.59	25.25 – 31.94	<0.001	28.59	25.25 – 31.94	<0.001
condition code	-0.49	-2.43 – 1.45	0.621	-0.49	-2.42 – 1.45	0.621
condition * phase	3.60	-0.87 – 8.07	0.115	4.57	0.71 – 8.44	0.020
condition * phase * STAI	0.17	-3.96 – 4.30	0.936			
condition * phase * gender	4.12	-5.24 – 13.48	0.388			
condition * phase * UPPS	-0.65	-4.58 – 3.28	0.747			
condition * phase * ICU				-2.37	-6.26 – 1.52	0.232
Random Effects						
σ^2		155.83			155.13	
τ_{00}		212.82 _{participant}			212.90 _{participant}	
ICC		0.58			0.58	
N		80 _{participant}			80 _{participant}	
Observations		640			640	
Marginal R^2 /		0.004 / 0.579			0.005 / 0.580	
Conditional R^2						

Note. as for Table 4.2.

Figure 4.4 *Renewal Model's Predictions of Individuals Responses*

Note. E9, E10 = modelled extinction bins; R1, R2 = modelled renewal bins. Coloured dots are observed values, whilst lines are values predicted by the “ICU test” model (Table 4.3). Plots are labelled with individual’s ICU scores.

Figure 4.5 *Renewal Model's Aggregate Prediction Across ICU Scores*

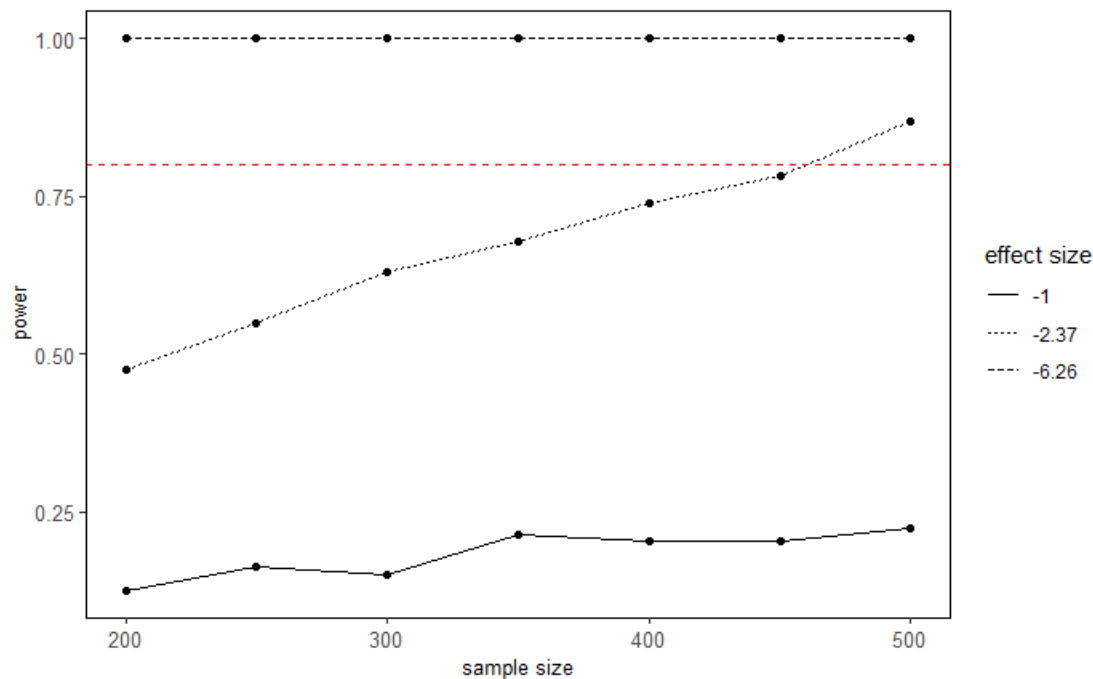
Note. Predicted values derived from the “ICU test” model in Table 3. High ICU is the maximum observed value, low ICU the minimum.

Simulation of the renewal model indicated that the first sample size to reach the criteria of .8 power for the observed value of the ICU interaction with renewal (-2.37) was that of $N = 465$ (Table 4.4). This estimate appeared to be highly sensitive to the size of the effect size used in its calculation. At the lower confidence interval bound for the size of this effect (-6.26) it was statistically significant for 100% of simulated samples of all tested values of N (200 – 500). However, if the coefficient for this effect was set at -1 it was not statistically observable at any tested sample sizes (Figure 4.6).

Table 4.4 *Power Analysis for the Renewal and ICU Score Relationship*

sample size	mean estimate	mean SE	power
450	-2.203	0.789	0.799
455	-2.188	0.784	0.774
460	-2.167	0.780	0.772
465	-2.199	0.776	0.813
470	-2.203	0.772	0.800
475	-2.208	0.768	0.815

Note. Mean estimate is the average simulated estimate for the interaction between condition, phase and ICU score (which was reported for observed data in Table 4.3). Mean SE is the average standard error of simulations of this effect. Power is the proportion of p values less than .05 in these simulations.

Figure 4.6 *Power Values for Simulations of the Effect of ICU scores on Renewal*

Note. Effect size values are simulated estimates of the interaction between condition, phase and ICU scores reported in the “ICU test” model in Table 3. The horizontal line corresponds to power of .8.

4.2 Discussion

Through this study I was able to achieve my aim of improving on the statistical models of extinction behaviour described in Chapter 3. This was the case as the model tested here could detect individual differences in the change in behaviour occurring between training and extinction, as well as across the course of extinction. This can be seen in Figure 4.2 where model predictions for low ICU individuals involved an initial decrease in behaviour between training and extinction, as well as ongoing reductions throughout extinction. Contrastingly, predicted behaviour for high ICU individuals was relatively constant between phases and across extinction. In predicting behaviour in this manner, this model demonstrated the capacity to avoid the issues caused by using separate models to test change in responses between phases and change in responding throughout extinction. Namely, potential increases in type one error, and the loss of statistical power caused by portioning statistical information into separate models.

The present method of modelling extinction behaviour exhibited statistically significant dependence of the extinction effect on individual ICU scores. This was consistent with the results of Chapter 3, as well as the hypothesis that increased CU traits are linked to extinction learning deficits. Furthermore, the present analysis demonstrates that this effect had high statistical power in simulated samples of 50 individuals. This result therefore reinforces that this extinction task represents a convenient, short duration procedure that can be used in small samples to make reliable observations of individual differences in extinction behaviour in future research.

I was also able to determine an appropriate sample size for testing the relationship between ICU scores and renewal effects. The renewal model fit here showed that a renewal effect was present in the current data. However, it did not meet the statistical criteria necessary for stating that this effect was dependent on individual levels of CU traits. This result was also

described in Experiment 3 of Chapter 3. Simulations of this effect showed that, based on the observed model parameters, a sample of 465 participants would have an 81% likelihood of meeting those statistical criteria. It was therefore unlikely that this effect would be observable in the sample size utilised in Experiment 3 of Chapter 3. Of course, it should be noted that this power analysis assumes the representativeness of the estimates produced in the renewal model, and such effects may not be present in the population.

Several other limitations in the renewal power analysis were notable. Figure 4.5 highlighted that the model coefficient for the relationship between ICU scores and renewal captured differences between responding in control and experimental conditions during the extinction phase (specifically in individuals with low ICU scores). Such differences were not anticipated, and it is unclear from this analysis whether this effect is reliable, or an artifact of the model specification. If this difference inflated the estimated renewal effect, then the present power analysis may have provided an underestimate.

A further factor that may limit the utility of this power analyses is the sensitivity of the model's power estimates to changes in the observed effect size. It was seen that the values within the 95% confidence interval for this effect changed the power estimate dramatically. Whilst any increase in the magnitude of the effect would lead to very high statistical power, reductions in its magnitude would make it unlikely to be statistically detected in a sample of 465 individuals. If this effect cannot be observed in a sample of nearly 500 participants it is unlikely that this methodology will be a useful means of measuring it, if it exists.

The present results indicated an improved model specification for the description of the effects of ICU scores on extinction behaviour reported in Chapter 3. Furthermore, they showed that this effect had high statistical power in small samples. Contrastingly, analysis of the renewal

effect indicates that a minimum of 465 participants was likely to be required in order to confirm the presence of an association between ICU scores and renewal effects. This analysis provided crucial justification for the sampling strategies of further studies attempting to address the aims of this thesis.

5 Cross Cultural Properties of the Inventory of Callous Unemotional Traits

Contents

5.0.1	Development of the ICU	153
5.0.2	Predictive validity of the ICU	155
5.0.3	Cross-Cultural Studies of the ICU	156
5.0.4	CU Traits and Collectivism.....	157
5.0.5	Aims and Hypotheses	158
5.1	Methods.....	160
5.1.1	Participants.....	160
5.1.2	Procedure	161
5.1.3	Materials	161
5.1.4	Data analysis	163
5.2	Results.....	166
5.2.1	Demographics and cultural factors	168
5.2.2	ICU Structure	173
5.2.3	ICU 11 Subscale Predictive Validity	178
5.3	Discussion	180

5.3.1 ICU Structure 180

5.3.2 ICU Validity 182

5.3.3 Conclusion 184

An unanticipated effect observed in Chapter 3 was the role of language background in moderating the relationship between the Inventory of Callous Unemotional Traits (ICU) scores and extinction learning. As the Australian international student population is predominantly of East Asian nationality (Department of Education Skills and Employment - Australian Government, n.d.), it was likely that differences between East Asian and local participants generated this effect. This chapter therefore addresses whether the interpretation and construct validity of the ICU is consistent across cultural variables present in these groups. In a recent synthesis Hauser et al. (2020) highlighted that like any psychopathology, the behavioural features of psychopathic traits have only attracted researchers' attention due to their deviance from culturally normalised behaviour. Therefore, the cultural context of western research has a bearing on the content of the accepted definitions and measures of CU and psychopathic traits (Hauser et al., 2020). However, even within the context of western research the understanding of the ICU's factor structure and construct validity remains unresolved. This study therefore sought to identify culturally independent components of the ICU and determine the association of these components with variables central to the CU trait construct. These goals did not directly correspond with the primary aims outlined for this thesis. However, this study was intended to strengthen the validity of the CU trait operationalisation used in this thesis, and thereby improve the accuracy of its results. In a broader research context, this study contributed to the critically important process of developing valid and consistent measures of CU traits.

5.0.1 Development of the ICU

The ICU was initially derived from the Antisocial Process Screening Device, a measure designed to allow parent, teacher, or self-assessment of children's levels of each of the PCL-R items. The APSD was found to have three factors: callous-unemotional, narcissistic, and

impulsive traits (Frick et al., 2000). After findings showing that the callous-unemotional dimension had the strongest predictive value for conduct problems, four items from this subscale were developed into the full 24 item Inventory of Callous Unemotional Traits (Essau et al., 2006). This measure was initially modelled in a community sample of adolescents with a three factor structure ‘callousness’, ‘uncaring’ and ‘unemotional’ (Essau et al., 2006). However, unidimensional and bifactor (with callous, uncaring and unemotional factors contributing to a single higher-order factor) models have been supported, and the ideal model of this measure is not established (Cardinale & Marsh, 2020).

The validity of the ICU measure is informed by the outcomes predicted by its subscales. Despite concerns about the utility of self-report measures of a construct which includes deceitfulness as a trait (Riopka et al., 2015), the (ICU) has shown predictive validity in its associations with conduct problems, youth offending, impulsive behaviour, low empathy and low prosocial behaviour (Cardinale & Marsh, 2020). These associations are particularly evident for the callous and uncaring subscales, rather than the unemotional subscale (Cardinale & Marsh, 2020). As this subscale is less sound in external validity, construct validity, and internal consistency, many researchers have argued for its exclusion (Cardinale & Marsh, 2020).

Consistent with these issues, S. W. Hawes et al. (2014) found that parent form ICU had poor psychometric properties in a sample of American children with conduct problems. To improve these properties these authors developed a short form version of the scale using six items from the callous factor, five from the uncaring factor and one from the unemotional scale. Colins et al. (2016) removed this unemotional item to create the “ICU 11”, which had improved properties in female incarcerated youth.

5.0.2 Predictive validity of the ICU

Despite showing greater validity than the unemotional dimension, the callous and uncaring subscales have been scrutinised due to their wording. These factors were intended to assess “a lack of empathy and remorse” in the callous subscale and “an uncaring attitude about performance on tasks and others’ feelings” in the uncaring subscale (Frick & Ray, 2015, pg. 713). However, the callous scale has positively worded items (e.g., “The feelings of others are unimportant to me”) where-as the uncaring scale is negatively worded (reverse coded) (e.g., “I am concerned about the feelings of others”). This wording difference, rather than the conceptual difference, may account for the loading of these items into two separate factors in factor analyses (Frick & Ray, 2015).

The differences between callous and uncaring subscales extend beyond their wording. For instance, the callous subscales has displayed greater associations with measures of aggression (Cardinale & Marsh, 2020). For instance, Kimonis et al. (2013) found, using a convenience sample of undergraduates, that antisocial behaviour measured by the Personality Diagnostic Questionnaire was predicted by the callous subscale of the ICU. Similarly (Ray et al., 2016) found that scores on the Reactive-Proactive Aggression Questionnaire were predicted most strongly by the callous dimension in incarcerated male youth. Although, White et al. (2015) observed that both callous and uncaring components of the ICU predicted pro-active aggression (on the Self Report of Aggression and Social Behavior Measure (SRASBM)) in female undergraduates (notably the unemotional scale of the ICU was not correlated with these aggression components).

Within shortened versions of the ICU (which include only shortened callous and uncaring subscales), the callous subscale specifically has been found to predict greater internalising (S. W.

Hawes et al., 2014) and externalising (Colins et al., 2016; S. W. Hawes et al., 2014) problems. Whereas the uncaring subscale of the ICU 11 has been linked to reduced internalising problems (Colins et al., 2016; S. W. Hawes et al., 2014), and reduced prosocial behaviour (S. W. Hawes et al., 2014). These varying associations suggest that the subscales of the ICU may detect distinct aspects of CU traits (S. W. Hawes et al., 2014).

5.0.3 Cross-Cultural Studies of the ICU

An important aspect of validly measuring psychological traits is determining the consistency of measures across populations. Recently, Allen et al. (2020) tested models of the ICU across Chinese and English school age students. They found that traditional unidimensional, bi-factor or three factor models of the ICU had poor fit, which was improved by the shortened ICU 11 model. This finding was supported by data from Chinese undergraduate students, which was well described by the ICU 11 model (Wang et al., 2017). This data indicated that removal of the unemotional scale is important when measuring CU traits in these samples, as it appears to generate inconsistency in responding between Western and Chinese samples.

A further methodology useful in determining the cross-cultural validity of psychometric tools is multi-group factor-analysis (Brown et al., 2017). This approach considers the correlations and relative levels of item endorsement between groups. These response patterns indicate whether different groups consider the same sets of items as related, and therefore, whether these groups are interpreting the items in a comparable way. Allen et al. (2020) utilised this methodology in their Chinese and English samples and observed differences in the loading of items onto factors, suggesting that the scale had divergent meanings across these populations. The most inconsistent item was “I do not care about doing things well”, which was more likely to be chosen as ‘slightly true for me’ by Chinese children and as ‘not true at all’ by British

children. These authors cited higher expectations for school performance in the Chinese cultural context (H. Chen & Lan, 1998) as an explanation for this difference. Whilst Allen et al. (2020) documented differences in the meaning of the ICU scale across cultures they noted that further research was necessary to determine if the predictive value of the ICU is cross-culturally consistent. Therefore, the cultural effects documented in Chapter 3 may have arisen from either of these processes: cultural differences in interpretation of the ICU, or cultural differences in the associations between CU traits and learning processes.

5.0.4 CU Traits and Collectivism

Allen et al. (2020) cited a range of specific cultural features in explaining variations in ICU responding, including cultural norms for apologizing and school achievement. However, researchers have characterized the primary cultural constructs that differ between East Asian and Western cultures as collectivism and individualism (Cooke, 1998). Furthermore it has been noted that the demands of contemporary individualistic cultures, typified by the USA, often promote behaviours assessed by measures of psychopathy, including remorselessness (Hare (1993) cited by Cooke (1998)) and interpersonal irresponsibility, which can allow for greater financial success in economies based on competition (Cooke, 1996). Contrastingly, collectivist cultures value personal interdependence, requiring individuals to care and support other in-group members (Shou et al., 2019), which may be inhibitory to the expression or development of psychopathic traits (Cooke, 1998).

Whilst various PT measures have validated the psychopathy construct within a collectivist context (predominantly China) (reviewed by Shou et al., 2019), these studies have not taken a comparative approach and therefore cannot identify culturally dependent components of these measures. Shou et al. (2019) completed the only study I know of to date, investigating

collectivism as a predictor of PT. These authors employed the triarchic psychopathy measure (TriPM), Levenson's self-report psychopathy scale (LSRP), and the Individualism and Collectivism Scale (ICS) in a range of populations (Australian students, American remunerated volunteers, Chinese students, and Singaporean remunerated participants). They found that individualism was positively related to the behavioural features of the devices (TriPM Disinhibition and LSRP Antisocial) and to a lesser extent the affective features (TriPM Meanness, LSRP Egocentricity, and LSRP Callousness). Collectivism was largely negatively related to the affective features of these PT measures.

Despite the traditional characterization of Asian cultures as collectivist and western cultures as individualist, Shou et al. (2019) did not show that average levels of ICS scores in eastern or western samples were consistent with this conceptualisation. This is not inconsistent with the construct that the ICS was designed to address, which may vary due to factors such as gender or economic status in addition to, or even beyond factors such as nationality or ethnic groupings (Triandis & Gelfand, 1998). Despite indicating that this cultural construct is an important determinant of responses to psychopathy measures, Shou et al. (2019) did not determine whether this effect is due to differences in scale interpretation or trait levels. It therefore remains unclear whether variation in collectivism and individualism affects interpretation of the ICU.

5.0.5 Aims and Hypotheses

As observed by Allen et al. (2020) in Chinese and English samples, I expected that the shortened ICU 11 measure would provide greater cross-cultural invariance than full scale models. To test this hypothesis, I aimed to recruit participants from culturally distinct populations (of East Asian and Western ethnicity – due to the relevance of these groups to findings in Chapter

3) and examine responses to the ICU scale across these groups. I also aimed to test other cultural features as determinants of ICU responding. These included the collectivism/individualism construct, due to the findings that these cultural values are associated with variation in psychopathic traits (Shou et al. 2019). I also considered English fluency, as I had observed effects of language background on ICU related behaviour in Chapter 3.

I also aimed to test the resulting, culturally consistent, ICU measure for convergent and predictive validity. Existing literature documents correlations of medium size ($r \sim 0.4$) between psychopathic traits measures and ICU callous and uncaring subscales (Cardinale & Marsh, 2020). I accordingly expected to observe similar effects in this study, thereby demonstrating convergent validity for the culturally consistent ICU measure. To examine the construct validity of this ICU measure, I hypothesised that ICU scores would be associated with measures of aggression and impulsivity, particularly for Callous subscale scores, as documented by Ray et al. (2016), and (S. W. Hawes et al., 2014). A further hypothesis, based on the findings of (S. W. Hawes et al., 2014), was that greater callous subscale scores would predict higher anxiety, whilst greater uncaring scores would predict lower anxiety. Finally, concordant with findings documented by Shou et al. (2019) I anticipated that ICU scores would be positively associated with individualism, and negatively associated with collectivism. By examining these hypotheses my overarching aim was to empirically validate a version of the ICU in the culturally diverse population sampled in my extinction research. This outcome would also contribute to existing understandings of the characteristics assessed by the ICU measure.

5.1 Methods

5.1.1 Participants

For this study two recruitment streams were created by selecting for participants with monocultural Western (“Anglo” or European) or East Asian (Chinese, Indian, Korean, Vietnamese, Filipino, Japanese, Malaysian, Thai, or other Asian) ethnicity. The sample size for this study was determined in the power analysis described in the previous chapter (see section 4.1.2). That analysis indicated that a sample of 465 participants with English as a first language would be sufficient to observe the renewal effect described in Experiment 3 of Chapter 3. It was assumed that the Western stream would correspond with the participants with English as a first language that were the subject of my power analysis. Therefore, I aimed to recruit 465 participants of this ethnicity, as well as a sample of individuals of East Asian ethnicity (the size of which was determined by logistical limitations). This sample was utilized in both the following psychometric study, as well as an experimental study pertained to by the power analysis.

Two recruitment sources were utilized in this study. Five hundred and twenty-one participants were recruited from an online platform with £3.34 (\$6.07 Australian) provided as reimbursement. Two hundred and thirty-eight participants were undergraduates from Sydney University who participated for course credit. A further 11 participants were recruited from the University of Sydney – school of psychology’s paid recruitment platform and were reimbursed with \$7.50. Recruitment and procedures were approved under the University of Sydney Human Research Ethics Committee protocol 2021/315 – “Culture and Learning in Personality”.

Over all 770 participants, 485 were female, 280 were male (5 did not specify) and the median age was 26 (21 – 39 interquartile range). My ethnically streamed recruitment strategy

produced 470 “Western” participants and 300 “East Asian” participants. Further demographic features of this sample are described in Table 5.2.

5.1.2 Procedure

After receiving participant information and consent materials, participants completed a behavioural task (analysed in Chapter 6). On completing the task all questionnaires were delivered online through the Qualtrics platform. Three attention screening questions were distributed throughout the questionnaires described below (Berinsky et al., 2014; Kung et al., 2018). Failure to respond accurately to two of these checks led to exclusion from the study.

5.1.3 Materials

The following scales were presented in a randomized order, with the order of items within scales also randomized.

5.1.3.1 Inventory of Callous Unemotional Traits – Youth Version (ICU)

The ICU (Frick, 2004) was the measure of interest for CU traits due to the existing work documenting its variation across western and non-western cultures (Allen et al. 2020). It is a 24 item self-report scale with responses collected on a likert scale ranging from 0 (not at all true) to 3 (definitely true). Responses on this measure were used to calculate scores on both the full ICU scale and the shortened ICU scale.

5.1.3.2 Levenson’s Self-Report Psychopathy Scale (LSRP)

The LSRP (Levenson et al., 1995) was used in this study as an established method of testing the convergent validity of the ICU (Kimonis et al., 2013; Cardinale & Marsh, 2020). It is a 26-item scale (with a 4-point Likert scale: 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree) designed as a self-report analogue of the PCL-R and therefore includes both

behavioural and affective factors. The construct validity of the behavioural factor has been questioned (Brinkley et al., 2001).

Whilst the two-factor model of this scale is the original conceptualization (Levenson et al., 1995), more recent work has examined a shortened three factor model (which splits the primary (affective) dimension into egocentricity and callous scales in a sample of Chinese undergraduates (Shou et al., 2017). These authors found that the three-factor model had improved fit as well as configural and partial metric invariance across Chinese and American undergraduates. This result indicates that whilst response patterns vary across these groups, the associations between this model of the LSRP and other variables can be validly compared across these cultural groups (Shou et al., 2017). As a result, I utilised these subscale definitions for all analyses.

5.1.3.3 Self Report of Aggression and Social Behavior Measure (SRASBM)

Nine items from the 39 item SRASBM were used to measure reactive and pro-active relational aggression. This scale was initially developed by Morales and Crick (1998) (published in Linder et al. (2002)) with relational and physical components assessed on a seven point Likert scale (never true – always true). Later studies grouped the relational items addressing pro-active and reactive aggression (Bailey & Ostrov, 2008). The internal consistency of these structures has been supported in factor analytic research in a community based twin sample (Murray-Close et al., 2010).

5.1.3.4 Individualism and Collectivism Scale (ICS)

The ICS (Triandis & Gelfand, 1998) is a self-report measure with 16 items. It contains four subscales; horizontal individualism -measuring personal independence, vertical individualism – measuring status values, horizontal collectivism – measuring collective values,

and vertical collectivism – desire to maintain in-groups. This scale predicts variation in PT and has been validated in Chinese and English samples (Shou, 2019).

5.1.3.5 Self-Reported Fluency of English Scale (SRFES)

This scale has three items measuring self-perceived English fluency and has been found to have good internal consistency (Yeh & Inose, 2003). Participants respond on a 5-point Likert scale to the following questions:

1. What is your present level of English fluency?
2. How comfortable are you communicating in English?
3. How often do you communicate in English?

5.1.3.6 Other Measures

Other measures utilised in this study have been described in Chapter 3, these included the STAI and UPPS, as well as the extinction task. Responding on the behavioural task was addressed in Chapter 6.

5.1.4 Data analysis

This analysis tested the fit of the three factor and single factor ICU models established by (Essau et al., 2006). I additionally tested the ICU 11 two factor model which was found to have the strongest fit statistics in Allen et al. 2020. I then tested the invariance of the best fitting model over cultural groups defined by the SRFES and ICS measures, as well as nominated ethnicities and sample source. Total subscale scores were then calculated from the items determined to be cross-culturally invariant. These were used for various tests of the predictive capacity of this model. To test the predictive validity of these subscales I calculated their partial and bivariate correlations with the LSRP, STAI, ICS, SRASBM, and UPPS. Details of these analytical methodologies are described below. Analysis scripts can be viewed at <https://osf.io/tfjqc/> (see

Chapter5FactorAnalysis.html), where validation of statistical assumptions was also carried out, along with additional descriptive analyses. Anonymised data from in this chapter is also available.

5.1.4.1 Pre-Processing

Missing data was present in responses on the SRFES (0.13%), UPPS (0.12%) and STAI (0.14%) scales. As responses on these scales appeared to be missing at random all missing data was imputed through Multivariate Imputation by Chained Equations (van Buuren & Groothuis-Oudshoorn, 2011).

5.1.4.2 Cluster Analyses

The multivariate cultural measures SRFES and ICS were converted to categorical variables via cluster analysis. This allowed for tests of the measurement invariance of the ICU over groupings of responders on these scales. Initially data was converted to z-scores, then the appropriate number of clusters was identified using scree (within-group sum of squares) and silhouette plots. Clusters were calculated using a k-means algorithm. Item level data was used for these analyses. In the case of the ICS, item responses, rather than subscale totals, were clustered as a CFA showed that these subscales were not consistent with the observed data (see Table 5.4).

5.1.4.3 Confirmatory Factor Analyses

Before fitting any models to scales in this study several assumptions were checked. Univariate normality of items in scales was tested using Anderson-Darling tests and the multivariate normality tested via Henze-Zirkler tests (Korkmaz et al., 2014). The multicollinearity of items within scales was checked through correlation matrices. Additionally, the presence of multivariate outliers for each scale was determined by calculating Mahalanobis distance for each individual's set of responses. These were then compared to a cut-off score

bounding 99.9% of a χ^2 distribution with degrees of freedom equal to the number of items on that scale. After model fitting, residuals were examined for normality using qqplots and histograms, and their heteroscedasticity was observed by plotting them against the model's fitted values. Instances where these assumptions were not met have been noted and robust statistics are reported.

CFA Models were fit using the Lavaan package (Rosseel, 2012). Scales measured on four-point Likert scales were treated as ordinal, i.e., models were fit on polychoric correlations using a diagonally weighted least squares estimator. Acceptable model fit was indicated by CFI and TLI statistics of $\geq .90$ and RMSEA ≤ 0.08 (following Allen et al., 2020). Satorra-Bentler χ^2 adjustment was utilised to make this statistic robust to assumption violations. The reliability of model factors was determined using Chronbach's alpha and McDonald's hierarchical omega (McNeish, 2018). Overall scale reliability was calculated using Mean Inter-item correlations (MIC) which were deemed acceptable in the range of 0.15 to 0.50 (Kline, 2015).

5.1.4.4 Measurement Invariance

To investigate the measurement invariance of the best fitting ICU model across grouping variables in this study, four levels of models were fit per grouping variable. At the first level the model was fit each group separately, allowing for indication of poor fit in specific subgroups. The second model tested whether the model structure was consistent across groups (configural invariance). The third model tested the invariance of items loadings onto factors across groups (metric invariance). The fourth tested whether model intercepts were comparable across groups (scalar invariance). Measurement invariance was established where there was a small change in model fit statistics between configural and metric models as well as between metric and scalar models. A small change was defined as $<.01$ for CFI and TLI (Cheung & Rensvold, 2002).

5.1.4.5 *External Correlations*

In order to understand relationships between ICU 11 subscales (represented as the sum of scores on relevant items) and external variables (the STAI, ICS, UPPS, LSRP and SRASBM) I followed the approach used by (S. W. Hawes et al., 2014). This involved calculating bivariate correlations between each ICU 11 subscale and each external variable. Partial correlations between external variables and each ICU 11 subscale were also calculated, accounting for variation in the other ICU 11 subscale. ICU 11 subscale scores were utilised due to the findings of the CFA (reported below). Total ICU 11 scores (across both subscales) were not considered in the multiple group CFA described above (as only the two-factor model has been cross-culturally examined (Allen et al., 2020)). As a result this ICU 11 total score was not included in the present analysis, despite its examination by S. W. Hawes et al. (2014).

5.2 Results

Descriptive statistics for each measure's subscales across the entire dataset are shown in Table 5.1.

Table 5.1 Descriptive Statistics for all Measured Scales

scale	mean	sd	min	max	skewness	kurtosis	alpha	MIC
ICU	21.87	8.39	2	56	0.40	3.32	0.82	0.16
ICU callous	6.06	4.18	0	25	1.32	5.13	0.73	0.21
ICU uncaring	7.81	3.87	0	20	0.24	2.59	0.75	0.28
ICU unemotional	7.99	3.34	0	15	-0.11	2.66	0.80	0.45
ICU 11 callous	2.32	2.61	0	15	1.60	5.93	0.68	0.28
ICU 11 uncaring	4.24	2.83	0	13	0.56	2.85	0.75	0.38
SRFES	13.81	2.24	5	15	-1.80	5.19	0.94	0.84
ICS Horizontal individualism	26.65	4.83	7	36	-0.37	3.05	0.70	0.37
ICS Vertical individualism	19.72	5.47	4	36	0.01	2.89	0.66	0.32
ICS Horizontal collectivism	25.58	5.08	6	36	-0.54	3.51	0.75	0.44
ICS Vertical collectivism	24.19	5.50	5	36	-0.28	2.97	0.71	0.36
LSRP egocentricity	18.42	5.59	10	37	0.53	2.65	0.86	0.38
LSRP callous	2.74	2.18	0	12	0.90	4.09	0.62	0.29
LSRP antisocial	10.11	2.84	5	20	0.39	2.69	0.65	0.27
SRASBM proactive	6.93	3.42	4	22	1.57	5.48	0.77	0.46
SRASBM reactive	9.92	4.57	5	30	1.12	4.23	0.79	0.44
UPPS premeditation	20.99	5.28	11	42	0.24	3.20	0.87	0.39
UPPS urgency	17.73	6.87	1	37	0.18	2.90	0.88	0.39
UPPS sensation	17.41	8.12	0	36	0.00	2.27	0.89	0.42
UPPS perseverance	19.09	5.30	8	36	0.17	2.87	0.86	0.39
STAI	38.41	12.50	11	69	0.17	2.48	0.94	0.43

Note. Alpha = Chronbach's alpha, MIC = mean spearman inter-item correlation

5.2.1 Demographics and cultural factors

Measured demographic features for each ethnic group, split by recruitment source are shown in Table 5.2. Average scores on the ICS and SRFES are shown in Table 5.1. English fluency reported on the SRFES was lowest for the East Asian Undergraduate sample. Overall, scores on the SRFES scale were highly negatively skewed (see Table 5.1). Therefore, k-means clustering was used to categorise low ($n = 162$) and high ($n = 608$) fluency responses on this scale. Demographics for these groups are shown in Table 5.3.

Before creating a categorical variable from the ICS, a confirmatory factor analysis was conducted to test the validity of the horizontal – vertical, collectivist – individualist model of the ICS. This model was found to fit the current data poorly (Table 5.4). It also exhibited heteroscedasticity and positive skew of residuals. Considering the poor fit of the ICS structure, responses on this scale were categorized by k-means clustering carried out on item level responses. This analysis identified three clusters: labelled “High responders” “Collectivists” and “Individualists” based on relative levels of mean scores on each of the ICS subscales. Descriptive statistics for demographic and cultural factors in each of the clustered cultural variables can be seen in Table 5.5.

Table 5.2 *Sample Demographics and Cultural Measures Split Across Ethnicity and Recruitment Source*

Variable	Western			East Asian		
	Overall, N = 470	community, N = 349	undergraduate, N = 121	Overall, N = 300	community, N = 183	undergraduate, N = 117
age	32 (22-45)	37 (27-50)	19 (19-21)	22 (20-29)	27 (23-33)	20 (19-21)
duration of residence	31 (22-44)	37 (27-50)	19 (18-20)	7 (1-23)	21 (10-27)	1 (0-2)
gender						
female	309 (66%)	222 (64%)	87 (72%)	176 (59%)	89 (49%)	87 (75%)
male	157 (34%)	123 (36%)	34 (28%)	123 (41%)	94 (51%)	29 (25%)
Not specified	4	4	0	1	0	1
ethnicity						
Anglo	418 (89%)	302 (87%)	116 (96%)			
European	52 (11%)	47 (13%)	5 (4.1%)			
Chinese				213 (71%)	98 (54%)	115 (98%)
Indian				11 (3.7%)	11 (6.0%)	0 (0%)
Korean				28 (9.3%)	28 (15%)	0 (0%)
Vietnamese				10 (3.3%)	10 (5.5%)	0 (0%)
Other Asian				38 (13%)	36 (20%)	2 (1.7%)
Student status						
International	2 (0.4%)	1 (0.3%)	1 (0.8%)	144 (48%)	27 (15%)	117 (100%)
Local	172 (37%)	52 (15%)	120 (99%)	52 (17%)	52 (28%)	0 (0%)
Non-student	296 (63%)	296 (85%)	0 (0%)	104 (35%)	104 (57%)	0 (0%)

	Western				East Asian	
SRFES English fluency	14.88 (0.60)	14.85 (0.69)	14.98 (0.13)	12.12 (2.78)	13.66 (2.07)	9.71 (1.88)
ICS						
horizontal individualism	26.81 (4.79)	26.89 (4.89)	26.60 (4.51)	26.40 (4.89)	26.86 (4.66)	25.68 (5.16)
vertical individualism	18.89 (5.47)	18.69 (5.58)	19.47 (5.11)	21.03 (5.24)	20.86 (5.05)	21.28 (5.53)
horizontal collectivism	26.39 (4.83)	25.93 (5.01)	27.73 (4.02)	24.30 (5.19)	25.07 (4.75)	23.09 (5.62)
vertical collectivism	24.86 (5.33)	25.45 (5.32)	23.17 (5.02)	23.14 (5.61)	23.84 (5.63)	22.03 (5.43)

Note. Values shown are median (25%-75%); n (%) for categorical measures, mean (SD) for continuous measures

Table 5.3 *Demographic Descriptives in SRFES Clusters*

Characteristic	Low Fluency, N = 162 ¹	High Fluency, N = 608 ¹	p-value ²
age	23.96 (8.24)	33.18 (13.77)	<0.001
gender			0.4
female	107 (66%)	378 (63%)	
male	54 (34%)	226 (37%)	
Not specified	1	4	
sample			<0.001
community	52 (32%)	480 (79%)	
undergraduate	110 (68%)	128 (21%)	
Ethnicity			<0.001
Western	6 (3.7%)	464 (76%)	
East Asian	156 (96%)	144 (24%)	
English fluency	9.79 (1.68)	14.88 (0.42)	<0.001
¹ Mean (SD); n (%)			
² Wilcoxon rank sum test; Pearson's Chi-squared test			

Table 5.4 Confirmatory Factor Analysis for ICS Measure

Component	S-B χ^2	df	CFI	TLI	RMSEA [90% CI]	α	ω
Model	533.583	98	0.826	0.787	0.084 [0.077 – 0.091]		
Horizontal individualism						0.70	0.72
Vertical individualism						0.66	0.66
Horizontal collectivism						0.75	0.75
Vertical collectivism						0.71	0.77

Note. S-B χ^2 = Satorra-Bentler Chi Square, df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis index, RMSEA = root mean square error of approximation, CI = confidence interval, α = Chronbach's Alpha, ω = McDonald's Heirarchical Omega. CFI, TLI and RMSEA were adjusted for robustness.

Table 5.5 *Demographic Descriptives in ICS Clusters*

Characteristic	Overall, N = 770	Collectivist, N = 291	High responders, N = 260	Individualist, N = 219	p-value
age	31.24 (13.34)	29.85 (12.56)	33.53 (14.90)	30.38 (12.04)	0.056
gender					0.089
female	485 (63%)	189 (66%)	171 (66%)	125 (57%)	
male	280 (37%)	99 (34%)	88 (34%)	93 (43%)	
not specified	5	3	1	1	
sample					0.038
community	532 (69%)	186 (64%)	192 (74%)	154 (70%)	
undergraduate	238 (31%)	105 (36%)	68 (26%)	65 (30%)	
culture					<0.001
Western	470 (61%)	176 (60%)	186 (72%)	108 (49%)	
East Asian	300 (39%)	115 (40%)	74 (28%)	111 (51%)	
English fluency	13.81 (2.24)	13.69 (2.35)	14.17 (1.77)	13.52 (2.53)	0.027
Horizontal individualism	26.65 (4.83)	22.39 (3.68)	29.19 (3.39)	29.30 (3.42)	<0.001
Vertical individualism	19.72 (5.47)	17.25 (4.62)	20.34 (5.64)	22.27 (4.94)	<0.001
Horizontal collectivism	25.58 (5.08)	25.21 (4.05)	29.48 (3.14)	21.43 (4.66)	<0.001
Vertical collectivism	24.19 (5.50)	23.01 (4.32)	28.73 (3.68)	20.37 (4.98)	<0.001
SRFES cluster					0.034
Low fluency	162 (21%)	67 (23%)	41 (16%)	54 (25%)	
High fluency	608 (79%)	224 (77%)	219 (84%)	165 (75%)	

Note. Displayed values are Mean (SD); n (%), p values from Kruskal-Wallis rank sum test or Pearson's Chi-squared test

5.2.2 ICU Structure

Both the three factor and single factor models of the 24 item ICU scale had fit statistics below the cut-offs for good fit of .09 TLI and CFI, and above the cut-off of .08 RMSEA.

However, fit statistics for the ICU 11 model met these cut-off criteria (Table 5.6). Loadings of

items onto this scale are shown in Table 5.7. The ICU 11 exhibited configural, metric and scalar invariance over the grouping variables of SRFES, sample source, and ethnicity, with reductions in CFI and TLI statistics of $< .01$ across these models when considering each of these grouping factors (Table 5.8). Change in CFI and TLI was $< .01$ between configural and metric models with ICS clusters as a grouping factor. However, between metric and scalar models the change in CFI was .012, whilst change in TLI was 0.005 (Table 5.8). Despite exceeding the CFI criterion for scalar invariance, the model at this level of constraint had extremely high fit with CFI = 0.963, TLI = 0.967, RMSEA = 0.056 (Table 5.8).

Table 5.6 *Confirmatory Factor Analysis for ICU Measures*

Component	S-B χ^2	df	CFI	TLI	RMSEA [95% CI]	α	ω
ICU 24 single factor	6,949.914	252	0.731	0.706	0.161 [0.158, 0.164]	0.82	0.97
ICU 24 three factor	3,675.296	249	0.864	0.850	0.115 [0.112, 0.118]		
callous						0.73	0.75
uncaring						0.75	0.87
unemotional						0.80	0.85
ICU 11	149.129	43	0.988	0.984	0.043 [0.036, 0.051]		
callous						0.68	0.72
uncaring						0.75	0.76

Note. S-B χ^2 = Satorra-Bentler Chi Square, df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis index, RMSEA = root mean square error of approximation, CI = confidence interval α = Chronbach's Alpha, ω = McDonald's Heirarchical Omega. CFI, TLI and RMSEA are adjusted for robustness.

Table 5.7 *ICU II Factor Loadings*

Latent Factor	Indicator	B	SE	Z	p-value	β
callous	4. I do not care who I hurt to get what I want	1.00	0.00			0.82
	9. I do not care if I get into trouble	0.71	0.06	11.75	<.001	0.58
	11. I do not care about doing things well	0.60	0.06	9.75	<.001	0.49
	12. I seem very cold and uncaring to others	0.73	0.06	12.94	<.001	0.60
	18. I do not feel remorseful when I do something wrong	0.82	0.06	13.07	<.001	0.67
	21. The feelings of others are unimportant to me	0.90	0.06	15.80	<.001	0.73
uncaring	5. I feel bad or guilty when I do something wrong (R)	1.00	0.00			0.59
	8. I am concerned about the feelings of others (R)	1.38	0.08	17.09	<.001	0.82
	16. I apologize (say 'I am sorry') to persons I hurt (R)	1.07	0.07	14.78	<.001	0.63
	17. I try not to hurt others' feelings (R)	1.16	0.07	15.51	<.001	0.69
	24. I do things to make others feel good (R)	1.11	0.07	15.60	<.001	0.66

Note. B = unstandardized coefficient, β = standardized coefficient, SE = standard error, Z = test statistic.

Table 5.8 ICU 11 Measurement Invariance Over Grouping Variables

Grouping Variable	model	S-B χ^2	df	CFI	TLI	RMSEA [95% CI]	$\Delta\chi^2$	Δdf	ΔCFI	ΔTLI
ethnicity	Western	124.806	43	0.981	0.975	0.05 [0.04, 0.061]				
	East Asian	118.392	43	0.980	0.975	0.059 [0.046, 0.071]				
	configural	243.380	86	0.981	0.975	0.054 [0.046, 0.062]				
	metric	282.751	95	0.973	0.968	0.06 [0.052, 0.069]	39.371	9	-0.008	-0.007
	scalar	299.695	115	0.974	0.975	0.054 [0.046, 0.061]	16.945	20	0.001	0.007
SRFES Clusters	low	101.263	43	0.973	0.966	0.071 [0.053, 0.089]				
	high	135.499	43	0.984	0.980	0.046 [0.038, 0.055]				
	configural	236.842	86	0.981	0.976	0.052 [0.045, 0.06]				
	metric	245.643	95	0.978	0.975	0.054 [0.046, 0.062]	8.801	9	-0.003	-0.001
	scalar	309.478	115	0.973	0.974	0.055 [0.047, 0.062]	63.835	20	-0.005	-0.001
ICS Clusters	collectivists	62.005	43	0.991	0.989	0.033 [0, 0.05]				
	high responders	48.570	43	0.996	0.994	0.018 [0, 0.039]				
	individualists	129.013	43	0.967	0.958	0.075 [0.06, 0.09]				
	configural	236.485	129	0.982	0.978	0.046 [0.037, 0.055]				
	metric	276.631	147	0.975	0.972	0.051 [0.042, 0.061]	40.146	18	-0.008	-0.006
	scalar	388.758	187	0.963	0.967	0.056 [0.048, 0.063]	112.127	40	-0.012	-0.005

Grouping Variable	model	S-B χ^2	df	CFI	TLI	RMSEA [95% CI]	$\Delta\chi^2$	Δdf	ΔCFI	ΔTLI
sample source	community	115.907	43	0.985	0.981	0.044 [0.035, 0.054]				
	undergraduate	98.469	43	0.987	0.984	0.054 [0.04, 0.068]				
	configural	215.636	86	0.986	0.982	0.047 [0.04, 0.055]				
	metric	227.669	95	0.983	0.980	0.05 [0.042, 0.059]	9	-0.03	-0.002	9
	scalar	297.481	115	0.978	0.979	0.052 [0.045, 0.06]	20	-0.05	-0.002	20

Note. See table 5.6 for statistics definitions. Additionally, Δ statistics show change from configural to metric, and metric to scalar invariance models.

5.2.3 ICU 11 Subscale Predictive Validity

Correlations between ICU 11 subscales and external psychometric measures are shown in Table 5.9. ICU 11 subscales correlated positively and significantly with the LSRP measure.

Numerical comparison of correlation coefficients showed that both uncaring and callous subscales predicted LSRP egocentricity to a similar degree. However, the callous dimension predicted LSRP antisocial to a greater extent than the uncaring subscale. Conversely, uncaring predicted LSRP callous more strongly than ICU 11 callous did. ICU 11 subscales also positively predicted both SRASBM subscale scores (pro-active and reactive aggression). Although, the callous subscale had a numerically larger relationship with these SRASBM subscales than the uncaring subscale.

UPPS subscale scores were higher in those with greater ICU 11 callous scores, particularly for the perseverance subscale. Relationships between the UPPS and the ICU 11 uncaring subscale were not present after partialling out variance shared with the callous subscale. Similarly, positive relationships were present between the STAI measure and the callous ICU 11 subscale, but not the uncaring subscale. ICS horizontal individualism was greater at higher levels of both ICU 11 subscales, whilst ICS collectivism subscales scores were lower in those with higher ICU 11 subscale scores (particularly for the uncaring subscale).

Table 5.9 *ICU 11 Subscales' Correlations with External Measures*

Measure	Subscale	ICU 11 uncaring				ICU 11 callous			
		Bivariate	p	partial	p	Bivariate	p	partial	p
UPPS	premeditation	0.117	0.001	0.045	0.216	0.192	<.001	0.159	<.001
	urgency	0.065	0.071	-0.012	0.731	0.189	<.001	0.178	<.001
	sensation seeking	0.019	0.603	-0.018	0.615	0.087	0.015	0.087	0.015
	perseverance	0.098	0.007	0.006	0.860	0.228	<.001	0.207	<.001
STAI		0.028	0.433	-0.044	0.224	0.168	<.001	0.171	<.001
LSRP	egocentricity	0.443	<.001	0.318	<.001	0.456	<.001	0.337	<.001
	callous	0.515	<.001	0.430	<.001	0.371	<.001	0.208	<.001
	antisocial	0.233	<.001	0.091	0.012	0.388	<.001	0.330	<.001
SRASBM	proactive	0.333	<.001	0.156	<.001	0.522	<.001	0.449	<.001
	reactive	0.317	<.001	0.166	<.001	0.448	<.001	0.368	<.001
ICS	horizontal individualism	0.018	0.623	-0.002	0.948	0.049	0.172	0.046	0.202
	vertical individualism	0.165	<.001	0.107	0.003	0.171	<.001	0.116	0.001
	horizontal collectivism	-0.565	<.001	-0.488	<.001	-0.373	<.001	-0.192	<.001
	vertical collectivism	-0.338	<.001	-0.277	<.001	-0.228	<.001	-0.106	0.003

Note. This table shows bivariate and partial correlations and their corresponding p values. Yellow text shows significant positive correlations, blue text shows significant negative correlations.

5.3 Discussion

The results of this study fulfilled my aim of validating a version of the ICU across community and undergraduate participants of East Asian and Western ethnicity. This was seen for in the culturally invariant structure of the ICU 11. Furthermore, they demonstrated that this measure possessed convergent validity with another measure of CU traits, and predictive validity for a number of measures historically associated with these traits. Specifically, ICU 11 subscales were both positively correlated with psychopathic traits and aggression. The callous subscale was also positively related to anxiety and impulsivity. Additionally, these both subscales predicted greater individualism, and diminished collectivism. Overall, these effects were consistent with the hypotheses of this study. The ICU 11 subscales were therefore interpreted invariantly across cultures and exhibited predictive capacity consistent with documented associations for these subscales. This study therefore supports a growing body of literature indicating the strong psychometric properties of the ICU 11.

5.3.1 ICU Structure

Confirmatory factor analysis of full-scale ICU models suggested that the correlation structure of items was not well described by the original three factor or single factor solutions reported by Essau et al. (2006) (in a sample of German children). In the present study, ICU 11 solution, which excludes unemotional subscale items, was found to have strong fit statistics. This finding is consistent with the prior cross-cultural analysis of English and Chinese children conducted by Allen et al. (2020), as well as that of Wang et al. (2017) in Chinese undergraduates. This corresponding result may indicate that the unemotional scale is not a consistent indicator of the CU trait construct in different cultures.

Even within the western literature, the unemotional scale exhibits issues in convergent validity and in prediction of externalising behaviour, as shown in a meta-analysis by Cardinale and Marsh (2020). These authors suggested that the best explanation for these validity issues was that the unemotional subscale captured broad deficits in emotional expression (with items such as “I do not show my emotions to others.”) whereas the deficits in emotional expression that correlate with CU traits tend to be specific to fear. Whilst the present study cannot address the specific causes for the problematic nature of this subscale, its results are consistent with suggestions that its exclusion can improve the consistency of the ICU as a measure (e.g. S. W. Hawes et al., 2014).

In the present data for the ICU 11 the weakest loading item was “I do not care about doing things well”. In Allen et al. (2020) this item was found to lead to cross-cultural non-invariance. The authors suggested that Chinese children may evaluate their concern for “doing things well” as low in contrast to their parents expectation for school achievement, which may be higher in Chinese than Western populations (H. Chen & Lan, 1998). In my data this item had an extremely low rate of endorsement, which could potentially indicate that when participating for money or for course credit, participants were unwilling to suggest that they were performing poorly. This item may therefore be the most problematic in the ICU 11 measure.

Whilst Allen et al. (2020) found that some ICU 11 items were not interpreted consistently across English and Chinese children, the present data were invariant across individuals with ethnicities nominated as East Asian or Western. The consistency of responding across these groups in this study may reflect a greater awareness of Western cultural norms in participants of East Asian ethnicity. Such awareness is part of acculturation processes that occur for migrants and international students (Smith & Khawaja, 2011). Such acculturation may have occurred for

community participants of East Asian ethnicity as they were recruited from an English platform, with no restrictions on immigration status. This meant that these participants may have been immigrants or descendants of immigrants. It appears likely that this sample had a higher level of experience with Western cultural settings than the East Asian undergraduate sample, which had a lower average self-reported English fluency. However, this undergraduate sample would be expected to have experienced some acculturation through contact with the Western educational system (Smith & Khawaja, 2011). Whilst these groups may have differing acculturational experiences (Smith & Khawaja, 2011), this study was not designed to compare such variables as possible determinants of responses to the ICU. Instead, the finding that the ICU 11 exhibited invariance across ethnicity, sample source and English language fluency shows that subscale scores on the English version of the ICU 11 can be validly compared across these populations. However, care should be taken in generalising these findings to populations of East Asian nationality.

It should be also noted that the invariance of the ICU 11 over ICS (Individualism and Collectivism Scale) groups met most, but not all statistical criteria. However, the validity of my observations on this scale was unclear as the horizontal-vertical, individualist-collectivist factor structure was not found consistent with my data. This means that the interpretability and generalisability of the ICU invariance over groups of responders on the ICS is limited.

5.3.2 ICU Validity

In addition to demonstrating strong psychometric properties, and cross-cultural invariance, the ICU 11 also showed convergent validity with the LSRP. ICU 11 subscales were predominantly medium sized predictors of LSRP subscales, as anticipated. Furthermore, the ICU 11 had predictive validity for the SRASBM measure of aggression. This was particularly evident

for the callous subscale, which was also more strongly associated with the aggressive subscale of the LSRP. The callous subscale was also predictive of impulsivity measured by the UPPS, whilst the uncaring subscale did not show any such relationship after its overlap with the callous scale had been accounted for.

Additional observations included increased individualism and reduced collectivism at higher levels of both ICU 11 subscales. This result was consistent with those reported by (Shou et al., 2019) for the LSRP measure of psychopathic traits. The present study is therefore novel in showing this result for the ICU. It is also informative that the uncaring subscale was particularly predictive of diminished collectivism. This may reflect findings showing that the uncaring subscale also predicts lower prosocial behaviour (S. W. Hawes et al., 2014), a construct conceptually relevant to collectivism (Triandis & Gelfand, 1998). Although, the present results for the collectivism and individualism measure were questionable due to noted problems with its factor structure.

Further differences in ICU 11 subscale associations were noted for trait anxiety, which was elevated for those with high callous scores, and unrelated to uncaring scores. In total these differences support the suggestion of S. W. Hawes et al. (2014) that these subscales may be differentially related to a latent factor occurring across individuals high in CU traits. In these authors data, as well as those reported here, individuals high on the Callous subscale appear to be anxious and exhibit greater externalising behaviour, whilst those high on the uncaring subscale are ostensibly socially disengaged.

These predictive differences across ICU subscales may also relate to their wording. The ICU 11 callous subscale is entirely comprised of positively worded items, while the ICU 11 uncaring subscale consists of negatively worded (reversed) items. It has been suggested that the

endorsement of these positively worded (callous subscale) items is a better predictor of the behavioural outcomes associated with the CU trait construct (e.g., aggression and impulsivity), as well as anxiety, because it involves endorsing undesirable statements (S. W. Hawes et al., 2014). The results described above are consistent with this account. However, this study was intended to evidence the validity of a culturally consistent ICU measure, rather than to detect differences in ICU subscales. Further studies of these differences should therefore be conducted in order to confirm their reliability.

5.3.3 Conclusion

The aim of the current study was to investigate the effect of cultural factors within common convenience samples utilised in Western research settings: undergraduates and general community members recruited online. In this context I was able to determine that the ICU 11 scale was interpreted consistently, despite variation in English fluency and ethnicity. Furthermore, the ICU 11 exhibited good construct validity, exhibiting expected correlations with another measure of psychopathic traits, and measures of aggression, impulsivity, anxiety, individualism and collectivism. As an unanticipated result of these associations, I was able to support a growing body of evidence showing discrepancies between uncaring and callous ICU subscales. Primarily, this study indicates that the ICU 11 should be the preferred model for research into the CU trait construct in adult undergraduate and community samples. These findings also contribute novel validating observations to the small pool of studies addressing the ICU 11.

6 Callous Unemotional Traits and Extinction Learning

Contents

6.0.1	Optimisation of Methodology for Testing Renewal	186
6.0.2	Revisions to the ICU	187
6.0.3	Hypotheses	189
6.1	Method	190
6.1.1	Participants	190
6.1.2	Procedure	190
6.1.3	Materials	191
6.1.4	Data Analysis	192
6.2	Results	197
6.2.1	Extinction	197
6.2.2	Renewal	202
6.3	Discussion	205
6.3.1	ICU Effects	206
6.3.2	Renewal	208
6.3.3	Covariate Effects	208
6.3.4	Conclusion	210

In this chapter, findings from the research described thus far were integrated in the design of an experimental procedure. This procedure was intended to address the primary aims of this thesis: to reinforce evidence of extinction learning deficits in the phenomenology of CU traits, and, to show congruent deficits in the contextual sensitivity of this learning (through the renewal phenomenon). To achieve this, I employed the experimental design that I developed in Chapter 3 as a measure of extinction learning processes. I also utilised existing data from that in a power analysis to inform the sampling strategy for this study (Chapter 4). Data resulting from implementation of this strategy, analysed in Chapter 5, indicated that the subscales of the ICU 11 would be the most valid measures for use as predictors of individual differences in CU traits. The present study therefore made use of all prior results of this thesis in order to provide the best empirical conditions for testing the hypothesis that CU traits are associated with deficits in extinction learning.

6.0.1 Optimisation of Methodology for Testing Renewal

My aim of demonstrating an association between CU traits and deficits in renewal of extinguished responding was motivated by the experimental utility of such a result, as well as its theoretical and interventional implications. As described in section 3.0.3, correspondence between diminished extinction and renewal responses would be expected given deficits in extinction learning (Wilson et al., 2014). However, critically, such findings would not arise from potential confounds of extinction behaviour, such as changes in overall task motivation. Therefore, a link between CU traits and renewal would validate existing observations of extinction deficits.

Further, renewal results would have significant implications for clinical interventions on these traits. Namely, that whilst children with CU traits may be resistant to BFIs (D. J. Hawes et

al., 2014), any benefit of such interventions may be better retained outside of the interventional context (as conjectured in section 1.2.3). These implications compelled the investigation of renewal effects reported in Chapter 3. Whilst that investigation produced a method of observing renewal, variation in this phenomenon was not statistically associated with CU traits, and it was unclear whether that study was sufficiently powered to detect such an association.

In light of prior renewal results, I carried out an analysis designed to determine an adequately powered sampling strategy. This analysis was necessarily specific to participants of Western ethnicity. This cultural specificity was determined by cultural impacts on extinction effects seen in Chapter 2. These impacts meant that renewal effects were only likely to be estimated validly in my existing data from a native English-speaking sample. Analysis of the relationship between ICU scores and renewal in this sample indicated that recruitment of 465 such participants would be necessary to confirm the presence of this relationship (Table 4.4). As a result, 470 western participants were recruited and psychometric data from these participants was analysed in the previous chapter, along with that of 300 East Asian participants. In this chapter I focussed on the responses of this sample to the extinction task developed in Chapter 3.

6.0.2 Revisions to the ICU

In Chapter 5, the ICU 11 (modelled as a two-factor scale) was found to be a valid measure of CU traits across cultural factors present in the sample. This was not the case for ICU total scores, indicating the possibility that differences in interpretation of the ICU measure led to differences in its association with extinction across language groups (documented in Chapter 3). Given this explanation, the culturally consistent ICU 11 would not be expected to show differential associations with extinction across Western and East Asian participants in the current sample. It was also possible that the observed dependence of the relationship between CU traits

and extinction on language background was due to differences in interpretation of the extinction task. Alternatively, CU traits themselves may have been differentially related to extinction in individuals of different language backgrounds. The current study was designed to disentangle these explanations, as cultural factors described in Chapter 5 (including ethnicity, Individualism and Collectivism Scale (ICS) responses, and Self-Reported Fluency of English Scale (SRFES) responses) could be tested as determinants of responses to the extinction task, and ethnicity could be tested as a moderator of associations between ICU scores and task responses.

In the previous chapter I also observed unexpected evidence for the diverging meanings of the callous and uncaring subscales of the ICU 11. In the full length ICU measure the callous subscale has been characterised as a lack of empathy, whilst the uncaring subscale was intended to represent a reduced concern for others (Essau et al., 2006). Both of these scales have been reliably linked to increases in aggression (Cardinale & Marsh, 2020). However, Ray et al. (2016) reported that the Callous subscale was more predictive of aggression severity whilst the uncaring subscale was a better predictor of age of onset of antisocial behaviour. Similarly, S. W. Hawes et al. (2014) observed that a shortened callous subscale was predictive of symptoms of oppositional defiant disorder, attention deficit hyperactivity disorder and conduct disorder, as well as anxiety. Contrastingly, a shortened uncaring subscale (equivalent to that of the ICU 11) predicted only lower levels of anxiety. Considering these findings S. W. Hawes et al. (2014) suggested that these differences may reflect a latent dimension of CU traits, with callous traits at one end detecting anxiety and antisocial behaviour, and uncaring traits detecting low anxiety and low prosocial behaviour at the other. The results of the prior chapter were consistent with this characterisation (see Table 5.9).

Despite these potential differences in the constructs assessed by ICU subscales, there is currently no literature assessing how each of these scales may be related to learning phenomena. Indeed, the systematic review in Chapter 2 indicated that no studies have investigated relationships between the ICU measure (and therefore its subscales) and extinction or reversal learning. The present study therefore raised an opportunity to test for such effects. However, the lack of prior research mean that it was not possible to anticipate how ICU 11 subscales would be associated with behaviour in the extinction task.

6.0.3 Hypotheses

The present sample was powered to detect the relationship between ICU scores and renewal estimated from participants of Western cultural background in Chapter 3 (section 3.7.2.2). This relationship was such that renewal was smaller at higher levels of CU traits. I therefore expected to observe this effect in such participants in this study, in addition to evidence for diminished extinction at higher levels of ICU total score. Further, after having demonstrated that the ICU 11 callous and uncaring subscales were consistent cross culturally in the present data set (see Chapter 5), I expected that these subscales would each validly demonstrate extinction and renewal effects across participants of both Western and East Asian ethnicity. I did not anticipate any differences in the effects of each of these subscales, as the systematic review in Chapter 2 did not indicate that data existed to support such differences. Further I did not expect ethnicity to moderate the effects of these ICU 11 subscales, due to their cross-cultural validity.

6.1 Method

6.1.1 Participants

This study utilised data from the sample whose recruitment was described in the previous chapter. It included 770 participants, 532 participants were paid community recruits from online platforms, and 238 were undergraduates from psychology courses at the University of Sydney, who were compensated with course credits. Across these recruitment sources, two streams of participants were selected. The first included 470 participants selected who identified as ethnically Western (Anglo or European). The size of this sample was determined by power analysis described in Chapter 4. The second included 300 participants identifying as ethnically East Asian (Chinese, Indian, Korean, Vietnamese, Filipino, Japanese, Malaysian, Thai, or other Asian). Four hundred and eighty-five participants were female, 280 were male, 5 were unspecified, and the median age was 26 (21 – 39 interquartile range). Thirteen of these individuals were not included in the following analysis as their task response data was not saved correctly by the task server.

6.1.2 Procedure

This study was conducted on data collected through the procedure described in Chapter 5. After completing the participant information statement and consenting to participate, participants completed the extinction learning task described below. This task was available as a browser based interactive web page. Participants were asked to utilize a mouse to input responses, and the task did not function on touch screen devices. Participants subsequently completed questionnaires in a randomised order.

6.1.3 Materials

This study utilized a range of measures described in previous chapters. These included the Inventory of Callous Unemotional Traits (ICU), the State Trait Anxiety Inventory (STAI), the Urgency, Premeditation (lack of), Perseverance (lack of), Sensation seeking (UPPS) impulsivity scale, which were described in Chapter 3. It also utilized the Individualism and Collectivism Scale (ICS) and Self-Reported Fluency of English Scale (SRFES) described in Chapter 5.

The extinction learning task described in section 3.7.1.4 was utilised in the present methodology. The structure of this task was reported in section 3.7.1.2 and is summarised in Table 6.1. This task was translated from the original MATLAB code to javascript to allow it to run online. Additionally, as a consideration for the online platform of this task, a manipulation check was included in the extinction task. This check was designed to determine whether subjects were responding to task instructions. It involved a 20 second period at the end of all other task phases. Before this period the following instructions were given: “Quentin has found a new pump but has stored enough water and will not need you to try to collect any more water for him. Please wait without lifting the lever until Quentin moves on.” Then, participants were shown a new coloured pump in a random context (day or night) which did not produce any water drops. Any responses in this period were recorded.

Table 6.1 *Structure of the Extinction task*

Condition		Training	Extinction phase	Renewal phase
Experimental	Reinforcement	4	Ø	Ø
	Context	A	B	A
Control	Reinforcement	4	Ø	Ø
	Context	A	A	A

Note. This table is taken from Chapter 3, Experiment 3. All numbers refer to the variable interval (VI) schedules (as number of seconds), and Ø indicates no reinforcement. Contexts A and B were each one of either night time or day time settings (visualized in Experiment 3 of Chapter 3)

6.1.4 Data Analysis

Analyses detail of analytical programming can be viewed at <https://osf.io/tfjqc/> (wherein the “Chapter6-ExtinctionAnalysis.html” file is located), along with validating plotting of statistical assumptions and anonymised data.

6.1.4.1 Task Data

Data from the task were binned and processed in the same manner as described for Experiment 3 of Chapter 3 (see section 3.7.1.5). Data from the additional manipulation period showed that large numbers of participants (265) responded during this period despite the instructions to the contrary. Therefore, this variable was effect coded as a binary variable (-.5 = non-responders, .5 = responders) identifying whether participants responded at all during this period and was treated as a covariate in the following analyses. Additionally, it was observed that some responders were able to respond at rates of up to 185 lever lifts per minute, which was much higher than anticipated. Therefore, individuals with excessively fast responding were identified as those whose maximum rate of responding was more than the upper quartile response rate plus 1.5 times the interquartile range in each period of analysed data (either that analysed for extinction or renewal effects). This variable was then investigated as a covariate in the following analyses. I also examined whether removal of individuals identified as excessive responders had any notable impact on results (results of this analysis can be viewed at the repository linked above). This procedure was designed to control for participants interacting with the task in a manner that I had not anticipated and was unable to observe as participation occurred remotely.

6.1.4.2 Variable Coding

ICU total scores were calculated in addition to totals of ICU 11 subscale items (Table 5.7). Categorical variables were effect coded, resulting in coded variables for ethnicity (Western = -.5, East Asian = .5), gender (male = -.5, female = .5), sample (undergraduate = -.5, community = .5). Categorical variables were constructed from SRFES and ICS responding using clustering methods (see section 5.1.4.2). This clustering was carried out for the SRFES as responses were highly skewed and so largely registered those fully confident in their English fluency, and those who were not. For its use as a predictor in this analysis, it was coded as (-.5 = low, .5 = high). The ICS measure was treated categorically as the factor structure informing its established subscales was not found to fit the present data well (Table 5.4), Three clusters were constructed from this variable, which were effect coded in two variables for the present analysis. The first variable “ICS high responder” was coded: ICS collectivist = -1 / 3, ICS high responder = 2 / 3, ICS individualist = -1 / 3. The second, “ICS individualist” was coded: ICS collectivist = -1 / 3, ICS high responder = -1 / 3, ICS individualist = 2 / 3. ICS collectivist was therefore the comparison group in this coding scheme. All continuous variables were converted to z scores for the following analyses.

6.1.4.3 Modelling

The following analyses were carried out as mixed effects models using the lmer package (Bates et al., 2015). Within-subjects effects were controlled by modelling random effects (intercepts) at the participant level of analysis. To test variables for their contribution to explained variance sequential regression methods were utilized, comparing model levels through analysis of deviance. To deal with any non-normality in their residuals, all models of ICU effects were repeated after exclusion of influential outliers. These outliers were identified as individuals

whose set of responses across relevant experimental periods led to a Cook's distance exceeding $4/n$, where n was the number of individuals in the dataset. These analyses were not reported unless they led to differences in reported effects (although their output can be viewed in the file linked above).

6.1.4.4 Covariates

Due to the larger number of measured individual difference features in this study compared to prior studies, I implemented a new strategy for identifying covariates for both extinction and renewal effects. This strategy utilized a series of sequential regressions. These models first tested for main effects of these covariates on responding, then considered whether covariates exhibited significant interactions with the effects of interest, beyond any identified main effects. The order of variable entry into blocks of these sequential regressions was designed to test whether more complex factors accounted for variance beyond more basic participant features.

In the case of cultural factors, I first tested recruitment source. This was considered the most important factor to control for, as study compensation varied across this variable, as did participants nationality and student status (See Table 5.2 for descriptions of these samples). Ethnicity was tested next, then self-reported English fluency, due to the relevance of these factors to prior extinction findings (see section 3.8). ICS responding groups were entered last as they were found to have validity issues in Chapter 5 (see section 5.3.2).

Other individual difference factors were considered in another sequential regression. This analysis first tested the effects of age, then gender, which have been considered as explanatory factors of behaviour in past studies of learning in CU traits (as reviewed in Chapter 2). Then, STAI scores followed by UPPS scores were tested as these variables had been considered as

covariates in Chapter 3 and were also found to have differing associations with ICU 11 subscales in Chapter 5. A further sequential regression tested the roles of several task response features. The first of these tested for differences in the responding of individuals who were able to achieve exceptionally fast rates of responding on the task interface, and the next made this comparison for individuals who responded during the manipulation check.

Finally, the unique variance predicted by each of the variables found to be significant predictors in the prior analyses was tested, initially entering task response factors, then individual difference factors, and finally cultural factors. Significant variables identified through this analysis were then controlled for in analysis of the effects of the ICU.

6.1.4.5 *Extinction*

Mixed effects models were fitted to responding across training, extinction, and renewal phases of the control condition in order to describe individual differences in responding to extinction manipulation. These models were fit as described in section 4.0.2.1. They included an interaction term for the difference in responding over bins of the experiment, dependent on whether bins were in training or not. This term captured average variation in responding due to the structure of the experiment. As the present hypotheses concerned the individual differences in responses to the extinction manipulation, I tested the interaction of individual difference variables with a term comparing average responding in training to the change in responding over bins of extinction. This “extinction” term was therefore relevant to both differences between training and extinction as well as change in responding over the course of extinction (as discussed in section 4.2).

To examine whether the present results were comparable to those reported earlier in this thesis, I initially tested whether the extinction effect was present as modelled in the power

analysis reported in Chapter 4. This power analysis was conducted on data from native English-speaking undergraduates. As the documented effect was culturally dependent in the original study (Chapter 3) I attempted to match the features of the sample analysed in this study, utilizing only participants of Western ethnic identity. As STAI was considered as a covariate in the original model, total scores on this variable were considered as moderators of the extinction term, as were total ICU scores.

Although the total ICU was used in the initial development of the extinction task, my subsequent findings indicated that this scale would not be a valid measure of CU traits in this sample. Therefore, to account for this validity issue a second set of models was fit to the present data, which examined the influence of each of the ICU 11 callous and uncaring subscales on responding. These models also took into account additional individual difference factors measured in this study and were fit to data from all participants. This analysis initially included a series of sequential regressions testing for effects of covariates, as described above. The main effects of ICU 11 callous and uncaring subscales (as totals of responses on relevant items) on overall responding were considered in a sequential regression, testing whether these variables explained variance beyond that of significant covariates. A further block of this regression tested the variance explained by the interaction between these ICU 11 subscale scores and the term for the extinction effect.

6.1.4.6 Renewal

The modelling of the renewal effect in this study was consistent with both the power analysis in Chapter 4 and Experiment 3 of Chapter 3. This analysis considered responding in the final two bins of responding in the extinction phases of the experiment as well as the two bins of the renewal phase during both control and experimental conditions. The model included terms

for the difference in responding between these conditions. The interaction of this difference with phase (extinction or renewal) was also tested to assess whether responding increased after the context change occurring during the experimental condition (indicating a renewal effect). The effect of individual differences on the size of this renewal effect was therefore determined by testing the moderation of this interaction term by relevant individual difference variables.

As in the extinction analysis, I initially examined renewal under the same model that was used in the power analysis reported in Chapter 4. This model considered the moderation of renewal by total ICU scores in participants of Western ethnicity. A further analysis tested the version of the ICU validated in Chapter 5 (the ICU 11). This analysis was conducted on the full sample and considered the moderation of the renewal effect by covariate factors measured in this study (as described above). It was comprised of a sequential regression which tested whether the ICU 11 callous and uncaring subscale scores predicted variation in overall responding beyond the effects of covariates. The final block of this model tested with these subscales predicted further significant variation in the renewal effect term.

6.2 Results

6.2.1 Extinction

The specified model of extinction responding was successfully fit to control condition data from Western participants (model fit is tabulated in Table 6.2). This model showed that the change in responding over bins was more negative in the extinction phase of the experiment than the training phase. STAI scores were not associated with differences in responding during extinction. Neither were ICU scores, although their effects approached significance, and described faster reductions in responding at higher levels of ICU scores.

Table 6.2 *Analysis of Total ICU Score as a Predictor of Extinction Responding in Western Participants*

Predictors	responses Per Min			
	Estimates	SE	CI	p
(Intercept)	26.89	0.6	25.71 – 28.07	<0.001
bin * phase	-0.69	0.02	-0.72 – -0.65	<0.001
bin * extinction phase * ICU	-0.02	0.01	-0.05 – 0.00	0.057
bin * extinction phase * STAI	0.00	0.01	-0.02 – 0.02	0.947
Random Effects				
σ^2			146.9	
τ_{00} participant			161.15	
ICC			0.52	
N participant			465	
Observations			11160	
Marginal R^2 / Conditional R^2			0.060 / 0.552	

Note. CI = 95% confidence interval. Marginal R^2 = variance accounted for by the fixed effects in the model, conditional R^2 = variance accounted for by entire model, σ^2 = model error term, τ_{00} = variance of random intercepts, ICC = intraclass correlation. The random effects term hierarchized data at the participant level of analysis.

Several covariates were identified as significant predictors of extinction responding and were therefore controlled during analysis of the effect of ICU 11 subscales. Statistical output of this analysis is depicted in Table 6.3. Covariates in this analysis (modelled in Block 1) included the excessive responder variable, which identified 12 individuals exhibiting extreme rates of responding. This group had higher average response rate and reduced their rate of responding more during extinction. Another covariate, sample source, showed an effect where community participants had a higher response rate than undergraduates overall and reduced their behaviour less during extinction. The 256 individuals who responded incorrectly to the task manipulation check exhibited less reduction of responding during extinction, as did individuals of Western ethnicity as compared to those of East Asian ethnicity. ICS individualists exhibited more reduction of responding during extinction than collectivists, whilst high responders showed no difference to collectivists.

ICU 11 callous and uncaring subscale scores did not affect overall responding during the control condition (Block 2 of the model, Table 6.3). However, the extinction term was more negative (reflecting greater reductions in behaviour in the extinction phase) at higher levels of the ICU 11 callous subscale (Block 3). Contrastingly, this effect was reversed for the uncaring subscale, which predicted less reduction in behaviour during extinction. The model's predictions for each of these subscales are visualised in Figure 6.1. There was no evidence that these effects were dependent on participant's ethnicity (Block 4).

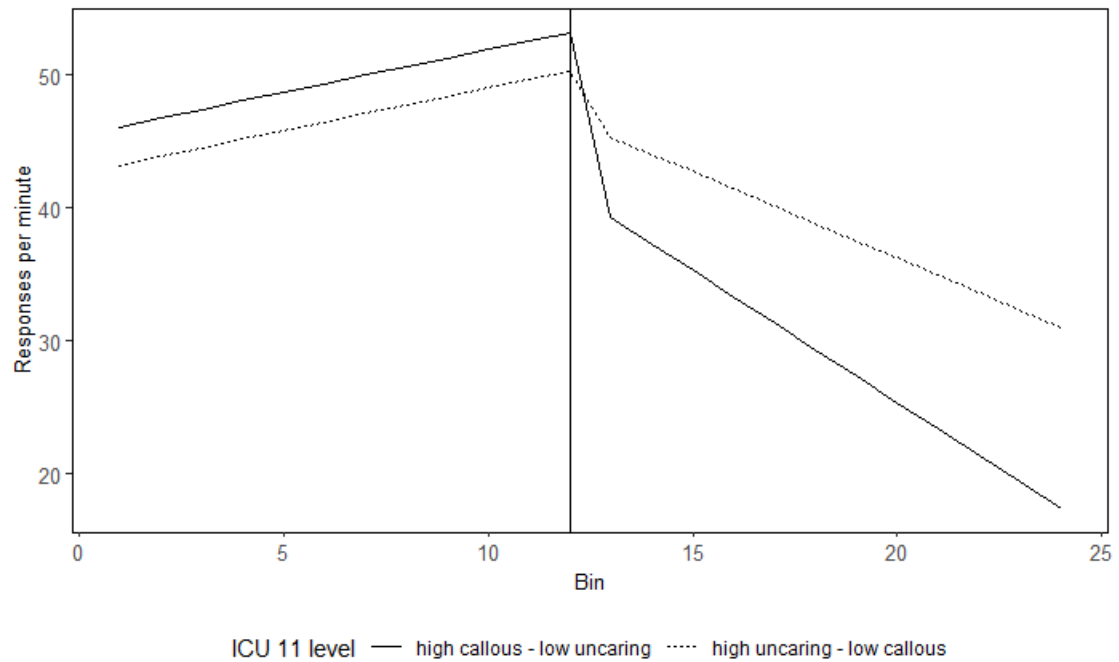
The divergence in extinction effects between ICU 11 subscales indicated that these effects were driven by non-shared variance in these scales. However, to statistically confirm this a post-hoc analysis was carried out in which shared variance in these scales was partialled out. To achieve this, each subscale (in z-score form) was regressed on the other, and residuals from these regressions were treated as 'partialled' variables (resulting effects on subscale overlap are visualised in Figure 6.2). These partialled variables were then entered in the model defined in Block 3 (Table 6.3), substituting the original subscale variables. Extinction effects consistent with original subscale scores were observed for both the partialled callous subscale ($\beta = -0.03$, 95% CI = [-0.06 - -0.01], SE = 0.01, $p = 0.007$) and the partialled uncaring subscale ($\beta = 0.06$, 95% CI = [0.04 - 0.09], SE = 0.01, $p < 0.001$).

Table 6.3 *Effect of ICU Subscales on Extinction Responding Controlling for Identified Covariates*

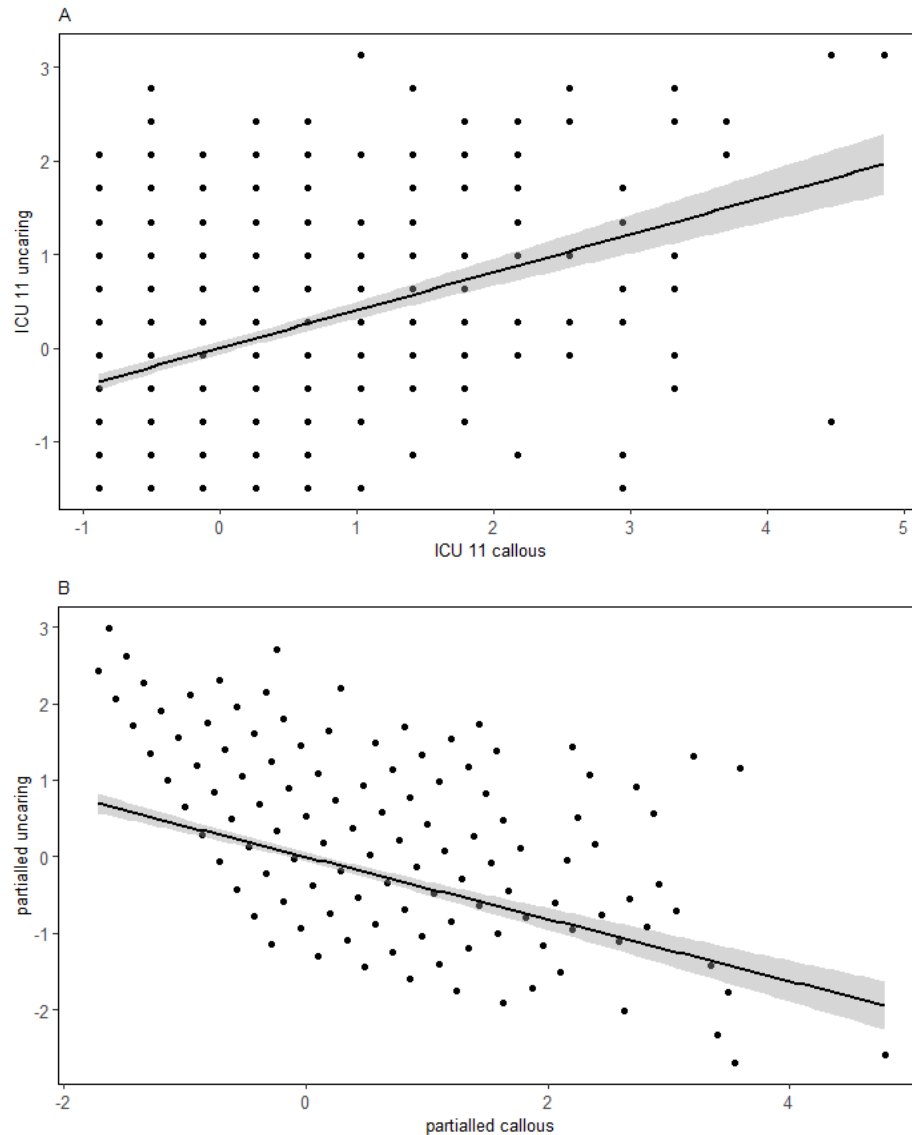
	Block 1				Block 2				Block 3				Block 4			
Predictors	Estimates	SE	CI	p	Estimates	SE	CI	p	Estimates	SE	CI	p	Estimates	SE	CI	p
(Intercept)	54.93	2	51.01 – 58.85	<0.001	55.02	2	51.10 – 58.94	<0.001	54.98	2	51.06 – 58.90	<0.001	54.98	2	51.06 – 58.90	<0.001
phase	19.7	0.79	18.16 – 21.25	<0.001	19.7	0.79	18.16 – 21.25	<0.001	19.7	0.79	18.16 – 21.25	<0.001	19.7	0.79	18.16 – 21.25	<0.001
bin	-0.48	0.03	-0.54 – -0.41	<0.001	-0.48	0.03	-0.54 – -0.41	<0.001	-0.47	0.03	-0.54 – -0.41	<0.001	-0.47	0.03	-0.54 – -0.41	<0.001
excessive responder	42.72	3.9	35.08 – 50.35	<0.001	42.84	3.89	35.21 – 50.47	<0.001	42.81	3.9	35.17 – 50.44	<0.001	42.81	3.89	35.17 – 50.44	<0.001
sample	2.76	1.05	0.71 – 4.82	0.009	2.59	1.06	0.51 – 4.66	0.015	2.75	1.06	0.67 – 4.83	0.01	2.75	1.06	0.67 – 4.82	0.01
bin * phase	-2.24	0.07	-2.37 – -2.11	<0.001	-2.24	0.07	-2.37 – -2.11	<0.001	-2.24	0.07	-2.37 – -2.10	<0.001	-2.23	0.07	-2.37 – -2.10	<0.001
bin * extinction phase	0.11	0.02	0.07 – 0.15	<0.001	0.11	0.02	0.07 – 0.15	<0.001	0.11	0.02	0.07 – 0.15	<0.001				
* incorrect													0.11	0.02	0.07 – 0.15	<0.001
manipulation check																
bin * extinction phase	-0.57	0.08	-0.72 – -0.41	<0.001	-0.57	0.08	-0.72 – -0.41	<0.001	-0.56	0.08	-0.72 – -0.40	<0.001	-0.56	0.08	-0.72 – -0.40	<0.001
* excessive responder																
bin * extinction phase	0.06	0.02	0.01 – 0.10	0.009	0.06	0.02	0.01 – 0.10	0.009	0.04	0.02	-0.01 – 0.08	0.094	0.03	0.02	-0.01 – 0.08	0.129
* sample																
bin * extinction phase	-0.15	0.02	-0.19 – -0.11	<0.001	-0.15	0.02	-0.19 – -0.11	<0.001	-0.16	0.02	-0.20 – -0.12	<0.001	-0.16	0.02	-0.20 – -0.12	<0.001
* ethnicity																
bin * extinction phase	0.01	0.02	-0.03 – 0.06	0.641	0.01	0.02	-0.04 – 0.06	0.663	0.04	0.02	-0.01 – 0.08	0.123	0.04	0.02	-0.01 – 0.08	0.127
* ICS high responder																
bin * extinction phase	-0.12	0.02	-0.17 – -0.08	<0.001	-0.12	0.02	-0.17 – -0.07	<0.001	-0.14	0.03	-0.19 – -0.09	<0.001	-0.14	0.03	-0.19 – -0.09	<0.001
* ICS individualist																
ICU 11 uncaring					-0.23	0.52	-1.26 – 0.80	0.661	-0.87	0.53	-1.91 – 0.18	0.104	-0.87	0.53	-1.91 – 0.18	0.104
ICU 11 callous					-0.75	0.53	-1.79 – 0.28	0.152	-0.19	0.54	-1.25 – 0.86	0.719	-0.19	0.54	-1.25 – 0.86	0.719
bin * extinction phase									0.08	0.01	0.05 – 0.10	<0.001	0.08	0.01	0.05 – 0.10	<0.001
* ICU 11 uncaring																
bin * extinction phase									-0.06	0.01	-0.08 – -0.04	<0.001	-0.06	0.01	-0.08 – -0.04	<0.001
* ICU 11 callous																
bin * extinction phase													-0.01	0.02	-0.05 – 0.04	0.807
* ICU 11 uncaring *																
ethnicity																
bin * extinction phase													-0.02	0.02	-0.06 – 0.02	0.375
* ICU 11 callous *																
ethnicity																
Random Effects																
σ^2		164.47				164.47				163.98				163.99		
τ_{00}		165.69	participant			165.37	participant			165.63	participant			165.56	participant	
ICC		0.5				0.5				0.5				0.5		
N		757	participant			757	participant			757	participant			757	participant	
Observations		18168				18168				18168				18168		
Marginal R ² /		0.149 / 0.576				0.152 / 0.577				0.154 / 0.579				0.154 / 0.579		
Conditional R ²																
Model comparison																
Deviance		146695.01				146691.701				146638.861				146637.681		
χ^2						3.309				52.84				1.18		
Δ df						2				2				2		
p						0.19				< .001				0.554		

Note. See Table 6.2 for statistic annotations. Additional statistics include χ^2 = change in deviance between models, Δ df = change in degrees of freedom.

Figure 6.1 *Model of Extinction Responding for Different Levels of ICU 11 Callous and Uncaring Subscales*



Note. This figure depicts the rate of responding during the control condition as predicted by Block 3 of the model presented in Table 3, the displayed values are therefore an aggregate and do not represent observed responses. The extinction phase begins at the vertical line at bin 12. High and low levels are the maximum and minimum observed values for the ICU 11 subscales. All covariates in the model are set to their mean values for predictions depicted here.

Figure 6.2 *ICU Subscale Correlations (Original and Partialled)*

Note. Displayed values are z-scores. Fitted lines represent linear regressions with a surrounding 95% confidence interval shaded in grey. Plot A shows original subscale z – scores, and plot B shows the relationship between these variables after partialling out shared variance.

6.2.2 Renewal

Our initial test of the responding of participants of Western ethnicity in the data analysed for renewal is shown in Table 6.4. Whilst responding increased more in the renewal phase of the experimental condition than in the control condition, it was not evident that this effect varied across total ICU scores.

Table 6.4 *Effect of Total ICU on Renewal Responding for Western Participants*

Predictors	Responses Per Minute			
	Estimates	std. Error	CI	p
(Intercept)	20.24	0.66	18.96 – 21.53	<0.001
condition	4.62	0.4	3.84 – 5.39	<0.001
condition * phase	7.41	0.81	5.83 – 8.99	<0.001
condition * phase * ICU	0.00	0.81	-1.59 – 1.60	0.996
Random Effects				
σ^2		145.23		
τ_{00} participant		181.81		
ICC		0.56		
$N_{\text{participant}}$		465		
Observations		3720		
Marginal R^2 / Conditional R^2		0.026 / 0.568		

Note. See table 6.2 for statistic annotations.

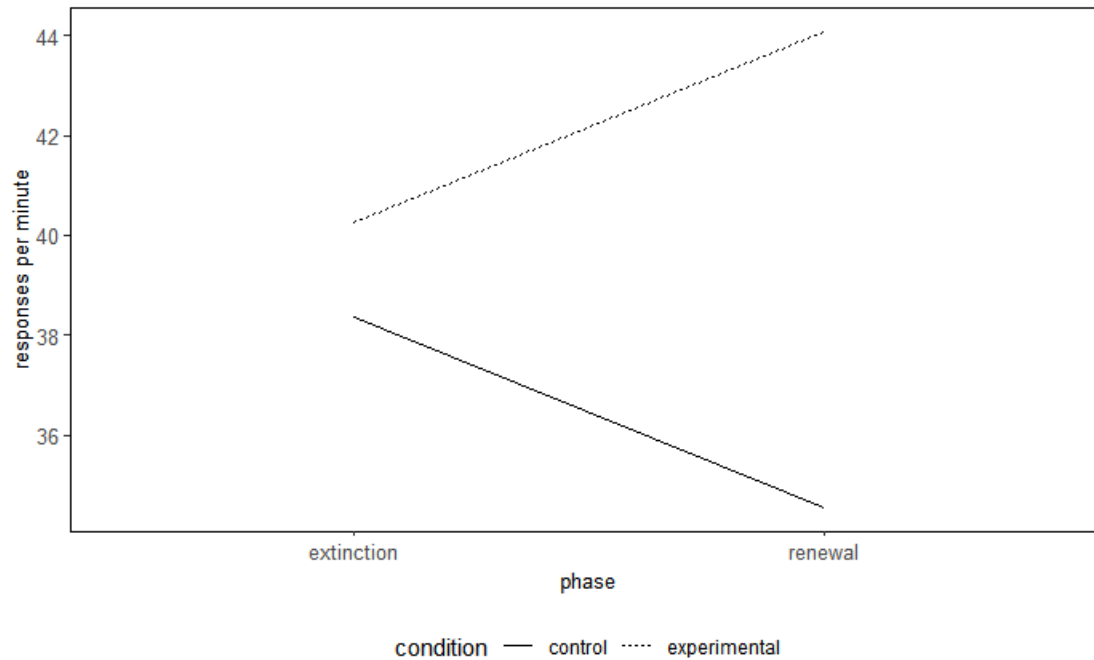
In the analysis of the full data set, it was evident that responding increased more between extinction and renewal phases in the Experimental Condition than the control condition.

Modelling of this effect is presented in Table 6.5, with the predicted effect of phase and condition visualised in Figure 6.3. This modelling indicated that this effect was larger in participants of East Asian ethnicity relative to Western ethnicity, and for participants with high English fluency compared to low language fluency (SRFES group). It was also evident that the overall rate of responding was higher in 20 individuals classed as extreme responders (Table 6.5, Block 1). Accounting for these covariates, it was evident that the overall responding was lower at higher levels of the ICU 11 callous subscale (Block 2). However, ICU 11 subscales were not associated with any differences between phases across control and experimental conditions (Block 3).

Table 6.5 *Effect of ICU Subscales and Covariates on Renewal Responding*

Predictors	Block 1				Block 2				Block3			
	Estimates	SE	CI	p	Estimates	SE	CI	p	Estimates	SE	CI	p
(Intercept)	39.59	2.21	35.26 – 43.92	<0.001	39.32	2.2	35.01 – 43.63	<0.001	39.32	2.2	35.01 – 43.63	<0.001
condition	5.71	0.33	5.06 – 6.36	<0.001	5.71	0.33	5.06 – 6.36	<0.001	5.71	0.33	5.06 – 6.36	<0.001
high responder	41.39	4.41	32.74 – 50.05	<0.001	40.83	4.4	32.21 – 49.45	<0.001	40.83	4.4	32.21 – 49.45	<0.001
condition * phase	7.61	0.84	5.98 – 9.25	<0.001	7.61	0.84	5.98 – 9.25	<0.001	7.58	0.84	5.93 – 9.23	<0.001
condition * phase	8.23	1.72	4.85 – 11.60	<0.001	8.23	1.72	4.85 – 11.60	<0.001	8.41	1.75	4.98 – 11.83	<0.001
* ethnicity												
condition * phase	8.02	2.06	3.97 – 12.06	<0.001	8.02	2.06	3.97 – 12.06	<0.001	8.22	2.08	4.13 – 12.31	<0.001
* SRFES												
ICU11 Uncaring					-0.12	0.55	-1.20 – 0.95	0.821	-0.12	0.55	-1.20 – 0.95	0.821
ICU11 Callous					-1.52	0.55	-2.59 – -0.44	0.006	-1.52	0.55	-2.59 – -0.44	0.006
condition * phase									-0.69	0.74	-2.13 – 0.76	0.353
* ICU11 uncaring												
condition * phase									0.59	0.74	-0.87 – 2.04	0.428
* ICU11 callous												
Random Effects												
σ^2		167.73				167.73				167.76		
τ_{00}		171.35	participant			169.39	participant			169.39	participant	
ICC		0.51				0.5				0.5		
N		757	participant			757	participant			757	participant	
Observations		6056				6056				6056		
Marginal R ² /		0.097 / 0.553				0.103 / 0.554				0.103 / 0.554		
Conditional R ²												
Model comparison												
Deviance		49878.784				49869.014				49867.923		
χ^2						9.77				1.091		
Δ df						2				2		
p						0.008				0.58		

Note. See Table 6.2 and 6.3 for statistic annotations.

Figure 6.3 Mean Predicted Values for Renewal Responses

Note. Visualised values are those predicted by the model in the second block of the sequential regression conducted on the full sample of this study (Table 6.5). Like Figure 6.1, these values represent aggregates rather than observations.

6.3 Discussion

Overall, the present results were largely inconsistent with hypotheses developed in this study. ICU total score was not a significant predictor of variation in extinction responding in Western participants. This finding meant that the association between the total ICU scores and extinction documented in Chapter 3, and modelled in Chapter 4, was inaccurate. However, it was evident that ICU 11 subscales had predictive capacity for extinction responding. Unexpectedly, these effects were opposite for callous and uncaring subscales. Individuals high on the uncaring subscale showed less reduction of behaviour during extinction, similar to results for total ICU scores described in Chapter 3. Conversely, responding decreased faster during extinction for individuals with high levels of callous subscale scores, a finding inconsistent with any prior results. The divergent effects of ICU 11 subscales provide an explanation for why the total ICU,

which aggregates these subscales, does not show consistent associations with extinction behaviour. I also did not observe the relationship between renewal effects and ICU total scores that had been estimated in the power analysis in Chapter 4. Similarly, ICU 11 subscale scores showed no relationship with renewal. As this sample was designed to be confirmatory for this effect, these findings indicate that associations between CU traits and renewal are not detectable using this methodology.

6.3.1 ICU Effects

With the analysis of the present data set I have made a comprehensive examination of responding to the extinction task developed in this thesis (in non-clinical populations). Total ICU scores were repeatedly observed to predict reduced extinction in Chapter 3, and the modelling of Chapter 4 suggested that such effects would be highly statistically significant in the present sample of Western participants. Instead, the observed effect was in the opposite direction and was not statistically significant. This could have been due to experimental error introduced by the online platform of the present study. This error was evident as some individuals' rates of responding on this platform substantially exceeded those observed in person. Additionally, the manipulation check was failed by a large proportion of participants. However, these factors do not account for the effects observable in this study. It was evident that the ICU 11 subscales were significantly related to extinction behaviour, even after controlling for these potential sources of measurement error.

The effects of ICU subscales explain the unreliability of the ICU total score. These subscale effects were in diverging directions, a result also seen after controlling for overlap in the variance of these subscales. These effects therefore indicate that the ICU total score was an invalid measure, as it aggregated over meaningful heterogeneity occurring within the ICU scale.

Differences in the relative levels of the traits measured by these subscales may have occurred across the samples measured here and in Chapter 3. These differences could therefore explain the absence of total ICU scores in this sample.

Diverging ICU 11 subscale effects on extinction are a novel and informative observation, as they have not been previously examined. It was evident that faster extinction occurred in individuals with high callous subscale scores (and low uncaring scores), whilst slower extinction occurred for individuals with high uncaring scores (and low callous scores). The effect of the uncaring subscale was consistent with hypotheses that CU traits are linked to poor extinction learning. Similar learning deficits are a component of the CU trait aetiology in multiple theoretical frameworks (Blair et al., 2005; Hamilton et al., 2015; Moul et al., 2012), as introduced in Chapter 1. In contrast, the callous subscale was associated with faster reductions in behaviour during extinction. In the narrative context of this task such behaviour may reflect a less prosocial attitude towards the quokka character. However, there is no indication that these individuals are less likely to help the character on average when considering both the training and extinction phases. Therefore, as this task was intended as an operationalisation of extinction learning rate, this result may reflect faster rates of such learning at higher levels of callous ICU 11 scores. Such a deviation in extinction behaviour across callous and uncaring ICU subscales has not been previously documented. Furthermore, the effects of unique variance in these scales adds weight to suggestions that a latent factor exists across individuals with high ICU scores, with callous and uncaring subscales of the ICU capturing different components of this factor (S. W. Hawes et al. 2014).

6.3.2 Renewal

In contrast to the results for extinction, ICU measures had no effect on renewal of extinguished behaviour. This indicates that the true association between ICU scores and renewal effects in this task is not as large as that used to justify the sampling strategy for this study. It is therefore unlikely that the current methodology can detect renewal effects. Indeed, such effects have never been observed and may not exist. However, as described in section 3.0.3, renewal effects are a useful means of tying task behaviour to extinction learning processes. Therefore, in the absence of renewal effects, the observed effects of ICU 11 subscales may be explicable through non-learning processes. Considering this outcome, future studies into renewal and CU traits may be warranted if they can overcome any factors that may have limited the sensitivity of the current experiment to such effects, some of which may be indicated by covariates in this study.

6.3.3 Covariate Effects

The present results indicated that cultural factors affected renewal responding in this study. Ethnicity predicted variation in both extinction and renewal behaviour, and the SRFES English fluency measure predicted renewal effects. Higher English fluency was associated with greater renewal, potentially indicating that this effect was contingent on task comprehension. However, as ethnicity determined task responding in both periods of interest, it was evident that other undetermined cultural features were potentially causal to task behaviour.

Although relevant to task responses, there was no evidence that cultural features moderated the relationship between ICU 11 responses and extinction behaviour. As the ICU 11 was found to be cross culturally consistent in Chapter 5, these results indicate that cultural differences in the interpretation of the ICU could account for culturally dependent effects

documented in Chapter 3. Whilst it is evident that I was able to control for differences in interpretation of the ICU in this study, cultural differences may have introduced measurement error to data from the behavioural task.

Cultural factors may have even contributed more to task responses than CU traits. This was evident in that ICU responses had less predictive value for extinction responding than responses to the ICS (based on standardized effect sizes). Whilst I cannot explain the dependence of responding in the present task on cultural factors empirically, I can suggest that the narrative content of the task may have introduced this cultural variation.

This narrative context was intended to motivate responding as a prosocial behaviour. This acted to contrast expected behaviour for individuals with extinction learning deficits and behaviour for antisocial individuals. However, the prosocial context may have led to the greater rates of responding observed during extinction in individuals with collectivism-oriented ICS responses (compared to those with individualism-oriented ICS responses). As an alternative account of this effect, higher levels of collectivist values may be associated with greater effort from individuals as research subjects, leading to maintenance of responding throughout the task. Consistent with this potential role of participant motivation; community participants who received monetary compensation showed less extinction than undergraduates whose participation contributed to their course requirements.

Regardless of the cause of effects of the ICS, their presence highlights a conceptual issue with the task. Specifically, that a highly prosocial individual may respond similarly to a high CU individual with poor extinction learning. Such observations would theoretically remove any extinction variation associated with the full ICU dimension. As a result, this task may be more

useful in an antisocial population, wherein variation in the task may be attributed to extinction learning processes, rather than prosocial behaviour.

6.3.4 Conclusion

The results of this study show that the ICU total score is an ineffective measure for determining relationships between CU traits and extinction responding. This appears to be due to conflicting effects of ICU subscales on extinction. These subscale effects add weight to growing evidence that the callous and uncaring ICU subscales capture a latent CU trait dimension (S. W. Hawes et al., 2014). Individuals high on the ICU 11 uncaring scale displayed the hypothesised relationship between CU traits and extinction learning. Contrastingly, those with high callous subscale scores exhibited faster extinction. It was noted that the absence of renewal effects, as well as the presence of cultural effects on task responding created some caveats for the present results. However, due to the sampling design of this study, a high degree of confidence can be attributed to the observed relationships between extinction behaviour and ICU subscales. The present results therefore comprised a significant contribution the current goal of examining extinction learning in CU traits.

7 General Discussion

Contents

7.0	Extinction and CU Traits	213
7.0.1	Aetiological Variants of CU Traits.....	214
7.0.2	CU Trait Variants and the ICU	215
7.0.3	CU Variants and Extinction Learning	218
7.1	Qualifying this Interpretation.....	221
7.2	Renewal and CU traits	222
7.3	CU Traits in the General Population.....	224
7.3.1	Affective deficits are not Deterministic to Disordered Behaviour	225
7.3.2	Importance of Dimensional Measures	226
7.4	Conclusion	227

This thesis was conducted in order to examine the hypothesis that CU traits are linked to deficits in extinction learning. The studies described here therefore aimed to examine existing evidence for this effect, and to provide validating empirical evidence to expand this existing literature. This initial aim was achieved through the systematic review reported in Chapter 2. I thereby observed that the reviewed studies were inconsistent in their methodology and results. This finding stressed the need for validating observations of variation in extinction learning across individual differences in CU traits. I proposed that a useful way of achieving this was by testing a novel behavioural hypothesis. Specifically, that renewal of extinguished responding would be diminished in individuals with high CU traits.

To test this hypothesis, I developed an extinction learning task over a series of six studies (reported in Chapter 3). Although these studies contributed to my aim of measuring relationships between extinction effects and CU traits, I was unable to substantiate these findings with validating observations of corresponding renewal effects. Fortunately, unexpected cultural influences in that study led me to conduct further investigation of the ICU measure of CU traits. A subsequent large-scale study showed that the full ICU measure was not consistent across cultural factors present in this sample (see Chapter 5). However, a shortened version of this measure, the “ICU 11”, with two subscales, “callous” and “uncaring” was found to have good consistency. In further analysis (Chapter 6) it was also revealed that variation across these subscales could account for unreliability in the total effect of the ICU. Although this study also failed to show any dependence of renewal on CU traits, the effects of ICU 11 subscales on extinction learning have never been previously observed. Further, they have highly novel implications for the role of learning processes in CU traits.

In the present chapter I will synthesise these findings and consider their implications. The different associates of the uncaring and callous subscales of the ICU 11 play a major role in this synthesis. I will emphasise that these associates are comparable to the differing characteristics of CU traits derived from either genetic or environmental processes. In this context I will argue that the learning effects seen in my results have implications for aetiologically distinct variants of CU traits. I will also qualify this discussion with the absence of renewal effects in the present research, and my use of samples from the general population. This discussion will consequently justify novel hypotheses concerning the aetiology of CU traits.

7.0 Extinction and CU Traits

This thesis constitutes the first evidence for variation in instrumental extinction learning concurrent with responses to the ICU measure. In Chapter 2 I identified that the variation of extinction or reversal learning across psychopathic traits has primarily been investigated using total scores on measures such as the APSD (see Budhani & Blair (2005)). This is important to note as these studies are unable to differentiate the roles of affective and behavioural components of these measures. Specifically the APSD (which includes impulsivity and grandiosity components) predicted higher longitudinal risk for conduct problems, aggression, and substance use than the ICU (Andershed et al., 2018). Only a small number of studies specifically correlated learning effects with the affective (CU trait) component of psychopathic traits (Fisher and Blair (1998), O'Brien et al. (1994), Mitchell et al. (2006) and X. Chen et al. (2020)). However, this thesis showed that learning effects diverged along subscales of the ICU 11. This result indicates that there is meaningful heterogeneity in learning within the CU trait construct, which has not been documented previously. This heterogeneity may contribute to the inconsistency in findings documented in Chapter 2 (see Figure 2.3). Therefore, the relevance of extinction learning to CU

traits cannot be fully understood until this heterogeneity has been further investigated. Additionally, this thesis shows that is imperative for such investigations to consider cultural variables, which I have found to influence both responses to the ICU and to the extinction task utilised here.

7.0.1 Aetiological Variants of CU Traits

Extinction learning in this thesis appeared to have diverging associations with the callous and uncaring subscales of the ICU 11. As prior learning studies have never utilised ICU subscales, this unexpected result has implications for theories that were not considered in the development of this thesis. Specifically, I will tentatively argue that this callous-uncaring scale reflects the primary and secondary aetiological variants of CU traits (collectively referred to as CU trait variants (Craig et al., 2021)), and that the present extinction findings have novel implications to the study of these variants.

In the case of the primary variant of CU traits, affective empathy deficits are primarily attributed to biologically derived processes. For example, epigenetic downregulation of oxytocin at birth has been seen to convey risk for the development of CU traits longitudinally to individuals aged 13 (Cecil et al., 2014). This view is concordant with evidence that genetic factors account for the stability of CU traits longitudinally (Fontaine et al., 2010). These genetic factors are thought to manifest as CU traits due to their impairing impact on social attention (Dadds et al., 2011). This impairment prevents encoding of the social responses that typically inhibit antisocial behaviour (in the “inhibitory model of moral socialisation” Kimonis et al., 2008) (Craig et al., 2021).

In addition to this genetic aetiology, CU traits can also be associated with environmental factors such as parental harshness (reviewed by Craig et al., 2021). This association can be

explained by a secondary aetiology where maltreatment leads to the development of avoidance of others' emotional behaviour (Porter, 1996). This avoidance may act to impair encoding of social responses to antisocial behaviour, thereby leading to the same failure of moral socialisation that occurs in the primary CU variant (Kimonis et al., 2008). Consistent with this explanation, secondary CU traits are typified by physiological hyper-arousal (Kostas A. Fanti et al., 2018) and avoidance of distressing stimuli (Sharf et al., 2014). Significantly, this variant does not exhibit the exhibit an endogenous affective empathy deficit, as individuals with secondary CU traits show intact recognition of other's emotions (Dadds et al., 2018). Overall, the features of this variant contrast those of primary CU traits, which are characterized by physiological hypo-arousal and poor awareness and responsivity to others emotion (Craig et al., 2021).

7.0.2 CU Trait Variants and the ICU

The trait features ascribed to primary and secondary CU traits are comparable to observed associates of the ICU callous and uncaring subscales. Hawes et al. (2014) reported that higher callous scale scores were linked to greater anxiety and disordered conduct (in the short form ICU 12²) This scale was also related to increased externalising behaviour by Colins et al. (2016). Similarly Ray et al. (2016) observed that this subscale (in the 24 item ICU measure) predicted increased severity of criminal behaviour. Contrastingly, Hawes et al. (2014) saw that higher uncaring subscale scores were uniquely associated with lower anxiety and lower prosocial behaviour and had no relationship with disordered conduct. Furthermore, Ray et al. (2016) found that this scale was associated with earlier onset of antisocial behaviour.

² Includes the parent/teacher report versions of items listed in Table 5.5 (comprising the ICU 11) with an additional item "Does not show emotions" from the ICU 24 Unemotional subscale.

Existing observation of callous and uncaring ICU subscales are therefore consistent with the correlations that I have observed for the ICU 11 (Table 5.9). Scores on the callous subscale were associated with higher trait anxiety, as well as impulsivity. Callous scores also showed stronger associations with measures of aggression than uncaring scores, which were not related to impulsivity or anxiety, and showed a stronger negative relationship with collectivism – a construct conceptually related to prosocial attitude (Triandis & Gelfand, 1998). The corresponding results observed here and in prior studies show that unique associates of the ICU uncaring scale are qualitatively comparable to the early onset, low empathy, low anxiety characteristics of primary CU traits. The associates of the Callous scale are similarly relatable to the high anxiety, socially dysregulated profile linked to secondary CU traits (as reviewed in Craig et al., 2021).

The varying associates of the ICU callous and uncaring subscales have been linked to the measurement properties of their items (Cardinale & Marsh, 2020). Those expressing callous but not uncaring traits were individuals who endorsed most of the items on the ICU 11, as callous items were positively coded (e.g., “I do not feel remorseful when I do something wrong”) and uncaring items were negatively (reverse) coded (“I feel bad or guilty when I do something wrong”). These individuals appear to characterize themselves as unempathetic but also emotional about their impact on others. This characterization aligns with a concept of secondary CU traits where anxiety drives deficits in empathy (Kimonis et al., 2008). Conversely, individuals high on uncaring but not callous items were individuals who predominantly disagreed with items on the scale. These individuals appear unwilling to admit to being callous or antisocial, but also unlikely to agree that they care about other’s emotions. This is consistent with the conceptualization of primary CU traits, where individuals are interpersonally disengaged and

have poor affective awareness (Kimonis et al., 2008). Therefore, in addition to the empirical parallels between the ICU uncaring and callous subscales and primary and secondary CU traits, I would propose that the content of the subscale items appears to possess face validity as measures of these variants.

Despite these indications, no prior research has used ICU subscales as a means of identifying individuals belonging to either CU trait variant. In existing research this has typically been achieved through clustering or mixture modelling analysis of CU trait measures and anxiety or maltreatment data. This process identifies groups with high CU traits, differentiated on levels of anxiety or maltreatment (Craig et al., 2021). However, no consistent practice exists for generating these clusters. Studies vary in the algorithms utilized, the number of calculated clusters/groups, and the use of additional cut-off criteria for defining high CU trait individuals, leading to unexplained differences in observed proportions of CU trait variants across samples of similar populations (Craig et al., 2021).

Problems with these operationalisations would have made their use in this thesis inappropriate. Further, use of these categories would have limited my investigation of associations between CU trait variants and other factors measured in this thesis. If CU trait data was treated as categorical, a loss of predictive power would occur as information is discarded during this discretisation (DeCoster et al., 2011; MacCallum et al., 2002). However, most significantly, categorization introduces several assumptions to an analysis: the categorical nature of the measured construct, the reliability of the measure, and the proportions of categories present in the population (DeCoster et al., 2009). In the case of primary and secondary psychopathy, the construct is only recently developed, and has no established categorical structure (Craig et al., 2021). These issues with identifying CU variants indicate that an

alternative analysis of their presence in my data would not improve the validity of the results I have reported.

7.0.3 CU Variants and Extinction Learning

In Chapter 3, I observed associations between total ICU scores and extinction learning. However, in Chapter 6, in a well powered sample, it was revealed that this effect was dependent on subscales of the ICU. After controlling for overlap in these scales, it was evident that individuals with high uncaring and low callous scores exhibited slower extinction, whilst individuals with high callous and low uncaring scores extinguished faster (see section 6.2.1). Variation observed for the uncaring subscale was consistent with the hypothesis developed in this thesis for CU traits as a whole: that they would associate with slower extinction of instrumental responses. This hypothesis was developed in relation to theories designed to link genetically derived variation in learning processes to the affective and interpersonal features of CU traits (Blair, 2017; Blair et al., 2005; Moul et al., 2012). As primary CU traits are thought to represent this genetic pathway to CU traits (Craig et al., 2021), the slow extinction hypothesis could be applied to this CU variant. The extinction result documented in Chapter 6 is therefore consistent with the concept that the ICU's uncaring dimension reflects the primary variant of CU traits.

Unexpectedly, individuals scoring highly on the Callous subscale exhibited faster reductions in responding during extinction. The unanticipated nature of this effect meant that the task was not designed to account for the influence of antisocial attitude on this form of behaviour. The greater reduction in responding observed for individuals high on the callous subscale could therefore reflect a less prosocial attitude in the narrative context of the extinction task. However, these traits were unrelated to overall levels of responding, indicating that there were no overall differences in motivation to engage with the task. It is therefore important to

consider that faster reductions in behaviour during extinction in the task may be explained by variation in extinction learning, especially given that this learning was the construct operationalised by the task.

If ICU callous and uncaring subscales do detect primary and secondary CU traits, then the relationship between these subscales and extinction responses may have relevance to the aetiology of these variants. Thus far, existing literature on extinction in CU variants has been carried out in fear learning paradigms. For example, Fanti et al. (2020) reported that secondary CU traits predicted less reduction of neurological threat responses (activity in the right amygdala measured through fMRI) during extinction, as compared to individuals classed as having primary CU traits or no CU traits. This result has also been observed for individuals high in intolerance of uncertainty (Morris et al., 2015).

Intolerance of uncertainty describes a trait level bias to interpret threat in uncertainty and is a transdiagnostic feature of anxiety (Grupe & Nitschke, 2013). A proposed process underlying this trait is that the failure to perceive predictors of events (uncertainty) is a driver of anxiety (Grillon, 2002), which in turn has an interfering effect on processing of predictable signals of threat (Grupe & Nitschke, 2013). This process has strong parallels to the proposed aetiology of secondary CU traits, where anxiety and hyperarousal drive disengagement from others' emotional behaviour, impairing moral socialization (Craig et al., 2021; Kimonis et al., 2008). The presence of such an uncertainty process could account for the inappropriate maintenance of threat responses observed by Fanti et al. (2020). Indeed, Brazil et al. (2017) observed that higher CU traits predicted greater estimates of uncertainty in models of amygdala responses to threat conditioning, although CU traits in this study were not differentiated into primary and secondary variants.

In contrast to these observations of impaired threat extinction, the data reported in this thesis concerns extinction of instrumental responding. It is possible to explain variation in this behaviour using a recent model developed by Moul et al. (2021). This model simulates uncertainty through a “learning window” term, which estimates the amount of historical information utilized during prediction of events. Moul et al. (2021) found that modelling of a wide learning window could predict reversal learning deficits seen in psychopathic traits (shown in Budhani et al. (Budhani et al., 2006)). In the context of the extinction task in the present study, slower extinction would be expected by individuals with a large learning window. This is because even after rewards are omitted, behaviour continues to be informed by a long “memory” of rewarded responses. Such behaviour was observed for individuals with higher uncaring traits than callous traits. As I have discussed, this response pattern was that which I hypothesized for CU traits, informed by studies such as Budhani et al. (Budhani et al., 2006). Therefore, in the context of the stated parallels between the uncaring subscale and primary CU traits, the present extinction results could suggest that stable predictive dispositions underpin primary CU traits.

The relevance of predictive stability to CU trait variants can be consistently applied to my observations regarding ICU subscales. The presence of a short learning window could explain the fast extinction of instrumental responding I observed in individuals high on the callous subscale in the results of Chapter 6. With a short learning window individuals’ behaviour would not take in to account a long history of instrumental reward after encountering the extinction contingency, thus quickly extinguishing. The extinction behaviour documented for this subscale is therefore consistent with the discussed evidence pointing to greater predictive uncertainty in the secondary CU traits (K.A. Fanti et al., 2020).

The relevance of predictive uncertainty to secondary CU traits fits with existing knowledge regarding the aetiology of these traits. Further, existing evidence for a maltreatment based mechanism for this aetiology have been unable to distinguish between the role of maltreatment and the effects of any pre-existing sensitivity to threat (Dadds et al., 2018). Therefore, it is possible that dispositionally high predictive uncertainty may contribute to anxiety responses (as described by Grillon (2002)) that are thought to impair empathic behaviour in secondary CU traits (under the model discussed by Craig et al. (2021) and Kimonis et al. (2008)). In concert with my suggestion that primary CU traits may be related to a stable predictive disposition, the current discussion points to latent trait dimension of predictive certainty that influences social traits. For highly uncertain individuals, others' emotions may create pathologically anxious and avoidant responses, where-as for those with an excessively stable predictive tendency, others' emotions do not garner appropriate attention.

7.1 Qualifying this Interpretation

I have argued for a hypothesis that learning processes (or predictive dispositions) may drive the emotional reactions characterising aetiological variants of CU traits. However, I cannot provide direct evidence of such a relationship. Furthermore, the magnitudes of associations between CU traits and extinction learning documented in Chapter 6 were small. It is therefore unclear whether learning processes would be influential enough to account for the variations in social behaviour associated with CU traits. None the less, aetiologies of CU traits often implicate non-additive processes. For example, Larsson et al. (2008) observed that early childhood behaviour problems could drive parental responses that caused further conduct problems. Such a process indicates that the basal causes of CU traits can initiate a positive feedback cycle of antisocial behaviour. Concordantly, Dadds et al. (2011) hypothesised that diminished attention to

emotional faces may inhibit social learning. Thereby, over the long sequence of social interactions occurring throughout development, these attentional features could cause broader the empathy deficits seen in CU traits. It is not presently known how excessively stable or unstable predictive tendencies would manifest in social development. However, given the cumulative effects of individual differences across development, the potential of such tendencies to manifest in a broader suit of behaviours should be considered.

It should also be acknowledged that this discussion assumes learning deficits may be a dispositional feature causal to social behaviour (Blair, 2017), whereas other researchers have attributed behavioural features such as response perseveration (which are cognate with poor extinction behaviour) to the environments experienced by individuals with CU traits (D. J. Hawes et al., 2014). Further, the implications of the present learning results are novel and may be a productive source of experimental hypotheses. However, much of this discussion relies on a correspondence between the ICU callous and uncaring subscales and CU variants, which is not empirically established. A study directly investigating the contribution of maltreatment, anxiety, and genetic variables to variation on these subscales would be informative, as it would substantiate its links to the ontogeny of primary and secondary CU variants. Whilst the present results cannot confirm the relationship between ICU subscales and aetiological variants of CU traits, they reinforce the importance of examining these subscales as continuous dimensions.

7.2 Renewal and CU traits

A further qualification of the current extinction results is the absence of corresponding individual differences in renewal. The extinction task examined in this thesis was found to be a novel methodology capable of measuring context dependent renewal of extinguished behaviour. This was a significant finding as existing studies of this effect in humans are limited (Crossley et

al., 2014). However, this methodology was unable to link individual differences in this behavioural phenomenon to CU traits. Renewal was considered in this thesis for two reasons. Primarily, because renewal is an identifying attribute of extinction learning, as described by Bouton et al. (2021). These authors define renewal as the return of a behaviour (such as reward seeking) after extinction learning (about the absence of the behaviours' goal) has occurred. This effect indicates that extinction learning is strongly linked to the context in which it occurs; if an organism leaves this context, it is imperative that it return to behaviours that have previously garnered rewarding outcomes (Bouton et al., 2021). Despite this imperative, the neurological capacity governing this process is subject to individual variation. If encoding of context is impaired during extinction, inhibitory learning will be delayed, and, will be less contextually sensitive (Wilson et al., 2014). These processes therefore signify that any deficit in extinction learning associated with CU traits should be accompanied by a deficit in renewal.

In the absence of renewal effects, existing studies of CU traits lack a key validating observation to support attributions of behaviour to extinction learning processes. This is significant because existing frameworks for understanding the cognitive mechanisms ontogenic to CU traits regularly reference extinction learning deficits (Blair et al., 2005; Moul et al., 2012), as well as generalised deficits in the encoding of contexts of behaviour (Hamilton et al., 2015). The relevance of the present result to extinction learning was also assumed in the theoretical application of the results developed above. None the less, the data reported here do not act to invalidate these accounts, as the analyses in which the absence of renewal effects was observed were not designed to be positive falsifying for the renewal hypothesis.

The absence of CU trait-based renewal deficits in this study cannot be compared to any prior results. The review reported in this thesis demonstrated that theories implicating extinction

learning in CU traits are not based on reliable evidence. Further, none of the reviewed research investigated renewal. This means that observations of diminished responses to extinction conditions in this literature (including those reported here) may be the result of processes other than extinction learning. Potential candidates for these processes are probably as varied as the procedures that have been used to investigate these effects (see sections 2.2.3 – 2.2.4). Therefore, the results of this thesis highlight the need for future systematic investigation of renewal effects to substantiate the role of extinction learning in CU trait aetiologies.

A secondary motivation for such research is the potential role of renewal in the efficacy of behavioural interventions for individuals with CU traits. As described in section 1.2, instrumental (goal directed) learning processes are relevant to common behavioural interventions utilised for conduct problems (Dadds & Hawes, 2006). One aim of these interventions is to enable parents to withhold responses that previously acted as reinforcers of antisocial behaviour in their children (Dadds & Hawes, 2006; D. J. Hawes et al., 2014). Whilst a range of factors have been considered as determinants of the success of these interventions (D. J. Hawes et al., 2014), the context sensitivity of the instrumental learning is not among them. This is despite the widespread relevance of renewal in behavioural interventions for other psychopathologies (Conklin & Tiffany, 2002; Hofmann, 2008). If reduced context encoding during inhibitory learning is indeed a component of the pathology of CU traits, then any such learning occurring during therapy may be better retained in individuals with these traits. Therefore, if this cognitive process can be evidenced in an experimental context, then it may warrant clinical attention.

7.3 CU Traits in the General Population

Whilst the motivation for the present research has implications for clinical populations, it is important to recognise that the populations sampled in the current studies were non-clinical.

Two factors were particularly proximal to this sampling decision. Firstly, general population samples are highly accessible for research, minimising the costs and logistics involved in recruitment. This was particularly important when recruiting the large sample described in chapters 5 and 6. Additionally, the accessibility of a sample population is limited by ethical considerations. It may not have been appropriate to test the current, highly theoretical renewal hypothesis in ethically protected populations. Therefore, provision of evidence for this effect in non-clinical populations would be useful for ethically justifying further research in clinical samples. The current sampling methodologies were justified in light of these practical considerations. However, by utilising non-clinical populations in this thesis, I have inherently assumed that the cognitive mechanisms underlying CU traits may not always lead to clinically relevant behaviours. This assumption naturally implies that there are processes mediating the manifestation of this mechanism as CU traits.

7.3.1 Affective deficits are not Deterministic to Disordered Behaviour

Significant aetiological features of CU traits are, in themselves, insufficient for development of clinically relevant behaviour. An existing model of CU traits' role in antisocial behaviour is the "inhibition model for moral socialization" (Kimonis et al., 2008). This model suggests that, typically, the emotional responses of others are sufficient to inhibit clinically relevant conduct or antisocial behaviour. Therefore, impairments in recognition of others' emotions (affective empathy), and in behavioural learning about those emotions, may lead to antisocial behaviour (Blair, 2017; Moul et al., 2017). However, this model requires that an individual expresses antisocial behaviour before such impairments become relevant determinants of whether that behaviour is learned. Environmental triggers of antisocial behaviour therefore

remain as important determinants of conduct problems in individuals with CU traits (D. J. Hawes et al., 2014).

In light of these processes it is evident that dispositional factors relevant to CU traits are contingent on environmental factors to manifest as antisocial behaviour. Indeed, genetic factors that amplify the behavioural effects of poor parenting have also been found to lead to improved behavioural regulation in the presence of good parenting (Belsky & Beaver, 2011). Hence, there is no reason to assume that processes relevant to aetiology of CU traits are not present in the general population. Indeed, other studies have noted the utility of examining these processes in non-clinical population, as they are not impacted by the environmental factors associated with antisocial behaviour (X. Chen et al., 2020). The present investigation of learning processes in CU traits was therefore not invalidated by the use of a general population sample.

7.3.2 Importance of Dimensional Measures

In addition to assuming that cognitive features relevant to CU traits are present in the general population, I have assumed that these features vary continuously across this population. My analytical approach in this thesis did not include clinical cut-offs or categorisations of individual differences, rather, I examined relationships between continuous variables. Such cut-offs imply that there is a particular level of these psychological processes which are relevant to disordered behaviour. Krueger et al. (2018) stress that discrete conceptualisations of such psychopathologies should be founded on good evidence, as the majority of psychopathologies are dimensional and not discrete. They also suggest that this dimensionality arises from the roles of multiple aetiological processes in causing convergent psychopathological profiles. This suggestion is applicable to studies of CU traits, given the establishment of aetiological variants in these traits (Craig et al., 2021). Further, given that little evidence exists to substantiate the role of

extinction learning deficits in CU traits, it is therefore appropriate to assume that this relationship is a dimensional construct, and that it occurs throughout the population.

7.4 Conclusion

Through the research conducted in this thesis I was able to achieve two of my three stated aims. Firstly, I was able to document existing research on the relationship between CU traits and extinction learning. Secondly, I provided novel data evidencing this relationship. However, the validity of the present data was limited by the absence of converging renewal effects. Nevertheless, this thesis was a significant contribution to the recent literature documenting heterogeneity within the CU trait construct. In this regard the present results suggest a previously unconsidered hypothesis. Specifically, that both highly stable and highly unstable predictive temperaments may be relevant to the aetiology of CU traits. Substantiation of this novel implication may offer clinicians useful insights into factors underlying abnormal instrumental behaviours in individuals with these traits.

8 References

- Allen, J. L., Shou, Y., Wang, M. C., & Bird, E. (2020). Assessing the Measurement Invariance of the Inventory of Callous-Unemotional Traits in School Students in China and the United Kingdom. *Child Psychiatry and Human Development*, 1–12.
<https://doi.org/10.1007/s10578-020-01018-0>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Andershed, H., Colins, O. F., Salekin, R. T., Lordos, A., Kyranides, M. N., & Fanti, K. A. (2018). Callous-Unemotional Traits Only Versus the Multidimensional Psychopathy Construct as Predictors of Various Antisocial Outcomes During Early Adolescence. *Journal of Psychopathology and Behavioral Assessment*, 40(1), 16–25. <https://doi.org/10.1007/s10862-018-9659-5>
- Bailey, C. A., & Ostrov, J. M. (2008). Differentiating forms and functions of aggression in emerging adults: Associations with hostile attribution biases and normative beliefs. *Journal of Youth and Adolescence*, 37(6), 713–722. <https://doi.org/10.1007/s10964-007-9211-5>
- Baliousis, M., Duggan, C., McCarthy, L., Huband, N., & Völlm, B. (2019). Executive function, attention, and memory deficits in antisocial personality disorder and psychopathy. *Psychiatry Research*, 278, 151–161. <https://doi.org/10.1016/j.psychres.2019.05.046>
- Balleine, B. W., Killcross, A. S., & Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *Journal of Neuroscience*, 23(2), 666–675.
<https://doi.org/10.1523/jneurosci.23-02-00666.2003>

- Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1–23.
- Barry, C. T., Frick, P. J., DeShazo, T. M., McCoy, M. G., Ellis, M., & Loney, B. R. (2000). The importance of callous-unemotional traits for extending the concept of psychopathy to children. *Journal of Abnormal Psychology*, 109(2), 335–340. <https://doi.org/10.1037//0021-843X.109.2.335>
- Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Belsky, J., & Beaver, K. M. (2011). Cumulative-genetic plasticity, parenting and adolescent self-regulation. *Journal of Child Psychology and Psychiatry*, 52(5), 619–626. <https://doi.org/10.1111/j.1469-7610.2010.02327.x>
- Bennett, D. C., & Kerig, P. K. (2014). Investigating the Construct of Trauma-Related Acquired Callousness Among Delinquent Youth: Differences in Emotion Processing. *Journal of Traumatic Stress*, 27(4), 415–422. <https://doi.org/10.1002/jts.21931>
- Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science*, 58(3), 739–753. <https://doi.org/10.1111/ajps.12081>
- Bętkowska-Korpala, B., Błaszczuk, E., Gierowski, J. K., Lickiewicz, J., Starowicz-Filip, A., & Szymlarz, A. (2013). Executive functions in psychopathy. The neuropsychological analysis of aggressive offenders. *Z Zagadnień Nauk Sądowych*, 96, 716–728. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84895468069&partnerID=40&md5=6707c21249f9fd0930ec78343ecf3204>

- Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., & Flor, H. (2005). Deficient Fear Conditioning in Psychopathy: A Functional Magnetic Resonance Imaging Study. *Archives of General Psychiatry*, 62(7), 799–805.
<https://doi.org/10.1001/archpsyc.62.7.799>
- Blair, R. J. R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4), 698–718. <https://doi.org/10.1016/j.concog.2005.06.004>
- Blair, R. J. R. (2017). Emotion-based learning systems and the development of morality. *Cognition*, 167, 38–45. <https://doi.org/http://dx.doi.org/10.1016/j.cognition.2017.03.013>
- Blair, R. J. R., & Cipolotti, L. (2000). Impaired social response reversal. A case of “acquired sociopathy”. *Brain: A Journal of Neurology*, 123(6), 1122–1141.
<https://doi.org/10.1093/brain/123.6.1122>
- Blair, R. J. R., Colledge, E., & Mitchell, D. G. V. (2001). Somatic markers and response reversal: Is there orbitofrontal cortex dysfunction in boys with psychopathic tendencies? *Journal of Abnormal Child Psychology*, 29(6), 499–511. <https://doi.org/10.1023/A:1012277125119>
- Blair, R. J. R., Mitchell, D., & Blair, K. (2005). *The psychopath: Emotion and the brain*. Blackwell Publishing.
- Blonigen, D. M., Hicks, B. M., Krueger, R. F., Patrick, C. J., & Iacono, W. G. (2006). Continuity and change in psychopathic traits as measured via normal-range personality: A longitudinal-biometric study. *Journal of Abnormal Psychology*, 115(1), 85–95.
<https://doi.org/10.1037/0021-843X.115.1.85>
- Bouton, M. E., Maren, S., & McNally, G. P. (2021). Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning. *Physiological Reviews*, 101(2), 611–681.

<https://doi.org/10.1152/physrev.00016.2020>

- Bradshaw, C. A., Freegard, G., & Reed, P. (2015). Human performance on random ratio and random interval schedules, performance awareness and verbal instructions. *Learning & Behavior*, 43(3), 272–288. <https://doi.org/10.3758/s13420-015-0178-x>
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436.
- Brazil, I. A., Maes, J. H. R., Scheper, I., Bulten, B. H., Kessels, R. P. C., Verkes, R. J., & de Bruijn, E. R. A. (2013). Reversal deficits in individuals with psychopathy in explicit but not implicit learning conditions. *Journal of Psychiatry & Neuroscience*, 38(4), E13–E20. <https://doi.org/10.1503/jpn.120152>
- Brazil, I. A., Mathys, C. D., Popma, A., Hoppenbrouwers, S. S., & Cohn, M. D. (2017). Representational Uncertainty in the Brain During Threat Conditioning and the Link With Psychopathic Traits. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, 2(8), 689–695. <https://doi.org/10.1016/j.bpsc.2017.04.005>
- Brinkley, C. A., Schmitt, W. A., Smith, S. S., & Newman, J. P. (2001). Construct validation of a self-report psychopathy scale: Does Levenson’s self-report psychopathy scale measure the same constructs as Hare’s psychopathy checklist-revised? *Personality and Individual Differences*, 31(7), 1021–1038. <https://doi.org/http://dx.doi.org/10.1016/S0191-8869%2800%2900178-1>
- Brouns, B. H. J., de Wied, M. A., Keijsers, L., Branje, S., van Goozen, S. H. M., & Meeus, W. H. J. (2013). Concurrent and prospective effects of psychopathic traits on affective and cognitive empathy in a community sample of late adolescents. *Journal of Child Psychology and Psychiatry*, 54(9), 969–976. <https://doi.org/10.1111/jcpp.12052>
- Brown, G. T. L., Harris, L. R., O’Quin, C., & Lane, K. E. (2017). Using multi-group

- confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, 40(1), 66–90. <https://doi.org/10.1080/1743727X.2015.1070823>
- Budhani, S., & Blair, R. J. R. (2005). Response reversal and children with psychopathic tendencies: Success is a function of salience of contingency change. *Journal of Child Psychology and Psychiatry*, 46(9), 972–981. <https://doi.org/10.1111/j.1469-7610.2004.00398.x>
- Budhani, S., Richell, R. A., & Blair, R. J. R. (2006). Impaired reversal but intact acquisition: Probabilistic response reversal deficits in adult individuals with psychopathy. *Journal of Abnormal Psychology*, 115(3), 552–558. <https://doi.org/10.1037/0021-843X.115.3.552>
- Burke, J. D., Loeber, R., & Lahey, B. B. (2007). Adolescent conduct disorder and interpersonal callousness as predictors of psychopathy in young adults. *Journal of Clinical Child and Adolescent Psychology*, 36(3), 334–346. <https://doi.org/10.1080/15374410701444223>
- Butter, C. M. (1969). Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in *Macaca mulatta*. *Physiology & Behavior*, 4(2), 163–171. [https://doi.org/10.1016/0031-9384\(69\)90075-4](https://doi.org/10.1016/0031-9384(69)90075-4)
- Cardinale, E. M., & Marsh, A. A. (2020). The Reliability and Validity of the Inventory of Callous Unemotional Traits: A Meta-Analytic Review. *Assessment*, 27(1), 57–71. <https://doi.org/10.1177/1073191117747392>
- Cecil, C. A. M., Lysenko, L. J., Jaffee, S. R., Pingault, J. B., Smith, R. G., Relton, C. L., Woodward, G., McArdle, W., Mill, J., & Barker, E. D. (2014). Environmental risk, Oxytocin Receptor Gene (OXTR) methylation and youth callous-unemotional traits: a 13-year longitudinal study. *Molecular Psychiatry*, 19(10), 1071–1077.

<https://doi.org/10.1038/mp.2014.95>

Chen, H., & Lan, W. (1998). Adolescents' perceptions of their parents' academic expectations: comparison of American, Chinese-American, and Chinese high school students.

Adolescence, 33(130), 385–385.

Chen, X., Osborne, L. A., & Reed, P. (2020). Role of psychopathic personality traits on the micro-structure of free-operant responding: Impacts on goal-directed but not stimulus-drive responses in extinction. *Personality and Individual Differences*, 163, 110055.

<https://doi.org/10.1016/j.paid.2020.110055>

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255.

https://doi.org/10.1207/S15328007SEM0902_5

Clark, L., Cools, R., & Robbins, T. W. (2004). The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain and Cognition*, 55(1), 41–53.

[https://doi.org/10.1016/S0278-2626\(03\)00284-7](https://doi.org/10.1016/S0278-2626(03)00284-7)

Cleckley, H. M. (1976). *The mask of sanity* (5th ed.). Mosby.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. . In *Current Directions in Psychological Science* (Issue 3). Academic press. [https://doi.org/10.1111/1467-](https://doi.org/10.1111/1467-8721.ep10768783)

[8721.ep10768783](https://doi.org/10.1111/1467-8721.ep10768783)

Colins, O. F., Andershed, H., Hawes, S. W., Bijttebier, P., & Pardini, D. A. (2016). Psychometric Properties of the Original and Short Form of the Inventory of Callous-Unemotional Traits in Detained Female Adolescents. *Child Psychiatry and Human Development*, 47(5), 679–690.

<https://doi.org/10.1007/s10578-015-0601-8>

Conklin, C. A., & Tiffany, S. T. (2002). Applying extinction research and theory to cue-exposure

- addiction treatments. *Addiction*, 97(2), 155–167. <https://doi.org/doi:10.1046/j.1360-0443.2002.00014.x>
- Cooke, D. J. (1996). Psychopathic personality in different cultures: What do we know? What do we need to find out? *Journal of Personality Disorders*, 10(1), 23–40.
<https://doi.org/10.1521/pedi.1996.10.1.23>
- Cooke, D. J. (1998). Psychopathy Across Cultures. In *Psychopathy: Theory, Research and Implications for Society* (pp. 13–45). Springer Netherlands. https://doi.org/10.1007/978-94-011-3965-6_2
- Cooke, D. J., & Michie, C. (2001). Refining the Construct of Psychopathy: Towards a Hierarchical Model. *Psychological Assessment*, 13(2), 171–188.
<https://doi.org/10.1037/111040-3590.13.2.171>
- Craig, S. G., Goulter, N., & Moretti, M. M. (2021). A Systematic Review of Primary and Secondary Callous-Unemotional Traits and Psychopathy Variants in Youth. *Clinical Child and Family Psychology Review*, 24(1), 65–91. <https://doi.org/10.1007/s10567-020-00329-x>
- Crego, C., & Widiger, T. A. (2016). Cleckley’s psychopaths: Revisited. *Journal of Abnormal Psychology*, 125(1), 75–87. <https://doi.org/10.1037/abn0000130>
- Crossley, M. J., Ashby, F. G., & Maddox, W. T. (2014). Context-dependent savings in procedural category learning. *Brain and Cognition*, 92, 1–10.
<https://doi.org/10.1016/j.bandc.2014.09.008>
- Crum, K. I., Waschbusch, D. A., Bagner, D. M., & Coxe, S. (2015). Effects of Callous–Unemotional Traits on the Association Between Parenting and Child Conduct Problems. *Child Psychiatry and Human Development*, 46(6), 967–980.
<https://doi.org/10.1007/s10578-015-0535-1>

- Dadds, M. R., Cauchi, A. J., Wimalaweera, S., Hawes, D. J., & Brennan, J. (2012). Outcomes, moderators, and mediators of empathic-emotion recognition training for complex conduct problems in childhood. *Psychiatry Research, 199*(3), 201–207.
<https://doi.org/10.1016/j.psychres.2012.04.033>
- Dadds, M. R., El Masry, Y., Wimalaweera, S., & Guastella, A. J. (2008). Reduced Eye Gaze Explains “Fear Blindness” in Childhood Psychopathic Traits. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(4), 455–463.
<https://doi.org/https://doi.org/10.1097/CHI.0b013e31816407f1>
- Dadds, M. R., & Hawes, D. (2006). *Integrated family intervention for child conduct problems: A behaviour-attachment-systems intervention for parents*. Australian Academic Press.
<https://psycnet.apa.org/record/2008-08563-000>
- Dadds, M. R., Hawes, D. J., Frost, A. D. J., Vassallo, S., Bunn, P., Hunter, K., & Merz, S. (2009). Learning to “talk the talk”: The relationship of psychopathic traits to deficits in empathy across childhood. *Journal of Child Psychology and Psychiatry, 50*(5), 599–606.
<https://doi.org/http://dx.doi.org/10.1111/j.1469-7610.2008.02058.x>
- Dadds, M. R., Jambrak, J., Pasalich, D., Hawes, D. J., & Brennan, J. (2011). Impaired attention to the eyes of attachment figures and the developmental origins of psychopathy. *Journal of Child Psychology and Psychiatry, 52*(3), 238–245. <https://doi.org/10.1111/j.1469-7610.2010.02323.x>
- Dadds, M. R., Kimonis, E. R., Schollar-Root, O., Moul, C., & Hawes, D. J. (2018). Are impairments in emotion recognition a core feature of callous-unemotional traits? Testing the primary versus secondary variants model in children. *Development and Psychopathology, 30*(1), 67–77. <https://doi.org/10.1017/S0954579417000475>

- Dadds, M. R., Perry, Y., Hawes, D. J., Merz, S., Riddell, A. C., Haines, D. J., Solak, E., & Abeygunawardane, A. I. (2006). Attention to the eyes and fear-recognition deficits in child psychopathy. *British Journal of Psychiatry*, 189(3), 280–281.
<https://doi.org/10.1192/bjp.bp.105.018150>
- Dargis, M., Wolf, R. C., & Koenigs, M. (2017). Reversal learning deficits in criminal offenders: Effects of psychopathy, substance use, and childhood maltreatment history. *Journal of Psychopathology and Behavioral Assessment*, 39(2), 189–197.
<https://doi.org/10.1007/s10862-016-9574-6>
- De Brito, S. A., Viding, E., Kumari, V., Blackwood, N., & Hodgins, S. (2013). Cool and hot executive function impairments in violent offenders with antisocial personality disorder with and without psychopathy. *PLoS ONE*, 8(6), e66556.
<https://doi.org/10.1371/journal.pone.0065566>
- DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592096511. <https://doi.org/10.1177/2515245920965119>
- DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). Best Practices for Using Median Splits, Artificial Categorization, and their Continuous Alternatives. *Journal of Experimental Psychopathology*, 2(2), 197–209. <https://doi.org/10.5127/jep.008310>
- DeCoster, J., Iselin, A. M. R., & Gallucci, M. (2009). A Conceptual and Empirical Examination of Justifications for Dichotomization. *Psychological Methods*, 14(4), 349–366.
<https://doi.org/10.1037/a0016956>
- Department of Education Skills and Employment - Australian Government. (n.d.). *International student data monthly summary*. Retrieved January 28, 2022, from

<https://www.dese.gov.au/international-data/data-visualisation-international-student-monthly-summary>

- Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, 380(6569), 69–72. <https://doi.org/10.1038/380069a0>
- Docherty, M., Boxer, P., Huesmann, L. R., O'Brien, M., & Bushman, B. (2017). Assessing Callous-Unemotional Traits in Adolescents: Determining Cutoff Scores for the Inventory of Callous and Unemotional Traits. *Journal of Clinical Psychology*, 73(3), 257–278. <https://doi.org/10.1002/jclp.22313>
- Dolan, M. (2012). The neuropsychology of prefrontal function in antisocial personality disordered offenders with varying degrees of psychopathy. *Psychological Medicine*, 42(8), 1715–1725. <https://doi.org/10.1017/S0033291711002686>
- Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, 13(4), 454–469. <https://doi.org/10.1177/1073191106287354>
- Fanti, K.A., Konikou, K., Cohn, M., Popma, A., & Brazil, I. A. (2020). Amygdala functioning during threat acquisition and extinction differentiates antisocial subtypes. *Journal of Neuropsychology*, 14(2), 226–241. <https://doi.org/http://dx.doi.org/10.1111/jnp.12183>
- Fanti, Kostas A., Kyranides, M. N., Petridou, M., Demetriou, C. A., & Georgiou, G. (2018). Neurophysiological markers associated with heterogeneity in conduct problems, callous unemotional traits, and anxiety: Comparing children to young adults. *Developmental Psychology*, 54(9), 1634–1649. <https://doi.org/10.1037/dev0000505>
- Fernández-Serrano, M. J., Perales, J. C., Moreno-López, L., Pérez-García, M., & Verdejo-García, A. (2012). Neuropsychological profiling of impulsivity and compulsivity in cocaine

dependent individuals. *Psychopharmacology*, 219(2), 673–683.

Finger, E. C., Marsh, A. A., Blair, K. S., Reid, M. E., Sims, C., Ng, P., Pine, D. S., & Blair, R. J.

R. (2011). Disrupted reinforcement signaling in the orbitofrontal cortex and caudate in youths with conduct disorder or oppositional defiant disorder and a high level of psychopathic traits. *American Journal of Psychiatry*, 168(2), 152–162.

<https://doi.org/10.1176/appi.ajp.2010.10010129>

Finger, E. C., Marsh, A. A., Mitchell, D. G. V, Reid, M. E., Sims, C., Budhani, S., Kosson, D. S.,

Chen, G., Towbin, K. E., Leibenluft, E., Pine, D. S., & Blair, R. J. R. (2008). Abnormal ventromedial prefrontal cortex function in children with psychopathic traits during reversal learning. *Archives of General Psychiatry*, 65(5), 586–594.

<https://doi.org/10.1001/archpsyc.65.5.586>

Fisher, L., & Blair, R. J. R. (1998). Cognitive impairment and its relationship to psychopathic

tendencies in children with emotional and behavioral difficulties. *Journal of Abnormal Child Psychology*, 26(6), 511–519. <https://doi.org/10.1023/A:1022655919743>

Flor, H., Birbaumer, N., Hermann, C., Ziegler, S., Patrick, H., & Christopher, J. A. I. (2002).

Aversive Pavlovian conditioning in psychopaths: Peripheral and central correlates. *Psychophysiology*, 39(4), 505–518. <https://doi.org/10.1017/S0048577202394046>

Fontaine, N. M. G., Rijdsdijk, F. V., McCrory, E. J. P., & Viding, E. (2010). Etiology of Different

Developmental Trajectories of Callous-Unemotional Traits. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(7), 656–664.

<https://doi.org/10.1016/j.jaac.2010.03.014>

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *The Hare Psychopathy Checklist: Youth*

Version. Multi-Health Systems.

- Frick, P. J., Bodin, S. D., & Barry, C. T. (2000). Psychopathic traits and conduct problems in community and clinic-referred samples of children: Further development of the psychopathy screening device. *Psychological Assessment, 12*(4), 382–393.
<https://doi.org/10.1037/1040-3590.12.4.382>
- Frick, P. J., & Hare, R. D. (2001). *The Antisocial Process Screening Device*. Multi-Health Systems.
- Frick, P. J., & Ray, J. V. (2015). Evaluating callous-unemotional traits as a personality construct. *Journal of Personality, 83*(6), 710–722.
- Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Can callous-unemotional traits enhance the understanding, diagnosis, and treatment of serious conduct problems in children and adolescents? A comprehensive review. *Psychological Bulletin, 140*(1), 1.
<https://doi.org/10.1037/a0033076>
- Frick, P. J., & White, S. F. (2008). Research review: The importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 49*(4), 359–375.
<https://doi.org/10.1111/j.1469-7610.2007.01862.x>
- Gao, Y., & Raine, A. (2010). Successful and Unsuccessful Psychopaths: A Neurobiological Model. *Behavioral Sciences & the Law, 28*(2), 194–210. <https://doi.org/10.1002/bsl.924>
- Gershman, S. J., Jones, C., Norman, K., Monfils, M.-H., & Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience, 7*(164). <https://doi.org/10.3389/fnbeh.2013.00164>
- Gillen, C. T. A., Lee, Z., Salekin, K. L., Iselin, A. M. R., Harrison, N. A., Clark, A. P., Colins, O. F., & Salekin, R. T. (2018). Psychopathic Traits in Adolescence: the Importance of

- Examining Components in Face Processing, Voice Processing, and Emotional Skill. *Journal of Psychopathology and Behavioral Assessment*, 40(1), 50–59.
<https://doi.org/10.1007/s10862-018-9656-8>
- 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>
- Gray, N. S., Weidacker, K., & Snowden, R. J. (2019). Psychopathy and impulsivity: The relationship of psychopathy to different aspects of UPPS-P impulsivity. *Psychiatry Research*, 272, 474–482.
- Gregory, S., Blair, R. J. R., Ffytche, D., Simmons, A., Kumari, V., Hodgins, S., & Blackwood, N. (2015). Punishment and psychopathy: A case-control functional MRI investigation of reinforcement learning in violent antisocial personality disordered men. *The Lancet Psychiatry*, 2(2), 153–160. <https://doi.org/10.1016/S2215-0366%2814%2900071-6>
- Grillon, C. (2002). Associative learning deficits increase symptoms of anxiety in humans. *Biological Psychiatry*, 51(11), 851–858. [https://doi.org/10.1016/S0006-3223\(01\)01370-1](https://doi.org/10.1016/S0006-3223(01)01370-1)
- Grupe, D. W., & Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: An integrated neurobiological and psychological perspective. In *Nature Reviews Neuroscience* (Vol. 14, Issue 7, pp. 488–501). Nature Publishing Group. <https://doi.org/10.1038/nrn3524>
- Hamilton, R. K. B., Hiatt Racer, K., & Newman, J. P. (2015). Impaired integration in psychopathy: A unified theory of psychopathic dysfunction. *Psychological Review*, 122(4), 770. <https://doi.org/10.1037/a0039703>
- Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences*, 1(2), 111–119. <https://doi.org/10.1016/0191->

8869(80)90028-8

- Hare, R. D. (1984). Performance of psychopaths on cognitive tasks related to frontal lobe function. *Journal of Abnormal Psychology*, 93(2), 133–140. <https://doi.org/10.1037/0021-843X.93.2.133>
- Hare, R. D. (1991). *The Hare Psychopathy Checklist - Revised*. Multi-Health Systems.
- Hare, R. D. (1996). Psychopathy: A clinical construct whose time has come. *Criminal Justice and Behavior*, 23(1), 25–54. <https://doi.org/10.1177/0093854896023001004>
- Harpur, T. J., Hare, R. D., & Hakstian, A. R. (1989). Two-Factor Conceptualization of Psychopathy: Construct Validity and Assessment Implications. *Psychological Assessment*, 1(1), 6–17. <https://doi.org/10.1037/1040-3590.1.1.6>
- Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *Hare psychopathy checklist: Screening version (PCL: SV)*. Multi-Health Systems.
- Hauser, L. L., Griffith, E. E. H., Buchanan, A., & Ruiz, P. (2020). Cultural Considerations of Psychopathy in the Forensic Context. In *The Wiley International Handbook on Psychopathic Disorders and the Law* (pp. 639–662). Wiley.
<https://doi.org/10.1002/9781119159322.ch76>
- Hawes, D. J., & Dadds, M. R. (2005). Oppositional and Conduct Problems. In *Psychopathology and the Family* (pp. 73–91). Elsevier Ltd. <https://doi.org/10.1016/B978-008044449-9/50006-X>
- Hawes, D. J., Price, M. J., & Dadds, M. R. (2014). Callous-Unemotional Traits and the Treatment of Conduct Problems in Childhood and Adolescence: A Comprehensive Review. *Clinical Child and Family Psychology Review*, 17(3), 248–267.
<https://doi.org/10.1007/s10567-014-0167-1>

- Hawes, S. W., Byrd, A. L., Henderson, C. E., Gazda, R. L., Burke, J. D., Loeber, R., & Pardini, D. A. (2014). Refining the parent-reported inventory of callous-unemotional traits in boys with conduct problems. *Psychological Assessment*, 26(1), 256–266.
<https://doi.org/10.1037/a0034718>
- Henry, J., Dionne, G., Viding, E., Petitclerc, A., Feng, B., Vitaro, F., Brendgen, M., Tremblay, R. E., & Boivin, M. (2018). A longitudinal twin study of callous-unemotional traits during childhood. *Journal of Abnormal Psychology*, 127(4), 374–384.
<https://doi.org/10.1037/abn0000349>
- Hofmann, S. G. (2008). Cognitive processes during fear acquisition and extinction in animals and humans: Implications for exposure therapy of anxiety disorders. *Clinical Psychology Review*, 28(2), 199–210. <https://doi.org/https://doi.org/10.1016/j.cpr.2007.04.009>
- Holland, P. C. (2004). Relations Between Pavlovian-Instrumental Transfer and Reinforcer Devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, 30(2), 104–117. <https://doi.org/10.1037/0097-7403.30.2.104>
- Ishikawa, S. S., Raine, A., Lencz, T., Bihrlé, S., & Lacasse, L. (2001). Autonomic stress reactivity and executive functions in successful and unsuccessful criminal psychopaths from the community. *Journal of Abnormal Psychology*, 110(3), 423–432.
<https://doi.org/10.1037/0021-843X.110.3.423>
- Jones, A. P., Happé, F. G. E., Gilbert, F., Burnett, S., & Viding, E. (2010). Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, 51(11), 1188–1197.
<https://doi.org/10.1111/j.1469-7610.2010.02280.x>
- Kahn, R. E., Frick, P. J., Youngstrom, E., Findling, R. L., & Youngstrom, J. K. (2012). The

- effects of including a callous-unemotional specifier for the diagnosis of conduct disorder. *Wiley Online Library*, 53(3), 271–282. <https://doi.org/10.1111/j.1469-7610.2011.02463.x>
- Kim, Y. Y., & Jung, Y. S. (2014). Reduced frontal activity during response inhibition in individuals with psychopathic traits: An sLORETA study. *Biological Psychology*, 97, 49–59. <https://doi.org/10.1016/j.biopsycho.2014.02.004>
- Kimonis, E. R., Branch, J., Hagman, B., Graham, N., & Miller, C. (2013). The psychometric properties of the Inventory of Callous-Unemotional Traits in an undergraduate sample. *Psychological Assessment*, 25(1), 84–93. <https://doi.org/10.1037/a0029024>
- Kimonis, E. R., Frick, P. J., Munoz, L. C., & Aucoin, K. J. (2008). Callous-unemotional traits and the emotional processing of distress cues in detained boys: Testing the moderating role of aggression, exposure to community violence, and histories of abuse. *Development and Psychopathology*, 20(2), 569–589. <https://doi.org/10.1017/S095457940800028X>
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.
- Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *R JOURNAL*, 6(2). <http://www.biosoft.hacettepe.edu.tr/MVN/>.
- Krakowski, M. I., & Czobor, P. (2017). Proneness to aggression and its inhibition in schizophrenia: Interconnections between personality traits, cognitive function and emotional processing. *Schizophrenia Research*, 184, 82–87. <https://doi.org/10.1016/j.schres.2016.11.038>
- Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., Bach, B., Bagby, R. M., Bornovalova, M. A., Carpenter, W. T., Chmielewski, M., Cicero, D. C., Clark, L. A., Conway, C., DeClercq, B., DeYoung,

- C. G., ... Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, 17(3), 282–293. <https://doi.org/10.1002/wps.20566>
- Kung, F. Y. H., Kwok, N., & Brown, D. J. (2018). Are Attention Check Questions a Threat to Scale Validity? *Applied Psychology*, 67(2), 264–283. <https://doi.org/10.1111/apps.12108>
- Lahey, B. B., Loeber, R., Quay, H. C., Applegate, B., Shaffer, D., Waldman, I., Hart, E. L., McBurnett, K., Frick, P. J., Jensen, P. S., Dulcan, M. K., Canino, G., & Bird, H. R. (1998). Validity of DSM-IV subtypes of conduct disorder based on age of onset. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37(4), 435–442. <https://doi.org/10.1097/00004583-199804000-00022>
- Lapierre, D., Braun, C. M. J., & Hodgins, S. (1995). Ventral frontal deficits in psychopathy - neuropsychological test findings. *Neuropsychologia*, 33(2), 139–151. [https://doi.org/10.1016/0028-3932\(94\)00110-B](https://doi.org/10.1016/0028-3932(94)00110-B)
- Larsson, H., Viding, E., & Plomin, R. (2008). Callous—Unemotional Traits and Antisocial Behavior. *Criminal Justice and Behavior*, 35(2), 197–211. <https://doi.org/10.1177/0093854807310225>
- Leistico, A. M. R., Salekin, R. T., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the hare measures of psychopathy to antisocial conduct. *Law and Human Behavior*, 32(1), 28–45. <https://doi.org/10.1007/s10979-007-9096-6>
- Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing Psychopathic Attributes in a Noninstitutionalized Population. *Journal of Personality and Social Psychology*, 68(1), 151–158. <https://doi.org/10.1037/0022-3514.68.1.151>
- Linder, J. R., Crick, N. R., & Collins, W. A. (2002). Relational Aggression and Victimization in Young Adults' Romantic Relationships: Associations with Perceptions of Parent, Peer, and

Romantic Relationship Quality. *Social Development*, 11(1), 69–86.

<https://doi.org/10.1111/1467-9507.00187>

Loeber, R., Stouthamer-Loeber, M., Van Kammen, W. B., & Farrington, D. P. (1989).

Development of a New Measure of Self-Reported Antisocial Behavior for Young Children: Prevalence and Reliability. In M. W. Klien (Ed.), *Cross-National Research in Self-Reported Crime and Delinquency. NATO ASI Series (Series D: Behavioural and Sciences)* (Vol. 50, pp. 203–225). Springer. https://doi.org/10.1007/978-94-009-1001-0_10

Lykken, D. T. (1957). A study of anxiety in the sociopathic personality. *Journal of Abnormal and Social Psychology*, 55(1), 6–10. <https://doi.org/10.1037/h0047232>

Lykken, D. T. (1995). *The antisocial personalities*. Psychology Press.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19–40.

<https://doi.org/10.1037/1082-989X.7.1.19>

Mackintosh, N. J. (1963). Extinction of a discrimination habit as a function of overtraining.

Journal of Comparative and Physiological Psychology, 56(5), 842–847. <https://oae-ovid-com.ezproxy.library.sydney.edu.au/article/00004726-196310000-00011/HTML>

Marshall, L. A., & Cooke, D. J. (1999). The childhood experiences of psychopaths: A retrospective study of familial and societal factors. *Journal of Personality Disorders*, 13(3), 211–225. <https://doi.org/10.1521/pedi.1999.13.3.211>

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43(3), 679–690.

<https://doi.org/10.3758/s13428-010-0049-5>

McDonald, R., Dodson, M. C., Rosenfield, D., & Jouriles, E. N. (2011). Effects of a Parenting

- Intervention on Features of Psychopathy in Children. *Journal of Abnormal Child Psychology*, 39(7), 1013–1023. <https://doi.org/10.1007/s10802-011-9512-8>
- McMahon, R. J., Witkiewitz, K., & Kotler, J. S. (2010). Predictive validity of callous-unemotional traits measured in early adolescence with respect to multiple antisocial outcomes. *Journal of Abnormal Psychology*, 119(4), 752–763. <https://doi.org/10.1037/a0020796>
- McNeish, D. (2018). Thanks coefficient alpha, We'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Mitchell, D. G. V, Avny, S. B., & Blair, R. J. R. (2006). Divergent Patterns of Aggressive and Neurocognitive Characteristics in Acquired Versus Developmental Psychopathy. *Neurocase*, 12(3), 164–178. <https://doi.org/10.1080/13554790600611288>
- Mitchell, D. G. V, Colledge, E., Leonard, A., & Blair, R. J. R. (2002). Risky decisions and response reversal: Is there evidence of orbito-frontal cortex dysfunction in psychopathic individuals? *Neuropsychologia*, 40(12), 2013–2022. <https://doi.org/10.1016/S0028-3932%2802%2900056-8>
- Mitchell, D. G. V, Fine, C., Richell, R. A., Newman, C., Lumsden, J., Blair, K. S., & Blair, R. J. R. (2006). Instrumental learning and relearning in individuals with psychopathy and in patients with lesions involving the amygdala or orbitofrontal cortex. *Neuropsychology*, 20(3), 280–289. <https://doi.org/10.1037/0894-4105.20.3.280>
- Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., & Robinson, O. J. (2017). Modeling Avoidance in Mood and Anxiety Disorders Using Reinforcement Learning. *Biological Psychiatry*, 82(7), 532–539. <https://doi.org/10.1016/j.biopsych.2017.01.017>
- Mol, B., Van Den Bos, P., Derks, Y., & Egger, Y. (2009). Executive functioning and the two-

- factor model of psychopathy: No differential relation? *International Journal of Neuroscience*, 119(1), 124–140. <https://doi.org/10.1080/00207450802324861>
- Molto, J., Poy, R., Segarra, P., Pastor, M. C., & Javier, M. (2007). Response perseveration in psychopaths: Interpersonal/affective or social deviance traits? *Journal of Abnormal Psychology*, 116(3), 632–637. <https://doi.org/10.1037/0021-843X.116.3.632>
- Moore, A. A., Blair, R. J. R., Hettema, J. M., & Roberson-Nay, R. (2019). The genetic underpinnings of callous-unemotional traits: A systematic research review. *Neuroscience and Biobehavioral Reviews*, 100, 85–97. <https://doi.org/10.1016/j.neubiorev.2019.02.018>
- Moran, P., Ford, T., Butler, G., & Goodman, R. (2008). Callous and unemotional traits in children and adolescents living in Great Britain Short report. *The British Journal of Psychiatry*, 192(1), 65–66. <https://doi.org/10.1192/bjp.bp.106.034876>
- Morriss, J., Christakou, A., & van Reekum, C. M. (2015). Intolerance of uncertainty predicts fear extinction in amygdala-ventromedial prefrontal cortical circuitry. *Biology of Mood & Anxiety Disorders*, 5(1), 1–13. <https://doi.org/10.1186/s13587-015-0019-8>
- Moul, C., Hawes, D. J., & Dadds, M. R. (2017). Mapping the developmental pathways of child conduct problems through the neurobiology of empathy. *Neuroscience & Biobehavioral Reviews*, 91, 34–50. <https://doi.org/doi.org/10.1016/j.neubiorev.2017.03.016>
- Moul, C., Killcross, S., & Dadds, M. R. (2012). A model of differential amygdala activation in psychopathy. *Psychological Review*, 119(4), 789–806. <https://doi.org/10.1037/a0029342>
- Moul, C., Robinson, O. J., & Livesey, E. J. (2021). Antisocial Learning: Using Learning Window Width to Model Callous-Unemotional Traits? *Computational Psychiatry*, 5(1), 54. <https://doi.org/10.5334/cpsy.68>
- Murray-Close, D., Ostrov, J. M., Nelson, D. A., Crick, N. R., & Coccaro, E. F. (2010). Proactive,

- reactive, and romantic relational aggression in adulthood: Measurement, predictive validity, gender differences, and association with Intermittent Explosive Disorder. *Journal of Psychiatric Research*, 44(6), 393–404. <https://doi.org/10.1016/j.jpsychires.2009.09.005>
- Newman, J. P., & Kosson, D. S. (1986). Passive avoidance learning in psychopathic and nonpsychopathic offenders. *Journal of Abnormal Psychology*, 95(3), 252–256. <https://doi.org/http://dx.doi.org/10.1037/0021-843X.95.3.252>
- Newman, J. P., Patterson, C. M., & Kosson, D. S. (1987). Response perseveration in psychopaths. *Journal of Abnormal Psychology*, 96(2), 145–148. <https://doi.org/10.1037/0021-843X.96.2.145>
- Nigg, J. T. (2000). On Inhibition/Disinhibition in Developmental Psychopathology: Views from Cognitive and Personality Psychology and a Working Inhibition Taxonomy. *Psychological Bulletin*, 126(2), 220–246. <https://doi.org/10.1037/0033-2909.126.2.220>
- O'Brien, B. S., & Frick, P. J. (1996). Reward dominance: Associations with anxiety, conduct problems, and psychopathy in children. *Journal of Abnormal Child Psychology*, 24(2), 223–240. <https://doi.org/10.1007/BF01441486>
- O'Brien, B. S., Frick, P. J., & Lyman, R. D. (1994). Reward dominance among children with disruptive behavior disorders. *Journal of Psychopathology and Behavioral Assessment*, 16(2), 131–145. <https://doi.org/10.1007/BF02232724>
- Pasalich, D. S., Dadds, M. R., Hawes, D. J., & Brennan, J. (2011). Do callous-unemotional traits moderate the relative importance of parental coercion versus warmth in child conduct problems? An observational study. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 52(12), 1308–1315. <https://doi.org/10.1111/j.1469-7610.2011.02435.x>
- Patrick, C. J. (2010). Operationalizing the triarchic conceptualization of psychopathy:

- Preliminary description of brief scales for assessment of boldness, meanness, and disinhibition. In *Unpublished test manual*. Florida State University.
- Patterson, G. R. (1982). *Coercive family process: A social learning approach* (Vol. 3). Castalia.
- Pavlov, I. P. (1927). *Conditioned Reflexes: an Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press.
- Pera-Guardiola, V., Batalla, I., Bosque, J., Kosson, D., Pifarré, J., Hernández-Ribas, R., Goldberg, X., Contreras-Rodríguez, O., Menchón, J. M., Soriano-Mas, C., & Cardoner, N. (2016). Modulatory effects of psychopathy on Wisconsin Card Sorting Test performance in male offenders with Antisocial Personality Disorder. *Psychiatry Research*, 235, 43–48. <https://doi.org/http://dx.doi.org/10.1016/j.psychres.2015.12.003>
- Pham, T. H., Vanderstucken, O., Philippot, P., & Vanderlinden, M. (2003). Selective Attention and Executive Functions Deficits Among Criminal Psychopaths. *Aggressive Behavior*, 29(5), 393–405. <https://doi.org/10.1002/ab.10051>
- Podlesnik, C. A., Kelley, M. E., Jimenez-Gomez, C., & Bouton, M. E. (2017). Renewed behavior produced by context change and its implications for treatment maintenance: A review. *Journal of Applied Behavior Analysis*, 50(3), 675–697. <https://doi.org/doi:10.1002/jaba.400>
- Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics* 2015 47:7, 47(7), 702–709. <https://doi.org/10.1038/ng.3285>
- Porter, S. (1996). Without conscience or without active conscience? The etiology of psychopathy revisited. *Aggression and Violent Behavior*, 1(2), 179–189. [https://doi.org/10.1016/1359-1789\(95\)00010-0](https://doi.org/10.1016/1359-1789(95)00010-0)

- Poythress, N. G., & Hall, J. R. (2011). Psychopathy and impulsivity reconsidered. *Aggression and Violent Behavior, 16*(2), 120–134. <https://doi.org/10.1016/j.avb.2011.02.003>
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. <https://www.r-project.org/>
- Ray, J. V., Pechorro, P., & Gonçalves, R. A. (2016). A Comparison of Self-Report Measures of Callous-Unemotional Traits Among Incarcerated Youth. *Criminal Justice and Behavior, 43*(10), 1293–1309. <https://doi.org/10.1177/0093854815628027>
- Ribes-Guardiola, P., Poy, R., Segarra, P., Branchadell, V., & Moltó, J. (2020). Response perseveration and the triarchic model of psychopathy in an undergraduate sample. *Personality Disorders: Theory, Research, and Treatment, 11*(1), 54–62. <https://doi.org/10.1037/per0000371>
- Riopka, S. J., Coupland, R. B. A., & Olver, M. E. (2015). Self-reported psychopathy and its association with criminal cognition and antisocial behavior in a sample of university undergraduates. *Canadian Journal of Behavioural Science, 47*(3), 216–225. <https://doi.org/10.1037/a0039075>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rothmund, Y., Ziegler, S., Hermann, C., Gruesser, S. M., Foell, J., Patrick, C. J., & Flor, H. (2012). Fear conditioning in psychopaths: Event-related potentials and peripheral measures. *Biological Psychology, 90*(1), 50–59. <https://doi.org/10.1016/j.biopsycho.2012.02.011>
- Salekin, R. T. (2017). Research review: What do we know about psychopathic traits in children? *Journal of Child Psychology and Psychiatry, 58*(11), 1180–1200. <https://doi.org/10.1111/jcpp.12738>

- Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The Triple P-Positive Parenting Program: A Comparison of Enhanced, Standard, and Self-Directed Behavioral Family Intervention for Parents of Children With Early Onset Conduct Problems. *Journal of Consulting and Clinical Psychology, 68*(4), 624–640. <https://doi.org/10.1037/0022-006X.68.4.624>
- Sawir, E. (2005). Language difficulties of international students in Australia: The effects of prior learning experience. *International Education Journal, 6*(5), 567–580. <http://iej.cjb.net>
- Schwenck, C., Mergenthaler, J., Keller, K., Zech, J., Salehi, S., Taurines, R., Romanos, M., Schecklmann, M., Schneider, W., Warnke, A., & Freitag, C. M. (2012). Empathy in children with autism and conduct disorder: Group-specific profiles and developmental aspects. *Journal of Child Psychology and Psychiatry, 53*(6), 651–659. <https://doi.org/10.1111/j.1469-7610.2011.02499.x>
- Sebastian, C. L., McCrory, E. J., Dadds, M. R., Cecil, C. A. M., Lockwood, P. L., Hyde, Z. H., De Brito, S. A., & Viding, E. (2014). Neural responses to fearful eyes in children with conduct problems and varying levels of callous-unemotional traits. *Psychological Medicine, 44*(1), 99–109. <https://doi.org/10.1017/S0033291713000482>
- Sharf, A., Kimonis, E. R., & Howard, A. (2014). Negative life events and posttraumatic stress disorder among incarcerated boys with callous-unemotional traits. *Journal of Psychopathology and Behavioral Assessment, 36*(3), 401–414. <https://doi.org/10.1007/s10862-013-9404-z>
- Sheffield, V. F. (1949). Extinction as a function of partial reinforcement and distribution of practice. *Journal of Experimental Psychology, 39*(4), 511–526. <https://doi.org/10.1037/h0057242>

- Shou, Y., Lay, S. E., De Silva, H. S., Xyrakis, N., & Sellbom, M. (2019). Sociocultural Influences on Psychopathy Traits: A Cross-National Investigation. *Journal of Personality Disorders*, 1–23. https://doi.org/10.1521/pedi_2019_33_428
- Shou, Y., Sellbom, M., & Han, J. (2017). Evaluating the Construct Validity of the Levenson Self-Report Psychopathy Scale in China. *Assessment*, 24(8), 1008–1023. <https://doi.org/10.1177/1073191116637421>
- Singewald, N., Schmuckermair, C., Whittle, N., Holmes, A., & Ressler, K. J. (2015). Pharmacology of cognitive enhancers for exposure-based therapy of fear, anxiety and trauma-related disorders. *Pharmacology & Therapeutics*, 149, 150–190. <https://doi.org/https://doi.org/10.1016/j.pharmthera.2014.12.004>
- Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.
- Skinner, B. F., & Morse, W. H. (1958). Fixed-interval reinforcement of running in a wheel. *Journal of the Experimental Analysis of Behavior*, 1(4), 371–379. <https://doi.org/doi:10.1901/jeab.1958.1-371>
- Smith, R. A., & Khawaja, N. G. (2011). A review of the acculturation experiences of international students. In *International Journal of Intercultural Relations* (Vol. 35, Issue 6, pp. 699–713). Pergamon. <https://doi.org/10.1016/j.ijintrel.2011.08.004>
- Takahashi, Y., Pease, C. R., Pingault, J., & Viding, E. (2021). Genetic and environmental influences on the developmental trajectory of callous-unemotional traits from childhood to adolescence. *Journal of Child Psychology and Psychiatry*, 62(4), 414–423. <https://doi.org/10.1111/jcpp.13259>
- Triandis, H., & Gelfand, M. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology*, 74(1), 118–

128. <https://doi.org/10.1037/0022-3514.74.1.118>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1–67.
<https://www.jstatsoft.org/v45/i03/>
- Veit, R., Flor, H., Erb, M., Hermann, C., Lotze, M., Grodd, W., & Birbaumer, N. (2002). Brain circuits involved in emotional learning in antisocial behavior and social phobia in humans. *Neuroscience Letters*, 328(3), 233–236. [https://doi.org/10.1016/S0304-3940\(02\)00519-0](https://doi.org/10.1016/S0304-3940(02)00519-0)
- Viding, E., & McCrory, E. (2019). Towards understanding atypical social affiliation in psychopathy. *The Lancet Psychiatry*, 6(5), 437–444. [https://doi.org/10.1016/S2215-0366\(19\)30049-5](https://doi.org/10.1016/S2215-0366(19)30049-5)
- Vitale, J. E., & Newman, J. P. (2001). Response perseveration in psychopathic women. *Journal of Abnormal Psychology*, 110(4), 644–647. <https://doi.org/10.1037/0021-843X.110.4.644>
- Waller, R., Gardner, F., & Hyde, L. W. (2013). What are the associations between parenting, callous-unemotional traits, and antisocial behavior in youth? A systematic review of evidence. *Clinical Psychology Review*, 33(4), 593–608.
- Waller, R., Hyde, L. W., Klump, K. L., & Burt, S. A. (2018). Parenting Is an Environmental Predictor of Callous-Unemotional Traits and Aggression: A Monozygotic Twin Differences Study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 57(12), 955–963. <https://doi.org/10.1016/j.jaac.2018.07.882>
- Wang, M.-C., Gao, Y., Deng, J., Lai, H., Deng, Q., & Armour, C. (2017). The factor structure and construct validity of the inventory of callous-unemotional traits in Chinese undergraduate students. *PLOS ONE*, 12(12), e0189003. <https://doi.org/10.1371/journal.pone.0189003>
- Weiner, H. (1962). Some effects of response cost upon human operant behavior. *Journal of the*

- Experimental Analysis of Behavior*, 5(2), 201–208. <https://doi.org/doi:10.1901/jeab.1962.5-201>
- White, B. A., Gordon, H., & Guerra, R. C. (2015). Callous-unemotional traits and empathy in proactive and reactive relational aggression in young women. *Personality and Individual Differences*, 75, 185–189. <https://doi.org/10.1016/j.paid.2014.11.031>
- Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, 30(4), 669–689. [https://doi.org/10.1016/S0191-8869\(00\)00064-7](https://doi.org/10.1016/S0191-8869(00)00064-7)
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Wilkinson, S., Waller, R., & Viding, E. (2015). Practitioner Review: Involving young people with callous unemotional traits in treatment-does it work? A systematic review. *Journal of Child Psychology and Psychiatry*, 57(5), 552–565. <https://doi.org/10.1111/jcpp.12494>
- Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron*, 81(2), 267–279. <https://doi.org/10.1016/j.neuron.2013.11.005>
- Wootton, J. M., Frick, P. J., Shelton, K. K., & Silverthorn, P. (1997). Ineffective parenting and childhood conduct problems: The moderating role of callous-unemotional traits. *Journal of Consulting and Clinical Psychology*, 65(2), 301–308. <https://doi.org/10.1037/0022-006X.65.2.292.b>
- Yang, Y., Raine, A., Colletti, P., Toga, A. W., & Narr, K. L. (2011). Abnormal structural correlates of response perseveration in individuals with psychopathy. *The Journal of Neuropsychiatry*

and Clinical Neurosciences, 23(1), 107–110.

<https://doi.org/http://dx.doi.org/10.1176/appi.neuropsych.23.1.107>

Yeh, C. J., & Inose, M. (2003). International students' reported English fluency, social support satisfaction, and social connectedness as predictors of acculturative stress. *Counselling Psychology Quarterly*, 16(1), 15–28. <https://doi.org/10.1080/0951507031000114058>