Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

The Neurocognitive Correlates of Compulsivity

Tricia X.F. Seow

School of Psychology and Trinity College Institute of Neuroscience

Trinity College Dublin, University of Dublin

A thesis submitted to

Trinity College Dublin, University of Dublin

For the degree of

Doctor of Philosophy

2020

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: 

Date: 30 March 2020

Summary

Progress in delineating the neurocognitive correlates of important aspects of mental health has been limited by the use of clinical taxonomies for research purposes. For example, the obsessive-compulsive disorder (OCD) literature implicates dysfunctions in goal-directed control, metacognition and error monitoring, but these effects are either inconsistent and/or unspecific, observed in disorders beyond OCD. The experiments in this thesis investigated the potential of a transdiagnostic approach to resolve these issues, focusing on a sub-component of OCD that has transdiagnostic relevance—compulsivity. Specifically, this thesis tested the extent to which dysfunctions in goal-directed control, metacognition and error monitoring are better captured by a transdiagnostic compulsive dimension and thereby aims to develop the current neurocognitive characterisation of compulsivity.

Chapter 2 investigated the premise that metacognitive deficits are implicated in compulsivity and may be obscured in case-control studies where patient groups present with, on-average, elevations in both depression and compulsive traits. The results were broadly supportive of the hypothesis, demonstrating that a host of metacognitive failures are characteristic of compulsivity, and that some effects associated with an anxious-depression dimension of mental health may serve to confound prior care-control work. These results are discussed with respect to how these impairments might lead to failures in the construction and update of mental

models of the world in compulsivity, potentially leading sufferers to rely excessively on habitual modes of thinking and behaving.

Chapter 3 built on this work and sought to reconcile metacognitive findings with a related literature on goal-directed control. Specifically, we used electroencephalography (EEG) to test if the well-documented deficits in goal-directed control commonly observed in compulsivity are associated with a failure to represent the mental model neurally, or if these deficits are solely explained by issues with the implementation of the model. We found evidence to suggest that the representation of the mental model was diminished in compulsivity, evidenced by a lack of sensitivity to state-state transitions in a well-studied two-step decision making task. This was apparent in parietal-occipital alpha power suppression and reaction times. We also observed that mid-frontal theta, a general marker of cognitive control, was lower in high compulsive individuals during the making of choice, but this bore a less clear relation to goal-directed behaviour. Together, the results suggest that compulsivity is likely characterised by dysfunctions in constructing an accurate mental model upon which to base goal-directed decisions, but not implementing goal-oriented actions themselves.

Finally, *Chapter 4* aimed to reconcile a set of transdiagnostic findings from a potentially related literature concerning error monitoring as indexed by error-related negativity (ERN). Alterations in ERN amplitude have been consistently observed in OCD and a multitude of other psychiatric disorders. Here, we tested if these neural

modulations might be better captured by a related transdiagnostic dimension pertaining to anxious-depression, rather than compulsivity itself. Contrary to our expectation, the data did not reveal any significant associations of ERN shifts with any psychiatric phenomena under study nor did transdiagnostic dimensions explain the data better.

In summary, the findings of the thesis support goal-directed and metacognitive dysfunctions, but not ERN abnormalities, as neurocognitive correlates of compulsivity.

List of publications & presentations

The experiments reported in this thesis are published or are currently under review in the following manuscripts:

Seow, X.F.T., Benoit, E., Dempsey, C., Jennings, M., Maxwell, A., McDonough, M. & Gillan, C. M. (2020). A dimensional study of error-related negativity (ERN) and self-reported psychiatric symptoms. *International Journal of Psychophysiology (in review).*

Seow, X.F.T. & Gillan, C. M. (2020). Transdiagnostic phenotyping reveals a host of metacognitive deficits associated with compulsivity. *Scientific* Reports **10**, 2883.

The following are presentations arising from this work:

Seow, X.F.T. & Gillan, C. M. (2019). Transdiagnostic phenotyping reveals a range of metacognitive deficits associated with compulsivity. Poster presented at the Society for NeuroEconomics 2019, Dublin, Ireland.

Seow, X.F.T. & Gillan, C. M. (2019). Impact of volatility on behaviour in trans-diagnostic psychiatric symptom dimensions. Poster presented at the British Association for Psychopharmacology 2019, Manchester, UK.

Seow, X.F.T. (2019). A dimensional approach to psychiatry. Invited talk given for the FriKo Seminar Series, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Seow, X.F.T. & Gillan, C. M. (2019). Confidence and action in trans-diagnostic psychiatric symptom dimensions. *Biological Psychiatry, **85**(10), S243.* Poster presented at the Society for Biological Psychiatry 2019, Chicago, Illinois, US.

Acknowledgements

They say it takes a village, and it cannot be truer.

First and foremost, I am indebted to my supervisor, Dr. Claire Gillan. Claire, I cannot thank you enough for how generous you have been with your time, advice and encouragement throughout these years. You are a huge inspiration and it has been amazing being under your tutelage. I truly cherish every step of the way.

To my board of examiners, Prof. Robert Whelan and Dr. Mike Browning, thank you for your expertise and patience. I hope this thesis will be a pleasure to read.

To Trinity College Dublin, thank you for financially supporting me under the Postgraduate Ussher Fellowship.

To my Year 1 and 2 appraisers, Dr. Redmond O'Connell and Prof. Shane O'Mara, thank you for the yearly guidance.

To the numerous more who have been integral to this body of work: Dr. Redmond O'Connell, thank you so much for your help whenever EEG perplexed me. Dr. Michael McDonough and the rest of the staff at St. Patrick's Mental Health Services, thank you for facilitating patient recruitment. Aoibheann Maxwell, Caoimhe Dempsey, Edith Benoit and Maeve Jennings, thank you for assisting with the humongous EEG dataset collection.

And of course, special gratitude goes to the OG core team of the Gillan Lab, both past and present members: Dr. Andy Pringle and Dr. Siobhan Harty, thanks for being so generous with advice and support throughout the years. You two have been

the best role-models. Kevin Lynch and Sean Kelly, thanks for adding colour into everyday lab life and seeing me through the dreary days. I'm absolutely going to miss all the handmade delicious snacks and evening chats about everything under the sun. Sharon Lee and Owen Lynch, thanks for cheering me on these last few months. Finally, Jonny Giordano, thanks so much for lending your eyes to the numerous words in this text.

There are also many kudos to extend to my friends and fandom family for giving me a life outside lines of experimental code. Yan, Nat, Taru: thanks for putting up with my anxieties and having unwavering faith that I'd pull through. Aoife: thanks for your steadfast prayers and hospitality throughout these years. Yen, Gab, David: thanks for checking on me time and time again despite being on the opposite end of the world. Gwen: thanks for always being the ear for science frustrations way back since undergrad. Luna, Dan: thanks for giving my mind a space to turn to when I needed to breathe.

Last but not least, I would not have ever stepped into this endeavour without Mom and Dad. Thanks for being my emotional, financial, and spiritual pillars. Most of all, thank you for your love.

Table of Contents

Chapter 1: General introduction

Disorder category: Obsessive-compulsive disorder (OCD)

When we think of *compulsivity*, which is the performance of persistent, repetitive, inappropriate actions despite no obvious goal-oriented function that often results in adverse consequences (Dalley et al., 2011), we normatively default to obsessive-compulsive disorder (OCD). The Diagnostic and Statistical Manual of Mental Disorders (DSM-V (APA, 2013)) defines OCD as a condition that presents obsessions and/or compulsions that are time-consuming, functionally impairing or distressing. Obsessions are recurring intrusive thoughts/images that often lead to anxiety and distress, such as repeated thoughts about fatal germ contamination. These false beliefs are often accompanied by compulsions—the habitual or ritualistic acts that follow rigid idiosyncratic rules—in attempts to prevent the perceived negative outcomes. Notably, these compulsive actions are either not logically associated with what they are intended to overcome or are extreme, leading to severely disabling conditions. For example, an OCD patient with cleaning compulsions may be continually compelled to wash their hands for prolonged periods, numerous times over the course of a day. Not only does the disproportionate amount of time spent at the sink impairs daily functioning, the excessive washing is also detrimental to skin health. Unfortunately, disability in OCD is often understated, with about 27% of patients experiencing debilitating conditions that render them unable to work (Pinto et al., 2006).

Various epidemiological studies purport that OCD affects approximately 2.8-3.5% of the population (Fineberg, Hengartner, Bergbaum, Gale, Rössler, et al., 2013; Ruscio et al., 2010) and features relatively equally between males and females (Kessler et al., 2005). The disorder has an early onset—about two-thirds of OCD cases present by the age of 22 (Fineberg, Hengartner, Bergbaum, Gale, Gamma, et al., 2013)—peaking in either early adolescence or early adulthood (Anholt et al., 2014). Neuroanatomically, cortical dysfunction in OCD is widespread. Functional imaging historically implicate abnormalities of the reciprocal connections between the orbital frontal cortex and the caudate (Whiteside et al., 2004) directly or indirectly through the thalamus; these are known as fronto-striatal “loops” (G. E. Alexander et al., 1986). Alterations in many other brain structures such as grey and white matter volumes in the posterior cingulate, temporal lobe, parietal cortex and limbic circuit regions are also thought to contribute to the pathophysiology of OCD (Fouche et al., 2017; Rotge et al., 2009, 2010). However, these cortical abnormalities are far from specific to OCD, for instance, dysfunction in the fronto-striatal loops are also thought to underlie the emergence of other clinical syndromes like schizophrenia (Simpson et al., 2010).

Heritability-wise, twin studies suggest that there is about 27-47% of genetic influence on OCD symptomology in adults (van Grootheest et al., 2005). Yet, despite the identification of over 60 candidate genes for OCD, none have achieved genome-wide significance (Hemmings & Stein, 2006). This is also applicable to most major psychiatric disorders: they are highly polygenic, have small effect sizes for their

individual polymorphisms and collaborative meta-analyses have not found support for any of their candidate genes to date (for instance, depression (Border et al., 2019; Culverhouse et al., 2018)). Efforts to determine specific genes to OCD and other clinical phenomena may be complicated by the questionable biological validity of disorder categories; this is explored in the later section.

Currently, the front-line treatments for OCD include both psychotherapeutic and pharmacological interventions (Bloch et al., 2010; Skapinakis et al., 2016). Exposure response therapy (ERP), a type of cognitive behavioural therapy (CBT), exposes patients to anxiety-provoking stimuli and challenges them to refrain from performing their compulsive behaviours. This is thought to be effective by not only challenging false beliefs (i.e. showing them the feared event will not occur), but also to extinguish automatic behavioural responses to triggering situations or stimuli (Foa et al., 2012; V. Meyer, 1966). As for pharmacotherapy, the choice medication is relatively high doses of selective serotonin reuptake inhibitors (SSRIs) that induce changes in serotonin (5-HT) function (Blier & El Mansari, 2007). These two treatment approaches are efficacious for about 50-68% of patients (Eddy et al., 2004; Fisher & Wells, 2005b; Pigott & Seay, 1999), however, there is still a large number of the clinical population who fail to see clinical benefit. The variability in treatment response is thought to partly reflect the heterogeneity in OCD diagnosis. Co-morbidity—where an individual's symptom profile passes the criteria for more than one disorder diagnosis—is present in the majority of OCD cases (Fineberg, Hengartner, Bergbaum, Gale, Rössler, et al., 2013; Ruscio et al., 2010). Indeed,

OCD co-morbidity with other psychiatric disorders such as depression, bipolar disorder, attention deficit hyperactivity disorder (ADHD) and anxiety disorders, have all been linked to poorer treatment outcomes (Abramowitz et al., 2000; Pallanti et al., 2011). Nonetheless, there are some hints of disorder-specific treatment effects. For instance, OCD with co-morbid tics similarly have worse response to SSRIs (Lochner & Stein, 2003) but they show a preferential therapeutic effect with the usage of dopamine antagonists adjuncts to SSRIs (Bloch et al., 2006). This suggests that there may be potential for effective individualised approaches to treatment.

Shifting towards a transdiagnostic perspective

The inconsistencies and complexities of not just treatment literature but also the neurobiological (e.g. neuroanatomical, genetic, etc.) studies of OCD suggest that what we know of the neural underpinnings of the disorder is still very much lacking. Unfortunately, this issue encapsulates all of psychiatry. For decades, the predominant taxonomic conceptualisations of mental health (e.g. the DSM-V (APA, 2013) and the International Classification of Diseases (ICD-10) (Organization, 1993)) have advanced psychiatry practice in profound ways by providing a standardised diagnostic rubric to assess mental health and guide treatment approaches worldwide. However, these diagnostic categories were never intended for research use. Through decades of research, it has become apparent that they lack clear biological grounding (McHugh, 2005) and in many cases, important psychometric properties. For instance, common mental health conditions like major depressive disorder and generalized anxiety disorder have questionable (kappa:

0.20–0.39) test-retest reliability (Regier et al., 2013).There is gathering agreement that while these taxonomic structures provided an important starting point, they may have reached the limits of their research and clinical utility (Insel et al., 2010). We highlight several grounds for this, taking the DSM as a key example.

Firstly, putatively distinct disorders in the DSM are poorly distinguished. This is particularly the case for OCD, where compulsive symptoms are not unique to the disorder. This is acknowledged by the DSM-V; many other disorder categories also exhibit compulsivity, such as body-dysmorphic disorder (BDD) (Simberlund & Hollander, 2017) and addiction (e.g. alcohol (Burchi et al., 2019), gambling (van Timmeren et al., 2018), substance use (Werner et al., 2019)). In these other disorders, the main characteristics of rigid behaviours remain except that they are expressed in different contexts. For instance, patients with BDD are pre-occupied with body image related obsessions and compulsions, while patients suffering from substance use addiction persist in compulsive-like drug abuse to neutralise cravings and withdrawals associated with addiction (Everitt & Robbins, 2005). Indeed, clinical commentary has suggested that these disorders may be better classified on an obsessive-compulsive spectrum given their overlapping symptom characteristics (Phillips et al., 2010).

Alongside the feature of similarity across disorders comes a broader issue of co-morbidity. In OCD, this is not confined to disorders of compulsion. Epidemiological studies have found a staggering 87.3-90% of OCD patients to be co-morbid with

another disorder, the most common co-occurring condition being anxiety disorders at 73-75.8% (Fineberg, Hengartner, Bergbaum, Gale, Rössler, et al., 2013; Ruscio et al., 2010). As mentioned in the earlier section, high levels of co-morbidity in OCD are associated with, and may in part explain, noisier response to treatment (Abramowitz et al., 2000; Pallanti et al., 2011). Studies have suggested that common risk factors likely account for some level of co-morbidity across different aspects of mental health (Andrews et al., 2009). Indeed, the Hierarchical Taxonomy of Psychopathology (HiToP) (Kotov et al., 2017), which is a data-driven hierarchical framework that groups observed co-varying symptoms together to construct psychopathological dimensions, is able to explain shared symptomatology between disorder categories to a certain extent. However, the pervasive issue of co-morbidity in psychiatry highlights the failure of the current diagnostic classification to distinguish distinct clinical phenotypes (van Loo & Romeijn, 2015).

Secondly, massive symptom variability and complexity are apparent *within* disorder categories. To be diagnosed with a disorder, the DSM outlines a set of criteria comprising of a cluster of symptoms. Only a subset of these symptoms is usually required to obtain diagnosis, and contradictory symptoms can also be equally contributable (e.g. major depressive disorder includes insomnia/hypersomnia as a criteria), leading to a multitude of symptom combinations. As a result, patients diagnosed with different psychiatric disorders may in fact have more similar clinical profiles than someone from their own disorder grouping. OCD is particularly known to be heterogenous in symptom presentation (Leckman et al., 2007; Lochner & Stein,

2003). Several groups have attempted to give more parsimonious descriptions to the heterogeneity of OCD symptoms by sorting them into more homogeneous groups utilising factor analysis in hopes of providing clearer basis for aetiology, pathogenesis and treatment. For instance, a prominent model sorts OCD symptomatology into four dimensions of symmetry/ordering, obsessions/checking, contamination/cleaning and hoarding (Mataix-Cols et al., 2005; van den Heuvel et al., 2009). These subtypes are in no way mutually exclusive, being able to co-exist in any blend, and are applicable to disorders beyond OCD. Though there are still mechanistic and neurobiological validity critiques to their conceptualisation (Bloch et al., 2008; McKay et al., 2004), there is some evidence that these symptom dimensions are distinguished by different neural systems (van den Heuvel et al., 2009), lifetime co-morbidities (Torres et al., 2006), heritability patterns (Leckman et al., 2003) and treatment response (Abramowitz et al., 2003; Mataix-Cols et al., 1999, 2002). OCD features also vary along other variables such as level of insight (Foa & Kozak, 1995; Matsunaga et al., 2002) and the ratio of obsessions to compulsions (Calamari et al., 2006; Taylor et al., 2006). For the latter, cluster studies suggest that a subgroup of OCD patients do not have elevated levels of dysfunctional beliefs. Co-morbidity may again explain this to some extent. For instance, OCD patients with co-morbid bipolar disorder present with a higher rate of sexual/religious obsessions and lower rate of checking rituals than OCD without bipolar co-morbidity (Perugi et al., 1997). Together, these studies paint a picture of the great etiologic and phenomenological heterogeneity in OCD.

These issues of the current classification system have led to calls for a shift towards transdiagnostic approaches that cut across the current psychiatric taxa to re-conceptualise mental health (Hyman, 2007). Global mental health initiatives including the National Institute of Mental Health's (NIMH) Research Domain Criteria (RDoC) initiative (Insel et al., 2010) and the European Commission's Roadmap for Mental Health Research (ROAMER) initiative (Haro et al., 2014) have even channelled substantial funding to boost psychiatry research toward this new direction in recent years. In particular, the RDoC project is structured around incorporating and integrating various analysis levels (e.g. from genes, molecules, cells, circuits, physiology, behaviour to self-report) dimensionally in six major bio-behavioural domains to define psychiatric phenomena beyond traditional diagnosis (Cuthbert & Insel, 2013). Indeed, many researchers in the last decade have highlighted how transdiagnostic approaches can advance our understanding of mental health through the definition of clinically-useful phenotypes grounded in biological systems (Fusar-Poli et al., 2019; Gillan et al., 2017; Huys et al., 2016). The idea is that if clinical definitions are more closely aligned to aetiology, there would be greater potential to develop objective tests and criteria (e.g. neurocognitive tools, genetic basis) for diagnosis and early identification of risk. With a more homogeneous clinical population, treatment interventions may also be more tightly linked to precisely defined mechanisms and thereby be more effective.

The rest of the chapter will outline how a transdiagnostic perspective has benefited or will benefit three existing neurocognitive models of OCD to provide the motivation

and context to the experimental hypotheses of this thesis investigating the neurocognitive correlates of compulsivity.

Neurocognitive correlate 1: Habit and goal-directed control

One prominent neuropsychological theory of compulsive behaviour implicates dysfunctions in behavioural control (Robbins et al., 2012). Dual system theories posit two modes of action control: the habit formation system and the goal-directed system (Balleine & Dickinson, 1998; Balleine & O'Doherty, 2010; Dickinson & Balleine, 1994). Habits are fast and automatic responses enacted not because of an expected outcome, but are instead reflexively triggered by acquired stimulus-response (S-R) associations (Thorndike, 1898). Connectivity between the putamen and pre-motor cortex is thought to "stamp in" S-R associations that support the formation and maintenance of habits (de Wit, Watson, et al., 2012; Tricomi et al., 2009; Yin, Ostlund, et al., 2005; Yin & Knowlton, 2006). Habits are computationally efficient, requiring minimal cognitive effort. Thus, relying on habitual associations frees cognitive resources to handle more demanding tasks. However, this comes at the cost of inflexibility where ingrained habits may become maladaptive in a fast-changing environment. In contrast, goal-directed actions, supported by the caudate (Yin et al., 2004, 2006; Yin, Knowlton, et al., 2005) and medial orbital frontal cortex (Gremel & Costa, 2013; O'Doherty, 2011; Yin et al., 2008), are oriented towards achieving desirable outcomes (Tolman, 1948). Successful goal-directed control requires many resources, including planning, simulation, reflection and the continued tracking and updating of action-outcome associations. As such, this mode

of action control is much more computationally expensive but is thought to be critical for cognitive flexibility especially when goals/environments are ever-changing.

Compulsive behaviour in OCD, where patients get ‘stuck’ in their repetitive action cycles, is hypothesised to arise from a shift away from goal-directed processing towards overreliance on habits (Robbins et al., 2012). Initially, OCD was conceptualised as a disorder of maladaptive habits, supported by the potential role of dysfunctional cortico-basal ganglia circuits in OCD (Graybiel & Rauch, 2000). Later, it emerged that this imbalance between the two modes of action control might be due to either an overactive habit system or an impaired goal-directed system. In examining the relative contribution of goal-directed versus habitual control in humans, empirical studies largely relied on outcome devaluation procedures (de Wit et al., 2007; de Wit, Standing, et al., 2012; Gillan et al., 2011). In these paradigms, participants learn stimulus-action-outcome contingencies by trial and error. In a subsequent phase of the task, one of the outcomes is reduced in value (i.e. devalued), which should reduce participants’ motivation to acquire it and therefore reduce responding. Participants’ continued response on trials with devalued outcomes suggests that the individual is unable to break strong S-R associations, an indication of habit dominance. On the other hand, successful inhibition on these trials suggests successful goal-directed control. The first study to use these techniques with OCD patients showed these individuals exhibited higher tendencies to keep responding to devalued outcomes (Gillan et al., 2011). A similar pattern of behaviour was also observed for avoidance of devalued shock consequences

(Gillan, Morein-Zamir, Urcelay, et al., 2014). For the latter study, OCD individuals showed no excess of fear responses to the stimuli as well as intact goal-directed contingency knowledge (i.e. shock expectancy). These persistent responses toward devalued outcomes despite having accurate action-outcome knowledge suggested that excessive habit formation may explain how actions become maladaptively repetitive in OCD. However, as acknowledged by most of these studies, the influence of habit and goal-directed control is not well dissociated in these devaluation techniques. Continued responses towards devalued outcomes can be driven by *both* increases in habit and impairments in goal-directed control.

Following on, it was suggested that faulty goal-directed control may be predominantly responsible for devaluation insensitivity in these paradigms instead of overactive habits (Friedel et al., 2014). Functional activity in OCD patients during the performance of avoidance habits support this theory—the caudate nucleus, a region involved in goal-directed control, exhibited hyperactivity that correlated with individual differences in devaluation sensitivity and self-reported urges to perform habitual responses, rather than habit associated regions (Gillan, Apergis-Schoute, et al., 2015). Additional evidence comes from OCD patients showing deficits in the ability to make goal-oriented choices prospectively in an economic choice paradigam (Gillan, Morein-Zamir, Kaser, et al., 2014). Furthermore, as more precise mechanistic descriptions of habit and goal-directed systems develop in recent years (Dolan & Dayan, 2013), the formalisation of 'model-based' planning allows the examination of goal-directed control apart from habit influences that confound the

results of devaluation paradigms (Daw et al., 2005, 2011). Model-based planning, which measures the extent by which individuals use state-action relationships in conjunction with reward history to guide choice, is a purported mechanism of goal-directed control and does converge on similar neural correlates such as activity in the caudate and medial orbital frontal cortex (Voon et al., 2015), the latter which is linked to prospective planning (Doll et al., 2015). In this framework, model-based planning is dissociated apart from 'model-free' learning where choices are simply enacted through the reinforcement of past rewards. This 'model-based'/'model-free' computation from a two-step reinforcement learning paradigm (Daw et al., 2005, 2011) was first utilised by Voon and colleagues to reveal that OCD patients indeed exhibit impairments in model-based planning (Voon et al., 2015). Conversely, model-free learning is intact in these individuals. Though model-free learning was initially theorised to reflect habitual strategies, there is currently little evidence for this (Friedel et al., 2014). Recent work have highlighted difficulty in studying habits in humans (de Wit et al., 2018) but a novel paradigm suggests that utilising extensive training and time pressure may overcome this issue (Luque et al., 2019).

The converging evidence from multiple methodologies as presented above suggests that OCD is characterised by goal-directed processing failures that might explain why patients are stuck in their maladaptive habitual cycles. Subsequently, as this theory developed, researchers began to question if impaired goal-directed control could also be core to a broader class of disorders with predominant compulsivity. This comes from the observation that goal-directed failures assessed by both

devaluation and model-based planning paradigms are not exclusive to OCD but also found in other related disorders such as alcohol addiction, drug addiction and binge-eating disorder (Ersche et al., 2016; Sjoerds et al., 2013; Voon et al., 2015). As such, the behavioural similarities of these compulsive type disorders are not merely superficial; they are also reflected in a transdiagnostic cognitive dysfunction (Robbins et al., 2012).

However, the candidacy of goal-directed deficits as a transdiagnostic marker of compulsivity is troubled by studies reporting that goal-directed control impairment is also observed in arguably non-compulsive psychopathology such as autism spectrum disorder (ASD) (Alvares et al., 2016), schizophrenia (Culbreth et al., 2016; R. W. Morris et al., 2015), social anxiety disorder (SAD) (Alvares et al., 2014) and Tourette's syndrome (C. Delorme et al., 2016). This is where the limitations of DSM disorder classification become starkly apparent. Case-control methodologies, i.e. comparing a diagnosed patient group to a healthy participant group, suffer from the difficulty in disentangling the confounding influences of co-morbid psychiatric disorders. For example, all the 'non-compulsive' psychiatric disorders which reported goal-directed learning deficits have significant co-morbidity and overlapping features with OCD (Bener et al., 2018; Carpita et al., 2019; Cath et al., 2008; Sheppard et al., 1999). As such, it is uncertain if decision deficits are truly attributable to compulsivity or to another co-occurring symptom.

Unfortunately, this problem is pervasive in psychiatric research; neurocognitive models seem to have no unique association to any particular psychiatric phenomena. For example, inhibitory control (Lipszyc & Schachar, 2010) and temporal discounting (Bickel et al., 2012) are also observed in a variety of disorders; now purported to be trans-disease processes. Researchers have attempted to control for these confounding variables by recruiting patients without co-morbidities e.g. only diagnosed with OCD. However, this approach raises a concern as it assumes patients of these 'pure' cases do not present other psychopathology simply because these symptoms do not reach the criteria threshold for diagnosis. In fact, these single diagnosis patients typically have substantially elevated levels of other symptomologies. While controlling for these influences are possible, it requires enough statistical power (i.e. larger sample sizes) that are already difficult to reach in case-control clinical research where sample sizes of N = 20-30 are common. To reconcile this, researchers have begun to explore new methods that can study the relationship between cognitive and clinical phenotypes at-scale.

Impaired goal-directed control and compulsivity

This thesis focuses on one such approach that leverages normal variation in the general population, dispensing with the standard case-control design that pervades psychiatry in favour of studying *dimensions* of illness. Gillan and colleagues first utilised this dimensional approach to investigate if goal-directed deficits were specifically associated with compulsivity rather than a generalised effect ubiquitous across many disorders (Gillan et al., 2016). Instead of a traditional case-control in-

person approach, the authors turned to internet-based testing (Gillan & Daw, 2016) and characterised psychiatric variation in a large general population sample (N = 1,413). These participants completed the two-step reinforcement learning task (Daw et al., 2005, 2011) measuring model-based planning and a varied set of nine self-report psychiatric questionnaires. The study first investigated the extent to which participants' individual differences in model-based planning was related, separately, to the different questionnaire scores. They found that OCD symptoms significantly correlated with goal-directed control deficiency (similarly observed in a devaluation paradigm: (Snorrason et al., 2016)). Notably, model-based planning deficits were also associated to other psychiatric questionnaire scores such as eating disorders, impulsivity and alcohol addiction, reflecting the lack of specificity of the dysfunction to OCD.

As the psychiatric questionnaire scores displayed both high collinearity with each other (e.g. trait anxiety and depression correlation: $r = 0.81$) and heterogeneity within questionnaire (e.g. schizotypy questionnaire measures positive/negative symptoms), Gillan and colleagues performed data-driven clinical phenotyping to investigate if a transdiagnostic dimensional psychiatric phenotype could explain this pattern of results. A factor analysis revealed a more parsimonious description of the psychiatric data; three dissociable transdiagnostic factors that were named to reflect their encompassing symptomology: 'anxious-depression', 'compulsive behaviour and intrusive thought' and 'social withdrawal'. Particularly, the 'compulsive behaviour and intrusive thought' dimension was of interest—this dimension encapsulated the loss

of control over behaviour seen in OCD, addiction, etc., which was hypothesized to be mechanistically explained by dysfunctions in the goal-directed system. Crucially, when the relationship of model-based control with these three dimensions was assessed, the compulsive dimension showed specificity to reduced model-based planning while the other two ‘non-compulsive' factors (anxious-depression and social withdrawal) showed no significant relation. This suggested that impaired goal-directed learning is a quantifiable transdiagnostic mechanism of compulsivity. This approach illustrates how the specificity of a neurocognitive model can be tested using rapidly acquired data from a general population sample. The method is also replicable; the same association of goal-directed deficits and the compulsive dimension was observed in another independent study (Patzelt et al., 2019). A common critique about this dimensional methodology, however, is that findings from these non-clinical groups may not generalise to diagnosed patients but a recent study has since proven otherwise—a compulsive dimension explained goal-directed deficits much better than disorder categories in a mixed OCD and generalised anxiety disorder patient group (Gillan et al., 2019).

The ‘habit’ hypothesis of OCD has come a long way since its conceptualisation two decades ago (Graybiel & Rauch, 2000), but more work is needed to characterise the precise cognitive dysfunction being captured by the expansive description of ‘goal-directed behaviour’ to move research in this area to the next phase of clinical translation. Goal-directed control, despite being often examined as a unitary construct, is an operation that relies on various cognitive processes. In broad

definition, there are two components necessary to perform goal-directed behaviour. Firstly, the knowledge of the environment, or a *mental model* of the world, is required to simulate or plan actions towards the intended objective (Doll et al., 2015). For instance, if an individual decides to stop by the pharmacy on their way to work, they will need to be aware of multiple relevant variables (e.g. spatial layout, opening hours of the store, etc.) in order to plan a successful trip to the pharmacy. Secondly, individuals would need the ability to *implement* this model into actions, to manipulate and contextualise these variables. As the literature has thus far been primarily concerned about the arbitration between habit and goal-directed systems underlying compulsive behaviour (Gillan & Robbins, 2014), how the aforementioned facets of goal-directed behaviour (i.e. having an accurate mental model versus executing it into behaviour) contribute to decision failures in compulsivity is unclear. Prior work with devaluation tasks have probed the mental model indirectly by examining if patients are able to report accurate action-outcome relationships. A number of studies observe that they can, suggesting that the mental model is intact in OCD (Gillan, Morein-Zamir, Urcelay, et al., 2014; Vaghi et al., 2019). However, these tasks are simple, only requiring the learning of straightforward stimulus-response associations. When more complex paradigms are used, OCD patients have shown impairments in their contingency knowledge that is correlated with devaluation failures (Gillan et al., 2011). In resolving this ambiguity, recent studies have turned to metacognition to investigate the state of the mental model in decision making in compulsivity.

Neurocognitive correlate 2: Metacognition

Metacognition is often described as “thinking about thinking” and “cognition about cognition”. Simply put, it refers to our ability to reflect on and evaluate our own behaviour (Flavell, 1979). How people track or assess their own knowledge plays a critical role in guiding reasoning and enacting behaviour (Fleming et al., 2012)—an impaired awareness with respect to one’s own performance can lead to persistent pathological decision-making exhibited by patients with psychiatric disorders (Hoven et al., 2019). Particularly for OCD, prominent cognitive-psychological models have long proposed dysfunctional metacognitive impairments as central to the disorder. In these clinical accounts, OCD patients overestimate the credibility or responsibility of their intrusive thoughts, and as such result in engagement of compulsive safety behaviours (Matthews & Wells, 2008; Rachman, 1997; Salkovskis, 1985; Salkovskis & McGuire, 2003). For instance, an OCD patient with contamination fears may overstate the danger and frequency of germ contamination, resulting in excessive handwashing. However, this hypothesis of inflated or exaggerated belief in obsessive thoughts seems to contrast with reports that OCD patients have insight; they recognise that their compulsive actions are irrational/excessive i.e. ‘ego-dystonic’. Some suggest that the insight into their beliefs/compulsions may be dependent on symptom severity, context (e.g. thoughts about contamination have different insight levels to those of aggressive/sexual nature) (Ryan, 2001) or reflect two separate metacognitive domains (Salkovskis et al., 1995).

Empirically, metacognition can be investigated via confidence judgements (metacognitive *bias*), which is the subjective feeling of being correct about a choice, decision or statement (Pouget et al., 2016). Confidence levels have been investigated in OCD in a variety of domains, but in particular, the memory domain has been under much focus. OCD checking behaviours are thought to be explained by under-confidence in memory resulting in a drive to repeatedly check if an action been completed (e.g. a stove turned off) to alleviate the obsessional metacognitive doubt (Fisher & Wells, 2008). One of the first studies that found support for this idea used an item recognition task and found that OCD patients exhibited lower confidence in their memory compared to healthy controls despite equal performance (McNally & Kohlbeck, 1993). This finding of under-confidence in memory re-collection has since been replicated in many other studies (Cougle et al., 2007; Ecker & Engelkamp, 1995; Foa et al., 1997; Macdonald et al., 1997; Moritz & Jaeger, 2018; Tolin et al., 2001). However, these confidence effects are inconsistent; a number of studies find no supporting evidence (Moritz et al., 2006, 2011; Moritz, Kloss, et al., 2009; Moritz, Ruhe, et al., 2009). Interestingly, a study proposed and showed that OCD checking behaviour can *cause* reductions in memory confidence (van den Hout & Kindt, 2003). Several subsequent studies have replicated the same results in both mental and real-life simulations (M. E. Coles et al., 2006; Radomsky et al., 2006; Radomsky & Alcolado, 2010). van den Hout and Kindt (2003) attribute this effect to increased familarity upon repeated checking, leading to reduced perceptual processing and thus less vivid recall of memories. Lowered confidence

levels have additionally been observed in perception, attention (Hermans et al., 2008) and interoceptive (Lazarov et al., 2014, 2015) tasks.

Linking this work with the decision making literature, a recent study by Vaghi and colleagues attempted to investigate how behaviour and confidence (an explicit report of belief and a proxy of the mental model) are 'coupled', i.e. tracking each other, using a change-point reinforcement learning task (Vaghi et al., 2017). In this paradigm, confidence is inversely related to the amount of behavioural adjustment (action) participants perform. In OCD patients, action and confidence was found to be 'decoupled' (in other words, not correlating as strongly), when compared to healthy controls. This divergence was accompanied by abnormalities in how action was updated with feedback. On the other hand, confidence reports and the sensitivity of confidence to feedback were not distinguishable to controls. As such, OCD individuals seem to have accurate meta-models but their action mechanisms fail to utilise the mental model for behaviour. However, other metacognitive impairments besides confidence bias have been linked to compulsivity in decision making. A study using a perceptual motion detection task found that non-clinical individuals on the high end of the OC symptom spectrum have lower metacognitive efficiencies (meta-d'/d') (Hauser, Allen, Rees, et al., 2017), indicating that their confidence ratings were less accurate in discriminating whether they made the right decisions.

Overall, the studies seem to be inconsistent whether the metacognitive model is impaired in compulsivity. It is probable that some of these effects are not associated to *compulsivity* per se, but instead linked to psychopathology that exist in elevated levels in OCD patients. For instance, depression is well-known to have a negative bias in information processing (Williams et al., 2009) and has empirically shown lowered confidence levels in various cognitive domains consistently (Fieker et al., 2016; Fu et al., 2005, 2012; Hancock, 1996). Indeed, some researchers have suggested that lowered confidence in memory could be accounted for by co-morbid depression (Moritz et al., 2006). Instead, the compulsive dimension may be associated to inflated confidence given that the phenotype as elucidated in Gillan et al. (2016) comprises of high loadings of schizotypy. Schizophrenic patients are overconfident, particularly in errors, and have lower metacognitive efficiencies than healthy controls (Gawęda et al., 2012; Kircher et al., 2007; Moritz et al., 2003, 2005, 2014). These metacognitive dysfunctions are hypothesized to underlie false beliefs endorsed by schizophrenic patients (Joyce et al., 2013; Moritz et al., 2005), which is suggested to be phenomenologically similar to OCD obsessionality (Sanders et al., 2006). Once again, a transdiagnostic perspective would present a welcome resolution in clarifying the relationship between metacognition and compulsivity.

Abnormal metacognition and compulsivity

Following the success of the transdiagnostic approach in Gillan et al. (2016), Rouault and colleagues (2018) used the same internet-based dimensional approach to investigate metacognitive associations to various aspects of psychopathology. With

a similar two-choice perceptual decision-making task to Hauser et al. (2017), they first examined the extent to which confidence shifts were associated with a set of nine psychiatric questionnaire scores (Rouault et al., 2018). They observed that OCD symptoms were not linked to any alterations in confidence, much like the prior studies in decision making (Hauser, Allen, Rees, et al., 2017; Vaghi et al., 2017), while two common OCD co-morbidities, anxiety and depression, were linked to reduced confidence. The authors then sought to test if using transdiagnostic dimensions, which could dissociate anxious-depression from compulsivity, would better explain the data. Despite a smaller sample size (N = 498), the same three-factor structure (anxious-depression, compulsive behaviour and intrusive thought, and social withdrawal) as the original study in goal-directed learning (N = 1,413) (Gillan et al., 2016) was replicated, highlighting the reproducibility of this method.

Importantly, specific and distinct relationships between metacognition and psychiatric dimensions were revealed: the anxious-depression dimension was associated to lower confidence levels, while the compulsive dimension was related to inflated confidence levels. There was also a trend of enhanced metacognitive efficiency in high anxious-depressive individuals and diminished metacognitive efficiency in high compulsive individuals. That is, subjects with anxious-depression symptoms were not simply less confident, they tended to also be more accurate in their judgements mirroring much of the ‘sadder but wiser’ literature from the late 70’s (Alloy & Abramson, 1979). What is most striking about this study is that it suggests that transdiagnostic phenotyping can reveal associations potentially hidden by co-

occurring symptomatology. This presents a point of consideration for investigating metacognitive associations in a OCD patient versus healthy control manner, as OCD has significant co-morbidity with anxiety and depression (Fineberg, Hengartner, Bergbaum, Gale, Rössler, et al., 2013; Ruscio et al., 2010). It is possible that reduced/null confidence effects previously observed in OCD patients were driven by anxious-depression symptomology, masking the true metacognitive dysfunctions associated to compulsivity.

Within the earlier sections, we highlighted how a transdiagnostic approach may resolve issues arising from disorder versus healthy control investigations. Firstly, goal-directed control deficits appeared pervasive across various disorders without compulsive features but was shown to map specifically onto compulsivity with the dimensional methodology. Secondly, we delved further into examining mechanisms potentially underlying these goal-directed control dysfunctions, such as those of metacognition which may be obscured within diagnostic categories by co-occurring levels of psychopathology. In this final section, we will discuss a third concern relating to the characterisation of compulsivity—how a neurocognitive profile appears to be linked to compulsivity but may instead be accounted for by other correlated dimensions of mental health.

Neurocognitive correlate 3: Error monitoring and the ERN

Error monitoring, a process by which we detect our mistakes, is also implicated in OCD. Errors are an important source of information as they signal that changes in

attentional focus or other strategic behavioural adjustments are needed. Detecting errors is essential for goal-directed behaviour by enabling flexible adaptations of behaviour when performance problems arise or when contingencies change. In humans, the most prominent electrophysiological index of error monitoring is the error-related negativity (ERN). This negative-going deflection following an error response was first independently observed by two research groups in the early nineties in studies of event-related brain potentials (ERPs) (Falkenstein et al., 1990, 1991; Gehring et al., 1993). The ERN is typically measured at midline fronto-central sites where its negativity is the largest and peaks at about 100ms from the time of the incorrect response. Though the debate about the functional significance of the ERN is unresolved (W. H. Alexander & Brown, 2011; Botvinick et al., 2001; Carter et al., 1998; Falkenstein et al., 1991; Gehring et al., 1993; Holroyd & Coles, 2002; Yeung et al., 2004), the different theories at least agree that the error signal serves to enact cognitive control to prevent future behavioural errors (Ridderinkhof et al., 2004; Ullsperger et al., 2014).

Clinically, OCD patients often report a constant feeling of erroneous or incomplete performance, or ‘not just right experiences’ that gives rise to the urge to perform “corrective” behaviours (M. E. Coles et al., 2003). Prominent theories suggest that these experiences might be explained by the generation of inappropriate or hyperactive error detection signals (Pitman, 1987). As abnormal error processing interferes with adaptive responses to goal-oriented outcomes, it may reasonably also contribute to impaired goal-directed behaviour in OCD. Additionally,

neuroimaging studies support the idea of dysfunctional performance monitoring architecture in OCD patients—they exhibit excessive activity in the orbital-frontal cortex, basal ganglia (Saxena et al., 1998) and the anterior cingulate cortex (ACC) (Ursu et al., 2003). Particularly relevant to the ERN is the ACC, as the structure is thought to be the neural generator of the error signal (Carter et al., 1998; Dehaene et al., 1994; Miltner et al., 2003; Ullsperger & Von Cramon, 2001).

In the first initial test of ERN abnormality in OCD two decades ago, Gehring and colleagues conducted the Stroop task, a speeded two-choice response paradigm, in OCD patients and observed that these individuals exhibited enhanced ERN amplitudes when compared to healthy controls (Gehring et al., 2000). This finding has been since been replicated in numerous other studies with a variety of other conflict tasks: in adult (Endrass et al., 2008, 2010; Grützmann et al., 2016; Johannes, Wieringa, Nager, Rada, et al., 2001; Klawohn et al., 2014; Riesel et al., 2011, 2015; Ruchsow et al., 2007) and paediatric OCD individuals (Carrasco, Harbin, et al., 2013; Hajcak et al., 2008; Hanna et al., 2018; Liu et al., 2014) as well as in subclinical OC populations (Gründler et al., 2009; Hajcak & Simons, 2002; Zambrano-Vazquez & Allen, 2014). However, some studies do fail in reporting this association (Agam et al., 2014; Mathews et al., 2016; Nieuwenhuis et al., 2005; Weinberg, Kotov, et al., 2015) and one study even observed reduced ERN amplitudes in high compulsive individuals (Gründler et al., 2009). Meta-analyses suggest that some of these non-significant or directionally reversed ERN effects are explained by task differences (Mathews et al., 2012; Riesel, 2019). For instance, ERN enhancement is typically

observed in paradigms primarily concerned with response-conflict, but not in probabilistic learning tasks (Gründler et al., 2009).

The ERN as an endophenotype for…?

Given its consistent association to OCD, ERN has been proposed to be an endophenotype for the disorder (Riesel, 2019). An endophenotype is a term to describe a stable phenotype of a disorder or illness that has a clear genetic connection, a valuable construct for any psychiatric syndrome. In order to be recognised as an endophenotype, the phenotype in question must be heritable, state-independent and found in unaffected family members of affected individuals at a higher rate than the general population (Gottesman & Gould, 2003). Many characteristics of ERN meet these criteria. Genetic analysis suggest that the ERN shows substantial heritability of about 50% (Anokhin et al., 2008). Enhanced ERN amplitudes can be observed independent of symptom severity in OCD patients (Endrass & Ullsperger, 2014; Riesel et al., 2014) and unaffected first-degree relatives of these individuals also show enhanced ERNs (Carrasco, Harbin, et al., 2013; Riesel et al., 2011, 2019). Moreover, larger ERNs persist in OCD patients even after cognitive behavioural therapy (CBT) despite a reduction in their symptom severity (Hajcak et al., 2008; Riesel et al., 2015). These observations of heritability and state-independency suggest that the ERN might represent a trait of underlying *vulnerability* for OCD.

However, whether an enhanced ERN is an endophenotype for OCD is challenged by the finding that other psychiatric disorders, particularly anxiety disorders, also exhibit elevated ERNs. For example, generalised anxiety disorder (GAD) (Carrasco, Hong, et al., 2013; Riesel et al., 2019; Weinberg et al., 2010; Weinberg, Klein, et al., 2012; Weinberg, Kotov, et al., 2015), health anxiety (Riesel et al., 2017) and social anxiety disorder (SAD) (Endrass et al., 2014) all exhibit larger ERNs than healthy controls. Enhanced ERN effects in these disorders are similarly found in unaffected first-degree relatives of anxiety patients (Riesel et al., 2019) and persist after successful treatment (Kujawa et al., 2016; Ladouceur et al., 2018). Moreover, the ERN has been shown to predict the onset of several anxiety disorders (Lahat et al., 2014; A. Meyer et al., 2015; A. Meyer & Klein, 2018). Benzodiazepines, which are commonly prescribed for anxiety, reduce ERN amplitudes (de Bruijn et al., 2004; Johannes, Wieringa, Nager, Dengler, et al., 2001; Riba et al., 2005). In terms of its functional importance for anxiety, the ERN is related to the priming of defensive responses (Hajcak & Foti, 2008) and avoidance behaviour bias (Frank et al., 2005). As such, the error signal exhibits good face validity as a model of the anxious phenotype. Given that OCD often presents with elevated levels of anxiety, it is possible that the enhanced ERN often observed in this disorder in fact reflects patients' co-morbid anxious symptoms, rather than any quintessential OCD feature such as obsessions or compulsions. However, some researchers disagree (Riesel, 2019) as other anxiety disorders characterised by anxious arousal like phobic anxiety (Hajcak et al., 2003a; Moser et al., 2005) or post-traumatic stress disorder (PTSD) (Rabinak et al., 2013) do not report enhanced ERNs. Some studies have

attempted to clarify the specificity of ERN alterations with direct comparisons between OCD and anxiety disorders, but the evidence is unfortunately ambiguous. Enhanced ERNs have been observed (i) only in OCD and not anxiety (Xiao et al., 2011), (ii) only in anxiety and not OCD (Weinberg, Kotov, et al., 2015) (iii) in both OCD and anxiety, with no significant difference in amplitudes (Carrasco, Hong, et al., 2013; Endrass et al., 2014). This thesis posits that this conundrum might be resolved using transdiagnostic methods.

Investigating the neurocognitive correlates of compulsivity

As outlined in the above sections, a transdiagnostic perspective may be critical for researchers to delineate the neurocognitive mechanisms that give rise to mental health phenomena. In this thesis, we consider three neurocognitive mechanisms that have been implicated in OCD and/or compulsivity and aim to use transdiagnostic methods to develop an integrated mechanistic view of how compulsions arise. To this end, we utilised behavioural, computational modelling and neuroimaging with electroencephalography (EEG) techniques plus the adoption of a recently developed dimensional approach that measures co-occurring self-report psychiatric symptoms in large general population samples (Gillan et al., 2016). This dimensional methodology shares many features with other transdiagnostic proposals briefly mentioned previously, such as the RDoC project (Cuthbert & Insel, 2013; Insel et al., 2010) and HiTOP framework (Kotov et al., 2017). Like the RDoC initiative, we first lean onto our current understanding of behaviour-brain relationships and attempt to relate them with clinical phenomena in a theory-driven

manner (Casey et al., 2013). Indeed, habit/goal-directed control and error monitoring are, in fact, two constructs highlighted in the RDoC framework. As for HiTOP, psychiatric dimensions from Gillan et al (2016) were similarly defined in a data-driven way, albeit with a less ambitious scope. Here, we chose to utilise the dimensional framework from Gillan et al (2016) as the original study specifically sought to produce a quantification of compulsivity that was juxtaposed by other non-compulsive dimensions of symptomatology with OCD relevance. This enables common co-occurring symptoms with compulsivity to be controlled for, of which is not necessarily methodologically clear to implement in the other frameworks and is an important consideration for ascertaining the specificity of our findings. Moreover, this approach has been validated in several reports (Gillan et al., 2016; Patzelt et al., 2019; Rouault et al., 2018) and as such is now poised to be used to advance our understanding of the core neurocognitive deficits that give rise to compulsivity. This was the central goal of this thesis.

The predominant habit hypothesis of OCD is both neurobiologically plausible and an intuitively straightforward way to explain how compulsive-like habits may arise independent of valence and context. Research in this area has primarily focused on delineating the direction of *shift* between the two modes of action control in compulsivity, whether compulsions arise from the dominance of habits *or* the impairment of goal-directed control. As the current state of the literature points to the latter hypothesis, this leaves the relatively unexplored question of *which* component(s) of this multifaceted concept of goal-directed control is responsible for

decision failures common to compulsivity. Here we tested if metacognitive processes, whether they are consciously reportable (e.g. confidence) or not (e.g. the ERN and other EEG signatures), are dysfunctional in compulsivity and whether they contribute to goal-directed deficits commonly observed in compulsive individuals. The central premise is that individuals need to be sensitive to the state of the environment (e.g. contingencies, errors) in order to mobilise resources in an adaptive, goal-directed fashion. If this is the case, is it possible that compulsivity is best described as a metacognitive problem, rather than a purely behavioural one?

The first specific hypothesis that will be examined in this study is that compulsivity is linked to abnormalities in metacognition. Prior work with a transdiagnostic approach suggests that compulsivity is associated to inflated confidence and a trend toward decreased metacognitive efficiency in the context of perceptual decision-making (Rouault et al., 2018). However, a case-control study with a reinforcement learning paradigm observed that metacognitive processes were not distinguishable between OCD patients and healthy controls (Vaghi et al., 2017). Here, we tested whether a transdiagnostic method is crucial to reveal metacognitive abnormalities that we hypothesized to be linked to compulsivity. Next, we attempted to relate alterations in the meta-model more directly to goal-directed control by investigating whether the failure in performing model-based plans is linked to an impairment in representing this mental model, assessed using a combination of behaviour and electrophysiology. Probing the mental model is important as it may form the basis by which we understand obsessive beliefs that are intimately connected to

compulsions and are hypothesised to drive reinforcing cycles of detrimental behaviour in OCD (Tolin et al., 2001, 2002). Lastly, we assessed the specificity of hyperactive error monitoring to compulsivity in the face of the evidence suggesting that enhanced ERNs are perhaps better explained by an anxious than compulsive phenotype. Together, these studies aim to further our understanding of the core mechanisms that give rise to compulsivity and in parallel to further probe the advantage that transdiagnostic methods might have over traditional case-control approaches.

Experiments and hypotheses

Experiment 1: Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity

Experiment 1 tested the hypothesis that metacognitive processes are dysfunctional in compulsivity. A large online population sample of 437 participants completed a predictive inference task previously studied in a case-control investigation that observed no confidence abnormalities in OCD patients (Vaghi et al., 2017). We tested if confidence, its relationship to behaviour and to environmental evidence, were associated with self-reported OCD symptoms and if they were common to an additional eight other psychiatric phenomena. We then investigated if utilising a dimensional methodology that accounts for other co-occurring psychopathology would reveal and specify metacognitive deficits associated to compulsivity.

Experiment 2: Encephalographic (EEG) correlates of reduced model-based planning in compulsivity

Experiment 2 examined which component of goal-directed control is dysfunctional in compulsivity. We tested 192 participants performing the two-step reinforcement learning task (Daw et al., 2005, 2011) with electroencephalography (EEG) and identified neural signals relating to the representation or the implementation of the mental model. Importantly, we asked if these representations were altered in the transdiagnostic compulsive phenotype linked to failures in model-based planning.

Experiment 3: A dimensional study of error-related negativity (ERN) and self-reported psychiatric symptoms

Experiment 3 tested the hypothesis that a dimensional framework could reveal specific transdiagnostic clinical manifestations of error processing dysfunctions that have long been implicated in OCD and various anxiety disorders. We obtained EEG recordings from 196 participants who performed the Flanker task (Eriksen & Eriksen, 1974) and observed if dysfunctional error monitoring indexed by error-related negativity (ERN) amplitudes shifts were, as per the literature, ubiquitous across various psychiatric phenomena including OCD and anxiety. Subsequently, we tested if the transdiagnostic approach was able to specify these amplitude shifts to a precise aspect of psychopathology—a dimensional anxious-depression or compulsive phenotype.

Chapter 2: Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity

Introduction

Intentional decisions are dependent on the interplay between behaviour and beliefs. Beliefs guide behaviour, and the consequences of our behaviour in turn update beliefs. Computational models of learning suggest that the strength of belief (i.e. 'confidence') governs the extent of its influence on action; the more confident we are, the more our behaviour is influenced by pre-existing beliefs, compared to new information (Behrens et al., 2007; Nassar et al., 2010). A breakdown in the relationship between action and belief is suggested to be characteristic of compulsive behaviours, e.g. in obsessive-compulsive disorder (OCD) or addiction. In these disorders, behaviour often appears autonomous (unguided by conscious control) and 'ego-dystonic', such as persistent drug use despite negative consequences (Everitt & Robbins, 2005) or out-of-control repetitive checking despite knowing the door is locked (Fineberg et al., 2010).

One potential cause of the divergence between intention and action in compulsive individuals is an impairment in the brain's goal-directed system, which links actions to consequences and protects against overreliance on rigid habits (Gillan, Otto, et al., 2015). Goal-directed planning deficits have been consistently observed in OCD (Gillan et al., 2011; Gillan, Apergis-Schoute, et al., 2015; Gillan, Morein-Zamir, Kaser,

et al., 2014; Gillan, Morein-Zamir, Urcelay, et al., 2014) and related disorders (Voon et al., 2015)—there is evidence to suggest this constitutes a transdiagnostic psychiatric trait linked to several aspects of clinically-relevant compulsive behaviour (Gillan et al., 2016).

The precise mechanism supporting this dysfunction is only partially understood as most employed tasks struggle to separate the construction of an internal model (e.g. action-outcome knowledge) from its implementation in behaviour. Those that have attempted this have yielded interesting, if equivocal, results. One study showed that OCD patients get stuck in habits even when they possess the requisite action-outcome knowledge to theoretically perform in a goal-directed fashion (Gillan, Morein-Zamir, Urcelay, et al., 2014). This suggests that the implementation of goal-directed behaviour is deficient in OCD, independent of the ability to construct the model. However, this does not mean the internal model is intact. Studies using more challenging tasks have found deficits in the acquisition of explicit action-outcome contingency knowledge in OCD patients (Gillan et al., 2011), suggesting that patients may have problems with both. On the other hand, these findings come from paradigms where instrumental action typically affects the kind of information that is gathered and thus are somewhat confounded and difficult to interpret.

Recently Vaghi and colleagues addressed this metacognitive question in OCD patients with more precision by using a paradigm that examined how patients make trial-wise adjustments to behaviour (i.e. implicit model) and confidence (i.e. explicit

model) in response to feedback (Nassar et al., 2016; Vaghi et al., 2017). They found that in OCD, the association between confidence and behavioural updating ('action-confidence coupling') was diminished—patients' behaviour did not align with their internal model. Further, while confidence estimates did not differ from healthy controls, OCD patients showed abnormalities in their learning rate, making more trial-wise adjustments in response to feedback than controls (Vaghi et al., 2017).

The finding of intact confidence in OCD is consistent with prior work in perceptual decision-making where individuals high versus low in OCD symptoms had no differences in their mean confidence (Hauser, Allen, Rees, et al., 2017)—results echoed by two large internet-based samples (N > 490) we conducted with the same task that also found no relationship to OCD symptoms (Rouault et al., 2018). A problem with this type of study design, however, is that it fails to capture the potentially competing influence of co-occurring disorders/symptoms in psychiatric populations. Even in studies where certain co-morbid diagnoses are explicitly excluded for, as in Vaghi et al. (2017), rates of depression and anxiety are greater than controls (Vaghi et al., 2017). Similarly, when self-report anxiety and depression severity are matched across groups by design, as in Hauser et al. (2017) (Hauser, Allen, Rees, et al., 2017), this may not accurately reflect the average OCD patient where co-morbidity is the rule, not the exception (e.g. >25% of OCD patients are co-morbid for ≥4 additional diagnoses (Gillan et al., 2017)). An alternative approach measures these relevant co-occurring symptoms in the same individuals and seeks to account for their (competing or inflating) influence on the cognitive measure of

interest. We took this approach in a prior study and found that confidence abnormalities in perceptual decision-making were reliably associated with two transdiagnostic psychiatric dimensions in opposing directions: ‘anxious-depression’ was associated with reduced confidence, while ‘compulsive behaviour and intrusive thought’ was linked to inflated confidence (Rouault et al., 2018). This finding was striking because confidence was not correlated with either OCD or depressive symptoms in the same sample. Given this, it is possible that true metacognitive abnormalities in OCD were obscured by the competing influence of co-occurring depression in this dataset, and potentially, this same issue is at play in the prior case-control study examining metacognition in OCD in the context of reinforcement learning.

To test this, here we used the same transdiagnostic methodology on an online sample of 437 participants who completed the same task from Vaghi and colleagues (Vaghi et al., 2017). We investigated the extent to which trial-wise action adjustments were disconnected from confidence reports with self-reported OCD symptoms, and whether this action-confidence decoupling is specific to OCD or also manifested in other psychiatric symptoms. We then tested if transdiagnostic phenotyping would reveal a more specific result—that only the compulsive dimension (as opposed to anxious-depression and social withdrawal) would be related to the decoupling of confidence and behaviour. Lastly, we investigated if the decoupling arose from failures in action-updating or confidence, and, with the same reduced Bayesian model used in the original study (McGuire et al., 2014; Nassar et

al., 2010, 2016; Vaghi et al., 2017), explored if there were abnormalities in the way compulsive individuals used information (e.g. recent outcomes, unexpected outcomes, environmental uncertainty and positive feedback) to update these behavioural measures.

Methods

Power Estimation. Previous research utilizing the predictive-inference task were constrained to small sample psychiatric populations (Vaghi et al., 2017). As such, we referred to earlier work that investigated confidence abnormalities in large general population cohorts with transdiagnostic symptom dimensions to determine an appropriate sample size (Rouault et al., 2018). The prior study reported an association of the anxious-depression dimension with lowered confidence level ($\beta = -0.20$, $p < 0.001$), an effect size suggesting that N = 295 participants were required to achieve 90% power at 0.001 significance level (significance threshold is corrected for multiple comparisons over the three psychiatric dimensions ('anxious-depression', 'compulsive behaviour and intrusive thought' and 'social withdrawal') investigated).

Participants. Data were collected online using Amazon's Mechanical Turk (N = 589). Participants were ≥18 years, based in USA and had >95% of their previous tasks on the platform approved. After reading the study information and consent pages, they provided informed consent by clicking the 'I give my consent' button. Participants were paid a base sum of 7 USD plus up to 1 USD bonus. Of the sample, 249 were female (42.3%) with ages ranging from 20-65 (mean = 36.3. SD = 10.2) years. All study procedures were approved by and carried out in accordance with regulations and guidelines of Trinity College Dublin School of Psychology Research Ethics Committee.

Exclusion criteria. Several pre-defined exclusion criteria were applied to ensure data quality (see ***A.I. Supplemental Methods*** for exclusion criteria details). In total, 153 participants (25.9%) were excluded, a rate typical for web-based experiments, leaving 437 participants for analysis. Of this, N = 20 (4.58%) of the current sample were the same participants from experiment 1 of a prior study where we examined confidence in perceptual decision-making (Rouault et al., 2018). An additional 10 participants (2.29%) of the current sample were included from experiment 2 of that same paper.

Predictive inference task. We adapted the predictive-inference task from Vaghi et al. (Vaghi et al., 2017) for web-based testing (***Figure 2.1***). Left and right arrow keys enabled response navigation while a spacebar press was used for decision confirmation (this is in contrast to a rotor controller used for response navigation in Vaghi et al.). The aim of the task presented to participants was to catch a particle flying from the centre of a large circle to its edge. To do so, participants positioned a 'bucket' (a free-moving arc) on the circle edge at the start of each trial. Once the bucket location was chosen, a confidence bar scaling 1 to 100 would appear below the circle after 500ms. The confidence indicator would begin randomly at either 25 or 75. Participants then indicated how confident they were that the particle would land in the bucket. After confirmation of the confidence report, a particle was then released from the centre to fly towards the edge of the circle 800ms later. If the particle landed within the boundaries of the bucket, the bucket would turn green for 500ms and the participant gained 10 points; else, the bucket turned red for 500ms

and lost 10 points. The number of points accumulated over the task was presented in the top right-hand corner for participants to track their performance. Payment was partially performance contingent; the more points earned, the higher amount of bonus they received at the end, up to a maximum of 1 USD on top of their flat fee of 7 USD. Confidence ratings were not incentivized.

On each trial, the particle's landing location on the circle edge was sampled independently and identically from a Gaussian distribution with SD = 12. As such, the particle landed in the same location with small variations determined by noise. The mean of this distribution did not change until a change-point trial was reached, where it was re-sampled from a uniform distribution U(1,360) (i.e. the number of points on the circle). Participants would therefore have to learn the mean of the new generative distribution after a change-point. The probability of a change-point occurring on each trial was determined by the hazard rate. In the task, there were two hazard rate conditions that varied the number of change-points in a stretch of 150 trials each: stable (hazard rate = 0.025, 4 change-points), and volatile (hazard rate = 0.125, 19 change-points). Hazard rate conditions were not relevant to the analyses of the current paper. The order of hazard rate conditions was randomly shuffled, as were the order of change-points within a condition. Participants completed 300 trials in total, divided into 4 blocks of 75 trials, with no explicit indication when a change in condition block occurred. Breaks were given between blocks which did not fall before the switch of a new hazard rate condition.

Before the start of the task, participants were instructed on the aim of the experiment and shown its layout. Participants then completed 10 practice trials that were excluded from the analysis and did not count for their final score. After the practice, they had to answer 5 quiz questions pertaining to the task instructions. If they answered any of the questions wrong, they would be brought back to the beginning of the instructions and taken through the practice block again. Additionally, in order to reduce the number of participants failing to utilize the confidence scale properly, the task was reset to the beginning if participants left their confidence ratings as the default score for more than 70% of the trials at the 20th and 50th trial mark. They would have to complete the instruction quiz again before proceeding with the task.

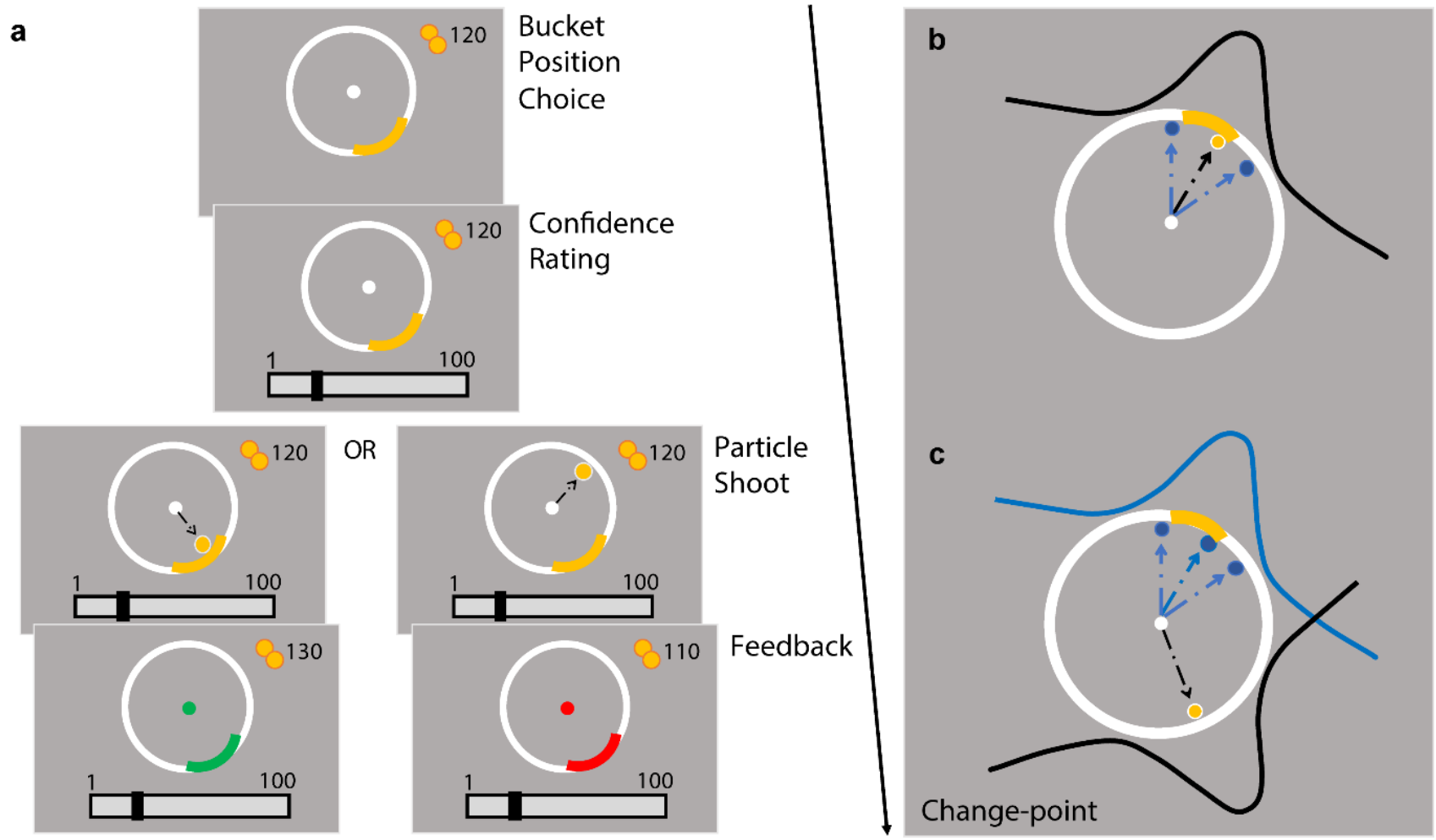


Figure 2.1. Predictive inference task.

Figure 2.1. Predictive inference task. (a) *Trial sequence. Participants were instructed to position a bucket (yellow arc on the circle edge) to catch a flying particle, and thereafter rated their confidence that they would catch the particle. Particles were fired from the centre of the circle to the edge. Points were gained when the particle was caught, and the bucket turned green; else, points were lost and the bucket turned red.* ***(b)*** *Particle trajectories. For every trial, landing locations were independently sampled from a Gaussian distribution. As such, particles landed around the same area with small variations induced by noise. For illustration purposes, dashed arrow lines represent particle trajectory of current (black) and past (blue) trials, which over trials allow subjects to generate a representation of the Gaussian.* ***(c)*** *Change-points. The mean of the distribution abruptly moves to another point on the circle when a "change-point" occurs. This new mean is then sampled in the same manner as* ***(b)*** *until the next change-point.*

Self-report psychiatric questionnaires & IQ. Participants completed a range of self-report psychiatric assessments after the behavioural task. To enable application of the transdiagnostic analysis with psychiatric dimensions described in previous studies (Gillan et al., 2016; Rouault et al., 2018), we administered the same nine questionnaires assessing: *alcohol addiction* using the Alcohol Use Disorder Identification Test (AUDIT) (Saunders et al., 1993), *apathy* using the Apathy Evaluation Scale (AES) (Marin et al., 1991), *depression* using the Self-Rating Depression Scale (SDS) (Zung, 1965), *eating disorders* using the Eating Attitudes Test (EAT-26) (Garner et al., 1982), *impulsivity* using the Barratt Impulsivity Scale (BIS-11) (Patton et al., 1995), *obsessive-compulsive disorder* (OCD) using the Obsessive-Compulsive Inventory - Revised (OCI-R) (Foa et al., 2002), *trait anxiety* using the trait portion of the State-Trait Anxiety Inventory (STAI) (Spielberger et al., 1983), *schizotypy* scores using the Short Scales for Measuring Schizotypy (Mason et al., 2005), and *social anxiety* using the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz, 1987). The administration order of these self-report assessments was fully randomized. Following the questionnaires, participants completed a Computerized Adaptive Task (CAT) based on items similar to that of Raven's Standard Progressive Matrices (SPM) (Raven, 2000) to approximate Intelligence Quotient (IQ).

Transdiagnostic factors (dimensions). Raw scores on the 209 individual questions that subjects answered from the 9 questionnaires were transformed into factor scores ('Anxious-Depression', 'Compulsive Behaviour and Intrusive Thought',

and ‘Social Withdrawal’), based on weights derived from a larger previous study (N = 1413) (Gillan et al., 2016) (***Supplemental Figure A.I.S6***). These factors are not orthogonal and therefore correlate moderately, with values ranging from r = 0.34 to 0.52.

Action-confidence coupling. Regression analyses were conducted using mixed-effects models written in R, version 3.5.1 via RStudio version 1.1.463 (http://cran.us.r-project.org) with the *lme4* package. We examined the coupling between trial-by-trial action updates (*Action*, the absolute difference of bucket position on trial *t* and *t*+1, as the dependent variable) and confidence (*Confidence*, confidence level on trial *t*+1, z-scored within-participant, as the independent variable) with age, gender and IQ as z-scored fixed effects co-variates. Within-subject factors (the intercept and main effect of *Confidence*) were taken as random effects (i.e. allowed to vary across subjects). To test if questionnaire total scores or transdiagnostic dimension severities were associated to changes in action-confidence coupling, the scores were included as z-scored between-subjects predictors in the basic model above. Separate regressions were performed for each individual questionnaire score due to high correlations across the different psychiatric questionnaires, whereas for the transdiagnostic analysis, we included all three psychiatric dimension scores in the same model, as correlation across variables was lessened in this formulation and thus more interpretable (only 3 moderately correlated variables $r = 0.34$ to 0.52, instead of 9 that ranged from $r = 0.13$ to 0.84). This allowed us to examine the association between CIT and

various task measures, after the relationships with other dimensions (AD and SW) were controlled for.

Action and confidence. A similar approach with mixed-effects models were used to analyse the basic relationship of questionnaire scores/transdiagnostic dimensions with *Action* or *Confidence* as dependent variables with the intercept as the random effect, controlled for age, gender and IQ.

Computation model describing behaviour dynamics. To employ model-based analysis, we followed the same analysis steps as Vaghi et al. (Vaghi et al., 2017). We calculated task prediction error (PE: distance between the particle landing location and the centre of the bucket) and human learning rate (LR^h: change in chosen bucket position from *t* to *t*+1 divided by PE on trial *t*) for each trial. Trials where LR^h exceeded the 99th percentile ($LR^h > 7.75$) and PE = 0 are thought to be unrelated to error-driven learning (Nassar et al., 2016), and were thus excluded from analyses with the model parameters (3.05% of total trials).

In the task, several evidence sources were available to participants (e.g. new information gained, surprise from unexpected outcomes and uncertainty of their belief) to estimate the mean of the generative distribution in order to position their bucket at where they hope to catch the greatest number of particles. We fitted a quasi-optimal Bayesian learning model, identical to the model specified in Vaghi et al. (Vaghi et al., 2017) using functions freely available online from the original study,

to particle landing location data in MATLAB R2018a (The MathWorks, Natick, MA). The model estimates parameters are thought to underlie task dynamics. This included PE^b (model prediction error, an index of recent outcomes), *CPP* (change-point probability, a measure representing the belief of a surprising outcome) and *RU* (relative uncertainty, the uncertainty owing to the imprecise estimation of the distribution mean; labelled as *(1-CPP)*(1-MC)* in Vaghi et al.). Similar to Vaghi et al., the correlations between model parameters were moderately correlated, with the largest correlation between PE^b and CPP: r_s = 0.68 (see ***Supplemental Table A.I.S1***). This is because large deviations necessarily induce a CPP of near 1 and small deviations a CPP near 0.

The reduced quasi-optimal Bayesian learner, in accordance with prior literature (McGuire et al., 2014; Nassar et al., 2010, 2016; Vaghi et al., 2017), uses a delta-rule to update its estimate of belief about the particle landing location distribution:

$$B_{t+1} = B_t + \alpha_t \times \delta_t$$

B is the new belief estimate on each trial *t*, which is equal to a point estimation of the mean of the Gaussian distribution where particle locations were sampled (i.e. 1 to 360). Its update is dependent on the learning rate α (LR^b) and model prediction error δ (PE^b). PE^b is calculated as the difference between the belief estimate B_t and the new particle landing location X_t and is a measure of information gained from the most recent trial.

$$\delta_t = X_t - B_t$$

As with common reinforcement learning models, LR^b determines how much new information (PE^b) will update the belief estimate. However, LR^b is dynamic in the current model i.e. can change on every trial. If $LR^b = 0$, new evidence has had no impact on the update of the belief estimate, while $LR^b = 1$ suggests that the new belief estimate is entirely determined by the most recent outcome. The magnitude of LR^b is dependent the statistics of environment with the equation:

$$\alpha_t = \Omega_t + (1 - \Omega_t)(1 - \nu_t)$$

The first term, the change-point probability Ω (CPP), represents an estimate of how likely a change in particle location distribution mean has occurred on a given trial. The second term, model confidence ν (MC), represents the uncertainty due to an inaccurate estimation of the mean. For regression analyses, $(1 - \Omega)(1 - \nu)$ was labelled as RU (as the additive inverse of MC is relative uncertainty). These two components allow the model to appropriately update belief according to (i) unexpected changes in the environment (change-points) and (ii) the uncertainty about the distribution mean—thus informing when to disregard outliers when the mean is certain. New outcomes are more influential when the model believes that the distribution mean has changed (i.e. CPP is large) or is less sure about the true distribution mean (i.e. MC is small).

The model generates CPP as the relative likelihood that a new particle location is sampled from a new distribution during a change-point (mean determined by a uniform distribution *U* over all 360 possible locations) or drawn from the same Gaussian (*N*) where the current belief estimate B_t is centered upon. These are influenced by the hazard rate *H*, the probability that the mean of the distribution has changed. We set *H* equal to the hazard rates of the task trials (which were either *H* = 0.025 or 0.125, depending on the block condition). When the probability of the new particle location coming from a new distribution is high, CPP will be close to 1.

$$\Omega_t = \frac{U(X_t \mid 1{,}360)H}{U(X_t \mid 1{,}360)H + N(X_t \mid B_t, \sigma_t^2)(1-H)}$$

σ_t^2 is the estimated variance of the predictive distribution, which consists of the variance of the generative Gaussian distribution σ_N^2 and the generative variance modulated by MC (ν). As the predictive distribution variance is dependent on MC, it is larger than the generative variance where MC is the smallest (i.e. after change-points, where uncertainty of the new distribution mean is the highest) and will slowly reduce towards the generative variance. Thus, the model describes particle locations occurring after a change-point as less likely sampled from another new distribution.

$$\sigma_t^2 = \sigma_N^2 + \frac{(1-\nu_t)\sigma_N^2}{\nu_t}$$

Lastly, MC is computed for the subsequent trial with a weighted average of the generative variance conditional on a change-point (first term), generative variance conditional on no change-point (second term), and variance due to the model's uncertainty of whether a change-point occurred (third term) in the numerator. The denominator includes the same terms plus just the generative distribution variance (σ_N^2) representing the uncertainty owing to noise. The full equation is as follows:

$$\nu_{t+1} = \frac{\Omega_t\sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t\nu_t)^2}{\Omega_t\sigma_N^2 + (1-\Omega_t)(1-\nu_t)\sigma_N^2 + \Omega_t(1-\Omega_t)(\delta_t\nu_t)^2 + \sigma_N^2}$$

Influence of parameters on action and confidence. Regressions were constructed as mixed-effect models with all of the model parameters (where PE^b is taken as its absolute) and a *Hit* categorical predictor (previous trial was a hit or miss) as within-subject regressors, controlled for age, IQ and gender. These regressions control for shared variance. For the regression on *Action*, following prior literature (McGuire et al., 2014; Nassar et al., 2010, 2016; Vaghi et al., 2017), all predictors except PE^b were implemented as interaction terms with PE^b. For *Confidence*, we used a similar regression model but without the interaction term with PE^b and with the predictors z-scored at participant level. We obtained similar regression estimates with Vaghi et al. (Vaghi et al., 2017), suggesting that action/confidence was

appropriately updated with these parameters describing belief updating (***Supplemental Table A.I.S2***). To investigate the relationship of the questionnaire scores and psychiatric dimensions with the influence of these parameters on action/confidence, we included these scores as z-scored fixed effect predictors in the basic model above (individual models were examined for each questionnaire score, while for the transdiagnostic analysis, all three psychiatric dimension scores were included together).

There was some evidence of heteroskedasticity in the association between psychiatric variables and task parameters. White's tests indicated that the model of RU on confidence with psychiatric dimensions was heteroskedastic (p = 0.04, but not the other parameters: p > 0.12). We therefore estimated heteroskedasticity-consistent standard errors for all coefficients reported by the *vcovHC* function from the *sandwich* package in R, detailed in the Supplement (***Supplemental Table A.I.S3***). The results do not diverge from those reported in the main paper.

For details of all regression equations, see ***A.I. Supplemental Methods***.

Data Availability. The code and data to reproduce the main analyses are freely available in an Open Science Framework (OSF) repository, at https://osf.io/2z6tw/.

Results

Action-confidence decoupling is linked to various psychiatric phenomena. In line with prior research, size of action updates (bucket position difference from trial *t* and *t*+1) were strongly related to confidence within-subjects (β = -8.85, *Standard Error (SE)* = 0.31, 95% Confidence Interval (CI) [-9.45, -8.25], $p < 0.001$) (***Figure 2.2a***), such that lower confidence was linked to larger updates, in the sample as a whole. Previous work by Vaghi et al. found that OCD patients exhibited reduced coupling between action and confidence compared to controls, which was correlated to the severity of self-reported OCD symptomology within the patient sample (Vaghi et al., 2017). We tested the latter in a general population sample and replicated this result; OCD symptom severity was associated with significantly lower action-confidence coupling (β = 1.30, *SE* = 0.21, 95% CI [0.89, 1.71], $p < 0.001$, corrected) (***Figure 2.2b***). However, we found that this relationship was profoundly non-specific—all nine questionnaire scores tested showed a similar pattern of reduced coupling. 6/9 questionnaires (alcohol addiction, depression, eating disorders, impulsivity, OCD and schizotypy) had significant decoupling at $p < 0.001$ corrected; the remaining three (apathy, social anxiety, trait anxiety) trended in the same direction, but did not survive Bonferroni correction for multiple comparisons.

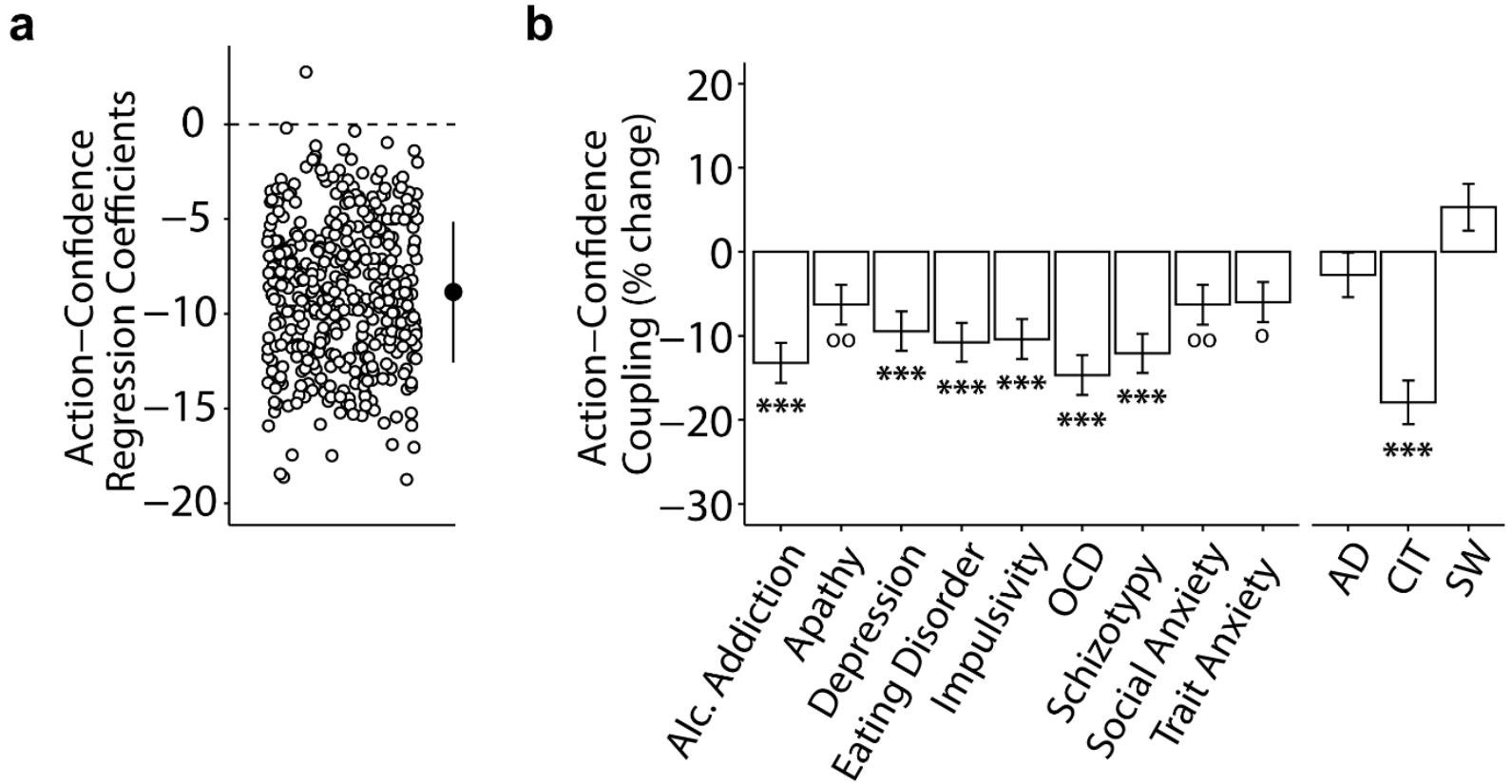


Figure 2.2. Action-confidence coupling and its relationship with questionnaire scores/dimensions (controlled for IQ, age and gender).

Figure 2.2. Action-confidence coupling and its relationship with questionnaire scores/dimensions (controlled for IQ, age and gender). *AD: anxious-depression, CIT: compulsive behaviour and intrusive thought, SW: social withdrawal.*

(a) *Regression model where action update is predicted by confidence. Individual coefficients are represented by circles. Marker indicates the mean and standard deviation. As expected, regression coefficients were negative, such that higher confidence was associated with smaller updates to the bucket position ('action').*

(b) *Associations between action-confidence coupling and questionnaire scores or psychiatric dimensions. All questionnaire scores predicted a decrease in action-confidence coupling. This decoupling relationship was strongest for the CIT dimension.*

Each questionnaire score was examined in a separate regression, while dimensions were included in the same regression model. The Y-axes shows the percentage change in the size of the action-confidence coupling effect as a function of 1 standard deviation increase of questionnaire/dimension scores. Error bars denote standard errors. $^{o}p < 0.05$, $^{oo}p < 0.01$ *uncorrected;* $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ *corrected. Results are Bonferroni corrected for multiple comparisons over number of psychiatric independent variables investigated. See also* ***Supplemental Figure A.I.S8*****.**

Transdiagnostic analysis shows a more specific pattern. When we refactored the data into three transdiagnostic dimensions defined previously in the literature, a profoundly different picture emerged. Compulsive behaviour and intrusive thought (CIT) was the only dimension to show decreased action-confidence coupling (β = 1.57, *SE* = 0.23, 95% CI [1.13, 2.01], $p < 0.001$, corrected) (***Figure 2.2b***). Thus, while reductions in action-confidence coupling show broad and non-specific relationships to all questionnaire scores studied here, this pattern is explained by a single transdiagnostic dimension.

Compulsivity is linked to inflated confidence, not aberrant action-updating. Prior work using this task in diagnosed patients found no confidence biases in OCD, but abnormalities in action-updating. Using our transdiagnostic method, we found a strikingly different pattern of results. CIT was associated with higher overall confidence levels (β = 6.74, *SE* = 1.02, 95% CI [4.75, 8.73], $p < 0.001$, corrected), and not changes in action-updating. In line with prior work, we found that anxious-depression (AD) was associated with lower confidence (β = -3.42, *SE* = 1.04, 95% CI [-5.45, -1.39], $p = 0.003$, corrected) (***Figure 2.3a***). Because OCD patients tend to have high levels of AD, this finding suggests that a transdiagnostic method may be necessary to reveal the role confidence plays in clinical phenotypes, which could otherwise be obscured within the heterogeneous diagnostic category. In terms of action-updating, only social withdrawal (SW) showed an association, such that participants scoring high in this dimension tended to move the bucket more (β = 0.89, *SE* = 0.28, 95% CI [0.34, 1.45], $p = 0.005$, corrected) (***Figure 2.3b***).

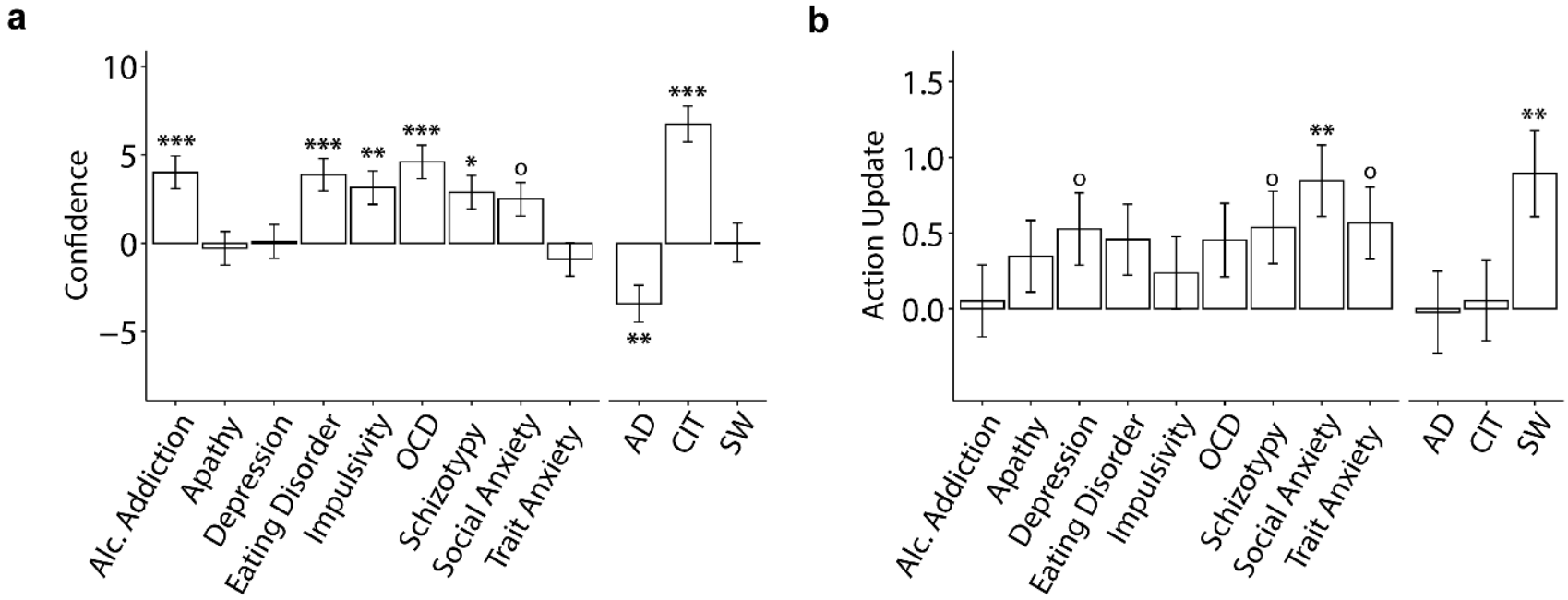


Figure 2.3. Associations between questionnaire scores, or dimensions, (controlled for IQ, age and gender) with confidence or action.

Figure 2.3. Associations between questionnaire scores, or dimensions, (controlled for IQ, age and gender) with confidence or action. *AD: anxious-depression, CIT: compulsive behaviour and intrusive thought, SW: social withdrawal.*

(a) *Associations with confidence rating on each trial. Most of the questionnaires scores were positively associated with confidence. However, refactoring into transdiagnostic dimensions revealed previously obscured bidirectional associations. AD was linked to decreased confidence, while CIT was linked to increased confidence.*

(b) *Associations with action updates (i.e. bucket movement from one trial to the next). Only social anxiety was associated with an increased tendency to move the bucket, and this was similarly captured by, and specific to, the SW dimension.*

The Y-axes shows the percentage decrease in the size of the confidence or action update as a function of 1 standard deviation increase of questionnaire/dimension scores. Error bars denote standard errors. $^{o}p < 0.05$, $^{oo}p < 0.01$ *uncorrected,* $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. *Results are Bonferroni corrected for multiple comparisons over the number of psychiatric independent variables investigated.*

Confidence in compulsivity is less sensitive to unexpected outcomes, environment uncertainty and positive feedback. The previous analyses suggest that confidence in compulsive individuals is both inflated and decoupled to behaviour. To understand the mechanism behind this, we tested the extent to which confidence estimates were sensitive to multiple factors that should drive belief-updating. Specifically, prior work has shown that trial-wise adjustments in behaviour are influenced by 1) information gained from the most recent change in particle location (PE^b; model prediction error), 2) surprising large particle location changes owing to change-points (CPP; change-point probability) and 3) uncertainty of one's belief about the particle landing location distribution mean (RU; relative uncertainty) (McGuire et al., 2014). To separate the contributions of these factors, we computed the three normative parameters with a quasi-optimal Bayesian model (McGuire et al., 2014; Nassar et al., 2010, 2016; Vaghi et al., 2017) (see ***Methods***) to the sequence of particle locations experienced by each participant.

We analysed trial-wise confidence using regression models with these parameters including a categorial Hit regressor (previous trial was a hit or miss), and controlled for age, gender and IQ. Overall, confidence was influenced by PE^b, CPP, RU and Hit (***Supplemental Table A.I.S2***). The CIT symptom dimension was associated with a significantly diminished influence of CPP (β = 0.05, *SE* = 0.01, 95% CI [0.03, 0.08], $p < 0.001$, corrected), RU (β = 0.05, *SE* = 0.01, 95% CI [0.03, 0.07], $p < 0.001$, corrected) and Hit (β = -0.03, *SE* = 0.01, 95% CI [-0.05, -0.01], $p = 0.003$, corrected) on confidence (***Figure 2.4*** and ***Supplemental Figure A.I.S5a***). In other words,

confidence estimates in CIT were less sensitive to unexpected outcomes, the uncertainty of the true distribution mean and whether the previous particle was caught (i.e. correct trial). These results suggest that confidence in highly compulsive individuals is not only inflated, it is also disconnected to several sources of environmental evidence. Interestingly, the failures in utilizing evidence do not explain away overall inflated confidence observed in CIT ($\beta = 6.78$, $SE = 1.02$, 95% CI [4.79, 8.76], $p < 0.001$, corrected), suggesting these are at least partially distinct metacognitive failures. There were no associations between AD or SW and the extent to which evidence influenced confidence (***Figure 2.4***).

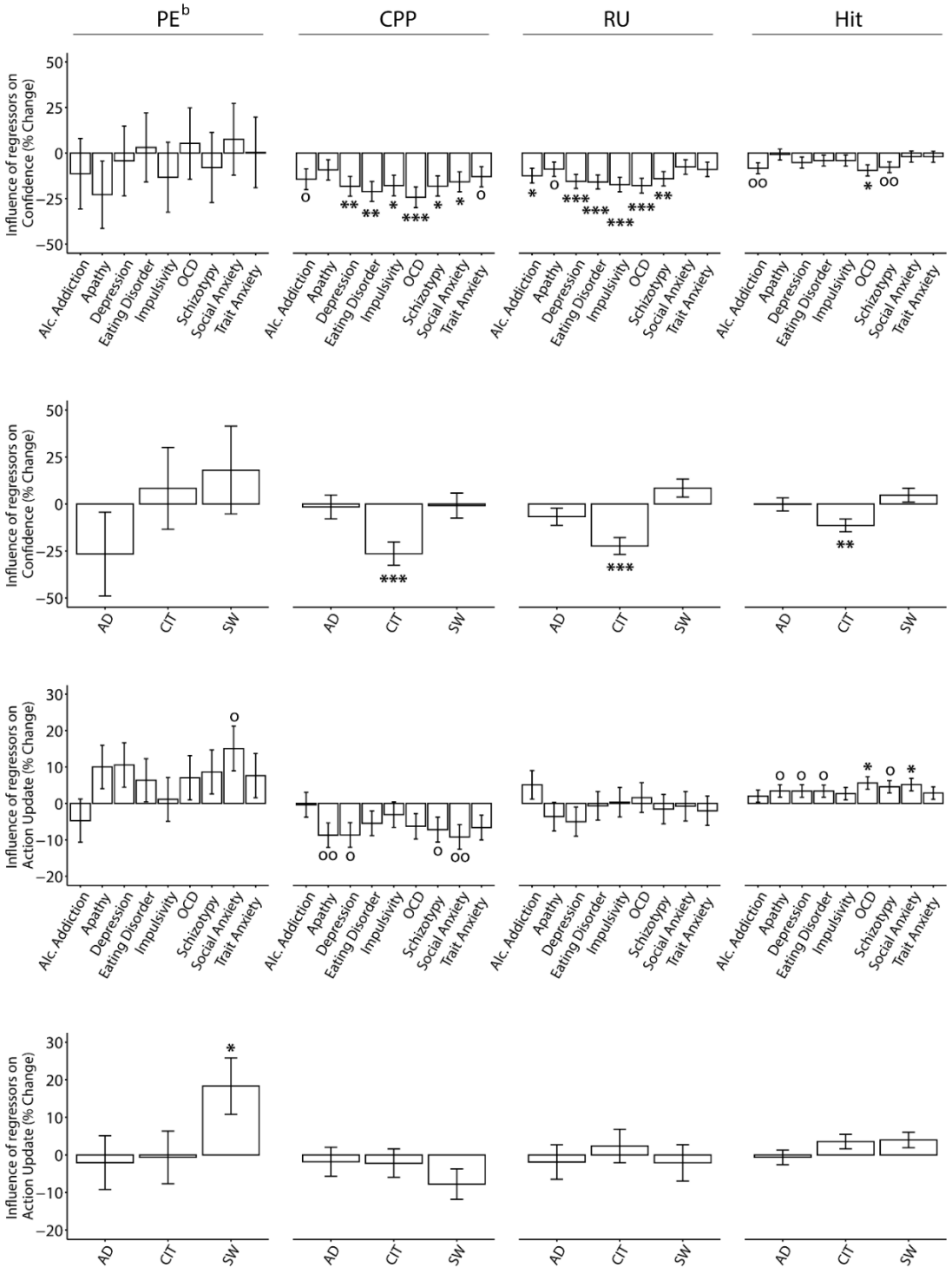


Figure 2.4. Regression analyses of trial-wise confidence and action adjustments with questionnaire scores/dimensions, controlled for age, IQ and gender.

Figure 2.4. Regression analyses of trial-wise confidence and action adjustments with questionnaire scores/dimensions, controlled for age, IQ and gender. *Predictors for confidence and action update regressions include model parameters PE^b (model prediction error), CPP (change point probability), RU (relative uncertainty) and Hit (previous trial was a hit). They are indicated at the top of the figure for each column inset. Each questionnaire score was examined in a separate regression, whereas dimensions were included in the same model (AD: anxious-depression, CIT: compulsive behaviour and intrusive thought, SW: social withdrawal). Error bars denote standard errors. The Y-axes indicate the percentage change in the size of the model parameter on confidence or action update effect as a function of 1 standard deviation of questionnaire/dimension scores. $^o p < 0.05$, $^{oo} p < 0.01$ uncorrected, *p < 0.05, ***p < 0.001. Results are Bonferroni corrected for multiple comparisons over the number of psychiatric independent variables investigated. See also* ***Supplemental Table A.I.S3*** *and* ***Figure A.I.S5****.*

Action updates in compulsivity respond appropriately to evidence. Using the same approach described for confidence, trial-wise action adjustments were influenced by all model parameters (***Supplemental Table A.I.S2***). In contrast to confidence, CIT was not linked to changes in the influence of any of the parameters on action (***Figure 2.4*** and ***Supplemental Figure A.I.S5b***). SW was related to a significant increased influence of PE^b, suggesting that individuals high in this trait had an increased tendency to update their action with every new outcome ($\beta = 0.06$, $SE = 0.02$, 95% CI [0.02, 0.09], $p < 0.05$, corrected) (***Figure 2.4***). There were no associations with AD. Additional analyses in ***Appendix I*** show that when demographics are not controlled for, some apparent associations between action-updating and compulsivity emerge that correspond to those reported previously in OCD (Vaghi et al., 2017) (***Supplemental Figure A.I.S7***).

Discussion

In this study, we demonstrated that a breakdown in the relationship between explicit belief (confidence) and behaviour is associated with a transdiagnostic psychiatric dimension—compulsive behaviour and intrusive thought (CIT). This decoupling arises from abnormalities in belief, rather than behaviour. Individuals high in CIT exhibited overall inflated confidence estimates and failures in utilising unexpected outcomes, belief uncertainty and positive feedback to inform their confidence levels appropriately. In contrast, action-updating in response to these factors did not differ as a function of CIT severity. Our findings suggest a dysfunctional metacognitive mechanism in compulsivity that implicates difficulty in updating the explicit model of the world in response to various sources of evidence.

Existing models of compulsivity propose that deficits in goal-directed control leave individuals vulnerable to establishing compulsive habits (Gillan & Robbins, 2014). Supporting evidence primarily come from behavioural tests, where OCD patients (and other compulsive disorders) have difficulty exerting control over well-trained habits when motivations change (i.e. a devaluation test) (Ersche et al., 2016; Gillan et al., 2011; Gillan, Apergis-Schoute, et al., 2015; Gillan, Otto, et al., 2015). Other tasks have shown that compulsive patients have deficits utilizing a world model to make choices prospectively (even when habits are not present), relying instead solely on reinforcement (i.e. feedback) to direct choice (Gillan, Morein-Zamir, Kaser, et al., 2014; Voon et al., 2015). Our current finding, that high compulsive individuals fail to update their world model in response to several types of evidence, is an

important extension of this literature. The challenge facing compulsive individuals has until now been presumed to be the implementation of the model rather than its generation and/or maintenance (Gillan & Robbins, 2014). This implicates not just our understanding of compulsive disorders, but also their treatment. Recent work has shown that metacognitive skills can be improved though adaptive training (Carpenter et al., 2019); there may be scope for trialling these treatments for psychiatric populations where compulsivity is an issue.

Confidence was not just unresponsive to various factors underlying learning, it was also inflated in compulsive individuals. This finding replicates prior work using this same transdiagnostic methodology that examined confidence in the context of perceptual decision-making (Rouault et al., 2018), showing elevated confidence in compulsivity extends to reinforcement learning—which is of course highly relevant for the behavioural aspects of compulsivity. However, this finding of increased confidence in compulsivity may seem at odds with prior research that overall suggests a decrease of confidence in OCD (Hoven et al., 2019). Critically, these studies have primarily been conducted with patient versus control comparisons and with respect to confidence in memory, rather than decision-making.

There are several factors that are important to consider as this field progresses. Firstly, many memory tasks have not controlled for differences in performance in the way that tasks in the perceptual/reinforcement learning domains have been able to, which introduces a confound to interpretation (Fleming & Lau, 2014). If subjects

perform worse at the memory task, untrue conclusions about under-confidence can arise. This is the case in some, but not all, tasks that have studied confidence in memory in OCD. Secondly, it has been demonstrated that metacognition is not a unitary phenomenon; there is specialisation in distinct brain regions for confidence in perceptual versus memory domains (Fleming et al., 2014). We therefore might reasonably expect different patterns of dysfunction in OCD, depending on the domain under study. Finally, and perhaps most importantly, these prior studies are based on group comparisons (usually patient versus control) and cannot account for the influence of co-morbid symptomatology (e.g. depression and anxiety) on confidence. Given depression is associated with decreases in confidence, the confluence of these symptoms might serve to mask the true direction of the relationship between compulsivity and confidence.

In instances when confidence diverges from action, prior work has suggested confidence estimates may be corrupted by noise, internal states or a continued/lack of evidence processing (Fleming & Daw, 2017; Meyniel et al., 2015). Coupled with the finding that confidence is less informed by several sources of evidence in high compulsive individuals, it is possible that inflated confidence in compulsivity observed here arose through some unmodeled form of information processing. In contrast, we found that actions were updated normally in response to feedback in high compulsive individuals, which accords with prior work showing that basic reinforcement learning in compulsive patient groups (i.e. ‘model-free’ learning) is intact (Gillan et al., 2016; Voon et al., 2015). That said, a previous study using this

task found increased action-updating tendencies in OCD (Vaghi et al., 2017). Here, the discrepancy does not appear to be explained by the superiority of a transdiagnostic approach per se, but our ability to control for some demographic confounds (see ***Supplemental Figure A.I.S7***). Instead, we found that social withdrawal (SW) was associated with a higher sensitivity to new information affecting action. Though this result was not hypothesized, it aligns with prior research suggesting that socially anxious people engage in greater performance monitoring (Judah et al., 2016) and have higher sensitivity to learning from feedback (Khdour et al., 2016).

Beyond the specific results of this study with respect to confidence and compulsivity, our data highlight the benefit of transdiagnostic dimensions over traditional modes of phenotyping. When we examined questionnaires that are ubiquitous but rarely compared to one another in clinical research, we found strikingly non-specific patterns of association with task variables. For example, all nine questionnaires showed an association with action-confidence coupling in the same direction (6/9 surviving strict correction). In contrast, the compulsive dimension was the only transdiagnostic dimension to show an association. In addition to resolving issues with collinearity across questionnaires, this approach also resolves issues associated with the heterogeneity within them. For example, severity of neither depression nor anxiety was associated with decreases in confidence using a standard clinical questionnaire, but the anxious-depression (AD) dimension was. In comparison to work with diagnosed patients, the benefits of the transdiagnostic

approach are the same. Prior work using this task found no difference in OCD patients' mean confidence ratings compared to healthy controls (Vaghi et al., 2017), while we found a strong a reproducible association between CIT and inflated confidence and AD and diminished confidence (Rouault et al., 2018), at least in the context of decision-making. Given that OCD is frequently co-morbid with anxiety disorders (over 75% (Ruscio et al., 2010)), which has an opposing relationship to confidence, it is no surprise that differences between OCD patients and controls are obscured when transdiagnostic dimensions are not considered. Together, these data suggest that transdiagnostic phenotyping may, at least in some domains, provide a closer fit to underlying brain processes than DSM distinctions.

This study was not without limitation. Our study was conducted online, thus experimenter control of the testing environment was virtually non-existent. Prior studies have however shown that cognitive data collect online, albeit noisier, is valid (Crump et al., 2013). Similarly, self-report psychiatric scores are on-par with the general population (Shapiro et al., 2013) and relationships between cognition and clinical measures are mirrored across testing modalities (Gillan et al., 2016; Snorrason et al., 2016). Additionally, as the task was adapted for web-based testing, response navigation was controlled by keyboard presses (right and left response keys to direct clockwise and counter-clockwise rotations) and not a rotor controller as in Vaghi et al. (Vaghi et al., 2017), which could plausibly feel less natural and thus increase noise in spatial update measure. However, we were able to reproduce basic main effects of model parameters on action from Vaghi et al., suggesting that

the different response modality did not affect our ability to study action updating behaviour. We observed similar basic main effects of model parameters on action updating as in the original paper (***Supplemental Table A.I.S2***). With respect to the psychiatric dimensions, we not only reproduced the factor structure from a prior paper with our current data (***Supplemental Figure A.I.S6***), we used the factor weights from this prior publication (Gillan et al., 2016) to transform raw questionnaire scores into transdiagnostic dimensions for analysis. This ensured independence and underscores the robustness and reproducibility of these dimensions and their association to cognition.

The extent to which these results are applicable to diagnosed patients is not something we can directly address here. However, it is notable that we replicated the association between OCD symptoms and action-confidence decoupling observed in a clinical sample that were tested in-person (Vaghi et al., 2017). The same applies to goal-directed planning, which is both deficient in patients tested in-person (Gillan et al., 2011) and correlated with OCD symptoms in the general population tested online (Gillan et al., 2016; Snorrason et al., 2016). Notably, recent work in a mixed generalised anxiety disorder and OCD patient sample that were tested online found that goal-directed deficits were more strongly associated with the compulsivity dimension than OCD diagnosis status, underscoring the importance of transdiagnostic methods for delineating specific associations between cognition and clinical phenomenology that can be masked when examining diagnostic status alone (Gillan et al., 2019). Future research is needed to investigate

if the association between inflated confidence and compulsivity is similarly evident in diagnosed patients, tested in-person. More concerted, multi-centre efforts are required to achieve the large samples necessary to undertake this work, if it is to take place in-person rather than online.

To conclude, we highlighted how a transdiagnostic methodology can be crucial for uncovering specific associations between pathophysiology and clinical symptoms. This method has several strengths; it directly addresses the issue of psychiatric co-morbidity, helps us to achieve higher statistical power and thus promotes reproducibility, and makes research faster, more efficient and even more representative (Gillan & Daw, 2016). The definition of compulsivity employed here was generated in an independent study and is not intended to be fixed or final. Rather, its application to this independent dataset is intended to show the general potential that transdiagnostic approaches have for dealing with the issues of co-morbidity and individual differences faced both in research and practice in psychiatry. We used this method to show that compulsive behaviour and intrusive thought is associated with reduced action-confidence coupling, inflated confidence and diminished influence of evidence on confidence estimates. Our findings suggest that compulsivity is linked to problems in developing an explicit and accurate model of the decision space, and this might contribute to broader class of problems these individuals face with goal-directed planning and execution.

Chapter 3: Encephalographic (EEG) correlates of reduced model-based control in compulsivity

Introduction

Compulsive behaviour manifests as actions that are autonomous, out-of-control and repetitive, often leading to adverse and functionally impairing outcomes (Robbins et al., 2012). This symptomology is characteristic of psychiatric disorders like obsessive-compulsive disorder (OCD) and addiction, and is thought to arise from an imbalance between the two modes of action control (Gillan & Robbins, 2014): (i) goal-directed learning that links actions to outcomes and enables adaptive behaviour (Balleine & Dickinson, 1998) and (ii) rigid habit formation that relies on reflexive stimulus-response mechanisms to guide decision making (Balleine & O'Doherty, 2010).

An accumulation of evidence suggests that the dominance of habits in compulsivity may result entirely from impairments in goal-directed control (Gillan & Robbins, 2014). Imaging work has shown that behavioural insensitivity to rapid changes in outcome value ("outcome devaluation") and self-reported habitual urges in OCD are associated with the hyperactivation of the caudate nucleus, a brain region associated with goal-directed control, but not habit-related regions (Gillan et al., 2015a). OCD patients are also poorer in prospective planning in making goal-oriented decisions compared to controls (Gillan et al., 2014). This view is further supported by studies utilising the two-step reinforcement task which computationally

frames goal-directed control as ‘model-based’ learning according to the extent that individuals use state-action relationships to guide choice (Daw et al., 2005, 2011). Individuals with OCD and other compulsive disorders are impaired in model-based planning (Voon et al., 2015). Building on this work, evidence suggests that compulsivity is a transdiagnostic dimension (rather than a category of mental illness) which manifests in failures in goal-directed control both in the general population (Gillan et al., 2016), as well as in diagnosed patients (Gillan et al., 2019).

Despite these advances in our understanding of the nature of the deficits experienced by compulsive individuals, it remains unclear *which component of* goal-directed control is impaired in compulsivity. Goal-directed behaviour is unarguably a multifaceted cognitive capacity; it depends upon several concurrent functions that include: (i) the construction and maintenance of an internal model (i.e. a representation of the environment, and more specifically, knowledge of relevant action-outcome relationships and state-state transitions), which is a pre-requisite for (ii) the implementation of this model in behaviour through prospective planning. Goal-directed failures could theoretically stem from problems in mechanisms underlying either component.

However, the general framing in the literature to date has focused largely on implementation; e.g. arbitration of the *balance* between competing habitual versus goal-directed decision systems which is suspected to occur at the time of choice (Gruner et al., 2016; Lee et al., 2014). This postulate has been supported in part by

the observation that OCD patients exhibit failures in devaluation sensitivity, even when they have accurate explicit knowledge of simple action-outcome contingencies (Gillan, Morein-Zamir, Kaser, et al., 2014). However, other studies have revealed problems in learning action-outcome associations in OCD and addiction alike using outcome devaluation paradigms that require subjects to learn more numerous and taxing contingency structures (Ersche et al., 2016; Gillan et al., 2011). Additionally, this lack of action-outcome contingency knowledge is correlated with devaluation sensitivity in OCD (Gillan et al., 2011).

Although lacking a direct link to goal-directed behaviour, studies of metacognitive processes in compulsive individuals have begun to mount evidence that rather than problems in *implementation* per se, the status of the internal model may be compromised in compulsivity. For example, in a perceptual decision making task, individuals high in compulsivity overall exhibit greater confidence in their decisions, but critically also showed a trend towards poorer accuracy of this mental model of their own performance—that is, they were less able to detect when they were correct versus incorrect (Rouault et al., 2018). In the previous chapter, we probed the link between reinforcement learning and metacognition. Notably, we found that compulsivity was again linked to inflated confidence, but also stark impairments in the ability to update confidence estimates in light of feedback (Seow & Gillan, 2020). This suggests that in compulsivity, goal-directed deficits may not simply reflect failures in 'cognitive control', or arbitration between decision systems, but stem from

issues in acquiring and maintaining an accurate internal model of the environment—upon which goal-directed control is predicated.

The present study aimed to test if compulsivity is characterised by a disruption in *constructing and maintaining* the mental model essential for goal-directed control or simply in the *use/implementation* of it. To do this, we used electroencephalography (EEG) to concurrently track neural responses as 192 subjects performed a two-step reinforcement learning task (Daw et al., 2005, 2011). We conducted single-trial regression analyses to characterise candidate neural correlates of the representation and implementation of the mental model and tested if these were linked to individual differences in compulsivity.

Methods

Power estimation. We determined a minimum sample size from a prior study that investigated the association of goal-directed control with OCI-R scores from non-clinical participants tested in-person (r = -0.26, p < 0.05) (Snorrason et al., 2016). The effect size suggested that N = 150 participants were required to achieve 90% power at 0.05 significance. Our final sample was larger than this to achieve the required power for a simultaneous investigation of other data collected from the same subjects; which is unrelated to the present study and reported in chapter 4 (Seow et al., 2019).

Participants. N = 234 participants were tested, of whom 138 were female (58.97%) with ages ranging from 18 to 65 (mean = 31.42, standard deviation (SD) = 11.48) years. Majority of the participants were from the general public, recruited via flyers and online advertisements, while a tiny subset (N = 8) were patients from St. Patrick's Mental Health hospital who were included to enrich our sample in self-report mental health symptoms. All participants were ≥18 years (with an age limit of 65 years) and had no personal/familial history of epilepsy, no personal history of neurological illness/head trauma nor personal history of unexplained fainting. Subjects were paid €20 Euro (€10/hr) upon completion of the study. All study procedures were approved by Trinity College Dublin School of Psychology Research Ethics Committee.

Procedure. Before presenting to the lab for in-person EEG testing, participants completed a brief at-home assessment via the Internet. They provided informed electronic consent, and submitted basic demographic data (age, gender),

information about any medication they might be taking for *a mental health issue* and completed a set of 9 self-report psychiatric questionnaires (see **Self-report psychiatric questionnaires, transdiagnostic dimensions & IQ**). During the experimental EEG session, participants completed two tasks: the modified Eriksen flanker task (Eriksen & Eriksen, 1974) and the two-step reinforcement learning task (Daw et al., 2005, 2011). Data from the former task data has no bearing on the results presented here (Seow et al., 2019), with the exception that we reported the basic behavioural association with compulsivity and model-based planning as a control measure in that paper. Once participants had completed both tasks, they completed a short IQ evaluation before debriefing. A subset of the participants (N = 110, 47%) completed a short psychiatric interview (Mini International Neuropsychiatric interview English Version 7.0.0; M.I.N.I.) (Sheehan et al., 1998) before the experimental tasks. Further medication and diagnosis details of the sample are in ***A.II. Supplemental Methods***.

Participant exclusion criteria. Several exclusion criteria were applied to ensure data quality. Participants were excluded if they failed any of the following on a rolling basis: Participants whose/who (i) EEG data were incomplete (N = 5) (i.e. recording was prematurely terminated before the completion of the task) or corrupted (N = 2), (ii) EEG data which contained excessive noise (i.e. >95% EEG epochs from the individual failing epoch exclusion criteria, see **EEG recording & pre-processing**) (N = 4), (iii) responded with the same key in stage one >90% (n > 135 trials) of the time (N = 10), (iv) probability of staying after common-rewarded trials was significantly worse than chance, defined as <5% probability of fitting a binomial distribution with 50% (chance) probability and the total

number of common-rewarded trials experienced by each subject (N = 11), (v) missed more than 20% of trials (n > 30 trials) (N = 3), and (vi) incorrectly responded to a “catch” question within the questionnaires: “If you are paying attention to these questions, please select ‘A little’ as your answer” (N = 7). Combining all exclusion criteria, 42 participants (17.95%) were excluded. N = 192 participants were left for analysis (115 females (59.90%), between 18-65 ages (mean = 31.55, SD = 11.75 years).

Two-step reinforcement learning task. We used the two-step reinforcement-learning task (Daw et al., 2005, 2011) to assess individual differences in model-based planning. Participants had to navigate two stages to learn reward probabilities associated with options presented, with the main goal of earning rewards. The paradigm was presented with a cover story (***Figure 3.1***). In the first stage, participants had to choose between two spaceships, each with a higher probability (‘common’ transitions: 70%) of leading to one of two planets (second stage states, represented by coloured blocks) that sometimes lead to the alternative (‘rare’ transitions: 30%) planet. Once on the planet, the participants then had to choose between two aliens to be probabilistically rewarded with ‘space treasure’, or unrewarded with ‘space dust’. Each alien, a total of four over two planets, had a unique probability of receiving ‘space treasure’, which drifted slowly and independently over time (always >0.25 or <0.75). Individuals performing goal-directed ('model-based') learning would make decisions based on the history of rewards and the transition structure of the task, while individuals performing basic temporal difference ('model-free') learning would simply make decisions solely on the history of rewards obtained.

The sequence of events was presented in the same manner as a prior study that conducted the two-step task in the EEG (Eppinger et al., 2017) except with differing transition probabilities (70/30% here versus 60/40% and 80/20%) and time given to make a choice (1500ms here versus 2000ms) (***Figure 3.1***). On each trial, participants were first presented with a fixation cross for 500ms, and then shown a choice between two spaceships. They had 1500ms to respond; after which, an outline over the chosen option would indicate their choice (feedback) for 500ms. A fixation cross was shown for 500ms before transition, where the transitioned planet was shown (a blank colour block) for 1000ms. Two aliens of that particular planet would then appear, with 1500ms for choice, with feedback of the chosen option subsequently shown for 500ms. Each of the aliens led to a probabilistic reward with a picture of 'space treasure', or no reward with 'space dust', that was presented for 1000ms. Responses were indicated using the left ('Q') and right ('P') keys. Colour of blocks behind rockets and those representing planets were randomised across all participants. Participants performed two blocks of 75 trials, i.e. 150 trials in total. Prior to the experimental task, participants completed a tutorial that explained the key concepts of the paradigm; the probabilistic association between the aliens and rewards (10 trials) and the probabilistic transition structure of rockets to planets (10 trials). After this practice phase, they had to answer a 3-item basic comprehension test regarding the key rules of the task. If participants failed to answer all questions correctly, the experimenter would reiterate the key concepts of the paradigm to the participant, allowing clarification.

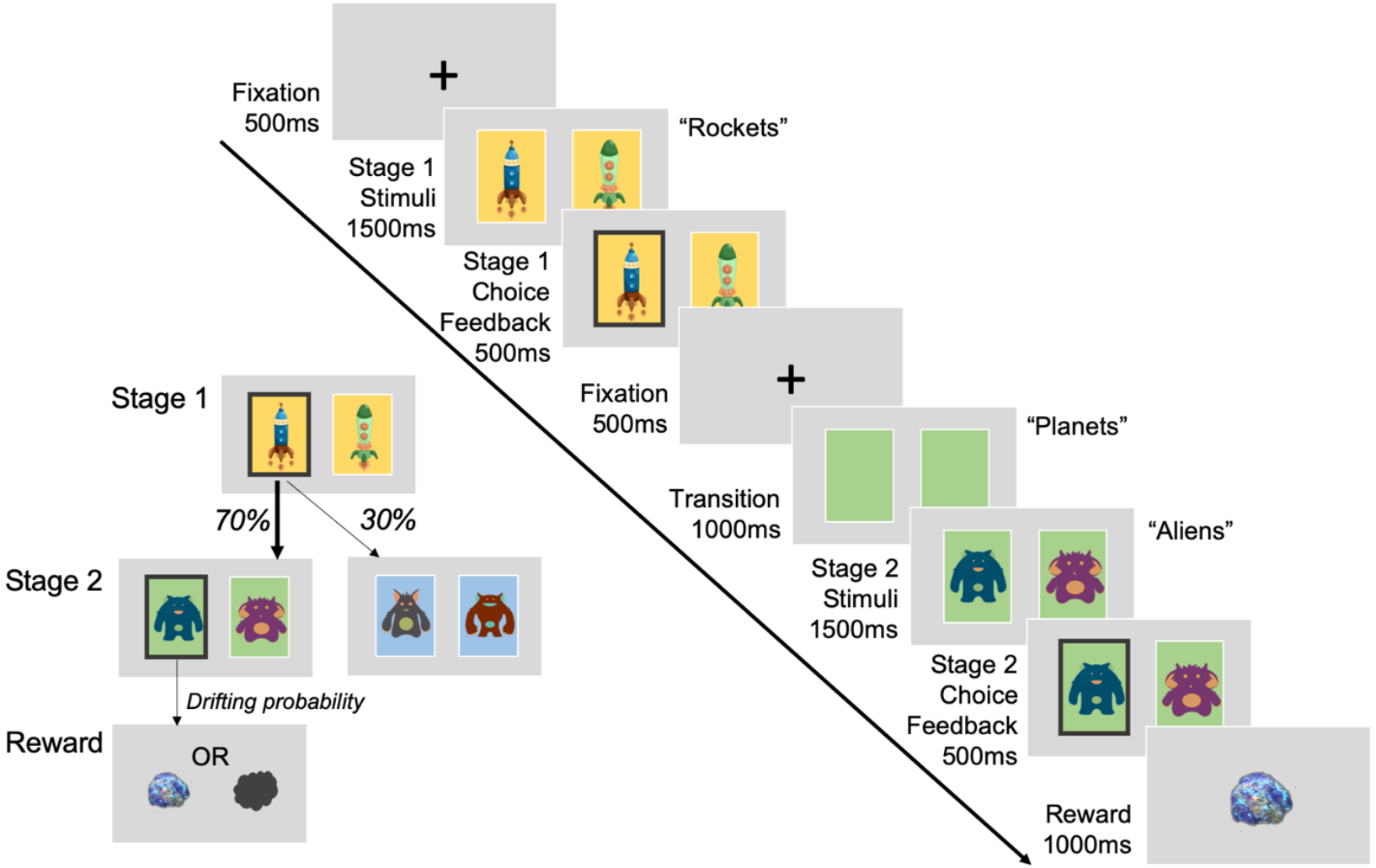


Figure 3.1. Two-step reinforcement learning task.

Figure 3.1. Two-step reinforcement learning task. *Paradigm consists of two stages where participants take a 'rocket' that has a common (70%) or rare (30%) transition to one of two second stage 'planets' (states). 'Aliens' on these 'planets' each have a unique probability of reward ('space treasure' (reward) or 'space dust' (non-reward)) that drifts slowly throughout the entire experiment. Participants have to take into consideration the task transition structure and their history of rewards to make choices that maximise gain. The sequence of events as presented for EEG is the same as that of Eppinger et al. (2017), except they included a manipulation of transition probabilities in their study (comparing 60/40% to 80/20%) and used a longer choice window (2000ms).*

Behavioural data pre-processing. Individual missed trials and trials with very fast (<150ms) reaction times at the first stage (indicating inattention or poor responding) were excluded from analyses. A total of 1082 trials (3.76%) were removed (per participant mean = 5.64 (3.76%) trials) across participants.

Self-report psychiatric questionnaires, transdiagnostic dimensions & IQ. In order to characterise our sample with a previously defined transdiagnostic dimension of compulsivity (Gillan et al., 2016), participants completed 9 self-report questionnaires (which were fully randomized) assessing: *alcohol addiction* using the Alcohol Use Disorder Identification Test (AUDIT) (Saunders et al., 1993), *apathy* using the Apathy Evaluation Scale (AES) (Marin et al., 1991), *depression* using the Self-Rating Depression Scale (SDS) (Zung, 1965), *eating disorders* using the Eating Attitudes Test (EAT-26) (Garner et al., 1982), *impulsivity* using the Barratt Impulsivity Scale (BIS-11) (Patton et al., 1995), *obsessive-compulsive disorder* (OCD) using the Obsessive-Compulsive Inventory - Revised (OCI-R) (Foa et al., 2002), *schizotypy* scores using the Short Scales for Measuring Schizotypy (SSMS) (Mason et al., 2005), *social anxiety* using the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz, 1987) and *trait anxiety* using the trait portion of the State-Trait Anxiety Inventory (STAI) (Spielberger et al., 1983). A proxy for IQ was also collected using the International Cognitive Ability Resource (I-CAR) (Condon & Revelle, 2014) sample test which included 4 item types of three-dimensional rotation, letter and number series, matrix reasoning and verbal reasoning (16 items total). See ***Supplemental***

Figure A.II.S1 and ***Table A.II.S1*** for more details of the distribution spread, correlations and reliability of the questionnaire scores.

We used weights derived from a previous study (Gillan et al., 2016) to transform the raw scores of the 209 individual items from the 9 questionnaires into dimension scores ('Anxious-Depression' (AD), 'Compulsive Behaviour and Intrusive Thought' (CIT; 'compulsivity'), and 'Social Withdrawal' (SW)). This was because our sample size had too low a subject-to-variable ratio (N = 192) for *de novo* factor analysis, as compared to the original study (N = 1413). Prior studies, including the previous chapter, have demonstrated the stability of the factor structure in new data (Rouault et al., 2018; Seow & Gillan, 2020). Consistent with prior work, the resulting dimension scores were moderately intercorrelated (r = 0.33 to 0.42) (***Supplemental Table A.I.S2***).

Quantifying model-based learning. Model-based estimates were estimated using mixed-effects models written in R, version 3.6.0 via RStudio version 1.2.1335 (http://cran.us.r-project.org) with the *glmer()* function from the *lme4* package, with Bound Optimization by Quadratic Approximation (bobyqa) with 1e5 functional evaluations. The basic model tested if participants' choice behaviour to *Stay* or switch relative to previous choice (stay: 1, switch: 0) was influenced by the previous trial's *Reward* (rewarded: 1, unrewarded: -1), *Transition* (common (70%): 1, rare (30%): -1) and their interaction (***Supplemental Figure A.II.S2***). Within-subject factors (the intercept, main effects of reward, transition, and their interaction) were

taken as random effects (i.e. allowed to vary across participants). In R syntax, the model was: Stay ~ Reward * Transition + (Reward * Transition + 1 | Subject). The extent to which model-based planning contributed to choice was indicated by the presence of a significant interaction effect between Reward and Transition (MB). Split half-reliability, where the data were split into two subsets (even versus odd trials) and correlated and adjusted with Spearman-Brown prediction formula, was estimated for model-based planning.

To test if the compulsive dimension was associated with goal-directed learning deficits, we included the total scores of all three dimensions (*AD*: anxious-depression, *CIT*: compulsive behaviour and intrusive thought, *SW*: social withdrawal) as z-scored fixed effect predictors into the basic model described above. The extent to which compulsivity is related to deficits in goal-directed learning was indicated by the presence of a significant negative Reward*Transition**CIT* interaction. Age and IQ tend to covary with model-based learning (Gillan et al., 2016); control analyses presented in ***A.II. Supplemental Methods*** demonstrate that these variables did not drive any of the results presented.

Sensitivity to task structure: Reaction time (RT). Recent work has shown that one effective way to index an individual's sensitivity to the structure of the task is via reaction times (RT) (Shahar et al., 2019). The logic is that someone who is aware of the task structure should, by right, expect a common transition (and the associated second stage choice options). As such, when a rare transition occurs, they require

more time to respond. This is presumably because they are relatively unprepared for the choice options presented to them following a rare transition. This appears to be the case; individuals with greater model-based planning have been shown to make faster second stage choices after common transitions, compared to rare transitions (Shahar et al., 2019). To test this, we conducted a mixed effect linear regression of transition type (*Transition* (common: -1, rare: 1) on second stage reaction time (S2-RT). In the syntax of R, the model was: S2-RT ~ Transition + (Transition + 1 | Subject). We asked if compulsivity was associated with a reduction in RT sensitivity to the transition structure by including the total scores of the three dimensions (*AD, CIT, SW*) as z-scored fixed effect predictors into the original model above.

EEG recording & pre-processing. EEG was recorded continuously using an ActiveTwo system (BioSemi, The Netherlands) from 128 scalp electrodes and digitized at 512 Hz. The data were processed offline using EEGLab (A. Delorme & Makeig, 2004) version 14.1.2 in MATLAB R2018a (The MathWorks, Natick, MA). Data were imported using A1 as a reference electrode, then downsampled to 250 Hz and band-pass filtered between 0.05 and 45 Hz. Bad channels were rejected with a criterion of 80% minimum channel correlation. All removed channels were interpolated, and the data were re-referenced to the average. To remove ocular and other non-EEG artefacts, ICA was run with runica, *pca* option on, and its components were rejected automatically with Multiple Artifact Rejection Algorithm (MARA) (Winkler et al., 2011), an EEGLab toolbox plug-in, at a conservative criterion

of >90% artefact probability. For all EEG analyses, other non-specific artefacts were removed after epoching using a criterion of any relevant electrode examined showing a voltage value exceeding ±100μV. If participants had a rate of >95% of total epochs failing this criterion, their data were excluded from all analyses (N = 4 as reported in **Participant exclusion criteria**). Each remaining participant had mean = 147.46 (SD = 2.98) epochs left.

Time-frequency analysis. EEG data were epoched for both first and second stages of the task for time-frequency analyses (alpha (9-13Hz) and theta (4-8Hz) power) detailed in the subsequent sections: -1700ms to 2200ms stimulus-locked at the first stage (rockets) as well as -2000ms to 3500ms stimulus-locked at second stage (aliens). Time-frequency calculations were computed using custom-written MATLAB (The MathWorks, Natick, MA) routines. The EEG time series in each epoch was convolved with a set of complex Morlet wavelets, defined as a Gaussian-windowed complex sine wave: $e^{(-i2*time*f)}\, e^{(-time\hat{}2/2\sigma\hat{}2)}$. Where *i* is the complex operator, *time* is time, *f* is frequency, which increased from 2 to 40 Hz in 40 logarithmically spaced steps. σ defines the cycle (or width) of each frequency band and was set to cycle/2πf, where cycle increased from 4 to 12 in 40 logarithmically spaced steps in accordance with each increase in frequency step. The variable number of cycles leverages the temporal precision at lower frequencies and increases frequency precision at higher frequencies. From the resulting complex signals of every epoch, we extracted estimates of power. Power is defined as the modulus of the resulting complex signal: $Z(time)$ (power time series: $\rho(time) = real[z(time)]^2 + imag[z(time)]^2$).

Stimulus-locked first stage epoch was baselined corrected to the average frequency power from -400ms to -100ms (corresponding to first stage fixation) while for stimulus-locked second stage epoch used -1400ms to -1100ms (corresponding to second stage fixation, before presentation of the coloured squares (planets)) as the baseline. For single-trial estimates of frequency power, as baselining with division induces spurious power fluctuations due to trial-to-trial fluctuations, power at each individual trial was baselined corrected with the linear subtraction method with its corresponding baseline activity: (power(*time*) – power(*baseline*)), at each frequency, at each channel. For visualisation purposes in the figures presented, power was normalized by conversion to a decibel (dB) scale: (10*log10[power(*time*)/power(*baseline*)]).

Single-trial analyses with EEG signals. All analyses described below were conducted with mixed effects models (regression model equations are in ***A.II. Supplemental Methods***). For every single-trial analysis, we excluded single-trial EEG estimates which were within ±5 SD away from the mean of the group. A maximum of <0.79% (n = 215) of the total trials across all participants were excluded for any measure. The regression model-based estimate (MB) was used as the individual between-subjects model-based estimate in all EEG analyses.

Sensitivity to task structure: P300 and transition type. The P300 component is a parietal positivity that occurs about 300ms after stimulus onset. Studies with odd-

ball paradigms observe that the signal is sensitive to stimulus probability; P300 amplitudes are larger as stimulus probability decreases i.e. gets more rare (Polich & Margala, 1997). Prior research in healthy humans thus hypothesised that the P300 may be a marker of sensitivity to state-state transition knowledge on this two-step task, albeit inconsistent direction of effect was found across studies (Eppinger et al., 2017; Sambrook et al., 2018; Shahnazian et al., 2019). However, the P300 component is time-locked to choice commitment (Twomey et al., 2015) and as such the amplitude of the averaged stimulus-locked signal will be partly determined by RT variability. To circumvent this issue, we tested if the response-locked signal was associated to transition types.

We first measured the P300 component at four parietal electrodes over the topography of the stimulus-locked peak (D16 (CP1), A3 (CPz), B2 (CP2), A4); ***Supplemental Figure A.II.S3***). Data were epoched from -500ms to 1700ms relative to the onset of the second stage stimulus (aliens presented) and baselined corrected from -200ms to 0ms. Stimulus-locked single-trial P300 amplitudes were estimated as the mean of ±100ms around the individual's averaged latency of their positive peak within a search window 250ms to 1000ms after stimulus onset. To eliminate amplitude biases owing to latency variances due to RT, we subsequently aligned the epochs (measured at A4, A5, A19 (Pz), A32, the response-locked peak; ***Supplemental Figure A.II.S5***) to the time of choice response execution. The response-locked P300 amplitude was quantified as the mean amplitude -100ms to 0ms before response. To investigate if the P300 was sensitive to rare versus

common transitions and whether this depended on model-based control/compulsivity, we regressed both stimulus- and response-locked P300 measures against transition type *Transition*: rare: 1, common: 0) interacting with z-scored model-based estimates (*MB*) or compulsivity (*CIT*, controlled for the other psychiatric dimensions *AD* and *SW*), taking *Transition* and the intercept as random effects.

Sensitivity to task structure: Alpha power and transition type. Neural oscillations are thought to play a role in orchestrating more distributed neural computations of the sort required in model-based planning and therefore might be better able to track sustained representations of the task's transition structure than short-latency evoked-related potentials (ERPs) like the P300 (Makeig & Onton, 2012). Alpha (9-13Hz) desynchronization (i.e. suppression) is associated with increases in BOLD activity in frontal cortical areas and is seen as a general marker of mental activity and attention (Laufs et al., 2003), which we reasoned might be modulated by the occurrence of unexpected state transitions in this task. Coupled with the fact that prior studies have found evidence for alterations in alpha power in OCD patients (Perera et al., 2019), we sought to test if alpha tracks transition structure of this task and if this is disrupted in compulsivity.

Alpha power was measured at five occipital-parietal electrodes (A18, A19 (Pz), A20, A21, A31; surrounding A20 electrode; ***Supplemental Figure A.II.S6***) in an epoch centered on the onset of the second stage stimuli (aliens) (see **Time-frequency**

analysis). Single-trial stimulus-locked alpha power estimates were measured as the mean power ±250ms around the average latency of the negative peak, specific for each individual, found within a search window 0ms to 1000ms after stimulus onset. We additionally obtained alpha power estimates quantified across four 1000ms rolling time bins by the mean amplitude within each time window. The time bins began from transition at -1000ms to 0ms locked to the stimulus (alien presentation), followed by three windows spanning choice to reward from 0ms to 1000ms, 1000ms to 2000ms, and 2000ms to 3000ms. The same approach of mixed effect models with **P300 and transition type** was used to examine the influence of model-based estimates/compulsivity on alpha power representation of rare versus common transitions, except for where *Transition* was coded differently (rare: -1, common: 1) for ease of interpreting the direction of interaction effects.

Behavioural control: theta power during choice. Mid-frontal theta (4-8Hz) power is a well-established EEG signature of cognitive control (Cavanagh & Frank, 2014; Sauseng et al., 2010), and for example is associated with exerting control over the influence of 'automatic' Pavlovian biases (Cavanagh et al., 2013). Cognitive control is thought to be essential in supporting the utilisation of goal-directed behaviour over more automatic (and less prospective) modes of action selection (Otto et al., 2014). We therefore probed if increased mid-frontal theta at the first stage of the task, as the crucial time where subjects make model-based or model-free choices, would reflect greater model-based performance and show disruption in compulsivity. For theta power (4-8Hz), power estimates were measured at four a priori frontal midline

electrodes (C21 (Fz), C22, C23 (FCz), A1 (Cz); see ***Supplemental Figure A.II.S9***) at the first stage (see **Time-frequency analysis**). The mean power ±250ms around the individual's average latency of the positive peak found within a search window 0ms to 500ms after stimulus onset was taken for every epoch. We tested if single-trial theta power was associated with model-based estimates (*MB*) or to compulsivity (*CIT*, controlled for *AD* and *SW*) by taking them as z-scored main regressors against theta power.

Specificity with psychiatric questionnaire scores versus transdiagnostic dimensions. Additionally, we visualised the advantages of utilising a transdiagnostic definition of compulsivity as opposed to examining single psychiatric questionnaires. We repeated the above time-frequency analyses (alpha and theta) with the individual total questionnaire scores (*QuestionnaireScore*, z-scored) replacing the three psychiatric dimensions (*CIT, AD, SW*) in their respective regression models detailed above. Separate mixed effects regression models were performed for each individual questionnaire as correlation across questionnaire scores ranged greatly from r = -0.09 to 0.79 as opposed to the transdiagnostic analysis where all three dimensions (that correlated moderately: r = 0.33 to 0.42) were included in the same model.

Supplemental analyses. We also explored the association of model-based planning/compulsivity with (i) alpha power at the first stage and (ii) theta power and

transition type at the second stage. These analyses are reported in ***Figure A.II.S6*** and ***Figure A.II.S9*** respectively.

Results

Compulsivity and model-based planning. Regression analysis of choice behaviour on the two-step task revealed a significant interaction between Reward and Transition (β = 0.20, *standard error (SE)* = 0.03, $p < 0.001$), indicating clear evidence for model-based planning in this sample. Individual subject coefficients for this interaction term were extracted and used as an individual difference measure for EEG analysis. Note that this 'model-based index' had good split half-reliability of r = 0.71. Consistent with prior work, there was also evidence for model-free learning, where subjects were more likely to repeat choices if they were followed by reward (main effect of Reward: β = 0.55, SE = 0.05, $p < 0.001$), and an overall tendency to repeat choices from one trial to the next (Intercept: β = 1.46, SE = 0.07, $p < 0.001$). Importantly, we found that individual differences in compulsivity and intrusive thought (hereafter: 'compulsivity') were associated with reduced goal-directed learning (β = -0.07, SE = 0.04, p = 0.05) (***Figure 3.2a***), while anxious-depression and social withdrawal were not.

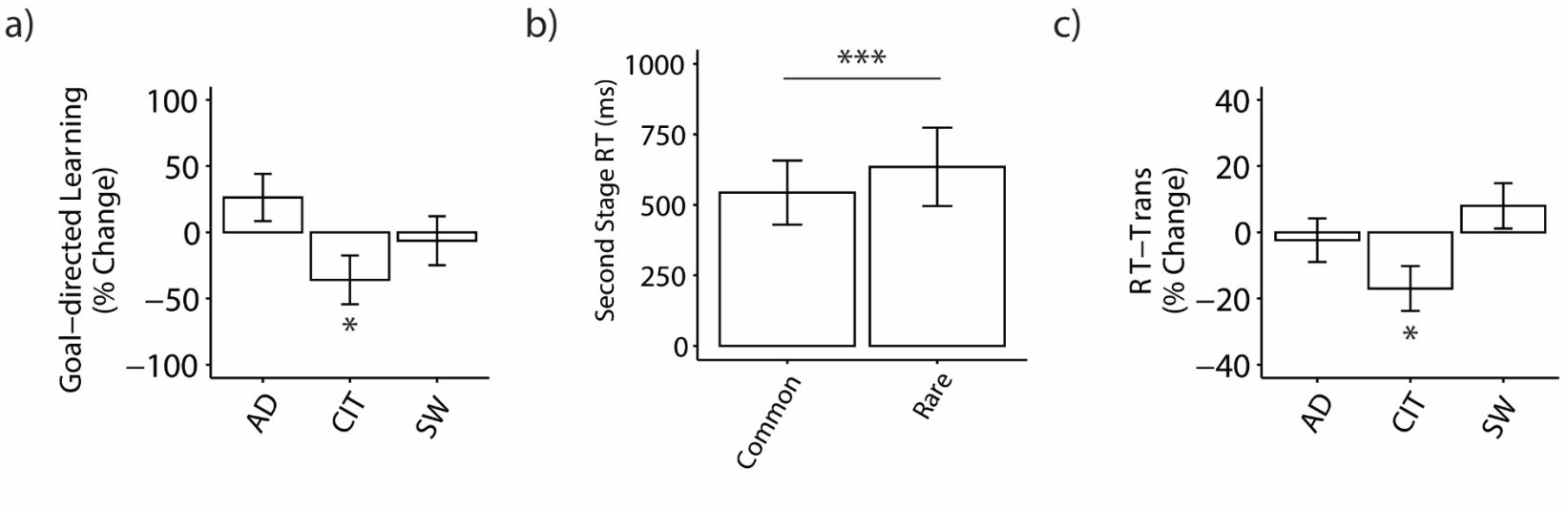


Figure 3.2. Goal-directed control and compulsivity. a) *Model-based control estimated by a logistic regression of choice behaviour with one-trial back reward and transition. Regressions were conducted in a model with all three dimensions: 'anxious-depression' (AD), 'compulsivity and intrusive thought' (CIT) and 'social withdrawal' (SW). Model-based control is reduced high compulsive individuals.* ***b)*** *Participants have on average a longer mean response time (RT) at second stage choice after a rare transition than a common one (t_{191} = 16.16, 95% Confidence Interval (CI) [79.85 102.05], p < 0.001).* ***c)*** *This RT difference between transition type (RT-Trans) is diminished in high compulsive individuals. For* ***b)****, error bars denote standard deviation. For* ***a)*** *&* ***c)****, error bars denote standard error. The Y-axes indicate the percentage change in goal-directed control/RT-Trans as a function of 1 standard deviation of psychiatric dimension scores. *p ≤ 0.05, ***p < 0.001.*

Reaction time (RT) sensitivity to task structure. In line with recent work, we hypothesised that participants would have a slower RT after a rare versus common transition and that this difference would be greater in more model-based participants. Indeed, we found that participants have a slower mean RT for rare versus common trials after transition (β = 47.15, *SE* = 2.85, $p < 0.001$) (***Figure 3.2b***). As expected, this effect (RT-Trans) was larger in those with higher levels of model-based control (β = 7.36, SE = 1.83, $p < 0.001$). Crucially, we found that this effect was reduced in high compulsive individuals (β = -8.02, *SE* = 3.19, p = 0.01) (***Figure 3.2c***). This is, to our knowledge, the first evidence that compulsivity is associated with a deficit in distinguishing rare from common transitions, a fundamental constituent of the mental-model required to engage in model-based planning.

P300 sensitivity to task structure. Previous studies found that stimulus-locked parietal positivity, like reaction time, was sensitive to the difference between rare and common transitions. Although this result has been inconsistent in direction, with some studies finding greater P300 amplitudes for rare versus common (Sambrook et al., 2018; Shahnazian et al., 2019) and another the opposite (Eppinger et al., 2017) we nonetheless thought this was a good place to start. We examined the second stage stimulus-locked P300 and found a significant main effect of transition type (β = 0.15, *SE* = 0.07, p = 0.03), consistent with Sambrook et al. (2018) and Shahnazian et al. (2019) whereby greater P300 amplitude was observed after rare versus common transitions (***Supplemental Figure A.II.S3***). However, to our surprise, this differential rare versus common signal was not larger in individuals high in model-based planning (β = 0.07, *SE* =

0.08, p = 0.35) nor did it show any association to compulsivity (β = 0.09, SE = 0.08, p = 0.24).

However, reaction times may confound the amplitude of the stimulus-locked peak (Twomey et al., 2015), and it is possible that differences in the P300 effects previously found reflected an RT-based peak shift. When we repeated the analysis using response-locked P300 (which mitigates this issue), we found that the transition effect was no longer significant and its direction was reversed (-100ms to 0ms) (β = -0.09, SE = 0.08, p = 0.23) (***Supplemental Figure A.II.S5***). Together with the lack of association with model-based planning, we concluded that the P300 may not provide the most reliable or sensitive measure of neural sensitivity to task structure.

Alpha power sensitivity to task structure. As ERPs principally reflect activity changes that are short-lived and strictly time-locked to particular events, we also investigated the possibility that occipital-parietal alpha power would provide a more sensitive index of task transitions. Specifically, we examined if occipital-parietal alpha power measured locked to the second stage stimulus was able to distinguish between rare and common transitions across a series of time bins in our task. This allowed us to ascertain not just if participants showed sensitivity to task structure following a transition, but for how long they sustained that representation. We found that alpha power was indeed able to differentiate transition types (β = 0.05, SE = 0.01, $p < 0.001$); occipital-parietal alpha was more suppressed after rare versus common transitions (***Figure 3.3***). Repeating the analysis with response-locked alpha also yielded a significant effect (β = 0.06,

SE = 0.01, $p < 0.001$) (***Supplemental Figure A.II.S8***). Unlike the P300, we found that over three rolling time bins beginning from the transition (planet) (-1000ms to 0ms: $\beta = 0.04$, *SE* = 0.02, $p < 0.05$) to the end of choice feedback (0ms to 1000ms: $\beta = 0.04$, *SE* = 0.01, $p = 0.005$; 1000ms to 2000ms: $\beta = 0.05$, *SE* = 0.02, $p = 0.01$), individuals high in model-based control showed the largest alpha power differences between the two transition types (***Figure 3.3***). Importantly, this same signature was negatively related to compulsivity, with significant associations observed starting from the time of state transition until the response was made (-1000ms to -0ms: $\beta = -0.06$, *SE* = 0.02, $p = 0.002$; 0ms to 1000ms: $\beta = -0.03$, *SE* = 0.02, $p = 0.05$) (***Figure 3.3***). Consistent with RT findings, these data suggest that individuals high in compulsivity have a diminished representation of the transition structure of the task.

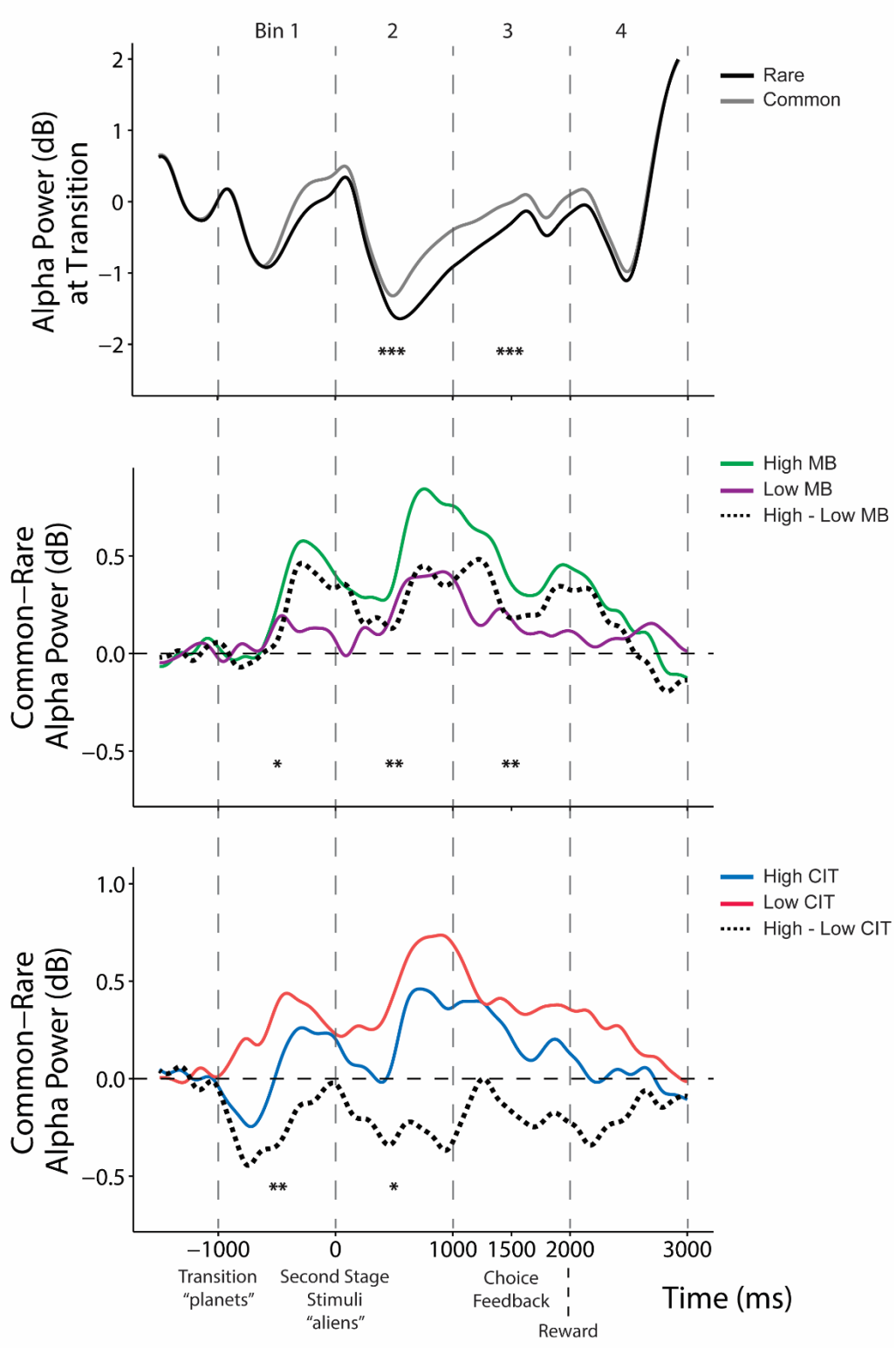


Figure 3.3. Stimulus-locked alpha power at transition.

Figure 3.3. Stimulus-locked alpha power at transition. *Alpha power was measured across 4 time bins of 1000ms each separated by vertical dashed lines, starting from the transition (-1000ms) until after reward (2000ms). Top inset shows grand average second stage alpha power waveforms between rare and common transitions. Significance indicates a significant main effect of larger alpha depression for rare than common transitions in single trial analysis with the full sample. Lower insets indicate alpha power difference between transitions (common minus rare) comparing top/bottom 40th percentile (N = 77 per group) of participants grouped by model-based estimates (MB) or compulsivity (CIT). Black dashed lines show that alpha difference is enhanced in time bins 1-3 for more model-based participants, while diminished in time bins 1-2 for high compulsive individuals. Significance indicates a significant transition*MB/CIT effect from single-trial regressions of alpha power (with all participants) on transition type and MB/CIT in single trial analysis with the full sample. *p ≤ 0.05, **p < 0.01, ***p < 0.001.*

Theta power as a marker of mental model implementation during choice. Moving beyond participants' awareness of the transition structure of the task, we next tested for evidence for a more general disruption in cognitive control at the time of choice, assayed as theta (4-8Hz) power during the first stage choice—the crucial time when model-based planning manifests in behaviour. We found that theta power was not significantly linked to model-based learning (β = 0.02, *SE* = 0.01, *p* = 0.11), but significantly lower theta was associated with individuals high in compulsivity (β = -0.03, *SE* = 0.01, *p* = 0.03) (***Figure 3.4***). Notably, greater theta power here was seen in participants who were more aware of state-state transitions (had larger differences in their RT after transition at second stage choice (RT-Trans)) (β = 0.03, *SE* = 0.01, *p* = 0.001). This suggests that those high in compulsivity may have deficits in the engagement of general cognitive control mechanisms when making their first stage choices which is linked to their awareness of transition types.

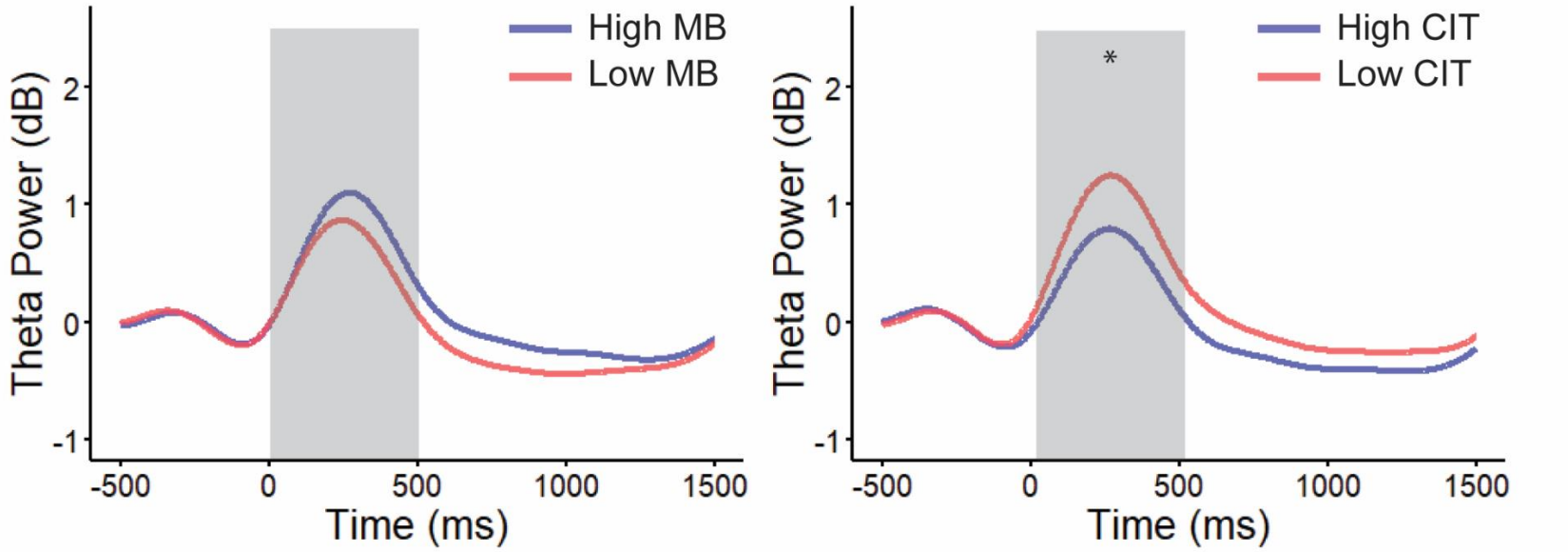


Figure 3.4. Stimulus-locked first stage theta power. *Grand average waveforms of theta power comparing the top/bottom 40th percentile (N = 77 per group) individuals based on their model-based (MB)/compulsivity (CIT) estimates. Single trial analyses (with all participants) indicate high compulsive individuals exhibit a decrease in theta power (β = -0.03, SE = 0.01, p = 0.03). Directional effect of increased theta power for model-based control was non-significant (β = 0.02, SE = 0.01, p = 0.11). Shaded grey area visualises an approximate time window where theta power was estimated (±250ms around the individual's average latency of the positive peak). *p < 0.05.*

Alpha and theta modulations are specific to compulsivity. We additionally ascertained the advantage of using a transdiagnostic definition of compulsivity in our study. When we examined how well alpha power suppression differentiated the transition types in the varied set of nine psychiatric questionnaire scores, diminished sensitivity to transition structure was linked to both OCD (β = -0.03, *SE* = 0.01, *p* = 0.02) and depression (β = -0.03, *SE* = 0.01, *p* = 0.03) scores (***Figure 3.5***). However, with the transdiagnostic analysis, the effect was shown to be specific to compulsivity (β = -0.03, *SE* = 0.02, *p* = 0.048). Similarly, reduced theta power at the first stage was linked to more than one questionnaire score—schizotypy (β = -0.03, *SE* = 0.01, *p* = 0.01), depression (β = -0.03, *SE* = 0.01, *p* = 0.02) and OCD (β = -0.02, *SE* = 0.01, *p* = 0.02). Again, the data was ultimately best explained by the compulsive dimension (β = -0.03, *SE* = 0.01, *p* = 0.03) (***Figure 3.5***).

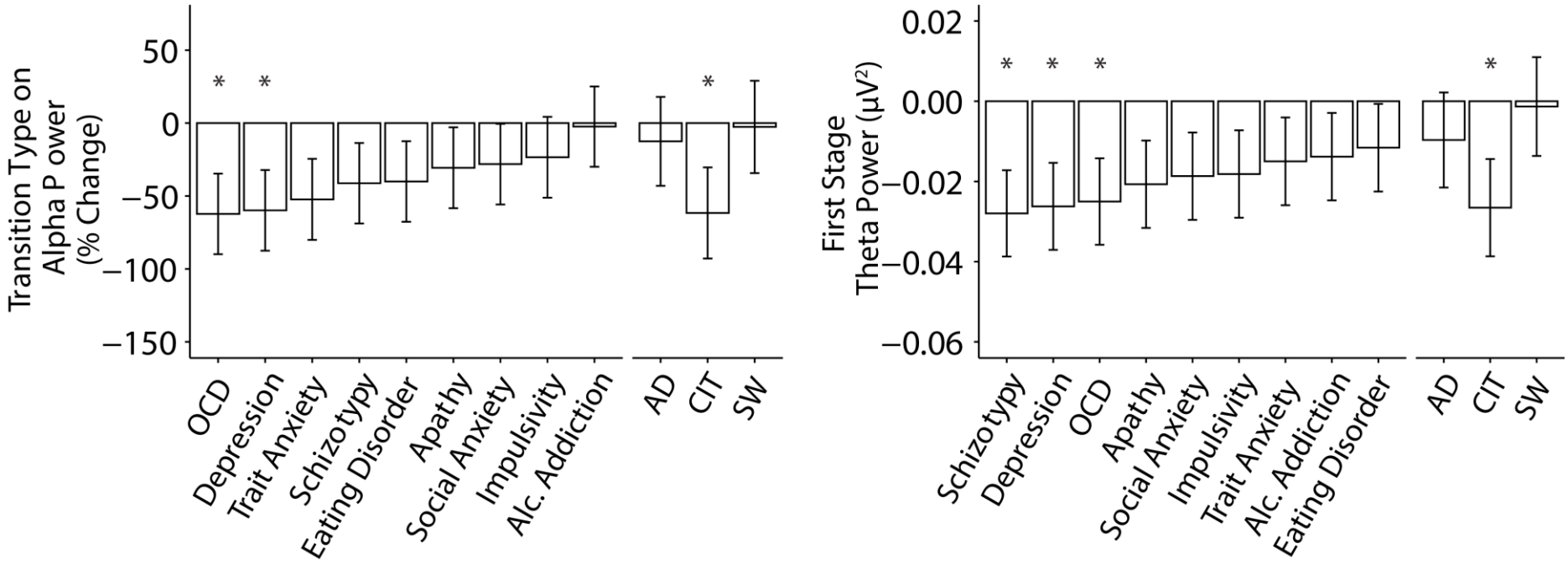


Figure 3.5. Alpha sensitivity to transition structure and theta power at first stage choice with total questionnaire scores and psychiatric dimensions.

Figure 3.5. Alpha sensitivity to transition structure and theta power at first stage choice with total questionnaire scores and psychiatric dimensions (AD: 'anxious-depression'; CIT: 'compulsive behaviour and intrusive thought', SW: 'social withdrawal'). *Alpha power differentiating rare versus common transitions was less distinguished with more than one questionnaire score, but the effect was shown to be specific to the compulsive dimension (CIT, as opposed to AD and SW). Similarly, reduced theta power at first stage choice was linked to several questionnaire scores but was ultimately specific to compulsivity. The Y-axes shows the percentage change in alpha power sensitivity to transition type (%) or change in theta power (μV^2) as a function of 1 standard deviation increase of psychiatric questionnaire/dimension scores. Error bars denote standard errors. *p < 0.05.*

Discussion

Goal-directed deficits are consistently observed in compulsivity (Gillan & Robbins, 2014) and this study was no exception—individuals high in compulsivity showed poorer model-based planning. Despite the consistency of these findings, little is known about the specific mechanisms underlying goal-directed deficits in compulsivity. Most of the theorising in this area has focused on the implementation of the model and specifically the balance/arbitration between competing model-based and model-free influences during choice (Gillan & Robbins, 2014; Gruner et al., 2016). What remains to be seen is whether compulsive individuals acquire an accurate representation of the 'model' itself, which is of course a prerequisite for engaging in goal-directed behaviour. In other words, *having* an accurate model of the world is of course necessary if one hopes to *use* this model to guide choice.

We sought to fill this gap by using electrophysiology to investigate if goal-directed deficits in compulsivity reflect issues with constructing and maintaining an accurate internal model of the task environment (i.e. state transition probabilities). An analysis of reaction times provided the first evidence that high compulsive individuals were less aware of state-state transitions. In line with prior research, subjects show longer RTs following rare transitions (Shahar et al., 2019), which is presumed to reflect the fact that one needs to adjust to this unlikely event and 're-plan' their next choice. In line with this account, we found that individuals who had higher levels of model-based planning performance showed the largest RT differences following rare versus common transitions. Crucially, the opposite was true of compulsivity, with the

most compulsive individuals showing the smallest difference in RT for these trial types. Thus, compulsivity is associated with having diminished awareness of the transition structure of the task, a prerequisite for engaging in model-based (goal-directed) planning on this task.

Moving beyond behaviour, analysis of alpha power following these state transitions revealed a strikingly similar picture. Much like reaction times, alpha desynchronization at the second stage of this task was sensitive to whether a rare or common transition occurred. Specifically, rare transitions were associated with greater alpha desynchronization compared to common trials, possibly reflecting the greater mental effort required on these trials to call to mind the action values associated with options the individual was not expecting to see. Consistent with this interpretation, previous studies using n-back paradigms have shown parieto-occipital alpha is more suppressed when working memory load increases (Pesonen et al., 2007; Stipacek et al., 2003). Individual difference analysis demonstrated that this difference in alpha desynchronization had important behavioural correlates. Those individuals who were highest in model-based planning showed the largest differences in alpha power for rare versus common transitions. Importantly, the effect for compulsivity was reversed—higher levels of compulsivity were associated with less of a distinction in alpha power for rare versus common transitions. Together with the reaction time data, this is the first neural evidence suggesting that compulsivity may be characterised by failures in representing the kind of causal state-state relations necessary to behave in a goal-directed manner.

These findings do not exclude the possibility that compulsive individuals also face issues with implementing model-based planning in situations when they might have the requisite state-state knowledge. Mid-frontal theta reflects a common mechanism for executing adaptive control in a variety of contexts (Cavanagh et al., 2012) and more specifically for selecting between competing options, including the suppression of distracting stimuli when focused attention is required (Nigbur et al., 2011). As theta was a good candidate neural signature of model-based deliberation, we thus tested if there was evidence for reduced theta power in compulsive individuals at the crucial time of first stage choice, when model-based and model-free action values putatively compete for control. We found that theta power was indeed reduced in high compulsive individuals during first stage choice, though evidence of a link between increased theta and increased model-based planning itself was not compelling in the present study. Prior research has shown that OCD patients exhibit lower theta power during tasks that require inhibitory regulation (Min et al., 2011), suggesting that reduced levels observed here in high compulsive individuals might reflect a failure to inhibit competing model-free action values. It is equally possible, however, that reduced theta might reflect a lack of the existence of competing signals. For instance, if patients fail to represent a model of the task environment, then one might argue that the need to engage cognitive control is necessarily reduced.

Previous EEG studies of the two-step task (Eppinger et al., 2017; Sambrook et al., 2018; Shahnazian et al., 2019) showed that the P300 was associated with state-state transitions. However, the inconsistent direction on the effects raises doubt as to how these differences should be interpreted. Recent literature conceptualises the P300 as a decision signal that builds towards a threshold at choice time (Twomey et al., 2015) and as such variances in RT will influence the latency of the stimulus-locked P300 amplitude peak (Kelly & O'Connell, 2015). Our results comparing stimulus-locked and response-locked analysis approaches suggest that the transition effect of the P300 may be attributed to RT variances. In each case however, we noted that the effects were not linked to individual differences in model-based planning. Shahnazian and colleagues also investigated theta (but not alpha) frequencies as potential neural correlates of state-state transitions related to model-based control on the two-step task (Shahnazian et al., 2019). We replicated their null results for theta here (***Supplemental Figure A.II.S9***).

In this study, we utilised a compulsive dimension born from transdiagnostic phenotyping in our analyses (Gillan et al., 2016) to ensure that we controlled for the collinearity of psychiatric symptoms across disorders. Indeed, we found that the modulations of alpha and theta power were non-specific when examined across the individual DSM-defined questionnaires, but the dimensional approach was able to assign specificity to our results: impairments in the representation of the mental model and cognitive control were solely linked to the compulsive dimension (and not anxious-depression or social withdrawal phenotypes).

Overall, our findings suggest that compulsive individuals have goal-directed difficulties that are not confined to the execution of the mental model, but also to its construction/maintenance. Understanding the mechanisms underlying these different processes will be vital for future research; in particular, understanding how the internal model becomes altered may be crucial for understanding how obsessive beliefs manifest and interact with the compulsive behaviour. It is still unclear how compulsions and obsessions synergise. Clinical cognitive models of OCD have long presumed that compulsions are performed to reduce anxiety induced by obsessive beliefs (Matthews & Wells, 2008; Salkovskis & McGuire, 2003), in contrast to a more recent hypothesis suggesting that obsessions are post-hoc rationalisations to explain the performance of compulsive behaviour (Gillan & Sahakian, 2015). These data may suggest that a more fluid distinction between obsessions and compulsions is instead warranted. It is possible that failures in linking actions to their consequences may be a common source of both compulsive habitual behaviours in OCD and also faulty metacognitive beliefs that form the basis of obsessions. Indeed, a parallel literature is emerging that suggests a range of impairments in metacognition are characteristic of compulsivity (Rouault et al., 2018; Seow & Gillan, 2020). These include inflated confidence in decisions, and deficits in the ability to use feedback to update metacognitive beliefs.

Chapter 4: A dimensional study of error-related negativity (ERN) and self-reported psychiatric symptoms

Introduction

Errors are a critically important information source. They allow us to monitor and continually adapt performance to changes in the environment, to slowly and incrementally improve skills, and to avoid large mistakes by having smaller ones attended to. Without this capacity, we might find ourselves repeating unproductive or damaging behaviours. Conversely, a hypersensitive error detection system might keep us from trying new things, from getting out of our comfort zone and experiencing the learning that comes from failure. Since the early nineties, the mainstay of error monitoring research has been the unique neural response to the commission of errors—the error-related negativity (ERN), a negative deflection of the event-related potential that peaks approximately 50-100ms after an error response (Falkenstein et al., 1991; Gehring et al., 1993). Fundamentally, the ERN represents a well-validated and reliable neurophysiological index of error processing (Holroyd & Coles, 2002) with the anterior cingulate cortex posited to be its neural generator (Debener, 2005; Grützmann et al., 2016; Miltner et al., 2003).

Impairments in error monitoring are phenomenologically characteristic of a range of psychiatric disorders (Ullsperger, 2006), and this has been supported by the frequent observation of alterations in the ERN in patient grous (Gillan et al., 2017;

Weinberg, Dieterich, et al., 2015). For example, studies have observed diminished ERNs in schizophrenia (Bates et al., 2002; Foti et al., 2012; Minzenberg et al., 2014; S. E. Morris et al., 2006; Simmonite et al., 2012), bipolar disorder (Minzenberg et al., 2014; Morsel et al., 2014) and substance use disorder (Franken et al., 2007; Sokhadze et al., 2008), while enhanced ERN amplitudes are consistently seen in obsessive-compulsive disorder (OCD) (Carrasco, Hong, et al., 2013; Endrass et al., 2008, 2014; Endrass & Ullsperger, 2014; Klawohn et al., 2014), social anxiety disorder (Endrass et al., 2014) and generalised anxiety disorder (Carrasco, Hong, et al., 2013; Weinberg et al., 2010; Weinberg, Klein, et al., 2012; Weinberg, Kotov, et al., 2015).

Though the precise functional role of the ERN is still highly debated (W. H. Alexander & Brown, 2011; M. G. H. Coles et al., 2001; Holroyd et al., 2005; Vidal et al., 2000; Yeung et al., 2004), there are several interpretations of the various ERN abnormalities observed in psychopathology. For diminished ERNs associated with bipolar disorder and schizophrenia, the phenomenon is hypothesised to reflect internal response monitoring deficits posited to underlie the generation of positive schizophrenia symptoms (Frith & Done, 1988; McGrath, 1991). As for disorders with enhanced ERN amplitudes (i.e. OCD, social anxiety and generalised anxiety), one commonality amongst these disorders is that they are fundamentally characterised by high levels of anxiety. Here, the enhanced ERN is thought to reflect an increased sensitivity to errors (Hajcak, 2012; Weinberg, Riesel, et al., 2012) which may be experienced as highly distressing (Dreisbach & Fischer, 2012; Hajcak & Foti, 2008;

Spunt et al., 2012) in anxiety. This is supported by a body of evidence showing exaggerated physiological changes associated with anxiety (e.g. enhanced startle reflex (Hajcak & Foti, 2008; Riesel et al., 2013), heart rate deceleration (Hajcak et al., 2003b, 2004) and skin conductance changes (Hajcak et al., 2003b, 2004)) are linked to larger ERNs. Given that ERN amplitude shifts are so pervasive in psychiatry, it has been suggested and recognised by the Research Domain Criteria Initiative (RDoC) (Insel et al., 2010) that these reflections of altered error processing may be a transdiagnostic phenomenon (Gillan et al., 2017; A. Meyer & Klein, 2018; Weinberg, Dieterich, et al., 2015) that holds potential as a biomarker of mental health.

In recent years, meta-analyses have proposed phenotypes beyond Diagnostic and Statistical Manual of Mental Disorders (DSM) categories that may underlie alterations in ERN amplitude and explain their ubiquity across psychiatric groups. Particularly for the enhanced ERN, anxious apprehension (Moser et al., 2013) or uncertainty (Cavanagh & Shackman, 2015) are key candidates, supported by studies in non-clinical samples demonstrating that increased levels of worry (Hajcak et al., 2003a; Moser et al., 2012; Zambrano-Vazquez & Allen, 2014) and threat sensitivity (Weinberg et al., 2016) are associated with larger ERNs. That said, a recent meta-analysis posits a higher enhanced ERN effect size for obsessive-compulsive symptomology versus anxiety (Pasion & Barbosa, 2019). As only a few studies have attempted to disentangle (and control for) intercorrelated symptoms within individuals in the same sample, and even fewer have done this in a sample

of sufficient size, it remains to be seen if enhancements in the ERN confer risk for anxiety or compulsive symptoms.

To test this, we used a dimensional approach whereby we measured co-occurring symptoms of a range of disorders within the same individuals and tested for associations with the ERN in their original form, as well as after they had been reduced to three dimensions—anxious-depression, compulsive behaviour and intrusive thought (hereafter 'compulsivity') and social withdrawal (Gillan et al., 2016). Using this method, we previously showed that a transdiagnostic compulsive dimension maps onto deficits in goal-directed control better than OCD symptoms (Gillan et al., 2016); a finding that has since been replicated (Patzelt et al., 2019). We also showed that this method can reveal associations that are hidden by categorical disorder groupings. For example, anxious-depression is linked to reduced confidence, while individuals high on the spectrum of compulsivity have elevated confidence (Rouault et al., 2018; Seow & Gillan, 2020). This finding might explain why group level effects in OCD (where patients have high levels of both compulsivity and anxious-depression) have not revealed confidence abnormalities (Hauser, Allen, Rees, et al., 2017; Vaghi et al., 2017). As such, the transdiagnostic method may be able to specify whether enhanced ERN amplitude shifts are truly related to anxious or compulsive symptomology. An additional advantage of using these previously defined transdiagnostic dimensions to ambiguate ERN relationships, as opposed to fitting new definitions of psychiatric phenotypes to our data here, is that it offers a clear extension to the several other cognitive phenomena

(i.e. goal-directed control and metacognition) related to these dimensions. Generalizing ERN effects to known cognitive mechanisms would foster better understanding of ERN amplitude shifts in psychopathology.

Following this methodology, we characterised participants in terms of a broad range of psychopathology (9 questionnaires in total) that have almost all been linked to the ERN in prior work; alcohol addiction, apathy, depression, eating disorders, impulsivity, OCD, schizotypy, social anxiety and trait anxiety. We hypothesised that an enhanced ERN would be associated with OCD, social anxiety and trait anxiety, but that this would be explained by a psychiatric dimension encapsulating high levels of anxiety, i.e. anxious-depression. While we expected OCD symptom severity to correlate with the ERN, we anticipated that the compulsivity dimension would not show an association as prior work has shown diminished ERN in addiction and schizophrenia (see review (Gillan et al., 2017)), both of which are strong contributors to the compulsivity dimension.

We related ERN amplitude to self-report psychiatric symptoms from 196 participants who completed the arrow-version of the Eriksen Flanker task (Eriksen & Eriksen, 1974). Contrary to our hypothesis, we found that none of the psychiatric symptoms nor the transdiagnostic dimensions were significantly associated to alterations in ERN amplitude. To contextualise the absence of ERN affects in the present sample (i.e. effect size), we report results from an additional cognitive task relating goal-directed learning (Daw et al., 2011) to dimensional phenotypes. Here, we did find

evidence for an association; replicating prior work (Gillan et al., 2016) showing that goal-directed learning was related to the compulsive behaviour and intrusive thought dimension.

Methods

Power estimation. An appropriate sample size was determined based on a previous study that reported an association of OCI-R scores and enhanced ERN amplitude that approached significance ($r = 0.32$, $p = 0.06$) (Gründler et al., 2009), an effect size suggesting that N = 155 participants were required to achieve 90% power at 0.005 significance (significance threshold is corrected for multiple comparisons over the nine psychiatric questionnaires investigated).

Participants. The majority of participants were recruited from the general public through university channels via flyers and online advertisements, and a small number were patients from St. Patrick's Mental Health hospital. We included these patients to enrich our sample for self-report mental health symptoms. They were all ≥18 years (with an age limit of 65 years) and had no personal/familial history of epilepsy, no personal history of neurological illness/head trauma nor personal history of unexplained fainting. After reading the study information and consent online, participants provided informed consent by clicking the 'I give my consent' button. They also gave written consent before the in-laboratory EEG session. They were paid €20 Euro (€10/hr) upon completion of the study. We collected data from N = 234 participants; N = 8 were patients starting group treatment for anxiety from a local clinic and the rest N = 226 were from the general public. Of the total sample, 138 were female (58.97%) with ages ranging from 18 to 65 (mean = 31.42, standard deviation (SD) = 11.48) years. All study procedures were approved by Trinity College Dublin, School of Psychology Research Ethics Committee and St. Patrick's Mental Health Services Research Ethics Committee.

Procedure. Before arriving to the lab for testing, participants navigated a webpage to provide informed consent, basic demographic data (age, gender), list any medications they were currently taking for *a mental health issue* (if so, to indicate the name, dosage and duration) and complete a set of 9 self-report psychiatric questionnaires. For a subset of the participants (N = 110, 47%), they completed a short psychiatric interview in-person on the day of testing (Mini International Neuropsychiatric Interview; M.I.N.I.) (Sheehan et al., 1998). During the experimental EEG session, participants completed two tasks: the modified Eriksen flanker task (Eriksen & Eriksen, 1974) and the two-step reinforcement learning task (Daw et al., 2011). The latter task was analysed in the previous experimental chapter and so the methods are not described in detail, however we report one basic behavioural result from this task to contextualise our ERN results. Once participants completed both tasks, they completed a short IQ evaluation before being debriefed and compensated for their time.

Exclusion criteria. Several exclusion criteria were applied to ensure data quality. Participants were excluded if they failed any of the following on a rolling basis. (i) Participants whose EEG data were corrupted (N = 2) or incomplete (i.e. recording was prematurely cut) were excluded (N = 4). (ii) Participants whose error response-locked epochs over four electrode sites examined failed a threshold criterion of ±50μV for 95% of epochs (see **Response-locked ERPs**) were excluded (N = 5). (ii) Participants who missed >20% of trials (n > 96) of the flanker task were excluded (N = 11). (iii) Participants who scored <55% accuracy were excluded (N = 9). (iv) Participants who incorrectly responded to a “catch” question

within the questionnaires: “If you are paying attention to these questions, please select ‘A little’ as your answer” were excluded (N = 7). Combining all exclusion criteria, 38 participants (16.24%) were excluded. 196 participants were left for analysis (115 females (58.67%), between 18-65 ages (mean = 30.82, SD = 11.53 years).

Disorder prevalence (M.I.N.I.). After exclusion, 87 participants (44.39%) completed the M.I.N.I., which was introduced part-way through the study. Of these participants, 38 (43.68%) presently met the criteria for one or more disorder. Broken down by recruitment arm, 8 (100%) from the clinical arm met criteria, while 30 (37.97%) from university channels met criteria. This rate is close to published reports on the prevalence of mental health disorders in college student samples (Auerbach et al., 2018; Evans et al., 2018). Of the total sample, 31 (15.82%) were currently medicated for a mental health issue. Broken down by recruitment arm, all individuals recruited from the clinic were medicated, while 23 (12.23%) of those recruited through normal channels were medicated. Further diagnostic information of the sample is summarised in ***Supplemental Table A.III.S4***.

Flanker task. Participants completed an arrow-version of the Eriksen Flanker task (Eriksen & Eriksen, 1974). Each trial consisted of either congruent (<<<<< or >>>>>) or incongruent (<<><< or >><>>) arrow stimuli presented in white on a grey background of a 32 x 24 cm computer monitor. Participants were instructed to respond as quickly and accurately as possible. Flanker stimulus were presented for 200ms and they had 1050ms to respond by pressing one of

two keyboard keys in order to identify the direction of the central arrow. Responses were indicated using the left ('Q') and right ('P') keys. There was a total of 480 trials split into two blocks, each with 240 stimuli (80 congruent, 160 incongruent) presented. At the end of the first block, if participants had >25% missed trials or had accuracy >90%, they were told to 'Please try to respond faster!' for the second block. If their accuracy was <75%, they were told 'Please try to respond more accurately!'. Otherwise they were told 'Great job!'. Participants completed 30 practice trials (10 congruent; 20 incongruent) of a slower version of the task prior to the beginning of the experimental task (stimulus presentation: 400ms, response time: 1000ms).

Behavioural data pre-processing. Missed trials were excluded from analysis. A total of 2275 trials (2.42%) were removed (per participant mean = 11.61 trials).

EEG recording & pre-processing. Scalp voltage was measured using 128 electrodes in a stretch-lycra cap (BioSemi, The Netherlands). EEG signals were sampled at 512 Hz. EEG data were processed offline using EEGLab (A. Delorme & Makeig, 2004) version 14.1.2 in MATLAB R2018a (The MathWorks, Natick, MA). Data were downsampled to 250 Hz and high-pass filtered at 0.5 Hz. Line noise was removed with CleanLine (Mullen, 2012) at frequencies 50, 100, 150, 200 and 250 Hz. Data were further pre-processed with Clean Rawdata plugin: bad channels were rejected with a criterion of 80% minimum channel correlation and continuous data were corrected using Artifact Subspace Reconstruction (ASR) (Mullen et al., 2013), with correction parameters set at 10 SD for burst criterion and 25% of contaminated channels for time window criterion. All

removed channels were interpolated, and the data were re-referenced to the average. To reject ocular and other non-EEG artefacts, we ran ICA with *runica* (pca option on) on unsegmented EEG data and rejected components automatically with Multiple Artifact Rejection Algorithm (MARA) (Winkler et al., 2011) at a threshold of >40% artefact probability.

Response-locked ERPs. To quantify ERN amplitudes, data were epoched response-locked from -400ms to 500ms and baseline adjusted using a -400ms to -200ms pre-response window on error trials. Epochs were rejected with a threshold criterion of ±50μV before being averaged within-in participant. A total of 36 epochs (0.35%) were removed (per participant mean = 0.18 epochs). The minimum number of epochs for any participant was n = 9, which was above the recommended n = 6 for a reliable ERN (Olvet & Hajcak, 2009b). We then used the adaptive mean method to estimate amplitude as it minimizes bias induced by individual-subject latency variability (Clayson et al., 2013). We searched for the largest negative peak within a window of -20ms to 120ms post-response and took the mean amplitude ±40ms of the negative peak's latency. Correct-related negativity (CRN) amplitudes were measured with the same approach as the ERN, but on correct response trials. We also report results using ERN activity over other electrodes (C22, C24 and D2, mean over 4 mid-frontal electrodes), other measurement methods (non-adaptive mean, minimum amplitude, trough to peak) and when controlled for CRN variation (ERN-CRN (ΔERN), residualised scores of ERN predicted by CRN (ERN_{resid})) in ***Appendix III***.

Reliability measures. Internal consistency (split-half reliability) was calculated for the ERN and CRN. Data were split into two subsets (even versus odd trials/epochs), correlated and adjusted with Spearman-Brown prediction formula.

Self-report psychiatric questionnaires & IQ. Participants completed self-report questionnaires assessing: *alcohol addiction* using the Alcohol Use Disorder Identification Test (AUDIT) (Saunders et al., 1993), *apathy* using the Apathy Evaluation Scale (AES) (Marin et al., 1991), depression using the Self-Rating Depression Scale (SDS) (Zung, 1965), *eating disorders* using the Eating Attitudes Test (EAT-26) (Garner et al., 1982), *impulsivity* using the Barratt Impulsivity Scale (BIS-11) (Patton et al., 1995), *obsessive-compulsive disorder* (OCD) using the Obsessive-Compulsive Inventory - Revised (OCI-R) (Foa et al., 2002), *schizotypy* scores using the Short Scales for Measuring Schizotypy (SSMS) (Mason et al., 2005), *social anxiety* using the Liebowitz Social Anxiety Scale (LSAS) (Liebowitz, 1987) and *trait anxiety* using the trait portion of the State-Trait Anxiety Inventory (STAI) (Spielberger et al., 1983). These self-report assessments were fully randomized within the psychiatric assessment component of the procedure and were chosen specifically to enable transdiagnostic analysis with psychiatric dimensions described in prior work (Gillan et al., 2016; Rouault et al., 2018). A proxy for IQ was also collected using the International Cognitive Ability Resource (I-CAR) (Condon & Revelle, 2014) sample test which includes 4 item types of three-dimensional rotation, letter and number series, matrix reasoning and verbal reasoning (16 items total). Correlations across questionnaires ranged highly (r = -0.05 to 0.75). Internal consistency for all questionnaires were high (Cronbach's alpha > 0.81). Further

details of correlations and reliability measures of questionnaires are in ***Supplemental Table A.III.S1***.

Transdiagnostic factors (dimensions). The current sample size was too small for *de novo* factor analysis (MacCallum et al., 1999). As such, raw scores of the 209 individual items from the 9 questionnaires were transformed into dimension scores (anxious-depression, compulsive behaviour and intrusive thought ('compulsivity'), and social withdrawal) based on weights derived from a larger previous study (Gillan et al., 2016) (N = 1413). These dimensions are not orthogonal and correlate moderately (r = 0.33 to 0.39). See ***Supplemental Table A.III.S2***.

Linear regressions. Regression analyses were conducted using linear models written in R, version 3.6.0 via RStudio version 1.2.1335 (http://cran.us.r-project.org) with the *lm()* function. We investigated if psychiatric questionnaire scores were related to ERN amplitude shifts by taking the total score for each questionnaire (*QuestionnaireScore*; z-scored) as a fixed effect predictor. Separate regressions were performed for each individual symptom due to high correlations across the different psychiatric questionnaires. The model was specified as: ERN ~ *QuestionnaireScore*. For the transdiagnostic analysis, we included all three dimensions in the same model, as correlation across variables was lessened in this formulation and thus more interpretable. We replaced *QuestionnaireScore* in the equation described previously with the three psychiatric dimensions scores (*anxious-depression, compulsivity, social*

withdrawal; all z-scored) entered as predictors. The model was: ERN ~ *Anxious-depression* + *Compulsivity* + *Social withdrawal.*

Goal-directed learning. Participants also completed a reinforcement learning task (Daw et al., 2011) that enabled individual estimations of goal-directed learning, which has previously been shown to be deficient in high compulsive individuals (Gillan et al., 2016) (see ***Figure 3.1***). Briefly, the task consisted of two stages; in the first stage, participants had to choose between two items that had different probabilities of transitioning (rare: 30% or common: 70%) to one of two possible second stages. In the second stage, participants again had to choose between another two items which were associated with a distinct probability of being rewarded that drifted over time. Individuals performing goal-directed learning ('model-based' learner) would make decisions based on the history of rewards and the transition structure of the task, as opposed to individuals who disregarded the transition structure and made decisions solely on the history of rewards ('model-free' learner). To quantify goal-directed learning, we implemented a logistic regression model testing if participants' choice behaviour was influenced by the reward, transition and their interaction of the previous trial. We then tested the relationship of psychiatric dimensions with goal-directed learning by including the three factors (*anxious-depression, compulsivity, social withdrawal*) into the basic model as z-scored predictors. Note that inclusion of age and IQ in the model did not change the pattern of results. See ***A.III. Supplemental Methods*** for further details.

Power calculation. Sample size calculation for a future study was calculated with *pwr.r.test* function in R utilising the correlation coefficient from the equation ERN ~ OCI-R scores.

Data Availability. The code and data to reproduce the ERN analyses of the paper are freely available at https://osf.io/vjda6/.

Results

Participants (N = 196) from a majority student sample completed an arrow-version of the Flanker task, a short IQ evaluation and a battery of self-report questionnaires assessing a range of psychiatric symptoms (see ***Methods***). Individual item-level responses on these questionnaires were transformed into scores for three transdiagnostic dimensions using weights defined in a prior study (Gillan et al., 2016); anxious-depression, compulsive behaviours and intrusive thought and social withdrawal.

Behavioural results. Across participants, mean error rates ranged from 1.97% to 38.24% (mean (M) = 11.55%, standard deviation (SD) = 7.64%) and mean response times (RT) ranged from 123.73ms to 472.55ms (M = 275.70ms, SD = 67.59ms). We observed basic behavioural patterns expected of the task. Mean error rates increased for incongruent trials (M = 15.58%, SD = 10.08%) relative to congruent trials (M = 3.61%, SD = 4.58%) (t_{195} = 19.97, 95% Confidence Interval (CI) [0.11, 0.13], $p < 0.001$). Mean RTs were shorter for congruent trials (M = 234.46ms, SD = 66.51ms) versus incongruent trials (M = 296.03ms, SD = 70.16ms) (t_{195} = -33.38, 95% CI [-0.07, -0.06], $p < 0.001$). Mean RTs were also shorter for error (mean = 212.47ms, SD = 75.20ms) as compared to correct (M = 283.23ms, SD = 66.14ms) trials (t_{195} = -22.41, 95% CI [-0.08, -0.06], $p < 0.001$). Lastly, post-error mean RTs (M = 294.44ms, SD = 88.47ms) were slower than post-correct mean RT (M = 274.51ms, SD = 68.38ms) (t_{195} = 6.13, 95% CI [0.01, 0.03], $p < 0.001$). Error rate and RT distributions are visualised in ***Supplemental Figure A.III.S1***.

Response-locked event related potentials (ERPs). Grand-average ERP waveforms at electrode FCz are presented in ***Figure 4.1a*** for the ERN and CRN. ERN waveforms contained an average of 51.88 (SD = 33.09) error trials per participant while the CRN waveform was constructed with average of 404 (SD = 58.78) correct trials. Across participants, we measured ERN amplitude with the adaptive mean method. The ERN exhibited an amplitude of -3.11μV (SD = 2.79μV) while the CRN had an amplitude of 0.30μV (SD = 1.89μV). Paired t-test indicated more pronounced negativities for the ERN than CRN (t_{195} = -16.66, 95% CI [-3.82, -3.01], $p < 0.001$) within-subject. Split half-reliability was high for both measures (ERN: $r = 0.90$; CRN: $r = 0.98$), confirming the suitability of this measure for between-subject analysis.

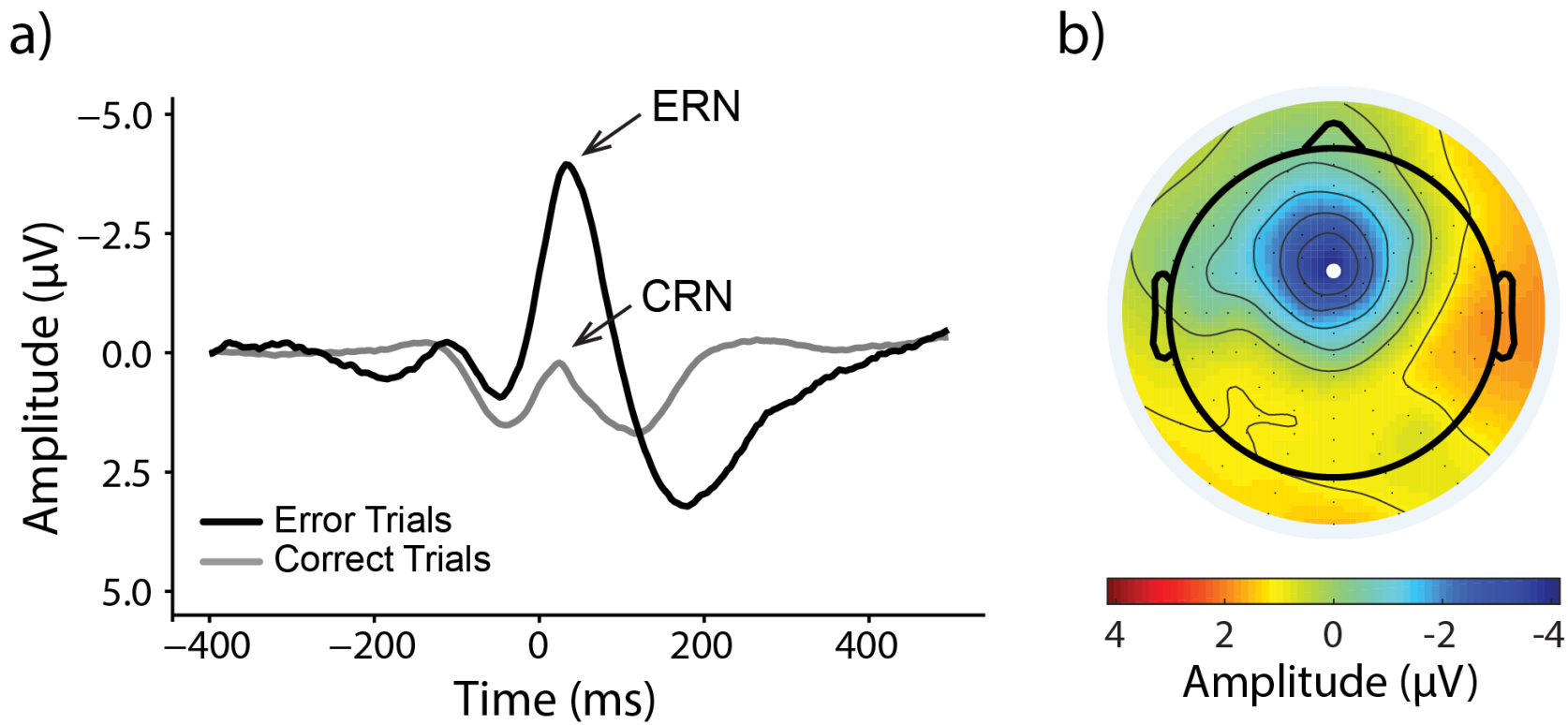


Figure 4.1. *Error-related negativity (ERN).* ***(a)*** *Response-locked grand average waveforms for error and correct responses at electrode FCz. Negative values are plotted upwards. Event-related potential components are labelled: ERN: error-related negativity; CRN: correct-related negativity.* ***(b)*** *Scalp map displays the voltage distribution at 37.61ms, the grand average latency of the most negative peak for error trials. Electrode FCz position is indicated with a white dot.*

ERN, questionnaire scores and transdiagnostic dimensions. We tested if ERN amplitudes were associated to the self-reported questionnaire scores. In contrast to our hypothesis, none of the psychiatric questionnaires showed a significant relationship to ERN amplitude (all $p > 0.15$, where $p < 0.005$ is the Bonferroni corrected significance threshold) (***Figure 4.2*** and ***Table 4.1***).

For the interested reader, we conducted unplanned, supplementary analyses to test for consistency of our findings across different methods of ERN quantification, electrode site and reaction times (speed-accuracy trade off (Arbel & Donchin, 2009; Gehring et al., 1993)). The patterns of results were remarkably similar: no symptom was significantly related to ERN amplitude in surrounding electrode sites (all $p > 0.08$, uncorrected) (***Supplemental Figure A.III.S6***) or with three other ERN quantification methods (all $p > 0.10$, uncorrected) (***Supplemental Figure A.III.S7***). Two other ERN measures that controlled for CRN variation (ΔERN and ERN_{resid}) also did not reveal any significant associations (all $p > 0.12$, uncorrected) (***Supplemental Figure A.III.S8*** and ***Table A.III.S3***). Inclusion of error rate, demographics or medication status did not affect the pattern of results (all $p > 0.09$, uncorrected) (***A.III. Supplemental Methods***).

Transdiagnostic phenotyping did not provide a better explanation for the data, with none of the transdiagnostic dimensions significantly associated to ERN amplitude (all $p > 0.18$, uncorrected) (***Figure 4.2*** and ***Table 4.1***).

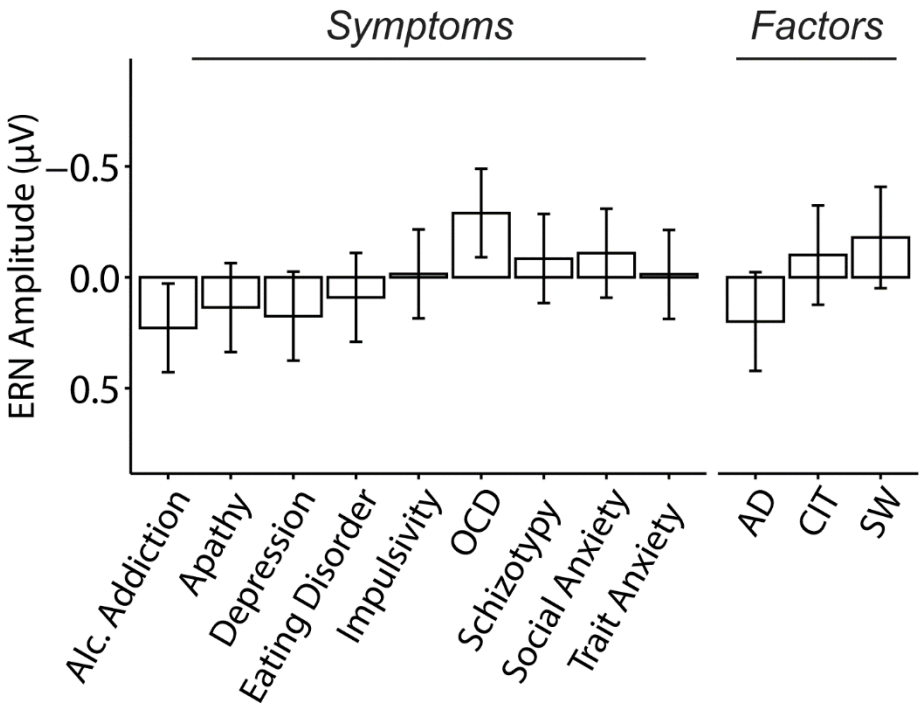


Figure 4.2. *Non-significant associations between ERN amplitude and self-reported psychopathology. Associations between ERN amplitude with questionnaire total scores or transdiagnostic dimension scores (anxious-depression (AD), compulsive behaviour and intrusive thought (CIT) and social withdrawal (SW)). Error bars denote standard errors. Each questionnaire score was examined in a separate regression, whereas dimensions were included in the same model. The Y-axis indicates the change in ERN amplitude as a function of 1 standard deviation (SD) increase of questionnaire or dimension scores. See* ***Table 4.1****.*

Psychiatric Questionnaire	***β (SE)***	***z-value***	***p-value***
Alcohol Addiction	0.23 (0.20)	1.14	0.25
Apathy	0.14 (0.20)	0.68	0.50
Depression	0.18 (0.20)	0.88	0.38
Eating Disorder	0.09 (0.20)	0.45	0.65
Impulsivity	-0.05 (0.20)	-0.08	0.94
OCD	-0.29 (0.20)	-1.45	0.15
Schizotypy	-0.08 (0.20)	-0.42	0.67
Social Anxiety	-0.11 (0.20)	-0.54	0.59
Trait Anxiety	-0.01 (0.20)	-0.07	0.95
Transdiagnostic Dimension	***β (SE)***	***t-value***	***p-value***
Anxious-depression	0.29 (0.22)	1.34	0.18
Compulsive behaviour and intrusive thought	-0.03 (0.22)	-0.14	0.86
Social withdrawal	-0.20 (0.22)	-0.91	0.36

Table 4.1. *Associations between ERN amplitude and total scores of self-report psychiatric questionnaires or transdiagnostic dimensions. SE = standard error. For psychiatric questionnaires, each row reflects the (uncorrected for multiple comparisons) results from an independent analysis where each psychiatric questionnaire score was regressed against ERN amplitude. For transdiagnostic dimensions, all three dimensions scores were included in the same regression model.*

Goal-directed control and compulsivity. For comparison purposes, we also assessed goal-directed learning in the same sample using the two-step reinforcement learning task (Daw et al., 2011). Split half-reliability was r = 0.71 for this measure. In prior work, the compulsivity dimension was associated with reduced goal-directed learning (Gillan et al., 2016). We replicated this finding (β = -0.07, SE = 0.04, p = 0.048) (***Figure 4.3***), suggesting that the dimension scores obtained from this general population sample were valid, providing a comparator for interpreting the effect size of ERN trends in the present work.

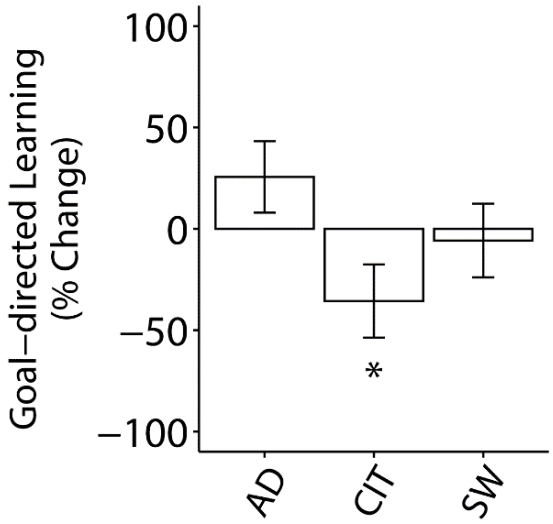


Figure 4.3. *Associations between goal-directed learning and psychiatric dimensions (anxious-depression (AD), compulsive behaviour and intrusive thought (CIT) and social withdrawal (SW)) (N = 196). Error bars denote standard errors. Factors were included in the same model. The Y-axis indicates the percentage change in goal-directed learning as a function of 1 SD increase of dimension scores. *p < 0.05.*

Discussion

In the present paper we investigated if ERN abnormalities commonly observed in a range of psychiatric disorders could be explained by a transdiagnostic dimension characterised by high levels of anxious-depression. Fundamental to this was the replication of existing associations of ERN amplitude shifts with the clinical phenotypes such as OCD and trait anxiety, but to our surprise we could not detect any significant associations of any symptoms with the ERN. Reformulating questionnaires into transdiagnostic dimensions did not improve signal.

We considered several possible explanations for the data. First, that the range of psychopathology sampled was insufficiently high to detect associations with the ERN. We intentionally enriched our sample by including 8 patients from a local anxiety clinic (who were starting group therapy) to protect against this possibility. This was not necessary; the rest of the sample exhibited high rates of psychopathology (***Supplemental Figure A.III.S3***), consistent with the documented characteristics of university students (Auerbach et al., 2018; Bayram & Bilgel, 2008; Evans et al., 2018). Excluding the 8 patients recruited from an anxiety disorder clinic, 37.97% of the sample who were assessed with a standard psychiatric interview (M.I.N.I., see ***Methods***) presently met criteria for at least one disorder. In terms of the range of self-report symptoms, 25.51% (N = 50) scored of ≥21 on the OCI-R and 54.59% (N = 107) scored of >41 on the STAI, the standard clinical threshold for OCD and anxiety for the respective instruments (Ercan et al., 2015; Foa et al., 2002).

Second, we ensured that our data, both self-report and electrophysiological, were valid. Internal consistency measures were high for all questionnaires (***Supplemental Table A.III.S1***). We note the transdiagnostic dimensions utilised here were defined from a prior study (Gillan et al., 2016) and not derived from our current data. The factor structure has been replicated in two other independent datasets (Rouault et al., 2018; Seow & Gillan, 2020) highlighting its reproducibility and validity. Finally, perhaps the strongest evidence for the validity of the transdiagnostic dimensions structure we employed here are two replications with respect to the specific association between compulsive behaviour and intrusive thought scores and goal-directed planning; one observed in the present study and another via a separate research group (Patzelt et al., 2019).

In terms of the ERN itself, we were able to reproduce all the expected behavioural (***Supplemental Figure A.III.S1*** and ***Figure A.III.S2***) and electrophysiological patterns this task was expected to elicit, suggesting that there were no issues with data quality. Our paradigm consisted of twice as many incompatible trials than compatible trials which was intended to induce higher conflict frequency to increase the number of errors made. However, conflict frequency has been shown to modulate performance monitoring ERPs such as N2 and CRN amplitudes (Bartholow et al., 2005; Grützmann et al., 2014). One study found that increasing task difficulty (by using shorter response times and poorer visual contrast) abolished ERN differences between groups of high/low OC symptoms (Kaczkurkin, 2013), raising the concern that our task may have induced a difficulty component that

served to dampen our ability to detect ERN-OCD associations. Evidence to the contrary however comes from the fact that the trial compatibility ratio in our task has been used previously in a clinical study and was able to detect larger ERN amplitudes in OCD patients versus healthy controls (Riesel et al., 2014).

Contextualising these data with the broader literature, ERN abnormalities may be more sensitive to the categorical comparison of patients versus controls than dimensional variation in the general population. Several OCD patient studies did not find any correlation with symptom severity and ERN amplitude within patient groups (Carrasco, Harbin, et al., 2013; Endrass et al., 2008; Riesel, 2019; Riesel et al., 2014, 2017). The ERN remains elevated in OCD despite successful treatment (Hajcak, 2006; Ladouceur et al., 2018; Schrijvers et al., 2009) and elevations are also observed in unaffected first-degree relatives of patients (Carrasco, Harbin, et al., 2013; Riesel et al., 2019). As such, the ERN has been couched as a psychiatric *vulnerability endophenotype* (Riesel, 2019). Nonetheless, our individual differences approach should have been able to pick up these trait effects along the continuum of scores, regardless of the subtleties of state-based fluctuations.

Perhaps the simplest explanation for these data is that ERN associations with psychopathology are smaller than previously assumed. Notably, effects of OCD symptoms trended in the predicted direction, where individuals who scored higher on this questionnaire had a larger ERN. Likewise, the trend was for alcohol addiction to be associated with a blunted ERN, consistent with the previous literature. Recent

reviews add support to this conclusion: two meta-analyses noted that overall effect sizes for anxiety or OCD traits for enhanced ERN were relatively small (Cavanagh & Shackman, 2015; Pasion & Barbosa, 2019) and another that assessed the effect size of ERN amplitude shifts in OCD (Riesel, 2019) noted that larger effect sizes were associated with smaller sample publications, suggesting publication bias. In terms of statistical power, our study has one of the highest sample numbers (N = 196) investigating the ERN in psychiatry. Despite the ERN's link to psychiatry for two decades (Gillan et al., 2017; Olvet & Hajcak, 2009a; Riesel, 2019; Weinberg, Dieterich, et al., 2015), there are only six studies with total N > 150 to date (Hanna et al., 2016, 2018; A. Meyer & Klein, 2018; Riesel et al., 2019; Weinberg et al., 2016; Weinberg, Kotov, et al., 2015). Our small effect size between OCD symptoms and the ERN suggest a sample size of N = 729 in order to have 80% power to detect a 0.05 significant association.

It is perhaps notable, nonetheless, that our transdiagnostic framework did not perform better in relative terms. Returning to our hypothesis, we found no evidence that anxious-depression might be responsible for the commonly observed enhancement of the ERN in anxiety disorders. In fact, the direction of this non-significant effect was in the opposite direction. This finding might reflect the fact that this dimensional framework is not apt to capture variation in the ERN. The transdiagnostic dimensions utilised in this study are not intended to be fixed and final. Future research might explore alternatives to the framework employed here to investigate if a dimensional structure exists that can explain the common ERN

patterns seen across psychiatric disorders. As these dimensions have previously shown specific associations with other cognitive deficits such as goal-directed planning and metacognition (Gillan et al., 2016; Rouault et al., 2018; Seow & Gillan, 2020), our findings suggest that ERN amplitude shifts either exhibit smaller effect sizes than was previously thought or are not underlain by the same psychopathology as other cognitive phenomena that are characteristic of OCD and related disorders.

This year, several authors have highlighted the potential for a transdiagnostic framework at reconciling the broad range of ERN patterns in the literature (Gillan et al., 2017; Pasion & Barbosa, 2019; Riesel, 2019). The present paper is timely, being the first study to apply an expansive and empirically robust transdiagnostic approach that directly addresses the issue of co-occurring symptoms in a large sample. To our surprise, despite being well-powered, we were unable to significantly replicate previously observed associations with various aspects of mental health and a transdiagnostic approach to quantifying mental health in the sample did nothing to remedy that. Future research in this area might agree that even larger samples than previously assumed are needed to delineate robust associations in general population samples.

Chapter 5: General discussion

Summary

Several features of the current psychiatric nosology, such as the similarity across supposedly distinct disorder classes and heterogeneity within these categories, have hampered advances in delineating mechanisms core to compulsive behaviour. For example, although it has been posited that the repetitive compulsive actions of OCD and related disorders arise from faulty goal-directed control systems in the brain (Gillan et al., 2016; Sjoerds et al., 2013; Voon et al., 2015), other research has also shown that disorders which are non-compulsive in nature exhibit the same impairment (Alvares et al., 2014, 2016; Culbreth et al., 2016; C. Delorme et al., 2016; R. W. Morris et al., 2015). At face value, this suggests that failures in goal-directed control are not specific to compulsivity, as opposed to other aspects of psychopathology. Another explanation, however, is that the methods we typically employ to study the neurocognitive basis of mental health, i.e. case-control comparison designs based on the DSM, are ill-equipped to delineate mechanistic explanations for specific psychiatric phenomena. A study by Gillan and colleagues from 2016 eschewed the traditional case-control design and instead examined whether transdiagnostic psychiatric dimensions, expressed in normal variation of general population samples, may provide an alternative pathway for mapping mental health to mechanisms (Gillan et al., 2016). Using this method, the authors found that deficits in goal-directed control failure were specific to variation along a spectrum of compulsivity (and were not linked to other dimensions of mental health including

anxious-depression and social withdrawal phenotypes). A subsequent study using this approach found novel metacognitive deficits in compulsivity, highlighting how these effects may otherwise be hidden by competing influences of co-occurring anxious-depression in patients versus healthy controls (Rouault et al., 2018). Together, these findings suggest that a transdiagnostic perspective may be critical in developing a robust, replicable and specific neurocognitive characterisation of compulsivity. This premise was the central focus of this thesis, which used a dimensional method to carry out a comprehensive investigation of the advantages of this approach in understanding the brain processes that go awry in compulsivity.

The first experimental chapter aimed to advance our understanding of metacognitive deficits in compulsivity by studying confidence formation in the context of reinforcement learning. This was already attempted in an OCD patient versus healthy control comparison study that utilised a predictive inference task and found no confidence abnormalities (Vaghi et al., 2017). Here, we tested the premise that adopting a transdiagnostic methodology would reveal important metacognitive deficits on this task (Gillan et al., 2016). We found that this was indeed the case; the dimensional approach revealed metacognitive dysfunctions linked to compulsivity which were not present in the prior case-control study. Specifically, we replicated the double dissociative confidence relationship between compulsivity and anxious-depression previously observed in perceptual decision making (Rouault et al., 2018): compulsivity was related to increased confidence levels while anxious-depression was linked to lowered confidence levels. We also found that high

compulsive individuals were less informed by several environmental evidence sources when adjusting their confidence reports. These metacognitive deficits suggest that high compulsive individuals have faulty metacognitive beliefs and were additionally impaired in updating these beliefs with surrounding evidence. With respect to one finding that we replicated from this paper, that OCD symptom severity was linked to a disconnection between confidence and behaviour ('decoupling'), we showed that the same deficit was also associated with five out of eight of the other psychiatric phenomena we measured (and all 8 when no correction for multiple comparisons was applied). This highlights the generalisability of this finding, which was circumvented when we applied the transdiagnostic analysis and obtained a more specific result: action-confidence decoupling was only significantly associated with the compulsive dimension. Overall, results from this experiment highlighted the advantage of using a dimensional method in disambiguating effects linked to different psychopathology that may confound each other in a case-control investigation and suggested that compulsivity is linked to deficiencies in constructing and maintaining the mental model.

One implication of having a faulty mental model is problems in enacting goal-directed behaviours that require this "cognitive map" for simulation and planning. In general, it is thought that goal-directed deficits can arise from either the failure in *constructing and maintaining* this model or simply in the *use/implementation* of it. In the second experimental chapter, we used EEG to examine the neural correlates of the representation of the mental model while participants performed the two-step

reinforcement learning task (Daw et al., 2005, 2011). We found behavioural and neural signatures of the mental model (state-state transition knowledge) in both reaction time and the desynchronization in parietal-occipital alpha power that were correlated with model-based planning ability. Notably, these signatures were diminished in high compulsive individuals. This suggests that high compulsive individuals have impaired representations of the mental model that is linked to failures in making goal-oriented choices. As for the implementation of the model, the data were not as clear. We found that the general marker of cognitive control, mid-frontal theta power, was diminished in high compulsive individuals during the making of choice that was linked to less awareness of state-state transition via reaction times. However, the measure did not correlate with individual differences in model-based planning; on a more conceptual level, it becomes challenging to measure the implementation of a model that is itself compromised. The data here, together with those of the prior chapter, suggest that compulsivity is characterised by an impoverished internal model of the task environment—a prerequisite for model-based (goal-directed) behaviour. Finally, we also observed that the modulations of both alpha and theta power were non-specific when single-disorder questionnaires were examined, but in support for the dimensional approach, the effects were only significantly linked to the compulsive dimension (and not anxious-depression or social withdrawal).

In the final experimental chapter, we attempted to address the conundrum whether hyperactive error monitoring, reflected as enhanced error-related negativity (ERN)

amplitudes, relate to compulsive or anxious symptomatology. ERN amplitude enhancement is observed in both OCD and anxious disorders (Moser et al., 2013; Riesel, 2019), but enhanced ERNs seem to make more functional sense for anxiety given the error signal's relation to startle response and avoidance (Frank et al., 2005; Hajcak & Foti, 2008). We thus utilised a modified Flanker task (Eriksen & Eriksen, 1974) and transdiagnostic phenotyping to test this hypothesis. Contrary to our expectations, we found that none of the nine psychiatric phenomena we investigated, including OCD and trait anxiety symptoms, were significantly associated with any ERN amplitude shifts. The dimensional analysis also did not explain the data better. This non-significant pattern of results was contrasted with a significant deficit in goal-directed control associated with compulsivity in the same sample. Our conservative interpretation suggests that ERN amplitude effect sizes in psychiatry may be much smaller than assumed in the literature. As of yet, it remains unclear if transdiagnostic approaches can provide a better fit to the ERN than classic disorder-based research designs.

Synthesis, limitations and future directions

The 'habit hypothesis' of OCD has proven itself to be a mechanistically and neurobiologically promising explanation for how compulsions arise. However, research thus far has focused mainly on delineating which general mode of action control goes awry: habitual or goal-directed (Robbins et al., 2012). As such, there is little evidence (or indeed, investigation) of *why* goal-directed control is deficient—whether it is because of the lack of an accurate world-model to rely upon and/or a

tendency to ignore that mental model in favour of an alternative and less taxing route to action selection. In this thesis, we observed that compulsive individuals exhibited several metacognitive dysfunctions and had a failure to acquire an accurate representation of task contingencies (evident in both reaction time and EEG). These studies suggest for the first time that failures in goal-directed control commonly observed in compulsivity may arise due to impairments in curating the meta-model for action planning.

More recent conceptualisations of OCD have begun to question if compulsive symptomatology is best explained in terms of behavioural inflexibility, citing that probabilistic reversal learning tasks have, on the contrary, found lower perseverative behaviours linked to compulsivity (Fradkin et al., 2020; Hauser, Iannaccone, et al., 2017). Instead, it has been suggested that compulsivity may be more accurately described by dysfunctions in altered feedback processing (Fradkin et al., 2018), more specifically that OCD patients are uncertain about state transition changes because they under-rely upon accumulated past knowledge (Fradkin et al., 2020). Our EEG findings align with this alternative hypothesis to a certain degree. We found that high compulsive individuals showed diminished neural representations of transition types and reduced awareness (by reaction times) to state-state transitions. Although the impaired transition knowledge suggests an issue with information processing, whether it is indeed caused particularly by an under-reliance of accumulated past evidence was not something we could distil in our experiment. Our results in chapter 2 suggest that the update of the mental model in compulsive

individuals is relatively insensitive to more than just major state changes in the environment, but also to positive versus negative feedback (hits/misses) and the uncertainty of the current state (relative uncertainty). It is important to note that our findings were confined to the mental model (i.e. confidence), rather than actual behavioural updates. This distinction may be important for understanding why behavioural deficits in compulsivity appear relatively confined to prospective, multi-step tasks (e.g. model-based planning) where the invocation of a mental model is required. Additionally, it may explain why trial-by-trial behavioural responses to feedback, like model-free reinforcement learning, remain intact in compulsivity in contrast (Gillan et al., 2016). Overall, our results describe a general impairment in evaluating and integrating information into the higher-order mental model in compulsivity, which may be the common source underlying both metacognitive deficits and decision failures observed in high compulsive individuals.

The dysfunctional metacognitive mechanisms we observed here may have important insights for understanding a clinical feature that often presents alongside compulsivity—obsessional belief. Inflated confidence levels and deficiencies in adjusting confidence in response to various feedback sources may contribute to the formation and reinforcement of more rigid beliefs. In tentative support to this idea, individuals with other disorder diagnoses who struggle with false beliefs such as schizophrenia are also overconfident, particularly for errors (Moritz et al., 2014). A cognitive model of schizophrenia has also outlined how delusional belief formation can arise from dysfunctional metacognitive processes (Joyce et al., 2013). In the

current transdiagnostic framework, loadings of obsessionality and schizotypy are captured within the same dimension as compulsivity (Gillan et al., 2016). As such, we are unable to probe if metacognitive dysfunction would form a transdiagnostic mechanism linked to dysfunctional beliefs if distinguished from compulsivity, though future work may attend to this.

Understanding how the mental model becomes dysfunctional in compulsivity may provide new insights regarding the nature of interaction between obsessions and compulsions in OCD, potentially providing new routes towards “breaking” the reinforcing cycle of the disorder. Conventional cognitive-behavioural clinical models of OCD believe that compulsions are performed in response to the primary issue of dysfunctional beliefs (Rachman, 1997; Salkovskis, 1985), e.g. as attempts to resolve uncertainties associated with an obsession (Tolin et al., 2001). On the other hand, some empirical cross-sectional data suggest the opposite; that obsessions are the by-products of compulsive behaviour (Gillan & Sahakian, 2015). Gillan and colleagues tracked the interaction of belief and compulsive habitual behaviour in OCD patients and healthy controls using a shock avoidance task with physiological arousal assessments (i.e. subjective report and skin conductance readings) (Gillan, Morein-Zamir, Urcelay, et al., 2014). Participants were first trained to perform actions to prevent experiencing shocks associated with stimuli. They were then tested if they continued to respond towards the options that were subsequently devalued i.e. show persistent avoidance habits. OCD patients’ behaviour with devalued stimuli were indistinguishable to controls at the first habit probe after a brief period of training,

however at the second habit probe after over-training, they exhibited enhanced avoidance habits in comparison to controls. Interestingly, it was only *after habit behaviour was expressed* at the second habit probe that OCD individuals reported irrational beliefs of shock threat for devalued options (e.g. “I thought it could still shock me”) even though they had accurate task contingency and shock expectancy knowledge. These observations suggest that obsessions are perhaps born to rationalise the compulsive behaviour. In this thesis, our data implicating the broad impairment of the mental model in metacognition and goal-directed control suggest a more nuanced view than either of these former models have suggested—that obsessions and compulsions might be different manifestations of a common metacognitive failure. Further work in this area will ultimately necessitate dynamic, longitudinal and causal investigations of the link between metacognitive and goal-directed processes in compulsivity.

Finally, we endeavoured to assess the specificity of hyperactive error monitoring to compulsivity, via the ERN, to further our understanding of the core mechanisms of this phenotype. Unfortunately, the ERN data from our last experimental chapter did not reveal any of our hypothesized effects, but instead presented a non-significant pattern of results. As such, we concluded that the ERN may be limited in its utility as a biomarker for mental health phenomena based on the small effect sizes observed. We highlight that ERN studies in the literature have predominantly utilised small sample sizes of around N = 20 to 30 cases versus controls and have massive variation in their ERN quantification methods; these issues may have led to false

positives increasing the effect size of the ERN in psychiatry. In contrast, we ensured we were well-powered (N = 196) and applied a rigorous analysis approach by pre-defining our ERN peak scoring method. Notably, our non-significant ERN effects were not an artefact of our analysis choice as all other possible ERN quantification methods we investigated post-hoc (in ***Appendix III***) showed the same pattern of results.

It was unfortunate that our current transdiagnostic framework did not explain the ERN data better. We used these pre-defined, replicable dimensions to allow a clear extension to prior work (Gillan et al., 2016; Rouault et al., 2018) and to the other experiments in this thesis that adopted the same approach. Additionally, this prevented issues arising from overfitting new psychiatric descriptions to our data (Gillan & Whelan, 2017; Whelan & Garavan, 2014). Notwithstanding the reproducibility issues we observed, it remains a possibility that another transdiagnostic, dimensional framework might be more sensitive to ERN abnormalities than the one we applied here. For instance, a popular two-factor conceptualisation of psychopathology that was born from paediatric psychiatry has been utilised in ERN research centres on internalising versus externalising dimensions (Krueger, 1999; Krueger et al., 1998). Internalising disorders are characterised by negative emotionality, encapsulating mood and anxiety disorders (e.g. depression, generalised anxiety disorder, phobias and OCD), whereas externalising disorders have prominent behavioural control issues, and comprise of aggressive, delinquent and impulsive disorders (e.g. conduct disorder, drug and

substance use disorders). As the directional amplitude shifts of the ERN seem to cluster in disorder groups that align with this framework (Olvet & Hajcak, 2008), several groups have tested (in paediatric samples) this hypothesis (Kessel et al., 2016; A. Meyer & Klein, 2018; Troller-Renfree et al., 2016). A recent meta-analysis of clinical and subclinical ERN studies generally supported internalising symptoms (but not depression) being linked to enhanced ERN amplitudes while externalising symptoms (except for alcohol abuse) were associated with lower ERN amplitudes (Pasion & Barbosa, 2019). Though this framework is attractive, it requires more stringent checks to promote validity, reliability and replicability. Across studies that have tested this framework in the ERN, there was no consensus as to how internalising or externalising phenotypes were defined. Notably, many studies do not account for heterogeneity within disorder groupings by simply comparing clusters of disorder categories which fall into either the internalising or externalising spectrum, which is susceptible to the usual criticisms of case-control approaches.

Clinical implications

Currently, many treatments for OCD exist—drug therapies (Del Casale et al., 2019), psychotherapy (Hezel & Simpson, 2019) and even brain activity augmentation interventions such as deep brain stimulation for treatment resistant cases (Rapinesi et al., 2019). Unfortunately, our understanding of the neurobiological basis of OCD and its treatments are still lacking. It is therefore unsurprising that these treatments suffer from an efficacy issue. For instance, the first line interventions for OCD patients such as selective serotonin reuptake inhibitors (SSRIs) and exposure

response prevention (ERP) report that approximately only 50-68% of patients have significant improvement in their symptoms post-treatment (Eddy et al., 2004; Fisher & Wells, 2005b; Stengler-Wenzke et al., 2006). Moreover, a large proportion of those patients who *respond* to treatment do not reach *remission*. Can computational psychiatry help us understand and leverage the mechanisms behind successful treatment for OCD? We acknowledge that the dysfunctional mechanisms we implicate in compulsivity in the outputs of this thesis (as well as those in the literature) are correlational in nature. As such, first probing the *causality* of these associations would be key towards translating these insights into the clinic. Nonetheless, we outline potential ways how our detailed characterisation of the metacognitive and goal-directed control mechanisms involved in compulsivity may impact treatment considerations.

Firstly, our findings reveal a new potential therapeutic target for compulsivity—metacognition. Metacognitive therapy (MCT) for OCD exists (Clark, 2005; Fisher & Wells, 2005a); but these treatments are currently focused on modifying intrusive thoughts and thereby the necessity of performing compulsive rituals. Our results pertain not to the content of one intrusion or another, but to the *general ability* of patients to construct accurate mental representations of environmental contingency, or in other words, cause-and-effect. It is possible that newer metacognitive treatments might focus on improving the patients' capability to globally discriminate between correct or incorrect inferences and as a result, might prove more beneficial in the long term than modifying specific idiosyncratic OCD beliefs that are context

dependent, subject to change or relapse. Promisingly, recent study has shown that domain-general metacognition can be improved by adaptive training (Carpenter et al., 2019), however whether that would be therapeutically relevant and lead to symptom improvement is an open question for future research to examine. Additionally, the administration of propranolol, which causes noradrenaline blockade, has been found to enhance metacognitive performance (Hauser, Allen, Purg, et al., 2017). This may potentially be a novel pharmacological therapeutic intervention for compulsivity. Though the use of propranolol is yet untested in OCD treatment, clomipramine (a common OCD medication) also impacts noradrenergic functions (Asakura et al., 1982)—this may be a mechanism by which its therapeutic relief manifests.

Secondly, our data offers new insights to cognitive paradigms that have the potential to investigate mechanisms underlying symptom improvement after treatment for compulsivity. As impaired goal-directed control is the current prominent neurocognitive model of compulsivity, the two-step reinforcement learning task has been trialled as a paradigm that could be of use to understand ERP therapy outcomes (Wheaton et al., 2019). However, it was observed that model-based planning performance does not improve after ERP therapy in OCD individuals despite an improvement in symptoms. It was suggested that ERP may not affect its efficacy through the goal-directed processing system or that goal-directed failures are a trait dependent deficit reflecting vulnerability for compulsivity. Given that we highlight that a deficiency in the mental model is a core feature for compulsivity,

future studies may consider focusing on testing whether the status of the mental model is augmented with therapy.

Thirdly and perhaps most importantly, these data suggest it may be time to start considering transdiagnostic approaches to treatment. For example, treatments might be more effective if issued on the basis of severity of *compulsivity* rather than discrete disorder categories where patients with the same diagnosis of OCD can differ markedly in their levels of anxiety, obsessionality and compulsivity. A reason why therapeutic methods have not been as efficacious as hoped is no doubt partly due to the lack of biological validity of both treatment interventions and the current psychiatric nosology (Abramowitz et al., 2000; Overbeek et al., 2002; Perugi et al., 2002). By describing patients along a continuum of compulsivity that has a well-defined neurobiological profile, it may be easier to find treatments that target the implicated neural mechanism and that are best suited for the individual.

Moreover, in principle, transdiagnostic interventions might be more cost-effective for healthcare systems as a smaller set of transdiagnostic protocols can replace an arguably much larger range of disorder-specific treatments. Currently, there are some emerging treatment protocols have that moved beyond focusing on single disorder groupings. A notable example is the Unified Protocol for Transdiagnostic Treatment of Emotional Disorders that addresses the hypothesized underlying mechanism of emotion dysregulation common to affective disorders such as anxiety disorders and depression (Barlow, Farchione, Sauer-Zavala, et al., 2017). There is

some evidence in favour of this approach, especially in its ability to replace multiple diagnosis-specific interventions without losing efficacy (Barlow, Farchione, Bullis, et al., 2017), though whether it is superior compared to other treatment modalities remains to be seen (Sakiris & Berle, 2019).

Conclusions

In summary, this thesis sought to understand the neurocognitive mechanisms of compulsivity. Over three experiments, we adopted a transdiagnostic perspective. In two of three experiments, we highlighted the clear advantages of the approach over classic disorder-based frameworks. In addition to yielding new insights into the core mechanisms posited to underlie compulsivity, we undertook a rigorous approach and established explicit replications of prior studies to all of our designs. We were able to reproduce associations between compulsivity and goal-directed deficits and metacognition. In the only study that did not provide clear support for the transdiagnostic method, we found that more general reproducibility issues were, too, at play. We propose that compulsivity is an important transdiagnostic symptom dimension that is characterised by failures in curating the meta-model. Our findings may explain the well-documented failures of behavioural control that have been previously observed in both OCD and compulsivity. These findings not only advance transdiagnostic theories of mental illness but it is hoped that they will open new therapeutic opportunities for consideration.

References

Abramowitz, J. S., Franklin, M. E., Schwartz, S. A., & Furr, J. M. (2003). Symptom presentation and outcome of cognitive-behavioral therapy for obsessive-compulsive disorder. *Journal of Consulting and Clinical Psychology*, ***71***(*6*), 1049.

Abramowitz, J. S., Franklin, M. E., Street, G. P., Kozak, M. J., & Foa, E. B. (2000). Effects of comorbid depression on response to treatment for obsessive-compulsive disorder. *Behavior Therapy*, ***31***(*3*), 517–528.

Agam, Y., Greenberg, J. L., Isom, M., Falkenstein, M. J., Jenike, E., Wilhelm, S., & Manoach, D. S. (2014). Aberrant error processing in relation to symptom severity in obsessive–compulsive disorder: a multimodal neuroimaging study. *NeuroImage: Clinical*, ***5***, 141–151.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, ***9***(*1*), 357–381.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, ***14***(*10*), 1338.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, ***108***(*4*), 441.

Alvares, G. A., Balleine, B. W., & Guastella, A. J. (2014). Impairments in goal-directed actions predict treatment response to cognitive-behavioral therapy in social anxiety disorder. *PLoS One*, ***9***(*4*), e94778.

Alvares, G. A., Balleine, B. W., Whittle, L., & Guastella, A. J. (2016). Reduced goal-directed action control in autism spectrum disorder. *Autism Research*, ***9***(*12*), 1285–1293.

Andrews, G., Goldberg, D. P., Krueger, R. F., Carpenter, W. T., Hyman, S. E., Sachdev, P., & Pine, D. S. (2009). Exploring the feasibility of a meta-structure for DSM-V and ICD-11: could it improve utility and validity?: Paper 1 of 7 of the thematic section:'A proposal for a meta-structure for DSM-V and ICD-11.' *Psychological Medicine*, ***39***(*12*), 1993–2000.

Anholt, G. E., Aderka, I. M., Van Balkom, A., Smit, J. H., Schruers, K., Van Der Wee, N. J. A., Eikelenboom, M., De Luca, V., & Van Oppen, P. (2014). Age of onset in obsessive–compulsive disorder: admixture analysis with a large sample. *Psychological Medicine*, ***44***(*1*), 185–194.

Anokhin, A. P., Golosheykin, S., & Heath, A. C. (2008). Heritability of frontal brain function related to action monitoring. *Psychophysiology*, ***45***(*4*), 524–534.

APA. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Arbel, Y., & Donchin, E. (2009). Parsing the componential structure of post-error ERPs: A principal component analysis of ERPs following errors. *Psychophysiology*, ***46***(*6*), 1179–1189.

Asakura, M., Tsukamoto, T., & Hasegawa, K. (1982). Modulation of rat brainα2- and β-adrenergic receptor sensitivity following long-term treatment with antidepressants. *Brain Research*, ***235***(*1*), 192–197.

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Green, J. G., Hasking, P., Murray, E., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Stein, D. J., Vilagut, G., Zaslavsky, A. M., & Kessler, R. C. (2018). WHO world mental health surveys international college student project: Prevalence and distribution of mental disorders. *Journal of Abnormal Psychology*, ***127***(*7*), 623.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, ***37***(*4–5*), 407–419.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, ***35***(*1*), 48–69.

Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., Bentley, K. H., Thompson-Hollands, J., Conklin, L. R., & Boswell, J. F. (2017). The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. *JAMA Psychiatry*, ***74***(*9*), 875–884.

Barlow, D. H., Farchione, T. J., Sauer-Zavala, S., Latin, H. M., Ellard, K. K., Bullis, J. R., Bentley, K. H., Boettcher, H. T., & Cassiello-Robbins, C. (2017).

Unified protocol for transdiagnostic treatment of emotional disorders: Therapist guide. Oxford University Press.

Bartholow, B. D., Pearson, M. A., Dickter, C. L., Sher, K. J., Fabiani, M., & Gratton, G. (2005). Strategic control and medial frontal negativity: Beyond errors and response conflict. *Psychophysiology*, ***42***(*1*), 33–42.

Bates, A. T., Kiehl, K. A., Laurens, K. R., & Liddle, P. F. (2002). Error-related negativity and correct response negativity in schizophrenia. *Clinical Neurophysiology*, ***113***(*9*), 1454–1463.

Bayram, N., & Bilgel, N. (2008). The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Social Psychiatry and Psychiatric Epidemiology*, ***43***(*8*), 667–672.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, ***10***(*9*), 1214–1221.

Bener, A., Dafeeah, E. E., Abou-Saleh, M. T., Bhugra, D., & Ventriglio, A. (2018). Schizophrenia and co-morbid obsessive-compulsive disorder: Clinical characteristics. *Asian Journal of Psychiatry*, ***37***, 80–84.

Bickel, W. K., Jarmolowicz, D. P., Mueller, E. T., Koffarnus, M. N., & Gatchalian, K. M. (2012). Excessive discounting of delayed reinforcers as a trans-disease process contributing to addiction and other disease-related vulnerabilities: emerging evidence. *Pharmacology & Therapeutics*, ***134***(*3*), 287–297.

Blier, P., & El Mansari, M. (2007). The importance of serotonin and noradrenaline in anxiety. *International Journal of Psychiatry in Clinical Practice*, ***11***(*Suppl 2*), 16–23.

Bloch, M. H., Landeros-Weisenberger, A., Kelmendi, B., Coric, V., Bracken, M. B., & Leckman, J. F. (2006). A systematic review: antipsychotic augmentation with treatment refractory obsessive-compulsive disorder. *Molecular Psychiatry*, ***11***(*7*), 622–632.

Bloch, M. H., Landeros-Weisenberger, A., Rosario, M. C., Pittenger, C., & Leckman, J. F. (2008). Meta-analysis of the symptom structure of obsessive-compulsive disorder. *American Journal of Psychiatry*, ***165***(*12*), 1532–1542.

Bloch, M. H., McGuire, J., Landeros-Weisenberger, A., Leckman, J. F., & Pittenger, C. (2010). Meta-analysis of the dose-response relationship of SSRI in obsessive-compulsive disorder. *Molecular Psychiatry*, ***15***(*8*), 850–855.

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry*, ***176***(*5*), 376–387.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, ***108***(*3*), 624.

Burchi, E., Makris, N., Lee, M. R., Pallanti, S., & Hollander, E. (2019). Compulsivity in Alcohol Use Disorder (and Obsessive Compulsive Disorder: Implications for Neuromodulation. *Frontiers in Behavioral Neuroscience*, ***13***, 70.

Calamari, J. E., Cohen, R. J., Rector, N. A., Szacun-Shimizu, K., Riemann, B. C., & Norberg, M. M. (2006). Dysfunctional belief-based obsessive-compulsive disorder subgroups. *Behaviour Research and Therapy*, ***44***(*9*), 1347–1360.

Carpenter, J., Sherman, M. T., Seth, A. K., Fleming, S. M., Lau, H., Kievit, R. A., Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, ***148***(*1*), 51–64.

Carpita, B., Muti, D., Petrucci, A., Romeo, F., Gesi, C., Marazziti, D., Carmassi, C., & Dell'Osso, L. (2019). Overlapping features between social anxiety and obsessive-compulsive spectrum in a clinical sample and in healthy controls: toward an integrative model. *CNS Spectrums*, 1–8.

Carrasco, M., Harbin, S. M., Nienhuis, J. K., Fitzgerald, K. D., Gehring, W. J., Hanna, G. L., Hong, C., Nienhuis, J. K., Harbin, S. M., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and unaffected siblings. *Neuroscience Letters*, ***30***(*1*), 39–46.

Carrasco, M., Hong, C., Nienhuis, J. K., Harbin, S. M., Fitzgerald, K. D., Gehring, W. J., & Hanna, G. L. (2013). Increased error-related brain activity in youth with obsessive-compulsive disorder and other anxiety disorders. *Neuroscience Letters*, ***541***, 214–218.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring

of performance. *Science*, ***280***(*5364*), 747–749.

Casey, B. J., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., & Ressler, K. J. (2013). DSM-5 and RDoC: progress in psychiatry research? *Nature Reviews Neuroscience*, ***14***(*11*), 810–814.

Cath, D. C., Ran, N., Smit, J. H., Van Balkom, A. J. L. M., & Comijs, H. C. (2008). Symptom overlap between autism spectrum disorder, generalized social anxiety disorder and obsessive-compulsive disorder in adults: a preliminary case-controlled study. *Psychopathology*, ***41***(*2*), 101–110.

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal theta overrides Pavlovian learning biases. *Journal of Neuroscience*, ***33***(*19*), 8541–8548.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, ***18***(*8*), 414–421.

Cavanagh, J. F., & Shackman, A. J. (2015). Frontal midline theta reflects anxiety and cognitive control: Meta-analytic evidence. *Journal of Physiology Paris*, ***109***(*1–3*), 3–15.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. B. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, ***49***(*2*), 220–238.

Clark, D. A. (2005). Focus on “cognition” in cognitive behavior therapy for OCD: Is it really necessary? In *Cognitive Behaviour Therapy*.

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology*, ***50***(*2*), 174–186.

Clayson, P. E., & Larson, M. J. (2011). Conflict adaptation and sequential trial effects: Support for the conflict monitoring theory. *Neuropsychologia*, ***49***(*7*), 1953–1961.

Coles, M. E., Frost, R. O., Heimberg, R. G., & Rhéaume, J. (2003). "Not just right experiences": perfectionism, obsessive–compulsive features and general psychopathology. *Behaviour Research and Therapy*, ***41***(*6*), 681–700.

Coles, M. E., Radomsky, A. S., & Horng, B. (2006). Exploring the boundaries of memory distrust from repeated checking: Increasing external validity and examining thresholds. *Behaviour Research and Therapy*, ***44***(*7*), 995–1006.

Coles, M. G. H., Scheffers, M. K., & Holroyd, C. B. (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological Psychology*, ***56***(*3*), 173–189.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, ***43***, 52–64.

Cougle, J. R., Salkovskis, P. M., & Wahl, K. (2007). Perception of memory ability and confidence in recollections in obsessive-compulsive checking. *Journal of*

Anxiety Disorders, ***21***(*1*), 118–130.

Crump, M. J. C., McDonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, ***8***(*3*), e57410.

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, ***125***(*6*), 777.

Culverhouse, R. C., Saccone, N. L., Horton, A. C., Ma, Y., Anstey, K. J., Banaschewski, T., Burmeister, M., Cohen-Woods, S., Etain, B., & Fisher, H. L. (2018). Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Molecular Psychiatry*, ***23***(*1*), 133–142.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, ***11***(*1*), 126.

Dalley, J. W., Everitt, B. J., & Robbins, T. W. (2011). Impulsivity, Compulsivity, and Top-Down Cognitive Control. *Neuron*, ***69***(*4*), 680–694.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, ***69***(*6*), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature*

Neuroscience, ***8***(*12*), 1704–1711.

de Bruijn, E. R. A., Hulstijn, W., Verkes, R. J., Ruigt, G. S. F., & Sabbe, B. G. C. (2004). Drug-induced stimulation and suppression of action monitoring in healthy volunteers. *Psychopharmacology*, ***177***(*1–2*), 151–160.

de Bruijn, E. R. A., Sabbe, B. G. C., Hulstijn, W., Ruigt, G. S. F., & Verkes, R. J. (2006). Effects of antipsychotic and antidepressant drugs on action monitoring in healthy volunteers. *Brain Research*, ***1105***(*1*), 122–129.

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, ***147***(*7*), 1043.

de Wit, S., Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, ***33***(*1*), 1.

de Wit, S., Standing, H. R., Devito, E. E., Robinson, O. J., Ridderinkhof, K. R., Robbins, T. W., & Sahakian, B. J. (2012). Reliance on habits at the expense of goal-directed control following dopamine precursor depletion. *Psychopharmacology*, ***219***(*2*), 621–631.

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control.

Journal of Neuroscience, ***32***(*35*), 12066–12075.

Debener, S. (2005). Trial-by-Trial Coupling of Concurrent Electroencephalogram and Functional Magnetic Resonance Imaging Identifies the Dynamics of Performance Monitoring. *Journal of Neuroscience*, ***25***(*50*), 11730–11737.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, ***5***(*5*), 303–305.

Del Casale, A., Sorice, S., Padovano, A., Simmaco, M., Ferracuti, S., Lamis, D. A., Rapinesi, C., Sani, G., Girardi, P., & Kotzalidis, G. D. (2019). Psychopharmacological Treatment of Obsessive-Compulsive Disorder (OCD). *Current Neuropharmacology*, ***17***(*8*), 710–736.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, ***134***(*1*), 9–21.

Delorme, C., Salvador, A., Valabregue, R., Roze, E., Palminteri, S., Vidailhet, M., de Wit, S., Robbins, T., Hartmann, A., & Worbe, Y. (2016). Enhanced habit formation in Gilles de la Tourette syndrome. *Brain*, ***139***(*2*), 605–615.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, ***22***(*1*), 1–18.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, ***80***(*2*), 312–325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015).

Model-based choices involve prospective neural activity. *Nature Neuroscience*, ***18***(*5*), 767–772.

Dreisbach, G., & Fischer, R. (2012). Conflicts as aversive signals. *Brain and Cognition*, ***78***(*2*), 94–98.

Ecker, W., & Engelkamp, J. (1995). Memory for actions in obsessive-compulsive disorder. *Behavioural and Cognitive Psychotherapy*, ***23***(*4*), 349–371.

Eddy, K. T., Dutra, L., Bradley, R., & Westen, D. (2004). A multidimensional meta-analysis of psychotherapy and pharmacotherapy for obsessive-compulsive disorder. *Clinical Psychology Review*, ***24***(*8*), 1011–1030.

Endrass, T., Klawohn, J., Schuster, F., & Kathmann, N. (2008). Overactive performance monitoring in obsessive-compulsive disorder: ERP evidence from correct and erroneous reactions. *Neuropsychologia*, ***46***(*7*), 1877–1887.

Endrass, T., Riesel, A., Kathmann, N., & Buhlmann, U. (2014). Performance monitoring in obsessive-compulsive disorder and social anxiety disorder. *Journal of Abnormal Psychology*, ***123***(*4*), 705–714.

Endrass, T., Schuermann, B., Kaufmann, C., Spielberg, R., Kniesche, R., & Kathmann, N. (2010). Performance monitoring and error significance in patients with obsessive-compulsive disorder. *Biological Psychology*, ***84***(*2*), 257–263.

Endrass, T., & Ullsperger, M. (2014). Specificity of performance monitoring changes in obsessive-compulsive disorder. *Neuroscience and Biobehavioral*

Reviews, ***46***, 124–138.

Eppinger, B., Walter, M., & Li, S.-C. C. (2017). Electrophysiological correlates reflect the integration of model-based and model-free decision information. *Cognitive, Affective, & Behavioral Neuroscience*, ***17***(*2*), 1–16.

Ercan, I., Ozkaya, G., Hafizoglu, S., Kirli, S., Akaya, C., & Yalcintas, E. (2015). Examining cut-off values for the state-trait anxiety inventory. *Revista Argentina de Clinica Psicologica*, ***24***(*2*), 143.

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

Ersche, K. D., Gillan, C. M., Simon Jones, P., Williams, G. B., Ward, L. H. E., Luijten, M., De Wit, S., Sahakian, B. J., Bullmore, E. T., & Robbins, T. W. (2016). Carrots and sticks fail to change behavior in cocaine addiction. *Science*, ***352***(*6292*), 1468–1471.

Evans, T. M., Bira, L., Gastelum, J. B., Weiss, L. T., & Vanderford, N. L. (2018). Evidence for a mental health crisis in graduate education. *Nature Biotechnology*, ***36***(*3*), 282.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, ***8***, 1481–1489.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1990). Effects of

errors in choice reaction tasks on the ERP under focused and divided attention. *Psychophysiological Brain Research*, ***1***, 192–195.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, ***78***(*6*), 447–455.

Falkenstein, M., Hoormann, J., & Hohnsbein, J. (2001). Changes of error-related ERPs with age. *Experimental Brain Research*, ***138***(*2*), 258–262.

Fieker, M., Moritz, S., Köther, U., & Jelinek, L. (2016). Emotion recognition in depression: An investigation of performance and response confidence in adult female patients with depression. *Psychiatry Research*, ***242***, 226–232.

Fineberg, N. A., Hengartner, M. P., Bergbaum, C. E., Gale, T. M., Gamma, A., Ajdacic-Gross, V., Rössler, W., & Angst, J. (2013). A prospective population-based cohort study of the prevalence, incidence and impact of obsessive-compulsive symptomatology. *International Journal of Psychiatry in Clinical Practice*, ***17***(*3*), 170–178.

Fineberg, N. A., Hengartner, M. P., Bergbaum, C., Gale, T., Rössler, W., & Angst, J. (2013). Lifetime comorbidity of obsessive-compulsive disorder and sub-threshold obsessive-compulsive symptomatology in the community: impact, prevalence, socio-demographic and clinical characteristics. *International Journal of Psychiatry in Clinical Practice*, ***17***(*3*), 188–196.

Fineberg, N. A., Potenza, M. N., Chamberlain, S. R., Berlin, H. A., Menzies, L.,

Bechara, A., Sahakian, B. J., Robbins, T. W., Bullmore, E. T., & Hollander, E. (2010). Probing compulsive and impulsive behaviors, from animal models to endophenotypes: A narrative review. *Neuropsychopharmacology*, ***35***, 591–604.

Fischer, A. G., Danielmeier, C., Villringer, A., Klein, T. A., & Ullsperger, M. (2016). Gender Influences on Brain Responses to Errors and Post-Error Adjustments. *Scientific Reports*, ***6***, 24435.

Fisher, P. L., & Wells, A. (2005a). Experimental modification of beliefs in obsessive-compulsive disorder: A test of the metacognitive model. *Behaviour Research and Therapy*, ***43***(*6*), 821–829.

Fisher, P. L., & Wells, A. (2005b). How effective are cognitive and behavioral treatments for obsessive-compulsive disorder? A clinical significance analysis. *Behaviour Research and Therapy*, ***43***(*12*), 1543–1558.

Fisher, P. L., & Wells, A. (2008). Metacognitive therapy for obsessive-compulsive disorder: A case series. *Journal of Behavior Therapy and Experimental Psychiatry*, ***39***(*2*), 117–132.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, ***34***(*10*), 906.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, ***124***(*1*), 91.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation,

biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1280–1286.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, ***8***, 443.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, ***137***(*10*), 2811–2822.

Foa, E. B., Amir, N., Gershuny, B., Molnar, C., & Kozak, M. J. (1997). Implicit and explicit memory in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, ***11***(*2*), 119–129.

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

Foa, E. B., & Kozak, M. J. (1995). DSM-IV field trial: Obsessive-compulsive disorder. *American Journal of Psychiatry*, ***152***(*1*), 90–96.

Foa, E. B., Yadin, E., & Lichner, T. K. (2012). *Exposure and response (ritual) prevention for obsessive compulsive disorder: Therapist guide*. Oxford University Press.

Foti, D., Kotov, R., Bromet, E., & Hajcak, G. (2012). Beyond the broken error-related negativity: Functional and diagnostic correlates of error processing in psychosis. *Biological Psychiatry*, ***71***(*10*), 864–872.

Fouche, J.-P., Du Plessis, S., Hattingh, C., Roos, A., Lochner, C., Soriano-Mas, C., Sato, J. R., Nakamae, T., Nishida, S., & Kwon, J. S. (2017). Cortical thickness in obsessive–compulsive disorder: Multisite mega-analysis of 780 brain scans from six centres. *The British Journal of Psychiatry*, ***210***(*1*), 67–74.

Fradkin, I., Ludwig, C., Eldar, E., & Huppert, J. D. (2020). Doubting what you already know: Uncertainty regarding state transitions is associated with obsessive compulsive symptoms. *PLOS Computational Biology*, ***16***(*2*), e1007634.

Fradkin, I., Strauss, A. Y., Pereg, M., & Huppert, J. D. (2018). Rigidly applied rules? Revisiting inflexibility in obsessive compulsive disorder using multilevel meta-analysis. *Clinical Psychological Science*, ***6***(*4*), 481–505.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, ***47***(*4*), 495–501.

Franken, I. H. A., van Strien, J. W., Franzek, E. J., & van de Wetering, B. J. (2007). Error-processing deficits in patients with cocaine dependence. *Biological Psychology*, ***75***(*1*), 45–51.

Friedel, E., Koch, S. P., Wendt, J., Heinz, A., Deserno, L., & Schlagenhauf, F. (2014). Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Frontiers in Human Neuroscience*, ***8***, 587.

Frith, C. D., & Done, D. J. (1988). Towards a neuropsychology of schizophrenia. *British Journal of Psychiatry*, ***153***(*4*), 437–443.

Fu, T. S.-T., Koutstaal, W., Fu, C. H. Y., Poon, L., & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, ***27***(*4*), 243–252.

Fu, T. S.-T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, ***43***(*2*), 699–704.

Fusar-Poli, P., Solmi, M., Brondino, N., Davies, C., Chae, C., Politi, P., Borgwardt, S., Lawrie, S. M., Parnas, J., & McGuire, P. (2019). Transdiagnostic psychiatry: a systematic review. *World Psychiatry*, ***18***(*2*), 192–207.

Garner, D. M., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: Psychometric features and clinical correlates. *Psychological Medicine*, ***12***(*4*), 871–878.

Gawęda, Ł., Moritz, S., & Kokoszka, A. (2012). Impaired discrimination between imagined and performed actions in schizophrenia. *Psychiatry Research*, ***195***(*1–2*), 1–8.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*, ***4***(*6*), 385–390.

Gehring, W. J., Himle, J., & Nisenson, L. G. (2000). Action-monitoring dysfunction in obsessive-compulsive disorder. *Psychological Science*, ***11***(*1*), 1–6.

Gillan, C. M., Apergis-Schoute, A. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2015). Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. *American Journal of Psychiatry*, ***172***(*3*), 284–293.

Gillan, C. M., & Daw, N. D. (2016). Taking psychiatry research online. *Neuron*, ***91***(*1*), 19–23.

Gillan, C. M., Fineberg, N. A., & Robbins, T. W. (2017). A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychological Medicine*, ***47***(*9*), 1528–1548.

Gillan, C. M., Kalanthroff, E., Evans, M., Weingarden, H. M., Jacoby, R. J., Gershkovich, M., Snorrason, I., Campeas, R., Cervoni, C., Crimarco, N. C., Sokol, Y., Garnaat, S. L., McLaughlin, N. C. R., Phelps, E. A., Pinto, A., Boisseau, C. L., Wilhelm, S., Daw, N. D., & Simpson, H. B. (2019). Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry*, ***77***(*1*), 77–85.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *ELife*, ***5***, e11305.

Gillan, C. M., Morein-Zamir, S., Kaser, M., Fineberg, N. A., Sule, A., Sahakian, B. J., Cardinal, R. N., & Robbins, T. W. (2014). Counterfactual processing of economic action-outcome alternatives in obsessive-compulsive disorder.

Biological Psychiatry, ***75***(*8*), 639–646.

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological Psychiatry*, ***75***(*8*), 631–638.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective and Behavioral Neuroscience*, ***15***(*3*), 523–536.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & De Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, ***168***(*7*), 718–726.

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, ***369***(*1655*), 20130475.

Gillan, C. M., & Sahakian, B. J. (2015). Which is the driver, the obsessions or the compulsions, in OCD? *Neuropsychopharmacology*, ***40***(*1*), 247–248.

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, ***18***, 34–42.

Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry*, ***160***(*4*),

636–645.

Graybiel, A. M., & Rauch, S. L. (2000). Toward a neurobiology of obsessive-compulsive disorder. *Neuron*, ***28***(*2*), 343–347.

Gremel, C. M., & Costa, R. M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature Communications*, ***4***(*1*), 1–12.

Gründler, T. O. J., Cavanagh, J. F., Figueroa, C. M., Frank, M. J., & Allen, J. J. B. (2009). Task-related dissociation in ERN amplitude as a function of obsessive-compulsive symptoms. *Neuropsychologia*, ***47***(*8–9*), 1978–1987.

Gruner, P., Anticevic, A., Lee, D., & Pittenger, C. (2016). Arbitration between action strategies in obsessive-compulsive disorder. *The Neuroscientist*, ***22***(*2*), 188–198.

Grützmann, R., Endrass, T., Kaufmann, C., Allen, E., Eichele, T., & Kathmann, N. (2016). Presupplementary Motor Area Contributes to Altered Error Monitoring in Obsessive-Compulsive Disorder. *Biological Psychiatry*, ***80***(*7*), 562–571.

Grützmann, R., Riesel, A., Klawohn, J., Kathmann, N., & Endrass, T. (2014). Complementary modulation of N2 and CRN by conflict frequency. *Psychophysiology*, ***51***(*8*), 761–772.

Hajcak, G. (2006). Error-related brain activity in pediatric obsessive-compulsive disorder and trichotillomania before and after cognitive-behavioral therapy. In

Dissertation Abstracts International: Section B: The Sciences and Engineering. University of Delaware.

Hajcak, G. (2012). What We've Learned From Mistakes: Insights From Error-Related Brain Activity. *Current Directions in Psychological Science*, ***21***(*2*), 101–106.

Hajcak, G., & Foti, D. (2008). Errors are aversive: Defensive motivation and the error-related negativity. *Psychological Science*, ***19***(*2*), 103–108.

Hajcak, G., Franklin, M. E., Foa, E. B., & Simons, R. F. (2008). Increased error-related brain activity in pediatric obsessive-compulsive disorder before and after treatment. *American Journal of Psychiatry*, ***165***(*1*), 116–123.

Hajcak, G., McDonald, N., & Simons, R. F. (2003a). Anxiety and error-related brain activity. *Biological Psychology*, ***64***(*1–2*), 77–90.

Hajcak, G., McDonald, N., & Simons, R. F. (2003b). To err is autonomic: Error-related brain potentials, ANS activity, and post-error compensatory behavior. *Psychophysiology*, ***40***(*6*), 895–903.

Hajcak, G., McDonald, N., & Simons, R. F. (2004). Error-related psychophysiology and negative affect. *Brain and Cognition*, ***56***(*2*), 189–197.

Hajcak, G., & Simons, R. F. (2002). Error-related brain activity in obsessive–compulsive undergraduates. *Psychiatry Research*, ***110***(*1*), 63–72.

Hancock, J. A. (1996). "Depressive Realism" assessed via Confidence in Decision-making. *Cognitive Neuropsychiatry*, ***1***(*3*), 213–220.

Hanna, G. L., Liu, Y., Isaacs, Y. E., Ayoub, A. M., Brosius, A., Salander, Z., Arnold, P. D., & Gehring, W. J. (2018). Error-related brain activity in adolescents with obsessive-compulsive disorder and major depressive disorder. *Depression and Anxiety*, ***35***(*8*), 752–760.

Hanna, G. L., Liu, Y., Isaacs, Y. E., Ayoub, A. M., Torres, J. J., O'Hara, N. B., & Gehring, W. J. (2016). Withdrawn/Depressed Behaviors and Error-Related Brain Activity in Youth With Obsessive-Compulsive Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, ***55***(*10*), 906–913.

Haro, J. M., Ayuso-Mateos, J. L., Bitter, I., Demotes-Mainard, J., Leboyer, M., Lewis, S. W., Linszen, D., Maj, M., Mcdaid, D., & Meyer-Lindenberg, A. (2014). ROAMER: roadmap for mental health research in Europe. *International Journal of Methods in Psychiatric Research*, ***23***(*S1*), 1–14.

Hauser, T. U., Allen, M., Purg, N., Moutoussis, M., Rees, G., & Dolan, R. J. (2017). Noradrenaline blockade specifically enhances metacognitive performance. *Elife*, ***6***, e24901.

Hauser, T. U., Allen, M., Rees, G., Dolan, R. J., Bullmore, E. T., Goodyer, I., Fonagy, P., Jones, P., Fearon, P., Prabhu, G., Moutoussis, M., St Clair, M., Cleridou, K., Dadabhoy, H., Granville, S., Harding, E., Hopkins, A., Isaacs, D., King, J., … Pantaleone, S. (2017). Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Scientific Reports*, ***7***(*1*), 6614.

Hauser, T. U., Iannaccone, R., Dolan, R. J., Ball, J., Hättenschwiler, J.,

Drechsler, R., Rufer, M., Brandeis, D., Walitza, S., & Brem, S. (2017). Increased fronto-striatal reward prediction errors moderate decision making in obsessive-compulsive disorder. *Psychological Medicine*.

Hemmings, S. M. J., & Stein, D. J. (2006). The current status of association studies in obsessive-compulsive disorder. *Psychiatric Clinics*, ***29***(*2*), 411–444.

Henderson, H., Schwartz, C., Mundy, P., Burnette, C., Sutton, S., Zahka, N., & Pradella, A. (2006). Response monitoring, the error-related negativity, and differences in social behavior in autism. *Brain and Cognition*, ***61***(*1*), 96–109.

Hermans, D., Engelen, U., Grouwels, L., Joos, E., Lemmens, J., & Pieters, G. (2008). Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behaviour Research and Therapy*, ***46***(*1*), 98–113.

Hezel, D. M., & Simpson, H. B. (2019). Exposure and response prevention for obsessive-compulsive disorder: A review and new directions. *Indian Journal of Psychiatry*, ***61***(*Suppl 1*), S85.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, ***109***(*4*), 679–709.

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General*, ***134***(*2*), 163.

Hoven, M., Lebreton, M. M. M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: an overview and future perspectives. *Translational Psychiatry*, ***9***(*1*), 1–18.

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, ***19***(*3*), 404–413.

Hyman, S. E. (2007). Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience*, ***8***(*9*), 725–732.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D., Quinn, K., Sanislow, C., Wang, P., Insel, et al., Cuthbert, & Garvey. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, 748–751.

Johannes, S., Wieringa, B. M., Nager, W., Dengler, R., & Münte, T. F. (2001). Oxazepam alters action monitoring. *Psychopharmacology*, ***155***(*1*), 100–106.

Johannes, S., Wieringa, B. M., Nager, W., Rada, D., Dengler, R., Emrich, H. M., Münte, T. F., & Dietrich, D. E. (2001). Discrepant target detection and action monitoring in obsessive–compulsive disorder. *Psychiatry Research: Neuroimaging*, ***108***(*2*), 101–110.

Joyce, D. W., Averbeck, B. B., Frith, C. D., & Shergill, S. S. (2013). Examining belief and confidence in schizophrenia. *Psychological Medicine*, ***43***(*11*), 2327–2338.

Judah, M. R., Grant, D. M., Frosio, K. E., White, E. J., Taylor, D. L., & Mills, A. C. (2016). Electrocortical evidence of enhanced performance monitoring in social anxiety. *Behavior Therapy*, ***47***(*2*), 274–285.

Kaczkurkin, A. N. (2013). The effect of manipulating task difficulty on error-related negativity in individuals with obsessive-compulsive symptoms. *Biological Psychology*, ***93***(*1*), 122–131.

Kelly, S. P., & O'Connell, R. G. (2015). The neural processes underlying perceptual decision making in humans: Recent progress and future directions. *Journal of Physiology Paris*, ***109***(*1–3*), 27–37.

Kessel, E. M., Meyer, A., Hajcak, G., Dougherty, L. R., Torpey-Newman, D. C., Carlson, G. A., & Klein, D. N. (2016). Transdiagnostic factors and pathways to multifinality: The error-related negativity predicts whether preschool irritability is associated with internalizing versus externalizing symptoms at age 9. *Development and Psychopathology*, ***28***(*4pt1*), 913–926.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, ***62***(*6*), 593–602.

Khdour, H. Y., Abushalbaq, O. M., Mughrabi, I. T., Imam, A. F., Gluck, M. A., Herzallah, M. M., & Moustafa, A. A. (2016). Generalized anxiety disorder and social anxiety disorder, but not panic anxiety disorder, are associated with higher sensitivity to learning from negative feedback: behavioral and

computational investigation. *Frontiers in Integrative Neuroscience*, ***10***, 20.

Kircher, T. T. J., Koch, K., Stottmeister, F., & Durst, V. (2007). Metacognition and reflexivity in patients with schizophrenia. *Psychopathology*, ***40***(*4*), 254–260.

Klawohn, J., Riesel, A., Grützmann, R., Kathmann, N., & Endrass, T. (2014). Performance monitoring in obsessive-compulsive disorder: A temporo-spatial principal component analysis. *Cognitive, Affective and Behavioral Neuroscience*, ***14***(*3*), 983–995.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., & Clark, L. A. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, ***126***(*4*), 454.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, ***56***(*10*), 921–926.

Krueger, R. F., Caspi, A., Moffitt, T. E., & Silva, P. A. (1998). The structure and stability of common mental disorders (DSM-III-R): a longitudinal-epidemiological study. *Journal of Abnormal Psychology*, ***107***(*2*), 216.

Kujawa, A., Weinberg, A., Bunford, N., Fitzgerald, K. D., Hanna, G. L., Monk, C. S., Kennedy, A. E., Klumpp, H., Hajcak, G., & Phan, K. L. (2016). Error-related brain activity in youth and young adults before and after treatment for generalized or social anxiety disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, ***71***, 162–168.

Ladouceur, C. D., Tan, P. Z., Sharma, V., Bylsma, L. M., Silk, J. S., Siegle, G. J., Forbes, E. E., McMakin, D. L., Dahl, R. E., Kendall, P. C., Mannarino, A., & Ryan, N. D. (2018). Error-related brain activity in pediatric anxiety disorders remains elevated following individual therapy: a randomized clinical trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, ***59***(*11*), 1152–1161.

Lahat, A., Lamm, C., Chronis-Tuscano, A., Pine, D. S., Henderson, H. A., & Fox, N. A. (2014). Early behavioral inhibition and increased error monitoring predict later social phobia symptoms in childhood. *Journal of the American Academy of Child & Adolescent Psychiatry*, ***53***(*4*), 447–455.

Larson, M. J., Clayson, P. E., Keith, C. M., Hunt, I. J., Hedges, D. W., Nielsen, B. L., & Call, V. R. A. (2016). Cognitive control adjustments in healthy older and younger adults: Conflict adaptation, the error-related negativity (ERN), and evidence of generalized decline with age. *Biological Psychology*, ***115***, 50–63.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., & Krakow, K. (2003). EEG-correlated fMRI of human alpha activity. *Neuroimage*, ***19***(*4*), 1463–1476.

Lazarov, A., Cohen, T., Liberman, N., & Dar, R. (2015). Can doubt attenuate access to internal states? Implications for obsessive-compulsive disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, ***49***, 150–156.

Lazarov, A., Liberman, N., Hermesh, H., & Dar, R. (2014). Seeking proxies for internal states in obsessive-compulsive disorder. *Journal of Abnormal*

Psychology, ***123***(*4*), 695–704.

Leckman, J. F., Pauls, D. L., Zhang, H., Rosario-Campos, M. C., Katsovich, L., Kidd, K. K., Pakstis, A. J., Alsobrook, J. P., Robertson, M. M., & McMahon, W. M. (2003). Obsessive-compulsive symptom dimensions in affected sibling pairs diagnosed with Gilles de la Tourette syndrome. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, ***116***(*1*), 60–68.

Leckman, J. F., Rauch, S. L., & Mataix-Cols, D. (2007). Symptom dimensions in obsessive-compulsive disorder: implications for the DSM-V. *CNS Spectrums*, ***12***(*5*), 376–387.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, ***81***(*3*), 687–699.

Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmapsychiatry*, ***22***, 141–173.

Lipszyc, J., & Schachar, R. (2010). Inhibitory control and psychopathology: a meta-analysis of studies using the stop signal task. *Journal of the International Neuropsychological Society*, ***16***(*6*), 1064–1076.

Liu, Y., Hanna, G. L., Carrasco, M., Gehring, W. J., & Fitzgerald, K. D. (2014). Altered relationship between electrophysiological response to errors and gray matter volumes in an extended network for error-processing in pediatric obsessive-compulsive disorder. *Human Brain Mapping*, ***35***(*4*), 1143–1153.

Lochner, C., & Stein, D. J. (2003). Heterogeneity of obsessive-compulsive disorder: a literature review. *Harvard Review of Psychiatry*, ***11***(*3*), 113–132.

Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. (2019). *Measuring habit formation through goal-directed response switching*.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, ***4***(*1*), 84.

Macdonald, P. A., Antony, M. M., Macleod, C. M., & Richter, M. A. (1997). Memory and confidence in memory judgments among individuals with obsessive compulsive disorder and non-clinical controls. *Behaviour Research and Therapy*, ***35***(*6*), 497–505.

Makeig, S., & Onton, J. (2012). ERP Features and EEG Dynamics: An ICA Perspective. In *The Oxford Handbook of Event-Related Potential Components*.

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatry Research*, ***38***(*2*), 143–162.

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, ***78***(*2–3*), 293–296.

Mataix-Cols, D., do Rosario-Campos, M. C., & Leckman, J. F. (2005). A multidimensional model of obsessive-compulsive disorder. *American Journal of Psychiatry*, ***162***(*2*), 228–238.

Mataix-Cols, D., Marks, I. M., Greist, J. H., Kobak, K. A., & Baer, L. (2002). Obsessive-compulsive symptom dimensions as predictors of compliance with

and response to behaviour therapy: results from a controlled trial. *Psychotherapy and Psychosomatics*, ***71***(*5*), 255–262.

Mataix-Cols, D., Rauch, S. L., Manzo, P. A., Jenike, M. A., & Baer, L. (1999). Use of factor-analyzed symptom dimensions to predict outcome with serotonin reuptake inhibitors and placebo in the treatment of obsessive-compulsive disorder. *American Journal of Psychiatry*, ***156***(*9*), 1409–1416.

Mathews, C. A., Perez, V. B., Delucchi, K. L., & Mathalon, D. H. (2012). Error-related negativity in individuals with obsessive–compulsive symptoms: toward an understanding of hoarding behaviors. *Biological Psychology*, ***89***(*2*), 487–494.

Mathews, C. A., Perez, V. B., Roach, B. J., Fekri, S., Vigil, O., Kupferman, E., & Mathalon, D. H. (2016). Error-related brain activity dissociates hoarding disorder from obsessive-compulsive disorder. *Psychological Medicine*, ***46***(*2*), 367–379.

Matsunaga, H., Kiriike, N., Matsui, T., Oya, K., Iwasaki, Y., Koshimune, K., Miyata, A., & Stein, D. J. (2002). Obsessive-compulsive disorder with poor insight. *Comprehensive Psychiatry*, ***43***(*2*), 150–157.

Matthews, G., & Wells, A. (2008). Rumination, Depression, and Metacognition: The S-REF Model. In *Depressive Rumination: Nature, Theory and Treatment*.

McGrath, J. (1991). Ordering thoughts on thought disorder. *British Journal of Psychiatry*, ***158***(*3*), 307–316.

McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally

dissociable influences on learning rate in a dynamic environment. *Neuron*, ***84***(*Figure 1*), 870–881.

McHugh, P. R. (2005). Striving for coherence: psychiatry's efforts over classification. *Jama*, ***293***(*20*), 2526–2528.

McKay, D., Abramowitz, J. S., Calamari, J. E., Kyrios, M., Radomsky, A., Sookman, D., Taylor, S., & Wilhelm, S. (2004). A critical evaluation of obsessive–compulsive disorder subtypes: symptoms versus mechanisms. *Clinical Psychology Review*, ***24***(*3*), 283–313.

McNally, R. J., & Kohlbeck, P. A. (1993). Reality monitoring in obsessive-compulsive disorder. *Behaviour Research and Therapy*, ***31***(*3*), 249–253.

Meyer, A., Hajcak, G., Torpey-Newman, D. C., Kujawa, A., & Klein, D. N. (2015). Enhanced error-related brain activity in children predicts the onset of anxiety disorders between the ages of 6 and 9. *Journal of Abnormal Psychology*, ***124***(*2*), 266.

Meyer, A., & Klein, D. N. (2018). Examining the relationships between error-related brain activity (the ERN) and anxiety disorders versus externalizing disorders in young children: Focusing on cognitive control, fear, and shyness. *Comprehensive Psychiatry*, ***87***, 112–119.

Meyer, A., Lerner, M. D., De Los Reyes, A., Laird, R. D., & Hajcak, G. (2017). Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology*, ***54***(*1*), 114–122.

Meyer, V. (1966). Modification of expectations in cases with obsessional rituals. *Behaviour Research and Therapy*, ***4***(*4*), 273–280.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, ***88***(*1*), 78–92.

Miltner, W. H. R., Lemke, U., Weiss, T., Holroyd, C., Scheffers, M. K., & Coles, M. G. H. (2003). Implementation of error-processing in the human anterior cingulate cortex: A source analysis of the magnetic equivalent of the error-related negativity. *Biological Psychology*, ***64***(*1–2*), 157–166.

Min, B. K., Kim, S. J., Park, J. Y., & Park, H. J. (2011). Prestimulus top-down reflection of obsessive-compulsive disorder in EEG frontal theta and occipital alpha oscillations. *Neuroscience Letters*, ***496***(*3*), 181–185.

Minzenberg, M. J., Gomes, G. C., Yoon, J. H., Swaab, T. Y., & Carter, C. S. (2014). Disrupted action monitoring in recent-onset psychosis patients with schizophrenia and bipolar disorder. *Psychiatry Research - Neuroimaging*, ***221***(*1*), 114–121.

Moritz, S., Jacobsen, D., Willenborg, B., Jelinek, L., & Fricke, S. (2006). A check on the memory deficit hypothesis of obsessive–compulsive checking. *European Archives of Psychiatry and Clinical Neuroscience*, ***256***(*2*), 82–86.

Moritz, S., & Jaeger, A. (2018). Decreased memory confidence in obsessive–compulsive disorder for scenarios high and low on responsibility: is low still too high? *European Archives of Psychiatry and Clinical Neuroscience*, ***268***(*3*), 291–299.

Moritz, S., Kloss, M., von Eckstaedt, F. V., & Jelinek, L. (2009). Comparable performance of patients with obsessive-compulsive disorder (OCD) and healthy controls for verbal and nonverbal memory accuracy and confidence:. Time to forget the forgetfulness hypothesis of OCD? *Psychiatry Research*, ***166***(*2–3*), 247–253.

Moritz, S., Ramdani, N., Klass, H., Andreou, C., Jungclaussen, D., Eifler, S., Englisch, S., Schirmbeck, F., & Zink, M. (2014). Overconfidence in incorrect perceptual judgments in patients with schizophrenia. *Schizophrenia Research: Cognition*, ***1***(*4*), 165–170.

Moritz, S., Rietschel, L., Jelinek, L., & Bäuml, K.-H. T. (2011). Are patients with obsessive–compulsive disorder generally more doubtful? Doubt is warranted! *Psychiatry Research*, ***189***(*2*), 265–269.

Moritz, S., Ruhe, C., Jelinek, L., & Naber, D. (2009). No deficits in nonverbal memory, metamemory and internal as well as external source memory in obsessive-compulsive disorder (OCD). *Behaviour Research and Therapy*, ***47***(*4*), 308–315.

Moritz, S., Woodward, T. S., & Ruff, C. C. (2003). Source monitoring and memory confidence in schizophrenia. *Psychological Medicine*, ***33***(*1*), 131–139.

Moritz, S., Woodward, T. S., Whitman, J. C., & Cuttler, C. (2005). Confidence in errors as a possible basis for delusions in schizophrenia. *The Journal of Nervous and Mental Disease*, ***193***(*1*), 9–16.

Morris, R. W., Quail, S., Griffiths, K. R., Green, M. J., & Balleine, B. W. (2015).

Corticostriatal control of goal-directed action is impaired in schizophrenia. *Biological Psychiatry*, ***77***(*2*), 187–195.

Morris, S. E., Yee, C. M., & Nuechterlein, K. H. (2006). Electrophysiological analysis of error monitoring in schizophrenia. *Journal of Abnormal Psychology*, ***115***(*2*), 239.

Morsel, A. M., Morrens, M., Temmerman, A., Sabbe, B., & de Bruijn, E. R. (2014). Electrophysiological (EEG) evidence for reduced performance monitoring in euthymic bipolar disorder. *Bipolar Disorders*, ***16***(*8*), 820–829.

Moser, J. S., Hajcak, G., & Simons, R. F. (2005). The effects of fear on performance monitoring and attentional allocation. *Psychophysiology*, ***42***(*3*), 261–268.

Moser, J. S., Moran, T. P., & Jendrusina, A. A. (2012). Parsing relationships between dimensions of anxiety and action monitoring brain potentials in female undergraduates. *Psychophysiology*, ***49***(*1*), 2–10.

Moser, J. S., Moran, T. P., Schroder, H. S., Donnellan, M. B., & Yeung, N. (2013). On the relationship between anxiety and error monitoring: a meta-analysis and conceptual framework. *Frontiers in Human Neuroscience*, ***7***, 466.

Mullen, T. (2012). CleanLine EEGLAB plugin. In *San Diego, CA: Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC)*.

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Cauwenberghs, G., & Jung, T. P. (2013). Real-time modeling and 3D visualization of source

dynamics and connectivity using wearable EEG. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2184–2187.

Nassar, M. R., Bruckner, R., Gold, J. I., Li, S. C., Heekeren, H. R., & Eppinger, B. (2016). Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nature Communications*, ***7***, 1–13.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, ***30***(*37*), 12366–12378.

Nieuwenhuis, S., Nielen, M. M., Mol, N., Hajcak, G., & Veltman, D. J. (2005). Performance monitoring in obsessive-compulsive disorder. *Psychiatry Research*, ***134***(*2*), 111–122.

Nigbur, R., Ivanova, G., & Stürmer, B. (2011). Theta power as a marker for cognitive interference. *Clinical Neurophysiology*, ***122***(*11*), 2185–2194.

O'Doherty, J. P. (2011). Contributions of the ventromedial prefrontal cortex to goal-directed action selection. *Annals of the New York Academy of Sciences*, ***1239***(*1*), 118–129.

Olvet, D. M., & Hajcak, G. (2008). The error-related negativity (ERN) and psychopathology: Toward an endophenotype. *Clinical Psychology Review*, ***28***(*8*), 1343–1354.

Olvet, D. M., & Hajcak, G. (2009a). The error-related negativity (ERN) and

psychopathology: Toward an Endophenotype. *Clin Psychol Rev.*, ***28***(*8*), 1343–1354.

Olvet, D. M., & Hajcak, G. (2009b). The stability of error-related brain activity with increasing trials. *Psychophysiology*, ***56***(*2*), 215–233.

Organization, W. H. (1993). *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research* (Vol. 2). World Health Organization.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2014). Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, ***27***(*2*), 319–333.

Overbeek, T., Schruers, K., Vermetten, E., & Griez, E. (2002). Comorbidity of obsessive-compulsive disorder and depression: prevalence, symptom severity, and treatment effect. *The Journal of Clinical Psychiatry*.

Pallanti, S., Grassi, G., Cantisani, A., Sarrecchia, E., & Pellegrini, M. (2011). Obsessive–compulsive disorder comorbidity: clinical assessment and therapeutic implications. *Frontiers in Psychiatry*, ***2***, 70.

Pasion, R., & Barbosa, F. (2019). ERN as a transdiagnostic marker of the internalizing-externalizing spectrum: A dissociable meta-analytic effect. *Neuroscience and Biobehavioral Reviews*, ***103***, 133–149.

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

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). Incentives boost

model-based control across a range of severity on several psychiatric constructs. *Biological Psychiatry*, ***85***(*5*), 425–433.

Perera, M. P. N., Bailey, N. W., Herring, S. E., & Fitzgerald, P. B. (2019). Electrophysiology of obsessive compulsive disorder: a systematic review of the electroencephalographic literature. *Journal of Anxiety Disorders*, ***62***, 1–14.

Perugi, G., Akiskal, H. S., Pfanner, C., Presta, S., Gemignani, A., Milanfranchi, A., Lensi, P., Ravagli, S., & Cassano, G. B. (1997). The clinical impact of bipolar and unipolar affective comorbidity on obsessive–compulsive disorder. *Journal of Affective Disorders*, ***46***(*1*), 15–23.

Perugi, G., Toni, C., Frare, F., Travierso, M. C., Hantouche, E., & Akiskal, H. S. (2002). Obsessive-compulsive-bipolar comorbidity: a systematic exploration of clinical features and treatment outcome. *The Journal of Clinical Psychiatry*, ***63***(*12*), 1129–1134.

Pesonen, M., Hämäläinen, H., & Krause, C. M. (2007). Brain oscillatory 4–30 Hz responses during a visual n-back memory task with varying memory load. *Brain Research*, ***1138***, 171–177.

Phillips, K. A., Stein, D. J., Rauch, S. L., Hollander, E., Fallon, B. A., Barsky, A., Fineberg, N., Mataix-Cols, D., Ferrão, Y. A., Saxena, S., Wilhelm, S., Kelly, M. M., Clark, L. A., Pinto, A., Bienvenu, O. J., Farrow, J., & Leckman, J. (2010). Should an obsessive-compulsive spectrum grouping of disorders be included in DSM-V? *Depression and Anxiety*, ***27***(*6*), 528–555.

Pigott, T. A., & Seay, S. M. (1999). A review of the efficacy of selective serotonin

reuptake inhibitors in obsessive-compulsive disorder. *The Journal of Clinical Psychiatry*, ***60***(*2*), 101–106.

Pinto, A., Mancebo, M. C., Eisen, J. L., Pagano, M. E., & Rasmussen, S. A. (2006). The Brown Longitudinal Obsessive Compulsive Study: Clinical features and symptoms of the sample at intake. *Journal of Clinical Psychiatry*, ***67***(*5*), 703.

Pitman, R. K. (1987). A cybernetic model of obsessive-compulsive psychopathology. *Comprehensive Psychiatry*, ***28***(*4*), 334–343.

Polich, J., & Margala, C. (1997). P300 and probability: comparison of oddball and single-stimulus paradigms. *International Journal of Psychophysiology*, ***25***(*2*), 169–176.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, ***19***(*3*), 366.

Rabinak, C. A., Holman, A., Angstadt, M., Kennedy, A. E., Hajcak, G., & Phan, K. L. (2013). Neural response to errors in combat-exposed returning veterans with and without post-traumatic stress disorder: A preliminary event-related potential study. *Psychiatry Research: Neuroimaging*, ***213***(*1*), 71–78.

Rachman, S. (1997). A cognitive theory of obsessions. *Behaviour Research and Therapy*, 209–222.

Radomsky, A. S., & Alcolado, G. M. (2010). Don't even think about checking: Mental checking causes memory distrust. *Journal of Behavior Therapy and*

Experimental Psychiatry, ***41***(*4*), 345–351.

Radomsky, A. S., Gilchrist, P. T., & Dussault, D. (2006). Repeated checking really does cause memory distrust. *Behaviour Research and Therapy*, ***44***(*2*), 305–316.

Rapinesi, C., Kotzalidis, G. D., Ferracuti, S., Sani, G., Girardi, P., & Del Casale, A. (2019). Brain stimulation in obsessive-compulsive disorder (OCD): a systematic review. *Current Neuropharmacology*, ***17***(*8*), 787–807.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and Time. *Cognitive Psychology*, ***41***(*1*), 1–48.

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

Riba, J., Rodríguez-Fornells, A., Münte, T. F., & Barbanoj, M. J. (2005). A neurophysiological study of the detrimental effects of alprazolam on human action monitoring. *Cognitive Brain Research*, ***25***(*2*), 554–565.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, ***306***(*5695*), 443–447.

Riesel, A. (2019). The erring brain: Error-related negativity as an endophenotype for OCD—A review and meta-analysis. *Psychophysiology*, ***56***(*4*), e13348.

Riesel, A., Endrass, T., Auerbach, L. A. ntoni., & Kathmann, N. (2015). Overactive Performance Monitoring as an Endophenotype for Obsessive-Compulsive Disorder: Evidence From a Treatment Study. *The American Journal of Psychiatry*, ***172***(*7*), 665–673.

Riesel, A., Endrass, T., Kaufmann, C., & Kathmann, N. (2011). Overactive error-related brain activity as a candidate endophenotype for obsessive-compulsive disorder: evidence from unaffected first-degree relatives. *American Journal of Psychiatry*, ***168***(*3*), 317–324.

Riesel, A., Goldhahn, S., & Kathmann, N. (2017). Hyperactive performance monitoring as a transdiagnostic marker: Results from health anxiety in comparison to obsessive–compulsive disorder. *Neuropsychologia*, ***96***, 1–8.

Riesel, A., Kathmann, N., & Endrass, T. (2014). Overactive performance monitoring in obsessive–compulsive disorder is independent of symptom expression. *European Archives of Psychiatry and Clinical Neuroscience*, ***264***(*8*), 707–717.

Riesel, A., Klawohn, J., Grützmann, R., Kaufmann, C., Heinzel, S., Bey, K., Lennertz, L., Wagner, M., & Kathmann, N. (2019). Error-related brain activity as a transdiagnostic endophenotype for obsessive-compulsive disorder, anxiety and substance use disorder. *Psychological Medicine*, ***49***(*7*), 1207–1217.

Riesel, A., Weinberg, A., Moran, T., & Hajcak, G. (2013). Time course of error-potentiated startle and its relationship to error-related brain activity. *Journal of Psychophysiology*, ***27***, 51–59.

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, ***16***(*1*), 81–91.

Rotge, J.-Y., Guehl, D., Dilharreguy, B., Tignol, J., Bioulac, B., Allard, M., Burbaud, P., & Aouizerate, B. (2009). Meta-analysis of brain volume changes in obsessive-compulsive disorder. *Biological Psychiatry*, ***65***(*1*), 75–83.

Rotge, J.-Y., Langbour, N., Guehl, D., Bioulac, B., Jaafari, N., Allard, M., Aouizerate, B., & Burbaud, P. (2010). Gray matter alterations in obsessive–compulsive disorder: an anatomic likelihood estimation meta-analysis. *Neuropsychopharmacology*, ***35***(*3*), 686–691.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, ***84***(*6*), 443–451.

Ruchsow, M., Reuter, K., Hermle, L., Ebert, D., Kiefer, M., & Falkenstein, M. (2007). Executive control in obsessive-compulsive disorder: event-related potentials in a Go/Nogo task. *Journal of Neural Transmission*, ***114***(*12*), 1595–1601.

Ruscio, A. M., Stein, D. J., Chiu, W. T., & Kessler, R. C. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Molecular Psychiatry*, ***15***(*1*), 53.

Ryan, C. (2001). *Ego-syntonic Obsessions*. University of Leicester.

Sakiris, N., & Berle, D. (2019). A systematic review and meta-analysis of the unified protocol as a transdiagnostic emotion regulation based intervention. *Clinical Psychology Review*, 101751.

Salkovskis, P. M. (1985). Obsessional-compulsive problems: A cognitive-behavioural analysis. *Behaviour Research and Therapy*, ***23***(*5*), 571–683.

Salkovskis, P. M., & McGuire, J. (2003). Cognitive-behavioural theory of OCD. *Obsessive Compulsive Disorder: Theory, Research and Treatment*, 59–78.

Salkovskis, P. M., Richards, H. C., & Forrester, E. (1995). The relationship between obsessional problems and intrusive thoughts. *Behavioural and Cognitive Psychotherapy*, ***23***(*3*), 281–299.

Sambrook, T. D., Hardwick, B., Wills, A. J., & Goslin, J. (2018). Model-free and model-based reward prediction errors in EEG. *NeuroImage*, ***178***, 162–171.

Sanders, J., Whitty, P., Murray, D., & Devitt, P. (2006). Delusions or obsessions: the same only different? *Psychopathology*, ***39***(*1*), 45–48.

Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction*, ***88***(*6*), 791–804.

Sauseng, P., Griesmayr, B., Freunberger, R., & Klimesch, W. (2010). Control mechanisms in working memory: a possible function of EEG theta oscillations. *Neuroscience & Biobehavioral Reviews*, ***34***(*7*), 1015–1022.

Saxena, S., Brody, A. L., Schwartz, J. M., & Baxter, L. R. (1998). Neuroimaging and frontal-subcortical circuitry in obsessive-compulsive disorder. *The British Journal of Psychiatry*, ***173***(*S35*), 26–37.

Schrijvers, D., De Bruijn, E. R. A., Maas, Y. J., Vancoillie, P., Hulstijn, W., & Sabbe, B. G. C. (2009). Action monitoring and depressive symptom reduction in major depressive disorder. *International Journal of Psychophysiology*, ***71***(*3*), 218–224.

Seow, T. X. F., Benoit, E., Dempsey, C., Jennings, M., Maxwell, A., McDonough, M., & Gillan, C. M. (2019). Null results from a dimensional study of error-related negativity (ERN) and self-reported psychiatric symptoms. *BioRxiv*, 732594.

Seow, T. X. F., & Gillan, C. M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports*, ***10***(*1*), 2883.

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, ***15***(*2*), e1006803.

Shahnazian, D., Ribas-Fernandes, J. J. F., & Holroyd, C. B. (2019). Electrophysiological correlates of state transition prediction errors. *BioRxiv*, 544551.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science*, ***1***(*2*), 213–220.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, ***59***(*Suppl 20*), 22–33.

Sheppard, D. M., Bradshaw, J. L., Purcell, R., & Pantelis, C. (1999). Tourette's and comorbid syndromes: obsessive compulsive and attention deficit hyperactivity disorder. A common etiology? *Clinical Psychology Review*, ***19***(*5*), 531–552.

Simberlund, J., & Hollander, E. (2017). The Relationship of Body Dysmorphic Disorder to Obsessive-Compulsive Disorder and the Concept of the Obsessive-Compulsive Spectrum. *Body Dysmorphic Disorder: Advances in Research and Clinical Practice*, 481.

Simmonite, M., Bates, A. T., Groom, M. J., Jackson, G. M., Hollis, C., & Liddle, P. F. (2012). Error processing-associated event-related potentials in schizophrenia and unaffected siblings. *International Journal of Psychophysiology*, ***84***(*1*), 74–79.

Simpson, E. H., Kellendonk, C., & Kandel, E. (2010). A possible role for the striatum in the pathogenesis of the cognitive symptoms of schizophrenia. *Neuron*, ***65***(*5*), 585–596.

Sjoerds, Z., de Wit, S., van den Brink, W., Robbins, T. W., Beekman, A. T. F., Penninx, B. W. J. H., & Veltman, D. J. (2013). Behavioral and neuroimaging

evidence for overreliance on habit learning in alcohol-dependent patients. *Translational Psychiatry*, ***3***(*12*), e337–e337.

Skapinakis, P., Caldwell, D., Hollingworth, W., Bryden, P., Fineberg, N., Salkovskis, P., Welton, N., Baxter, H., Kessler, D., Churchill, R., & Lewis, G. (2016). A systematic review of the clinical effectiveness and cost-effectiveness of pharmacological and psychological interventions for the management of obsessive–compulsive disorder in children/adolescents and adults. *Health Technology Assessment*, ***20***(*43*), 1–392.

Snorrason, I., Lee, H. J., de Wit, S., & Woods, D. W. (2016). Are nonclinical obsessive-compulsive symptoms associated with bias toward habits? *Psychiatry Research*, ***241***, 221–223.

Sokhadze, E., Stewart, C., Hollifield, M., & Tasman, A. (2008). Event-related potential study of executive dysfunctions in a speeded reaction task in cocaine addiction. *Journal of Neurotherapy*, ***12***(*4*), 185–204.

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

Spunt, R. P., Lieberman, M. D., Cohen, J. R., & Eisenberger, N. I. (2012). The phenomenology of error processing: The dorsal ACC response to stop-signal errors tracks reports of negative affect. *Journal of Cognitive Neuroscience*, ***24***(*8*), 1753–1765.

Stengler-Wenzke, K., Müller, U., Barthel, H., Angermeyer, M. C., Sabri, O., &

Hesse, S. (2006). Serotonin transporter imaging with [123I] β-CIT SPECT before and after one year of citalopram treatment of obsessive-compulsive disorder. *Neuropsychobiology*, ***53***(*1*), 40–45.

Stipacek, A., Grabner, R. H., Neuper, C., Fink, A., & Neubauer, A. C. (2003). Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. *Neuroscience Letters*, ***353***(*3*), 193–196.

Taylor, S., Abramowitz, J. S., McKay, D., Calamari, J. E., Sookman, D., Kyrios, M., Wilhelm, S., & Carmin, C. (2006). Do dysfunctional beliefs play a role in all types of obsessive-compulsive disorder? *Journal of Anxiety Disorders*, ***20***(*1*), 85–97.

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, ***2***(*4*), i.

Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive-compulsive disorder. *Behaviour Research and Therapy*, ***39***(*8*), 913–927.

Tolin, D. F., Abramowitz, J. S., Przeworski, A., & Foa, E. B. (2002). Thought suppression in obsessive-compulsive disorder. *Behaviour Research and Therapy*, ***40***(*11*), 1255–1274.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, ***55***(*4*), 189.

Torres, A. R., Prince, M. J., Bebbington, P. E., Bhugra, D., Brugha, T. S., Farrell, M., Jenkins, R., Lewis, G., Meltzer, H., & Singleton, N. (2006). Obsessive-compulsive disorder: prevalence, comorbidity, impact, and help-seeking in the British National Psychiatric Morbidity Survey of 2000. *American Journal of Psychiatry*, ***163***(*11*), 1978–1985.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, ***29***(*11*), 2225–2232.

Troller-Renfree, S., Nelson, C. A., Zeanah, C. H., & Fox, N. A. (2016). Deficits in error monitoring are associated with externalizing but not internalizing behaviors among children with a history of institutionalization. *Journal of Child Psychology and Psychiatry*, ***57***(*10*), 1145–1153.

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'Connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience*, ***42***(*1*), 1636–1643.

Ullsperger, M. (2006). Performance monitoring in neurological and psychiatric patients. *International Journal of Psychophysiology*, ***59***(*1*), 59–69.

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, ***94***(*1*), 35–79.

Ullsperger, M., & Von Cramon, D. Y. (2001). Subprocesses of performance monitoring: A dissociation of error processing and response competition

revealed by event-related fMRI and ERPs. *NeuroImage*, ***14***(*6*), 1387–1401.

Ursu, S., Stenger, V. A., Shear, M. K., Jones, M. R., & Carter, C. S. (2003). Overactive action monitoring in obsessive-compulsive disorder: evidence from functional magnetic resonance imaging. *Psychological Science*, ***14***(*4*), 347–353.

Vaghi, M. M., Cardinal, R. N., Apergis-Schoute, A. M., Fineberg, N. A., Sule, A., & Robbins, T. W. (2019). Action-Outcome Knowledge Dissociates From Behavior in Obsessive-Compulsive Disorder Following Contingency Degradation. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, ***96***(*2*), 348-354.e4.

van den Heuvel, O. A., Remijnse, P. L., Mataix-Cols, D., Vrenken, H., Groenewegen, H. J., Uylings, H. B. M., Van Balkom, A. J. L. M., & Veltman, D. J. (2009). The major symptom dimensions of obsessive-compulsive disorder are mediated by partially distinct neural systems. *Brain*, ***132***(*4*), 853–868.

van den Hout, M., & Kindt, M. (2003). Repeated checking causes memory distrust. *Behaviour Research and Therapy*, ***41***(*3*), 301–316.

van Grootheest, D. S., Cath, D. C., Beekman, A. T., & Boomsma, D. I. (2005). Twin studies on obsessive–compulsive disorder: a review. *Twin Research and Human Genetics*, ***8***(*5*), 450–458.

van Loo, H. M., & Romeijn, J.-W. (2015). Psychiatric comorbidity: fact or artifact?

Theoretical Medicine and Bioethics, ***36***(*1*), 41–60.

van Timmeren, T., Daams, J. G., Van Holst, R. J., & Goudriaan, A. E. (2018). Compulsivity-related neurocognitive performance deficits in gambling disorder: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, ***84***, 204–217.

Vidal, F., Hasbroucq, T., Grapperon, J., & Bonnet, M. (2000). Is the "error negativity" specific to errors? *Biological Psychology*, ***51***(*2–3*), 109–128.

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

Weinberg, A., Dieterich, R., & Riesel, A. (2015). Error-related brain activity in the age of RDoC: A review of the literature. *International Journal of Psychophysiology*, ***98***(*2*), 276–299.

Weinberg, A., Klein, D. N., & Hajcak, G. (2012). Increased error-related brain activity distinguishes generalized anxiety disorder with and without comorbid major depressive disorder. *Journal of Abnormal Psychology*, ***121***(*4*), 885.

Weinberg, A., Kotov, R., & Proudfit, G. H. (2015). Neural indicators of error processing in generalized anxiety disorder, obsessive-compulsive disorder, and major depressive disorder. *Journal of Abnormal Psychology*, ***124***(*1*), 172.

Weinberg, A., Meyer, A., Hale-Rude, E., Perlman, G., Kotov, R., Klein, D. N., & Hajcak, G. (2016). Error-related negativity (ERN) and sustained threat: Conceptual framework and empirical evaluation in an adolescent sample. *Psychophysiology*, ***53***(*3*), 372–385.

Weinberg, A., Olvet, D. M., & Hajcak, G. (2010). Increased error-related brain activity in generalized anxiety disorder. *Biological Psychology*, ***85***(*3*), 472–480.

Weinberg, A., Riesel, A., & Hajcak, G. (2012). Integrating multiple perspectives on error-related brain activity: The ERN as a neural indicator of trait defensive reactivity. *Motivation and Emotion*, ***36***(*1*), 84–100.

Werner, C. T., Gancarz, A. M., & Dietz, D. M. (2019). Mechanisms Regulating Compulsive Drug Behaviors. *Neural Mechanisms of Addiction*, 137–155.

Wheaton, M. G., Gillan, C. M., & Simpson, H. B. (2019). Does cognitive-behavioral therapy affect goal-directed planning in obsessive-compulsive disorder? *Psychiatry Research*.

Whelan, R., & Garavan, H. (2014). When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, ***75***(*9*), 746–748.

Whiteside, S. P., Port, J. D., & Abramowitz, J. S. (2004). A meta–analysis of functional neuroimaging in obsessive–compulsive disorder. *Psychiatry Research: Neuroimaging*, ***132***(*1*), 69–79.

Williams, L. M., Gatt, J. M., Schofield, P. R., Olivieri, G., Peduto, A., & Gordon, E. (2009). 'Negativity bias' in risk for depression and anxiety: Brain–body fear

circuitry correlates, 5-HTT-LPR and early life stress. *Neuroimage*, ***47***(*3*), 804–814.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, ***7***(*1*), 30.

Xiao, Z., Wang, J., Zhang, M., Li, H., Tang, Y., Wang, Y., Fan, Q., & Fromson, J. A. (2011). Error-related negativity abnormalities in generalized anxiety disorder and obsessive–compulsive disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, ***35***(*1*), 265–272.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, ***111***(*4*), 931–959.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, ***7***(*6*), 464–476.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, ***19***(*1*), 181–189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. *European Journal of Neuroscience*, ***22***(*2*), 505–512.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral

striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behavioural Brain Research*, ***166***(*2*), 189–196.

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *European Journal of Neuroscience*, ***28***(*8*), 1437–1448.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, ***22***(*2*), 513–523.

Zambrano-Vazquez, L., & Allen, J. J. B. (2014). Differential contributions of worry, anxiety, and obsessive compulsive symptoms to ERN amplitudes in response monitoring and reinforcement learning tasks. *Neuropsychologia*, ***61***(*1*), 197–209.

Zijlmans, J., Bevaart, F., van Duin, L., Luijks, M. J. A., Popma, A., & Marhe, R. (2019). Error-related brain activity in relation to psychopathic traits in multi-problem young adults: An ERP study. *Biological Psychology*, ***144***, 46–53.

Zung, W. W. (1965). A self rating depression scale. *Archives of General Psychiatry*, ***12***(*1*), 63–70.

Appendices

Appendix I: Supplemental information for Chapter 2

A.I. Supplemental Methods

Exclusion criteria. Participants were excluded if they failed any of the following: (i) In the behavioural task, the confidence scale indicator would always start at either 25 or 75 on every trial. Participants who left their confidence rating as the default score for more than 60% of the trials (n > 180 trials) were excluded (N = 42). (ii) The task was also reset from the beginning if confidence ratings were left as the default score for >70% of the first 50 trials (56 participants (9.82%) restarted the task at least once). Those who had their task reset >5 times were excluded (N = 6). (iii) Participants who had more than 50% correlation between the default score and their selected confidence rating were excluded (N = 109). (iv) Participants with a lower mean confidence where the previous trial was correct than incorrect were excluded (N = 66). (v) Participants who incorrectly responded to a "catch" question within the questionnaires: "If you are paying attention to these questions, please select 'A little' as your answer" were excluded (N = 16).

Medication status. Participants were asked if they were currently taking medication for a mental health issue, and if so, to indicate the name, dosage and duration. 41 (9.38%) participants were currently medicated.

Action-confidence coupling. First, we measured the coupling between action updates (i.e. the tendency to move the bucket) and confidence. *Action* (the absolute difference of bucket position on trial *t* and *t*+1) was the dependent variable and *Confidence* (confidence level on trial *t*+1) was the independent variable in a trial-by-trial regression analysis with age, gender and IQ as fixed effects co-variates (as with all subsequent analyses). Within-subject factors (the intercept and main effect of *Confidence*) were taken as random effects (i.e. allowed to vary across subjects). *Confidence* was z-scored within-participant, while the fixed effect predictors were z-scored across participant. If action and confidence are appropriately coupled, participants should move the bucket more (larger *Action*) when their confidence levels were low, producing a significant negative main effect of *Confidence* on *Action*. In the syntax of the *lmer* function, the regression was: Action ~ Confidence * (Age + IQ + Gender) + (1 + Confidence | Subject).

We then tested if psychiatric symptom severity was associated to changes in action-confidence coupling by including the total score for each questionnaire (*QuestionnaireScore*, z-scored) as a between-subjects predictor in the model above. Separate regressions were performed for each individual symptom due to high correlations across the different psychiatric questionnaires. The extent to which

questionnaire total scores contribute to changes in action-confidence coupling is indicated by the presence of a significant *Confidence*QuestionnaireScore* interaction. A positive interaction effect indicates decreased action-confidence coupling (i.e. decoupling), while a negative interaction effect indicates greater action-confidence coupling. The model was specified as: Action ~ Confidence * (*QuestionnaireScore* + Age + IQ + Gender) + (1 + Confidence | Subject). For the transdiagnostic analysis, we included all three dimensions in the same model, as correlation across variables was lessened in this formulation and thus more interpretable (only 3 moderately correlated variables r = 0.34 to 0.52, instead of 9 that ranged from *r* = 0.13 to 0.84). We replaced *QuestionnaireScore* in the model formula described previously with three psychiatric dimensions (*AD, CIT, SW*) entered as z-scored fixed effect predictors. The model was: Action ~ Confidence * (AD + CIT + SW + Age + IQ + Gender) + (1 + Confidence | Subject).

Action and confidence. To analyse the basic relationship between task-related variables and psychiatric dimensions, the analysis approach was the same, but simpler. Dependent variables were: 1) Size of bucket updates (*Action*) and 2) reported confidence (*Confidence*). The models were simply: Task Variable ~ AD + CIT + SW + Age + IQ + Gender + (1 | Subject).

Computation model describing behaviour dynamics. In the behavioural task, participants were required to learn the mean of the underlying generative distribution in order to position their bucket at where they hope to catch the greatest number of

particles. Their belief on where the particle landing distribution mean could be guided by 1) information gained from the most recent outcome (i.e. moving the bucket with every small shift in particle location), 2) surprising large changes signalling a change in mean distribution (i.e. change-points) and 3) their uncertainty of the distribution mean based on particle landing location experience over trials. To separate these contributions, a quasi-optimal Bayesian computational learning model was used to estimate these parameters thought to underlie task dynamics with MATLAB R2018a (The MathWorks, Natick, MA) using functions from Vaghi et al. (Vaghi et al., 2017). This included PE^b (model prediction error, an index of recent outcomes), *CPP* (probability that a trial was a change-point, a measure representing the belief of a surprising outcome) and *RU* (relative uncertainty, the uncertainty owing to the imprecise estimation of the distribution mean; labelled as *(1-CPP)*(1-MC)* in Vaghi et al. (Vaghi et al., 2017)). These parameters (where PE^b is taken as its absolute) together with a *Hit* categorical predictor (previous trial was a hit or miss) were used to regress participant adjustments against the benchmark Bayesian model to investigate participant adjustments in reported confidence (*Confidence*; z-scored confidence level on trial *t*) and bucket movements (*Action*) according to the particle landing locations experienced.

Influence of parameters on action and confidence. For the regression on *Action*, following Vaghi et al. (Vaghi et al., 2017) and prior literature (McGuire et al., 2014; Nassar et al., 2010, 2016), all predictors except PE^b were implemented as interaction terms with PE^b. For Confidence, we used a similar regression model but

without the interaction term with PE^b and with the regressand and predictors z-scored at participant level. Regressions were constructed as mixed-effect models controlled for age, IQ and gender, with the interaction term and main effect of regressors as random effects. The model syntax was written as: Dependent Variable ~ (PE^b + CPP + RU + Hit)*(Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject).

To include psychiatric symptom severity in the same analysis model, we entered each psychiatric questionnaire score as an additional z-scored fixed effect predictor into the basic model above, where the equation was: Dependent Variable ~ (PE^b + CPP + RU + Hit)*(*QuestionaireScor*e + Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject). For confidence, a positive interaction between a symptom score and *PE^b, CPP, RU* indicates that higher scores on that symptom are associated with a decrease in influence of these parameters on confidence. The converse was applicable for significant *Hit*QuestionaireScore* interactions (as main effect of *Hit* on *Confidence* is opposite signed). For action, as main effect of the parameters on *Action* is inverse from the main effects on *Confidence*, significant parameter**QuestionaireScore* interactions are interpreted in reverse. For the transdiagnostic analysis, we included all three dimensions in the same model by replacing *QuestionnaireScore* with three psychiatric dimensions (*AD, CIT, SW*) entered as z-scored fixed effect predictors. The model was: Dependent Variable ~ (PE^b + CPP + RU + Hit)*(AD + CIT + SW + Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject).

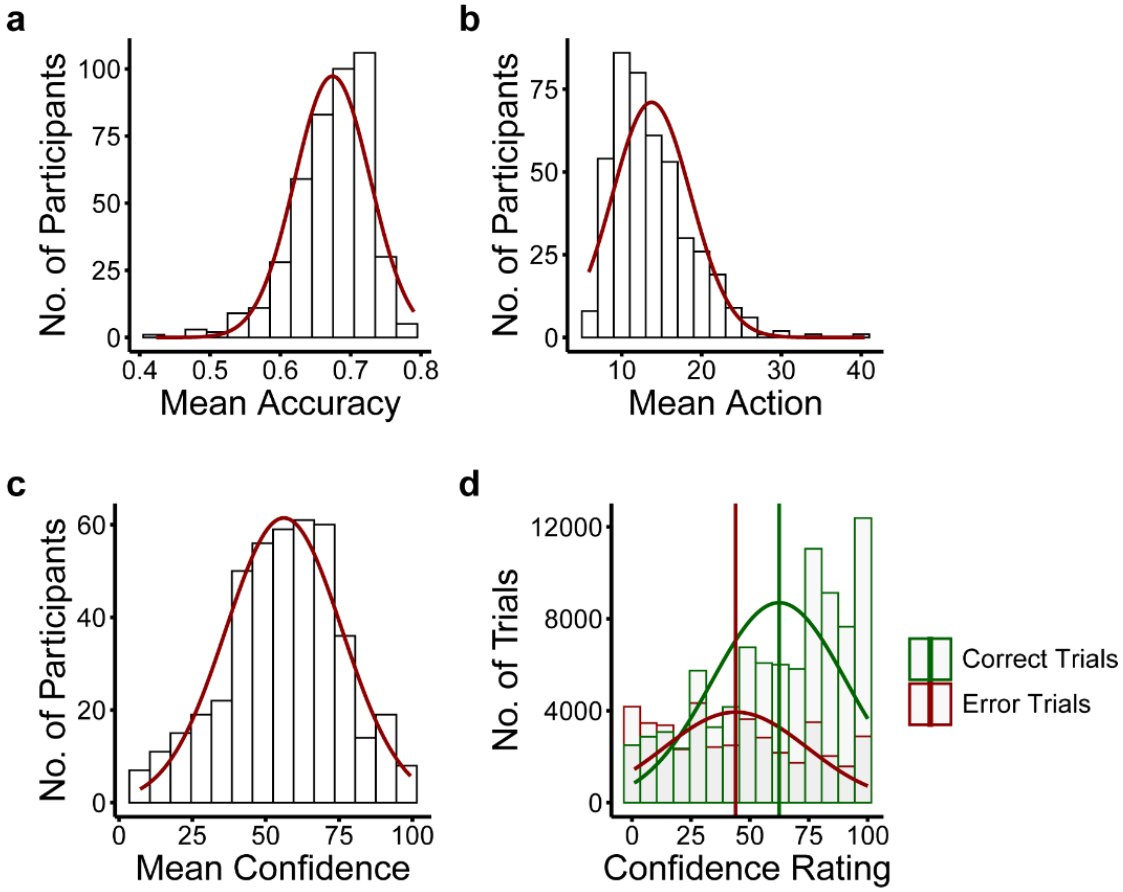
For visualization purposes, the main effects of the four predictors were correlated with CIT severity, where Spearman's correlation was used to measure the association between symptom dimension severity and the influence of the learning parameters on action update/confidence (***Supplemental Figure A.I.S5***).

Influence of metacognitive parameters on action-confidence coupling in compulsivity. We investigated how confidence bias and participants' sensitivity to feedback on confidence were related to action-confidence coupling. We obtained individual beta coefficients from the basic regression model of the model parameters (*PE^b, CPP, RU and Hit*) on confidence from the mixed model equation: Confidence ~ (PE^b + CPP + RU + Hit)*(AD + CIT + SW + Age + IQ + Gender) + (1 + PE^b + CPP + RU + Hit | Subject), individual beta coefficients regression of action on confidence from the equation: Action ~ Confidence * (Age + IQ + Gender) + (1 + Confidence | Subject) and participants' mean confidence level. We regressed each subjects' coefficients for the effect of model parameters on confidence and their mean confidence level against action-confidence in a linear regression, with all regressors taken as z-scored fixed effect predictors. The equation was: Action on Confidence ~ PE^b on Confidence + CPP on Confidence + RU on Confidence + Hit on Confidence + Mean Confidence. To specifically examine how these factors were related to action-confidence coupling in compulsivity, we compared the main effect of CIT on action-confidence coupling in a model with above metacognitive factors: Action on Confidence ~ PE^b on Confidence + CPP on Confidence + RU on Confidence + Hit

on Confidence + Mean Confidence + AD + CIT + SW and without the above metacognitive factors: Action on Confidence ~ AD + CIT + SW. Heteroskedasticity-consistent standard errors for all coefficients are reported by the *vcovHC* function from the *sandwich* package in R.

As expected, action-confidence coupling was significantly related to PE on confidence: $\beta = 1.91$, $SE = 0.18$, $p < 0.001$, CPP on confidence: $\beta = 4.50$, $SE = 0.40$, $p < 0.001$, RU on confidence: $\beta = -1.21$, $SE = 0.37$, $p = 0.001$, Hit on confidence: $\beta = -1.53$, $SE = 0.14$, $p < 0.001$) and marginally to confidence bias ($\beta = -0.13$, $SE = 0.07$, $p = 0.07$). When we included compulsivity in the model above, we found that the original effect of compulsivity on action-confidence coupling was reduced but remained significant (CIT: $\beta = 0.32$, $SE = 0.09$, $p = 0.002$, corrected), suggesting that decreased action-confidence coupling is only partially explained by the multiple metacognitive parameters of the task.

A.I. Supplemental Figures and Tables



Supplemental Figure A.I.S1. *Behavioural results. Across participants, the distribution of:*

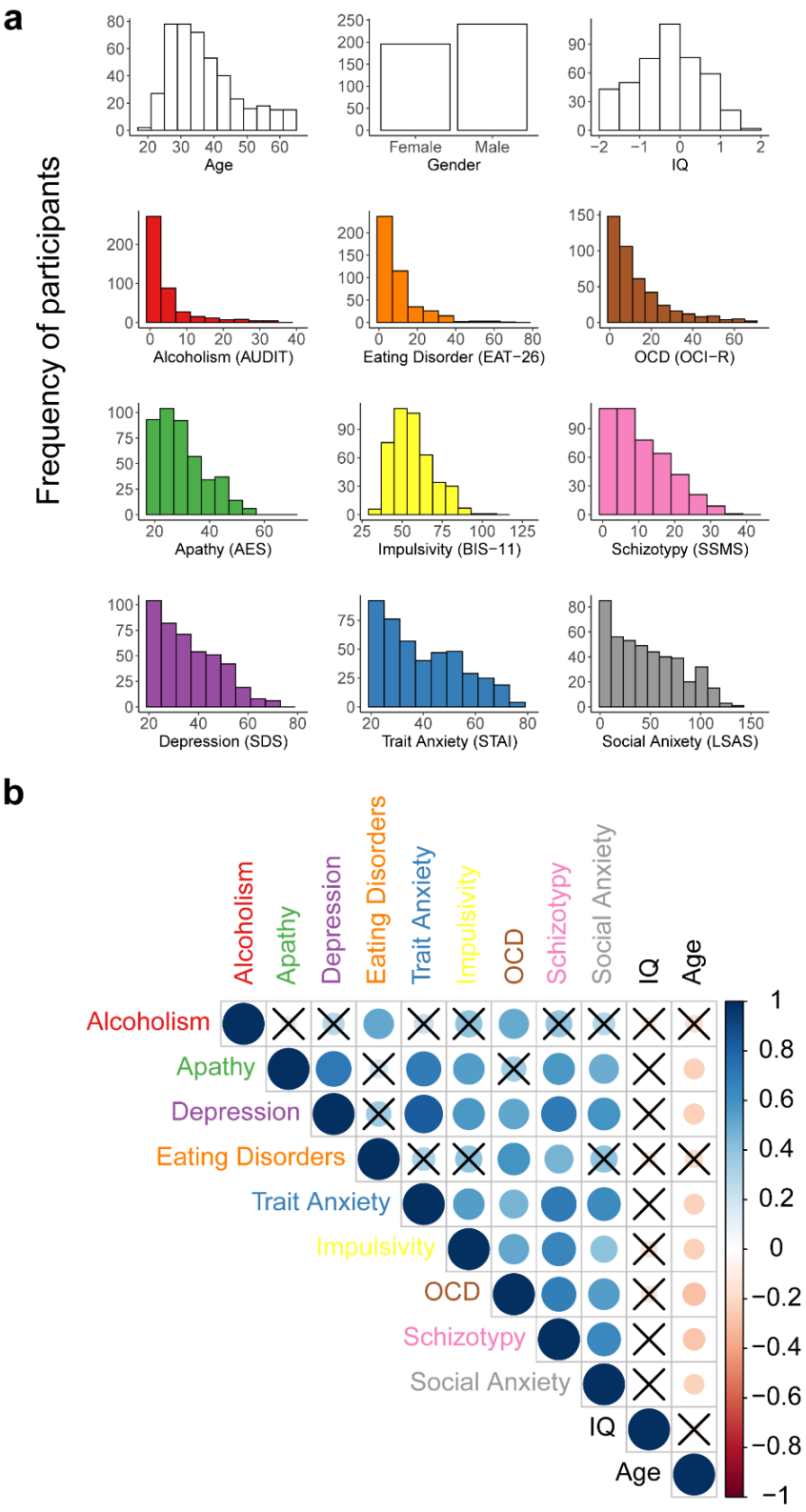
(a) *Mean accuracy.*

(b) *Mean action (the tendency to move the bucket).*

(c) *Mean confidence level.*

(d) *Confidence ratings for correct (green) and incorrect (red) trials. Vertical lines denote mean confidence level for respective trial type.*

Across participants, mean accuracy ranged from 42.33% to 79.00% (mean = 67.42%, SD = 5.38%; ***Figure A.I.S1a***), mean action (tendency to move bucket position) ranged from 5.88 to 40.44 (mean = 13.74, SD = 4.91, ***Figure A.I.S1b***) and mean confidence level ranged from 7.21 to 99.39 across participants (mean = 56.19, SD = 19.85; ***Figure A.I.S1c***). Performance accuracy accounted for only 1.7% of the variance in confidence levels (between-subject correlation: $r = 0.13$, $p < 0.009$). Participants were using the confidence scale appropriately, giving higher confidence after correct trials (mean = 62.42, SD = 28.53), and lower confidence after incorrect trials (mean = 43.98, SD = 30.45) (***Figure A.I.S1d***).



Supplemental Figure A.I.S2.

Supplemental Figure A.I.S2. *Demographics and self-reported psychopathology spread.*

(a) *Age, IQ and questionnaire score distributions across participants.*

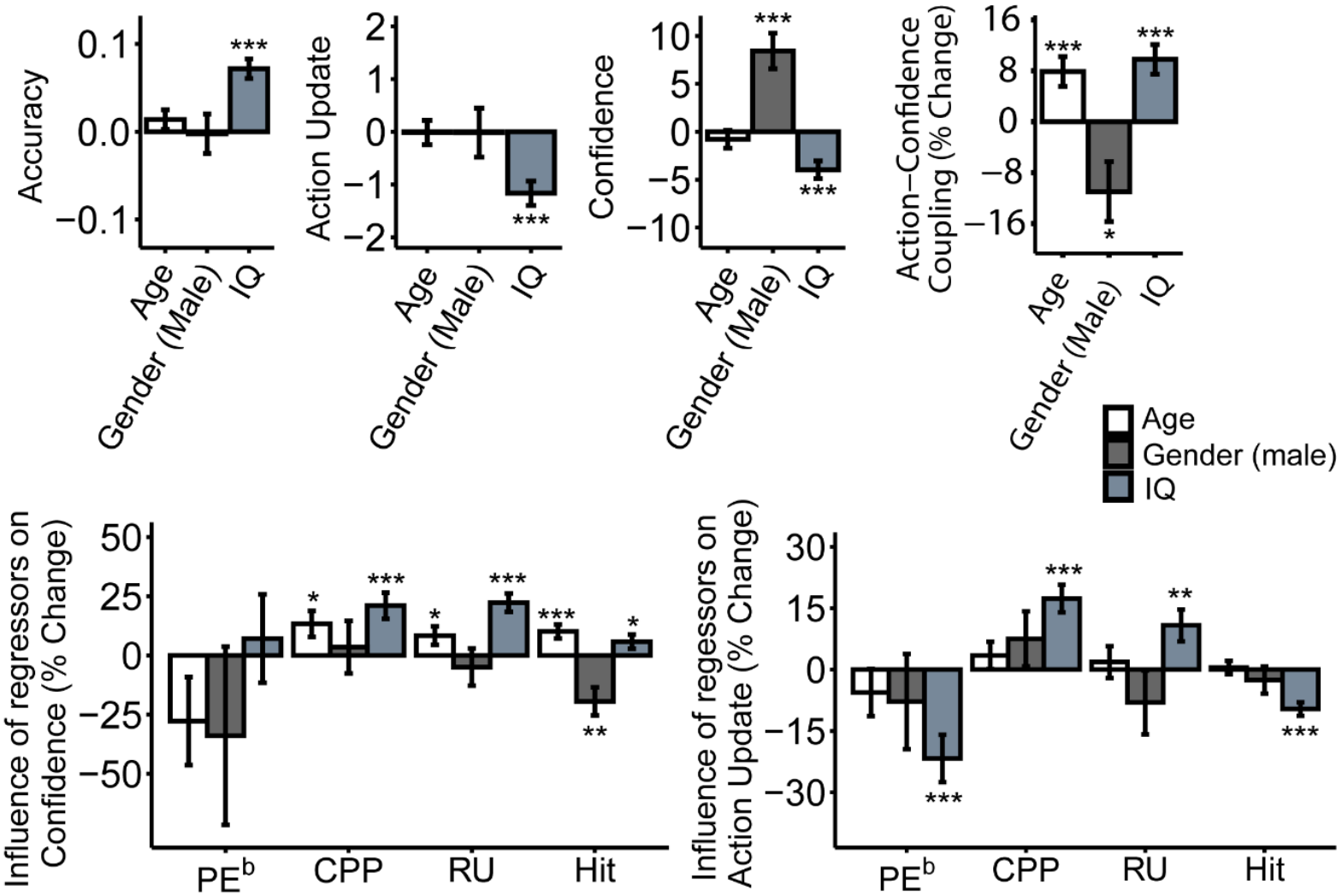
(b) *Correlation matrix of mean scores of the nine questionnaires, age and IQ. Colour scale indicates correlation coefficient, size of colour patch indicates significance. X denotes correlation fails 95% Confidence Interval.*

	PE^b	CPP	RU	Hit
PE^b	1			
CPP	0.68	1		
RU	0.09	0.46	1	
Hit	-0.55	-0.44	-0.14	1

Supplemental Table A.I.S1. *Spearman's correlation between Bayesian Model Parameters (and Hit).*

Predictor	*β (SE)*	*95% CI*	*t-value*	*p-value*
	Regression on Action			
PE[b]	*0.33 (0.02)*	*[0.27, 0.38]*	*11.61*	*< 0.001 ****
CPP	*0.46 (0.02)*	*[0.41, 0.50]*	*20.06*	*< 0.001 ****
RU	*1.37 (0.08)*	*[1.21, 1.52]*	*17.25*	*< 0.001 ****
Hit	*-0.77 (0.02)*	*[-0.81, -0.73]*	*-40.54*	*< 0.001 ****
	Regression on Confidence			
PE[b]	*-0.04 (0.01)*	*[-0.06, -0.02]*	*-3.57*	*< 0.001 ****
CPP	*-0.20 (0.02)*	*[-0.24, -0.17]*	*-12.16*	*< 0.001 ****
RU	*-0.24 (0.01)*	*[-0.27, -0.21]*	*-17.15*	*< 0.001 ****
Hit	*0.26 (0.01)*	*[0.23, 0.29]*	*22.84*	*< 0.001 ****

Supplemental Table A.I.S2. *Effects of Bayesian Model Parameters on Action and Confidence. SE = standard Error, CI = confidence interval.*



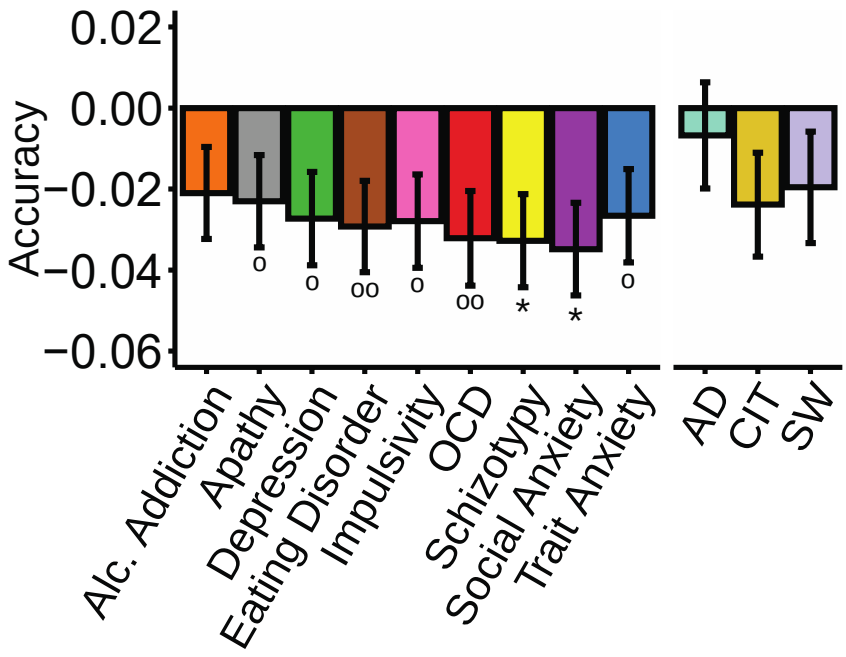
Supplemental Figure A.I.S3.

Supplemental Figure A.I.S3. *Associations between age, gender and IQ with accuracy, action update, reported confidence, action-confidence coupling or the influence of the model predictors (PE^b, CPP, RU) and Hit on confidence/action update. Error bars denote standard errors. The Y-axes indicates the change/percentage change in each dependent variable as a function of 1 standard deviation increase of demographic scores. *p < 0.05, **p < 0.01, ***p < 0.001.*

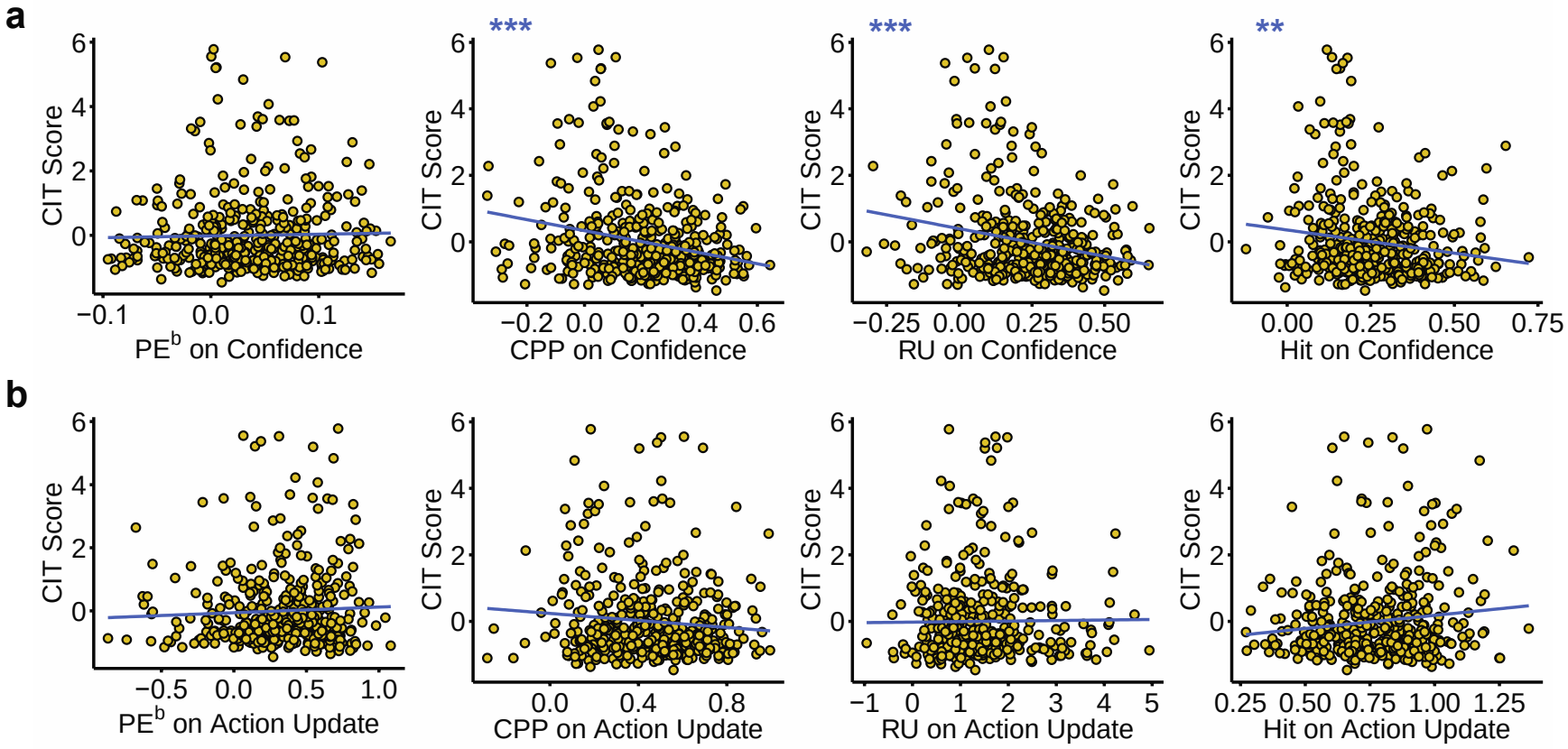
We tested in an exploratory fashion for relationships of task accuracy, action and confidence with age, IQ and gender (***Figure A.I.S3***). IQ was found to predict better performance (β = 0.07, *SE* = 0.01, 95% CI [0.05, 0.09], p < 0.001), lower action updating, (β = -1.16, *SE* = 0.23, 95% CI [-1.62, -0.71], *p* < 0.001) and lower confidence (β = -3.97, *SE* = 0.92, 95% CI [-5.77, -2.17], *p* < 0.001). Additionally, gender (male) was associated with higher confidence (β = 8.43, *SE* = 1.85, 95% CI [4.81, 12.06], *p* < 0.001).

IQ, age and gender were controlled for in all analyses. Increased action-confidence coupling was associated to age (β = -0.70, *SE* = 0.21, 95% CI [-1.10, -0.29], *p* < 0.001), and IQ (β = -0.87, *SE* = 0.21, 95% CI [-1.27, -0.46], *p* < 0.001) while decreased in males (β = 0.97, *SE* = 0.42, 95% CI [0.16, 1.78], *p* = 0.02). For the model-based trial-wise analyses, age was related to an increased influence of CPP (β = -0.03, *SE* = 0.01, 95% CI [-0.05, -0.01], *p* = 0.02), RU (β = -0.02, *SE* = 0.01, 95% CI [-0.04, -0.002], *p* = 0.03) and Hit (β = 0.02, *SE* = 0.01, 95% CI [0.01, 0.04], *p* = 0.03) on confidence. Males were associated to an increased influence of

Hit (β = -0.05, *SE* = 0.02, 95% CI [-0.08, -0.02], p = 0.001) on confidence, while IQ predicted increased influence of CPP (β = -0.05, *SE* = 0.01, 95% CI [-0.06, -0.02], p < 0.001), RU (β = -0.05, *SE* = 0.01, 95% CI [-0.07, -0.03], p < 0.001) and Hit (β = 0.02, *SE* = 0.01, 95% CI [0.0003, 0.03], p = 0.05) on confidence. For action update, only IQ effects were significant – it was related to an increase in CPP (β = 0.08, *SE* = 0.02, 95% CI [0.05, 0.11], p < 0.001) and RU (β = 0.15, *SE* = 0.05, 95% CI [0.04, 0.25], p = 0.006) influence, and decreased PE^b (β = -0.07, *SE* = 0.02, 95% CI [-0.11, -0.03], p < 0.001) and Hit (β = 0.07, *SE* = 0.01, 95% CI [0.05, 0.10], p < 0.001) influence on action update.



Supplemental Figure A.I.S4. *Associations between accuracy (hit (1) or miss (0)) with questionnaire scores or transdiagnostic dimensions, controlled for age, IQ and gender. Error bars denote standard errors. The Y-axis indicates the change in accuracy as a function of 1 standard deviation of questionnaire/dimension scores.* $^{o}p < 0.05$, $^{oo}p < 0.01$ *uncorrected,* $*p < 0.05$. *Results are Bonferroni corrected for multiple comparisons over number of questionnaires/dimensions.*



Supplemental Figure A.I.S5.

Supplemental Figure A.I.S5. *Confidence level/action update was predicted by absolute model prediction error (PE^b), change-point probability (CPP), relative uncertainty (RU) and hit/miss categorial regressor (Hit), controlled for IQ, age and gender. Coefficient estimates from the model were correlated with 'compulsive behaviour and intrusive thought' (CIT) severity.*

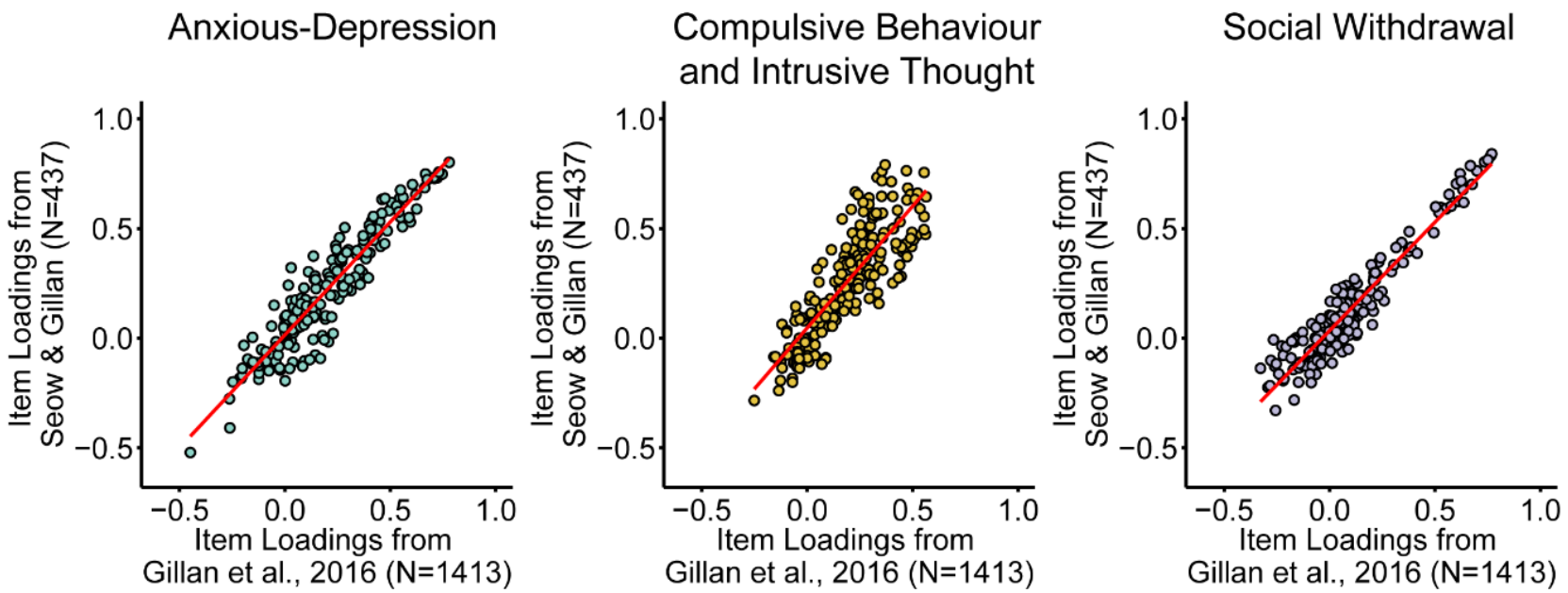
(a) *CIT was found to be associated with significantly diminished influence of CPP, RU and Hit on z-scored confidence. PE^b, CPP and RU on confidence coefficients are inverted to illustrate direction of effects. PE^b: $r_s = 0.003$, $p = 1.00$; CPP: $r_s = -0.19$, $p < 0.001$; RU: $r_s = -0.17$, $p < 0.001$; Hit: $r_s = -0.15$, $p = 0.004$.*

(b) *In contrast, CIT was found not linked to changes in the influence of any of model parameters on action update. For plotting purposes, we show the association of parameter and compulsivity without controlling for AD and SW. PE^b: $r_s = 0.05$, $p = 0.99$; CPP: $r_s = -0.10$, $p = 0.14$; RU: $r_s = 0.01$, $p = 1.00$; Hit: $r_s = 0.09$, $p = 0.17$.*

*Circles represent coefficients of individual participants for model parameters from a basic mixed model of confidence/action update ~ regressors*demographics + (1 + regressors|subject) (x-axis), against their CIT score (y-axis) (see Methods). Hit on action update coefficients are inverted to illustrate direction of effects, such that CIT is linked to an increase influence of hits on action-updating (which is negative in direction). CI = Confidence interval. $^o p < 0.05$, uncorrected, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Correlations are Spearman's rank correlations and results are Bonferroni corrected for multiple comparisons over the three dimensions. See also* ***Figure 2.4****.*

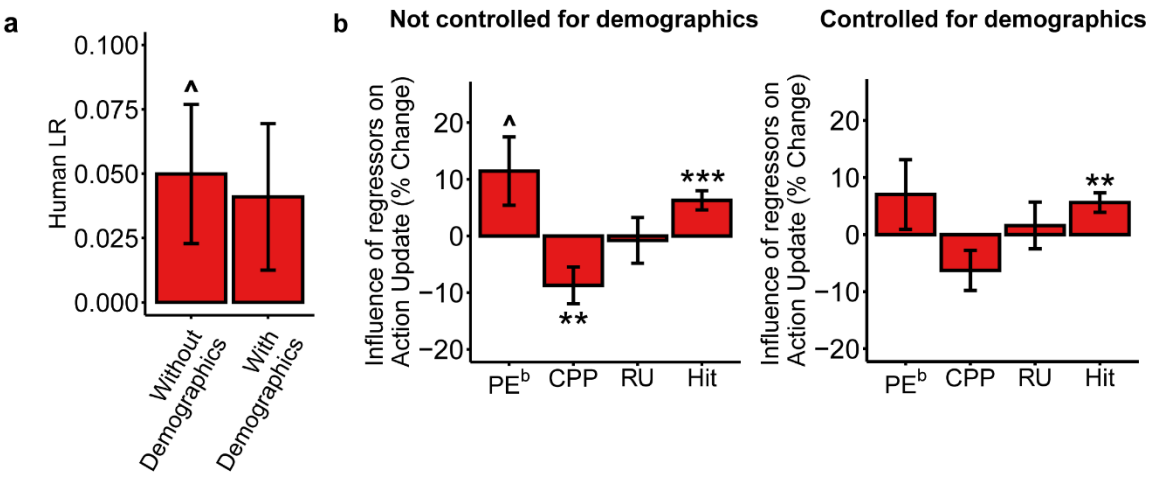
Predictor	β (SE)	SE	t-value	p-value
On Action				
PE[b]	*-0.0005*	*0.02*	*-0.03*	*0.98*
CPP	*-0.01*	*0.01*	*-0.79*	*0.43*
RU	*0.03*	*0.04*	*0.82*	*0.41*
Hit	*-0.01*	*0.01*	*-1.27*	*0.21*
On Confidence				
PE[b]	*-0.001*	*0.002*	*-0.58*	*0.57*
CPP	*0.04*	*0.007*	*6.07*	*< 0.001 ****
RU	*0.04*	*0.007*	*6.36*	*< 0.001 ****
Hit	*-0.23*	*0.006*	*-3.95*	*< 0.001 ****

Supplemental Table A.I.S3. *Effects of 'compulsive behaviour and intrusive thought' (CIT) severity on Bayesian Model Parameters Coefficients on Action and Confidence, with heteroskedasticity-consistent standard errors. Model parameters on action or confidence coefficients were extracted from the basic mixed model action update/confidence ~ regressors*demographics + (1 + regressors|subject) and then regressed in a single linear model by all three psychiatric dimensions anxious-depression (AD), CIT and social withdrawal (SW) for each model parameter. Heteroskedasticity-consistent standard errors were estimated for each model by the vcovHC function from the sandwich package in R. Only CIT effects are reported here. In effect, results are similar to* ***Supplemental Figure A.I.S5****, but with all dimensions scores included in the same model. SE = standard Error, CI = confidence interval.*



Supplemental Figure A.I.S6. *Correlations between item loadings obtained from the factor analysis in Gillan et al. (2016) and the present study for each psychiatric symptom dimension. Questionnaire item loadings were highly correlated for all three factors (Anxious-depression: r = 0.94; Compulsive behaviour and intrusive thought: r = 0.85, Social withdrawal: r = 0.95), supporting the reproducibility of the psychiatric symptom dimensions.*

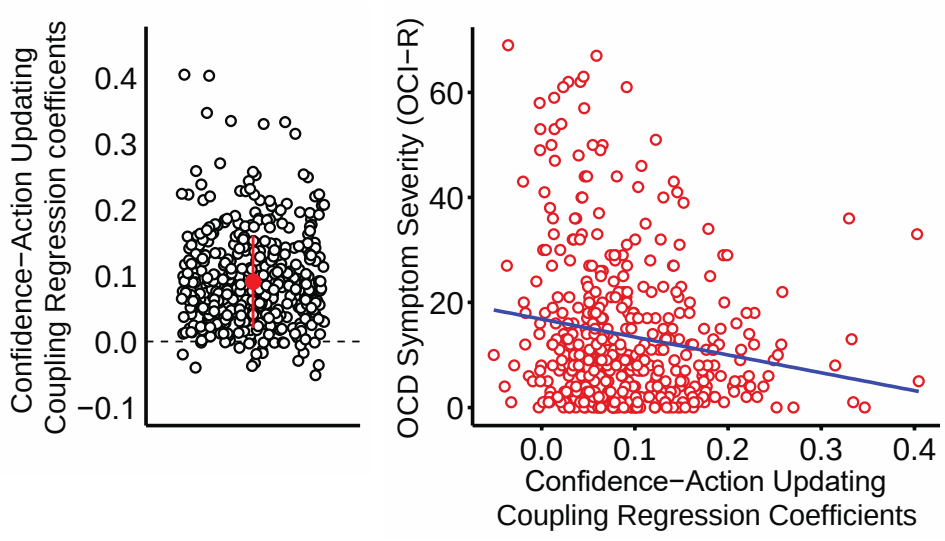
Transdiagnostic symptom dimensions are reproducible. Transdiagnostic dimension scores ('Anxious-depression', 'Compulsive behaviour and intrusive thoughts', 'Social withdrawal') in the present study were derived from weights obtained from a prior larger study (N = 1413) (Gillan et al., 2016). This 3-factor structure was previously reproduced in a smaller independent sample (N = 497) (Rouault et al., 2018), and here we again replicated similar psychiatric dimensions with our current data (N = 437) with the factor analysis (***Figure A.I.S6***). For further details of the factor analysis methodology, see Gillan et al. (Gillan et al., 2016).



Supplemental Figure A.I.S7. *Regression analyses of* ***(a)*** *human learning rate (ratio of bucket movement and task prediction error) and* ***(b)*** *action adjustments in OCD, in a model that controlled for age, IQ and gender and in a model that did not. Error bars denote standard errors. The Y-axes indicate the change/percentage change in dependent variable as a function of 1 standard deviation of OCD symptom scores. ^p < 0.07, **p < 0.01, ***p < 0.001. Results are not Bonferroni corrected for multiple comparisons.*

Action updating effects in OCD with/without controlling for demographics. Vaghi et al. (Vaghi et al., 2017) reported that OCD patients exhibited a higher mean learning rate and that their action updates were more strongly influenced by recent information (PE^b) and less to large unexpected environmental changes (CPP). In the course of exploring the source of this discrepancy with our data, we found that when we repeated our analysis without controlling for age, gender and IQ, some of their effects were recovered here. OCD symptoms were associated with changes in learning and sensitivity to both PE^b and CPP in action updating. Specifically, LR^h

(human learning rate) (β = 0.05, *SE* = 0.03, 95% CI [-0.003, 0.10], *p* = 0.07, uncorr.) and the influence of PE^b on action showed a trend towards a positive association with OCD symptoms (β = 0.04, *SE* = 0.02, 95% CI [-0.001, 0.07], *p* = 0.06, uncorr.) and the influence of CPP on action showed a negative association with OCD symptoms (β = -0.04, *SE* = 0.02, 95% CI [-0.07, -0.01], *p* = 0.007, uncorr.). These discrepancies suggest that demographic characteristics perhaps partially explain the pattern of action updating effects observed in the prior patient study (***Figure A.I.S7***).



Supplemental Figure A.I.S8. *Regression model where confidence updating was predicted by action updating. Dots represent coefficient estimates for individual participants. Red marker indicates mean and SD. These coefficients were correlated with OCD symptom severity, where confidence-action updating coupling was observed to decrease with increasing OCD symptom severity (r = -0.18, p < 0.001).*

Action-confidence decoupling analysis. Although this has no bearing on our results (or theirs), we note that Vaghi et al. (Vaghi et al., 2017) defined action-confidence coupling slightly differently to how we chose to define it in the present paper – they used confidence *updating* (i.e. absolute difference between z-scored confidence from trial *t* and *t*-1), instead of the reported confidence level on trial *t*. We suggest that z-scored confidence ratings (rather than their change from trial to trial) are more appropriate because this accounts better for instances where a person has several relatively large PEs in a row (as they figure out where to place the bucket), and should thus not rationally 'change' their confidence rating in response to these PEs, but maintain it at a low level. Although we flag this for the interested reader, we

underscore that the two measures are correlated and indeed when we use their definition, we similarly show that self-reported OCD symptom severity predicts confidence-action updating decoupling (r = -0.17, t = -3.58, 95% CI [-0.26, -0.07], p < 0.001, ***Figure A.I.S8***).

Appendix II: Supplemental information for Chapter 3

A.II. Supplemental Methods

Disorder prevalence (M.I.N.I.). After exclusion, 80 participants (41.67%) completed the M.I.N.I., which was introduced part-way through the study. Of these participants, 35 (43.75%) presently met the criteria for one or more disorder. Broken down by recruitment arm, 7 (100%) from the clinical arm met criteria, while 28 (38.36%) from university channels met criteria. This rate is close to published reports on the prevalence of mental health disorders in college student samples (Auerbach et al., 2018; Evans et al., 2018). Of the total sample, 33 (17.19%) were currently medicated for a mental health issue. Broken down by recruitment arm, all individuals recruited from the clinic were medicated, while 26 (14.05%) of those recruited through normal channels were medicated. Further diagnostic information of the sample is summarised in ***Supplemental Table A.II.S3***.

P300 and transition type. We tested if P300 amplitude could differentiate rare versus common transitions (*Transition*: rare: 1, common: 0), and whether the difference in amplitude between transition type was related to z-scored model-based estimates (of the logistic regression: *MB*) and to compulsivity (*CIT*). controlled for other psychiatric dimensions (anxious-depression: *AD* and social withdrawal: *SW*). The equations were:

EEG ~ Transition*MB + (Transition + 1 | Subj), and

EEG ~ Transition*(CIT + AD + SW) + (Transition + 1| Subj)

The extent to which model-based control/compulsivity is related to the difference in P300 amplitude of rare versus common transitions was indicated by the presence of a significant Transition*_MB_ or Transition*_CIT_ interaction.

Alpha power and transition type. Similar to the approach of **P300 and transition type**, we tested if alpha power could differentiate rare versus common transitions (*Transition*: rare: -1, common: 0), and whether the difference in power between transition type was related to z-scored model-based estimates (*MB*) and to compulsivity (*CIT*), controlled for other psychiatric dimensions (*AD* and *SW*).

EEG ~ Transition*MB + (Transition + 1 | Subj), and

EEG ~ Transition*(CIT + AD + SW) + (Transition + 1| Subj)

The extent to which model-based control/compulsivity is related to the difference in alpha power of rare versus common transitions was indicated by the presence of a significant Transition*_MB_ or Transition*_CIT_ interaction.

Theta and behavioural control. We tested if single-trial theta power was associated with model-based estimates (*MB*) or to compulsivity (*CIT*, controlled for *AD* and *SW*) by taking them as z-scored main regressors in the models:

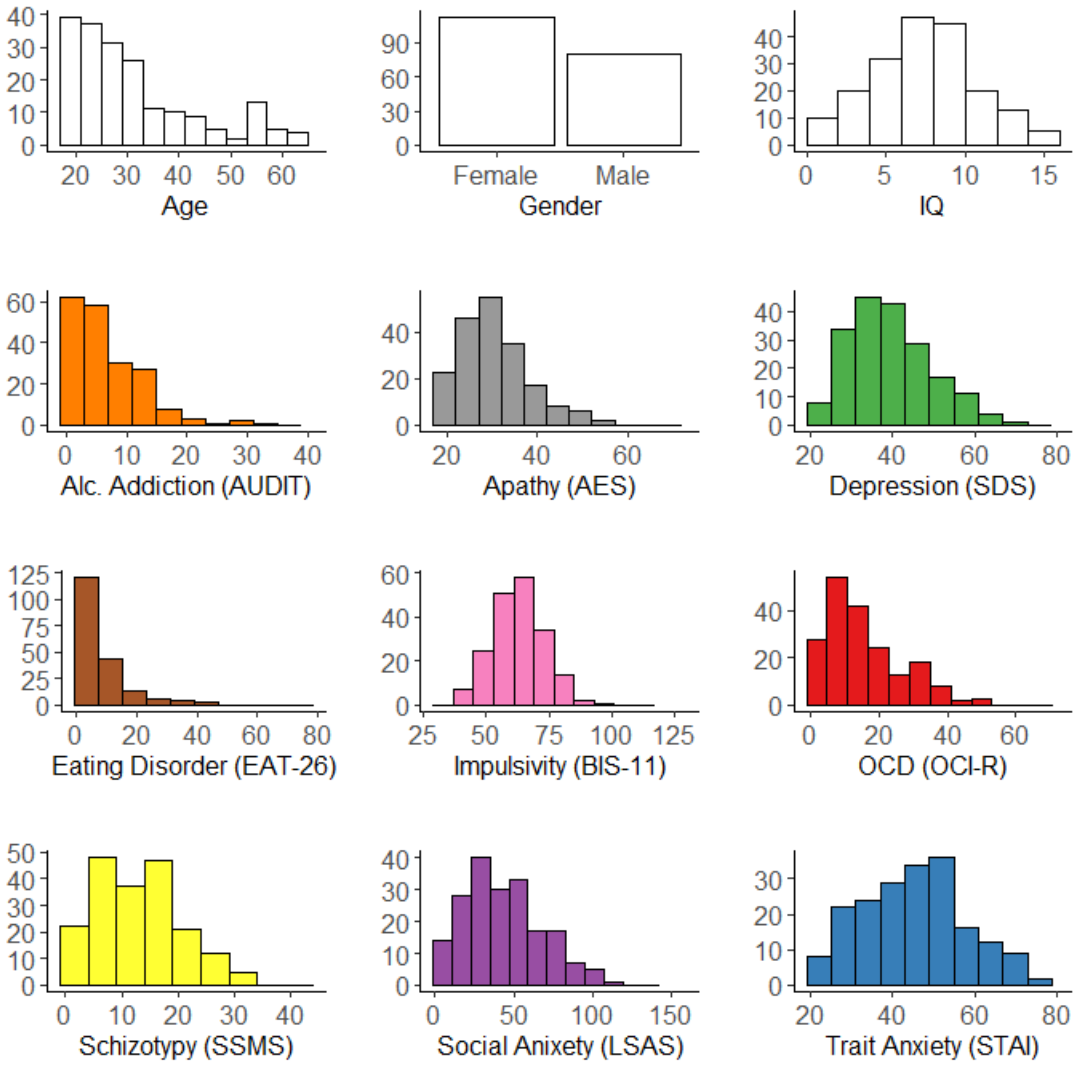
EEG ~ MB + (1 | Subj), and

EEG ~ (CIT + AD + SW) + (1| Subj)

The extent to which model-based control/compulsivity was related to theta in first stage choice was indicated by the presence of a significant *MB* or *CIT* main effect.

Controlling for age and IQ. Age and IQ are known to covary with model-based learning (Gillan et al., 2016). Here, only IQ was significantly associated model-based learning (β = 0.12, *SE* = 0.03, $p < 0.001$); age was not (β = -0.04, *standard error (SE)* = 0.03, p = 0.30). Nonetheless, inclusion of both variables did not change the pattern of our main findings. Reduced goal-directed control was linked to compulsivity (β = -0.08, *SE* = 0.04, p = 0.03) and individuals high in model-based control showed larger alpha power differences between the two transition types over the three rolling time bins beginning from the transition (planet) (-1000ms to 0ms: β = 0.03, *SE* = 0.02, p = 0.06) to the end of choice feedback (0ms to 1000ms: β = 0.04, *SE* = 0.01, p = 0.009; 1000ms to 2000ms: β = 0.05, *SE* = 0.02, p = 0.01).

A.II. Supplemental Figures and Tables



Supplemental Figure A.II.S1. *Histogram of demographics (age, gender and IQ) and total questionnaire scores across participants. Y-axes of each plot indicates the number of participants.*

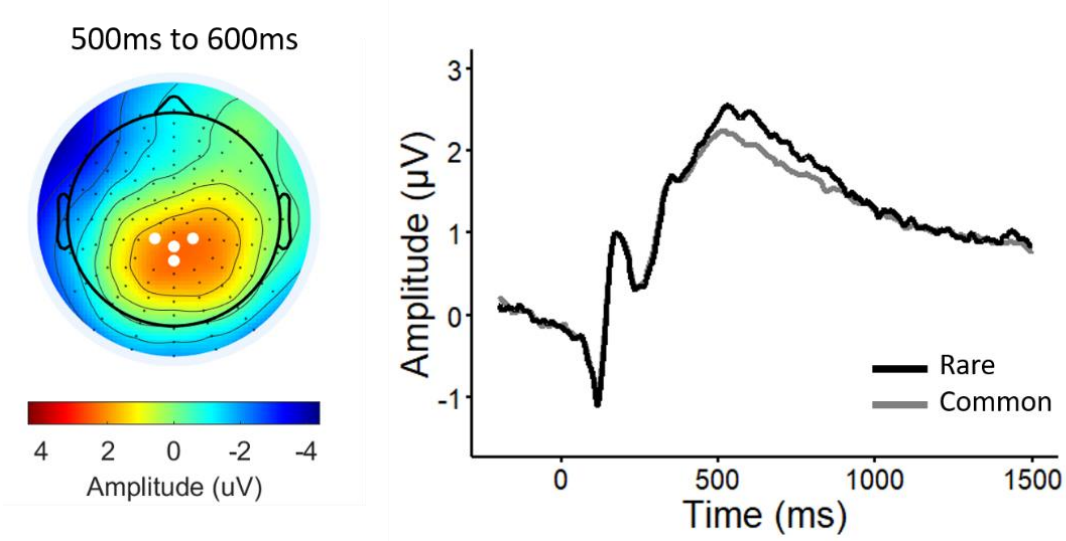
	Alcohol Addiction	Apathy	Depression	Eating Disorder	Impulsivity	OCD	Schizotypy	Social Anxiety	Trait Anxiety	Age	IQ
Alcohol Addiction											
Apathy	0.13										
Depression	0.09	0.64									
Eating Disorder	0.07	-0.008	0.17								
Impulsivity	0.22	0.47	0.36	0.19							
OCD	-0.08	0.31	0.43	0.31	0.30						
Schizotypy	0.04	0.49	0.61	0.30	0.55	0.56					
Social Anxiety	-0.09	0.29	0.43	0.26	0.19	0.47	0.50				
Trait Anxiety	0.04	0.57	0.79	0.14	0.34	0.47	0.66	0.46			
Age	-0.17	-0.03	-0.15	-0.24	-0.10	-0.14	-0.18	-0.13	-0.11		
IQ	-0.10	-0.03	-0.05	0.03	-0.24	-0.02	-0.12	-0.007	0.006	-0.19	
Reliability	0.87	0.86	0.86	0.87	0.81	0.90	0.86	0.92	0.93		

Supplemental Table A.II.S1. *Pearson's correlations between total questionnaire scores and demographics (age and IQ). Cronbach's Alpha was used to calculate reliability for questionnaires.*

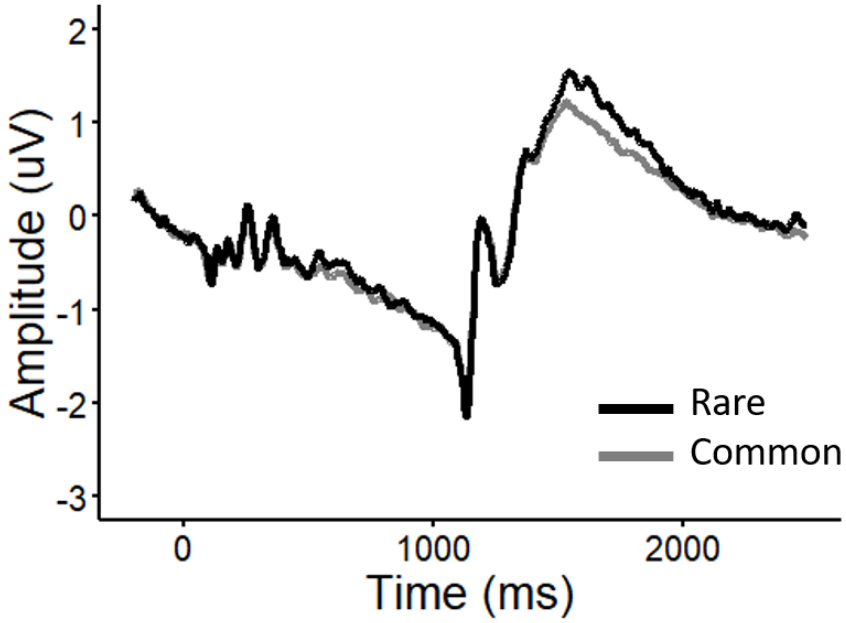
	AD	CIT	SW
AD			
CIT	0.33		
SW	0.37	0.42	

Supplemental Table A.II.S2. *Pearson's correlations between transdiagnostic dimensions scores (AD: 'anxious-depression', CIT: 'compulsive behaviour and intrusive thought', SW: 'social withdrawal').*

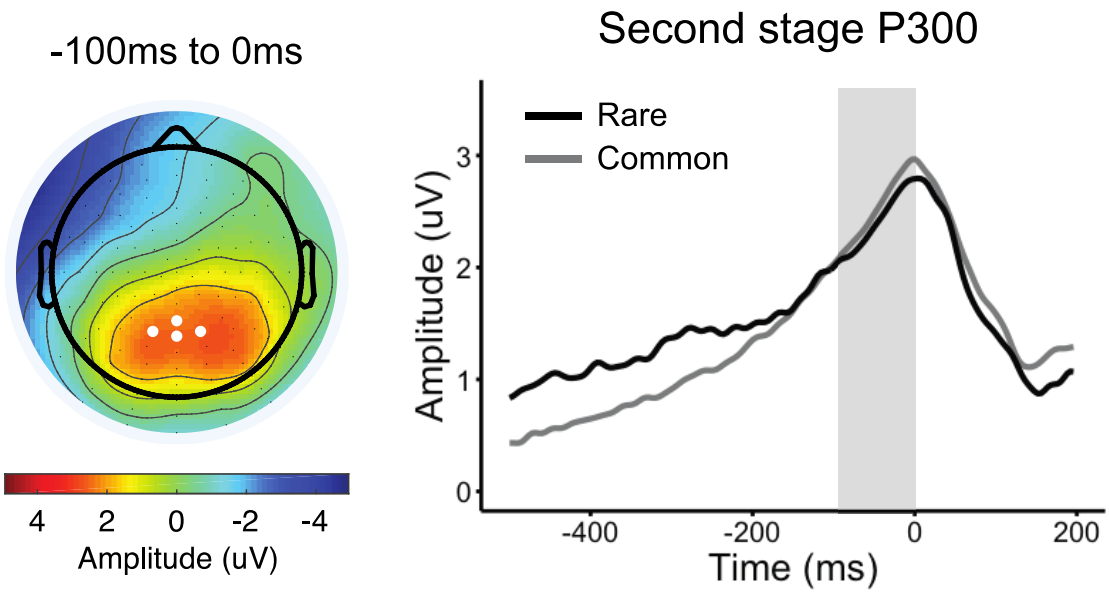
Supplemental Figure A.II.S2. First stage stay probabilities. *Model-based behaviour is reflected as the probability of repeating the first stage choice (stay) as a function of the occurrence of a transition from the previous trial (common: 70%, rare: 30%) and whether a reward was received (reward, non reward). In a purely model-free learner, stay probabilities after reward should be higher than when no reward was presented regardless of transition type. In a purely model-based learner, stay probabilities after common-reward and rare-non reward should be higher than common-non reward and rare-reward. In our empirical data here, the stay probabilities obtained across conditions is a mix of both model-based and model-free behaviour. Error bars reflect standard errors of mean.*



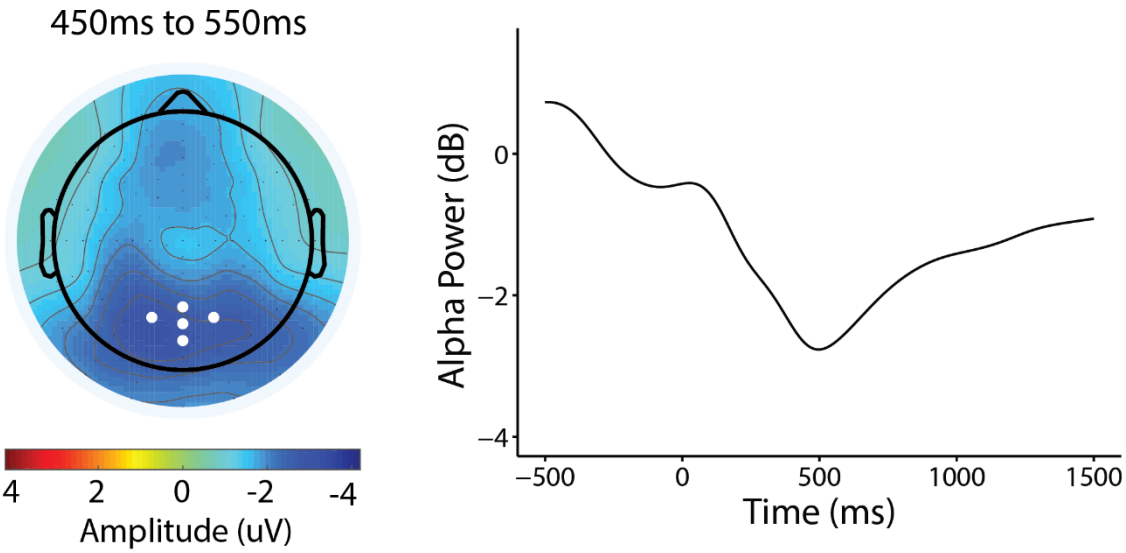
Supplemental Figure A.II.S3. Second stage stimulus-locked P300 and transition type. *Grand average waveforms of rare and common trials locked to second stage stimuli (aliens). Waveform is baselined -200ms to 0ms. The mean amplitude for stimulus-locked P300 was obtained over 4 centro-parietal electrodes (D16 (CP1), A3 (CPz), B2 (CP2), A4) as indicated by the white dots in the topography plot.*



Supplemental Figure A.II.S4. Second stage transition-locked grand average ERP waveforms comparing rare versus common transitions. *Amplitude was measured from a mean over 4 centro-parietal electrodes (D16 (CP1), A3 (CPz), B2 (CP2), A4). At 0ms, transitioned states (planets) appeared, followed by second stage stimuli (aliens) at 1000ms. Parietal ERP did not differ between rare versus common transitions after the states (planet) were presented and before the aliens appeared, unlike parietal-occipital alpha (see Figure 3). Waveform in this plot is similar to that of Figure S3, except that it is baseline-corrected to -200 to 0ms before transition onset.*

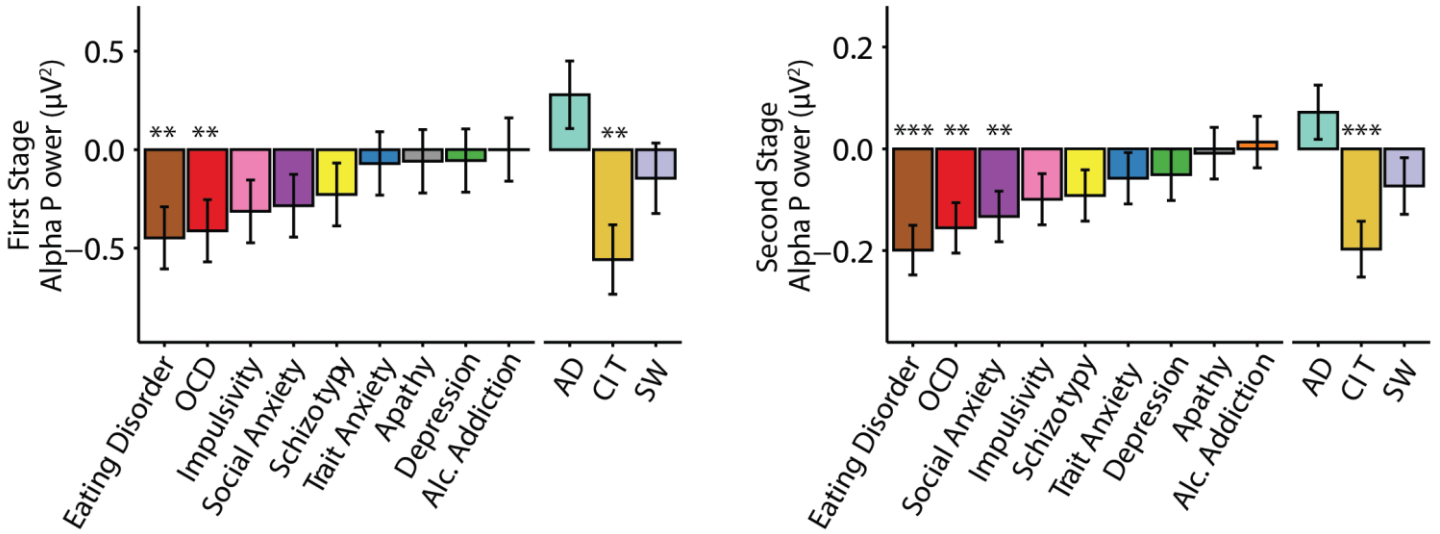


Supplemental Figure A.II.S5. Response-locked P300 and transition type. *Topography plot represents the P300 component -100ms to 0ms before second stage response. White dots indicate parietal electrode sites (A4, A5, A19 (Pz), A32) where the positive component was measured. Grand average second stage P300 is plotted response-locked comparing the waveforms following rare versus common transitions. Single-trial analyses indicate that the P300 amplitude, measured as the mean amplitude -100ms to 0ms (shaded grey), does not distinguish transition type (β = -0.09, SE = 0.08, p = 0.23).*



Supplemental Figure A.II.S6. First stage alpha power. *Topography and line plot (locked to first-stage rockets) show alpha depression during the making a choice at the first stage. White dots on the topography plot indicate occipital electrode sites (A18, A19 (Pz), A20, A21, A31) where alpha was measured for both first and second stages.*

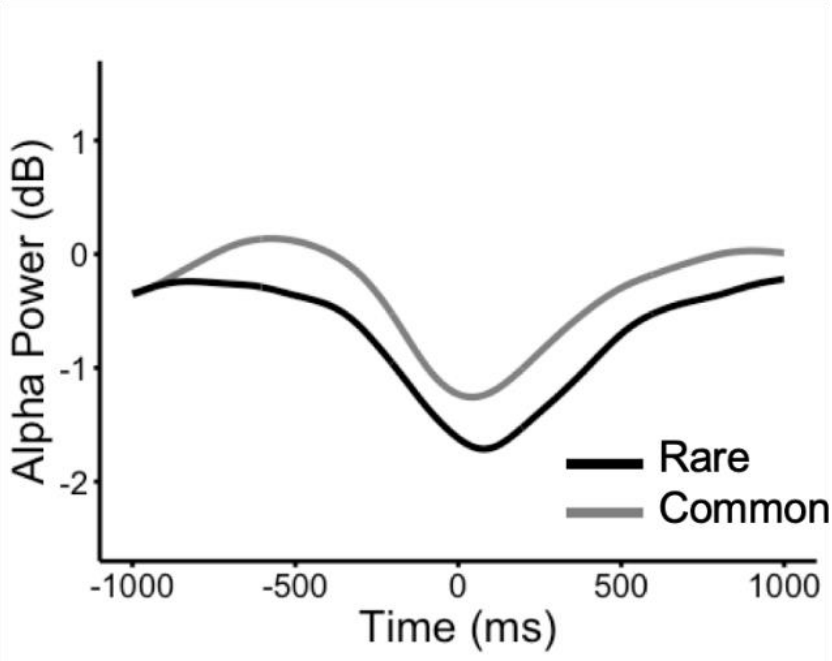
We found that alpha power at first stage (See ***Figure A.II.S7*** for alpha power quantification details) was more suppressed in high compulsive individuals (β = -0.56, *SE* = 0.03, p = 0.002). However, this effect was not associated to model-based planning (β = 0.13, *SE* = 0.02, p = 0.43) nor RT differences in transition types (β = -0.01, *SE* = 0.16, p = 0.94).



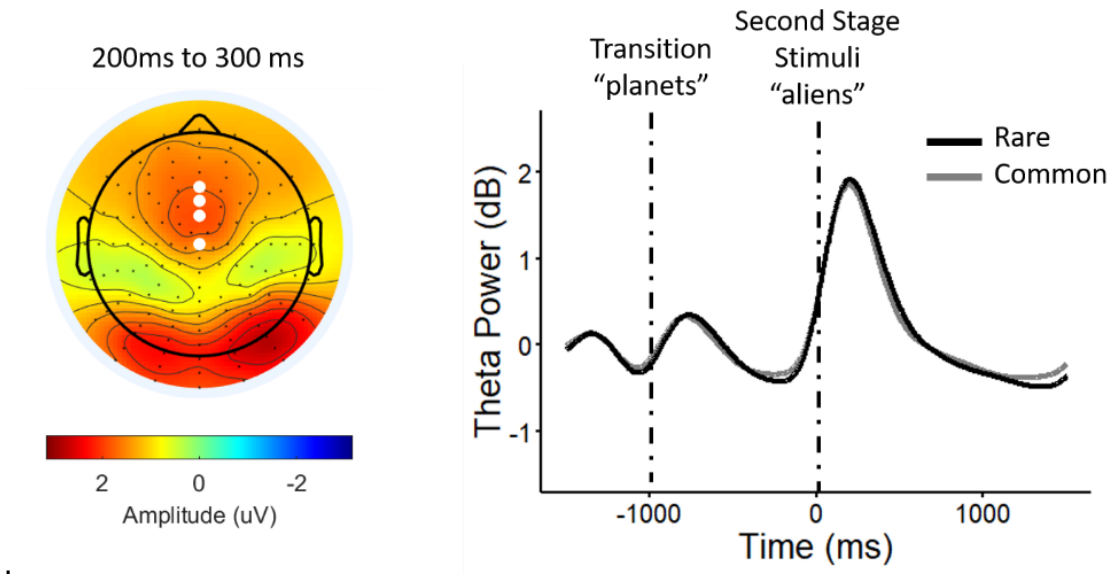
Supplemental Figure A.II.S7. First and second stage alpha power and their relationship with total questionnaire scores and psychiatric dimensions (AD: 'anxious-depression'; CIT: 'compulsive behaviour and intrusive thought', SW: 'social withdrawal'). *In both stages, alpha decrease was associated to more than one questionnaire but was ultimately specific to the compulsive dimension (CIT, as opposed to AD and SW). The Y-axes shows the change in alpha power (μV²) as a function of 1 standard deviation increase of psychiatric questionnaire/dimension scores. Error bars denote standard errors. *p < 0.05, **p < 0.01, ***p < 0.001.*

Single-trial stimulus-locked alpha estimates was measured as the mean power ±250ms centered around the average latency of the negative peak, specific to each individual, and found within a search window defined from 0ms to 1000ms after stimulus (for first stage epoch: rockets, for second stage epoch: aliens) onset. We tested if alpha power changed as a function of total questionnaire scores (*QuestionnaireScore*, z-scored). Separate mixed effects regression models were performed for each individual questionnaire (as correlation across questionnaire scores ranged greatly from r = -0.09 to 0.79), taking the intercept as random effect. For the transdiagnostic analysis, we included all three dimensions in the same model. We replaced *QuestionnaireScore* in the previous model equation with three psychiatric dimensions (*AD, CIT, SW*) entered as z-scored fixed effect predictors together.

We found that for both stages, there was a decrease in alpha power associated with compulsivity (first stage: β = -0.56, SE = 0.18, p = 0.002; second stage: β = -0.20, SE = 0.05, p < 0.001). However, this effect was not related to model-based control in either stages (first stage: β = 0.12, SE = 0.17, p = 0.43; second stage: β = 0.04, SE = 0.05, p = 0.42).



Supplemental Figure A.II.S8. Second stage response-locked alpha power and transition type. *Grand average waveforms of rare versus common transitions were plotted. A significant association was found between single-trial alpha estimates (measured as the mean of ±100ms centered around each participant's averaged latency of the negative peak) and transition type (β = 0.06, SE = 0.01, p < 0.001). Similar to stimulus-locked alpha, rare transitions showed greater depression of alpha during choice selection for rare versus common transitions.*



.

Supplemental Figure A.II.S9. Second stage stimulus-locked theta power. *Topography plot shows theta power increase after stimulus-onset at the mid-frontal scalp. White dots indicate electrode sites (C21 (Fz), C22, C23 (FCz), A1 (Cz)) where theta power was measured. No difference was observed between the grand average waveforms obtained across rare versus common transitions.*

Single-trial second stage theta power estimates were measured as the mean amplitude ±250ms around the average latency of the positive peak, specific to each individual, and found within a search window from 0ms to 500ms after stimulus onset. These were regressed against *Transition* (rare: 1, common: 0) as the intercept and as random effects. The model was: Theta ~ Transition + (Transition + 1 | Subj).

In contrast to alpha power, theta power at second stage was not associated to transition types (β = 0.03 SE = 0.02, p = 0.22) (***Figure A.II.S9***). Theta was also not associated to compulsivity (β = -0.0001, SE = 0.03, p < 1) nor model-based planning

(β = 0.01, SE = 0.03, p = 0.57) nor any transition interaction effects with compulsivity (β = -0.03, SE = 0.03, p = 0.31) or model-based planning (β = 0.03, SE = 0.02, p = 0.23).

Disorder	*Diagnosis*
Mood disorders	
Major depressive disorder	18
Suicide behavior disorder	1
Bipolar Disorder	1
Anxiety disorders	
Panic Disorder	12
Agoraphobia	5
Generalised Anxiety Disorder	13
Social Anxiety Disorder	10
Obsessive-Compulsive Disorder	4
Posttraumatic Stress Disorder	0
Substance use disorders	
Alcohol Use Disorder	6
Substance Use Disorder (Non-alcohol)	9
Psychotic Disorders	
Psychotic Disorders	0
Eating Disorders	
Anorexia Nervosa	0
Bulimia Nervosa	2
Binge-Eating Disorder	4
Other disorders	
Antisocial Personality Disorder	2

Supplemental Table A.II.S3. *Mini International Neuropsychiatric interview (M.I.N.I.) diagnostic information summary for participants who presently met the criteria for at least one DSM-V disorder (N = 35).*

Appendix III: Supplemental information for Chapter 4

A.III. Supplemental Methods

ERN, demographics and error rate. In existing work, age, gender and IQ (Falkenstein et al., 2001; Fischer et al., 2016; Larson et al., 2016; Zijlmans et al., 2019) yield various relationships with ERN amplitude. We explored their effects on ERN in our data. IQ (β = -0.38, *SE* = 0.20, p = 0.06) and age (β = 0.39, *SE* = 0.20, p = 0.05) showed a trending effect with ERN amplitude shifts, while gender not associated (both $p < 1$). We also observed that error rate was related to the ERN (β = 0.40, *SE* = 0.20, p = 0.04). However, inclusion of age and IQ nor error rate did not change the effect of questionnaires scores on ERN amplitude (all $p > 0.13$, uncorrected).

ERN and medication status. The ERN has also been previously influenced by various psychotropic medication (Bates et al., 2002; de Bruijn et al., 2006; Endrass et al., 2008; Henderson et al., 2006; Riba et al., 2005). Only 31 (15.82%) participants were currently medicated for a mental health issue, which was too small a sample to conduct analyses divided by medication type. Nonetheless, we investigated if medication status was related to ERN amplitude. It was not (β = 0.80, *SE* = 0.55, p = 0.15), and neither did inclusion of medication status significantly modulate the effect of questionnaires scores on ERN amplitude (all $p > 0.09$, uncorrected).

ERN amplitude measures. In the literature, there are various ways to quantify ERN amplitude (Clayson et al., 2013). Here we report the supplementary analyses showing that the main results were not due to our chosen analysis approach—whether it was from electrode site (***Supplemental Figure A.III.S6***) or ERN quantification method (***Supplemental Figure A.III.S7***, method details below).

For non-adaptive mean, ERN amplitude calculated as the mean of ±40ms at 37.61ms post-response, which was the mean latency of the most negative peak across participants. For peak, the most negative peak was identified, and amplitude was extracted, for each participant by searching for the largest preceding negativity within -20ms to 120ms post-response. For trough-peak, the trough was identified for each participant by searching for the largest preceding positivity within -100ms before the peak. The amplitude of this positive peak was then subtracted from the negative peak amplitude.

ERN controlled for CRN variation. A common method thought to isolate activity specific to error monitoring is calculated by the subtraction of the CRN from the ERN i.e. ERN-CRN (ΔERN) (Gehring et al., 1993). However, using the subtraction method is conceptually problematic as the ERN and CRN are highly correlated across individuals (here, ERN and CRN correlate: $r = 0.30$, $p < 0.001$). This is because difference scores are not independent from the constituent measures (i.e. ΔERN is not an error processing measure independent of the CRN) and may conflate effects relating to either signal (A. Meyer et al., 2017). An alternative approach to control for variation of the CRN is to use the variation

left over from a regression of CRN predicting ERN (ERN_{resid}) as the ERN amplitude measure. ERN_{resid} was correlated to ERN (r = 0.95, p < 0.001) but not to the CRN (r = ~0, p = 1), suggesting that it specifically indexes error-related activity and is a more interpretable measure. We report the associations of these two different ERN measures, ΔERN and ERN_{resid} with questionnaire scores (***Supplemental Figure A.III.S8*** and ***Table A.III.S3***), and note both findings do not reveal any significant effects (all p > 0.12, uncorrected).

ERN, depression and anxiety. Previous studies have suggested that depression can reduce the increased ERN amplitudes effect associated with anxiety (Weinberg et al., 2016; Weinberg, Klein, et al., 2012; Weinberg, Kotov, et al., 2015). We tested if this was true in our data by regressing depression and anxiety total scores against ERN estimates in the same model. Both effects remained non-significant; but the direction of effects was perhaps more representative of the literature with anxiety leaning towards a larger ERN (β = -0.33, SE = 0.30, p = 0.27) and depression towards a smaller ERN (β = 0.43, SE = 0.30, p = 0.16).

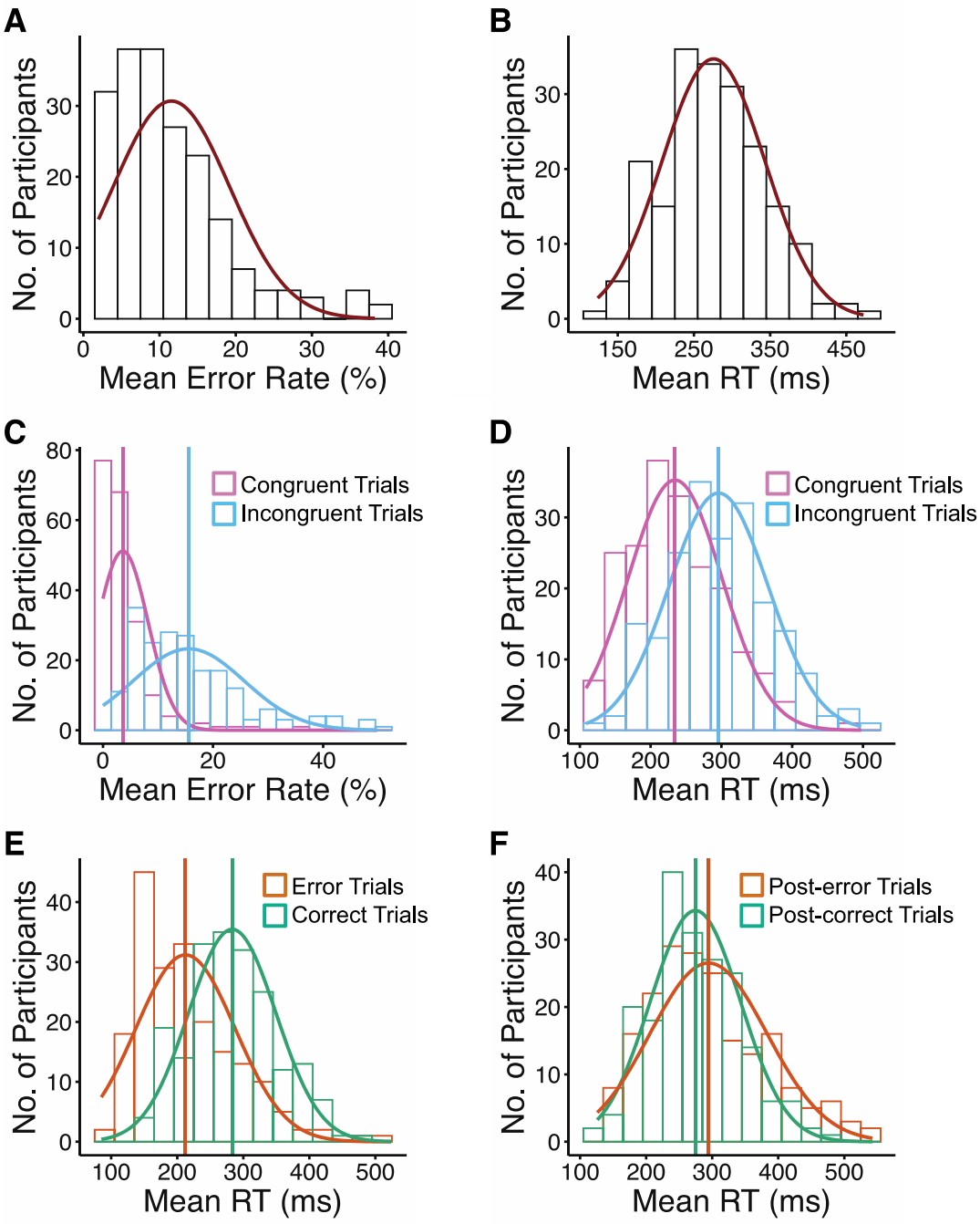
Goal-directed learning. The same sample of participants (N = 234) completed the two-step reinforcement learning task (Daw et al., 2011). Several exclusion criteria were applied to ensure data quality, on a rolling basis. i) Participants who responded with the same key in stage one >90% (n = 135) of the time (N = 10). ii) Participants whose probability of staying after common, rewarded trials was less than 5% likely to be at chance, based on a binomial distribution with 50% (chance) probability and the total number of common-rewarded trials

experienced by each participant (N = 11). iii) Participants who missed >20% (n = 30) of the trials were excluded (N = 3). (iv) Participants who incorrectly responded to a “catch” question within the questionnaires: “If you are paying attention to these questions, please select ‘A little’ as your answer” were excluded (N = 7). (v) As we intend to analyse the EEG data collected for this task, we additionally excluded participants whose EEG data were incomplete (N = 5) or corrupt (N = 2) from the analysis. 38 participants (16.24%) were excluded in total, leaving 196 participants for analysis. To clean the task data, we excluded individual trials with very fast reaction times (<150ms) reflecting inattention or poor responding. Including missed trials, a total of 1114 (3.77%) trials were excluded.

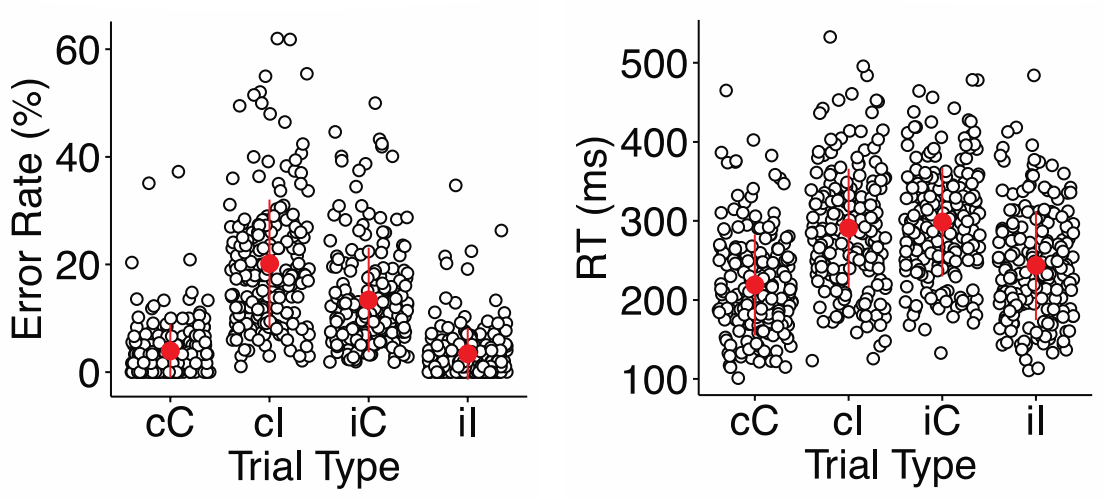
To estimate goal-directed learning, we performed logistic regression via mixed-effects models with the *lme4* package in R, with Bound Optimization by Quadratic Approximation (bobyqa) with 1e5 functional evaluations. The basic model tested if participants’ choice behaviour to *Stay* or switch relative to previous choice (stay: 1, switch: 0) was influenced by the previous trial’s *Reward* (rewarded: 1, unrewarded: -1), *Transition* (common (70%): 1, rare (30%): -1) and their interaction, with age, gender and IQ as z-scored fixed-effects covariates. Within-subject factors (the intercept, main effects of reward, transition, and their interaction) were taken as random effects (i.e. allowed to vary across participants). In syntax of R, the model was: Stay ~ Reward * Transition + (Reward * Transition + 1 | Subject). The interaction effect between Reward and Transition was significant, indicating a contribution of goal-directed learning to choice behaviour ($\beta = 0.20$, $SE = 0.03$, $p < 0.001$). To test if symptom dimensions were associated with goal-directed learning deficits, we included the total scores

of the three dimensions (anxious-depression, compulsive behaviour and intrusive thought ('compulsivity'), social withdrawal) as z-scored fixed effect predictors into the basic model described above. The model was: Stay ~ Reward * Transition + (*Anxious-depression* + *Compulsivity* + *Social withdrawal*) + (Reward * Transition + 1 | Subject). The extent to which a dimension is related to deficits in goal-directed learning was indicated by the presence of a significant Reward*Transition**Dimension* interaction. In prior work, age and IQ were associated to model-based planning (Gillan et al., 2016). Inclusion of these demographics did not change the pattern of effect to compulsivity ($\beta = -0.08$, $SE = 0.04$, $p = 0.04$).

A.III. Supplemental Figures and Tables

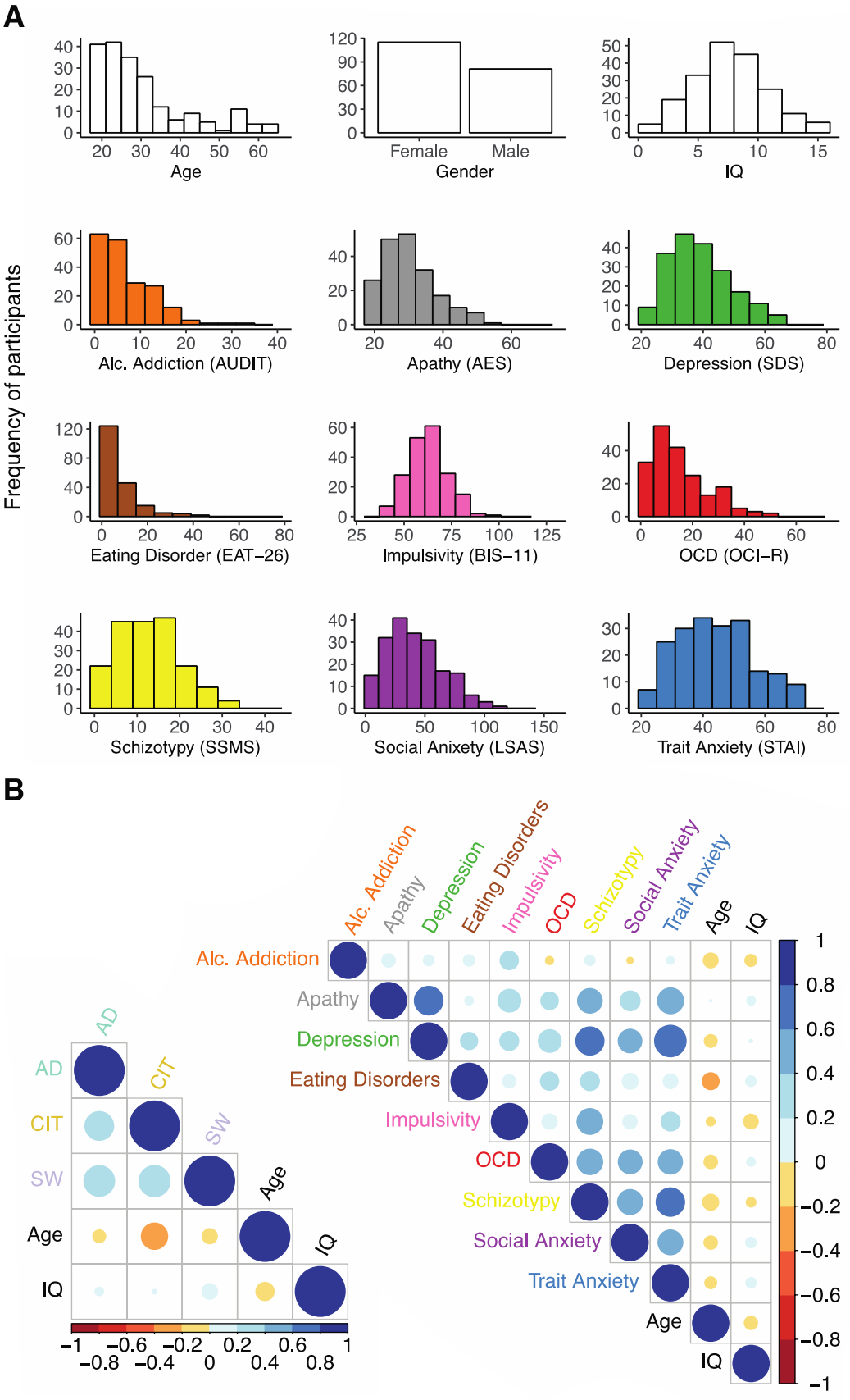


Supplemental Figure A.III.S1. *Across participants, the distribution of:* ***(A)*** *Mean error rate.* ***(B)*** *Mean response time (RT).* ***(C)*** *Mean RT by trial congruency.* ***(D)*** *Mean RT by trial congruency.* ***(E)*** *Mean RT by trial accuracy.* ***(F)*** *Mean RT by post-trial accuracy. Vertical lines denote mean error rate/RT for respective trial type.*



Supplemental Figure A.III.S2. *Mean error rate and response times (RT) for various trial types. cC: congruent trials preceded by a congruent trial, cI: incongruent trial preceded by a congruent trial, iC: congruent trial proceeded by an incongruent trial), iL: incongruent trial preceded by an incongruent trial). White dots represent individual participants, red marker indicates mean and SD.*

Conflict adaptation. Conflict adaptation effects refer to the phenomenon wherein previous-trial congruency affects current-trial performance, which have consistently been shown as behavioural adjustment in error rates and RTs in Flanker tasks (Clayson & Larson, 2011; Larson et al., 2016). We replicate these effects, where mean error rates were smaller for iI than for cI trials (t_{195} = -22.08, 95% CI [-0.18, -0.15], p < 0.001) and for cC relative to iC trials (t_{195} = -15.76, 95% CI [-0.11, -0.08], p < 0.001). Additionally, mean RTs were shorter for iI compared to cI (t_{195} = -24.24, 95% CI (Confidence Interval) [-0.05, -0.04], p < 0.001) and for cC relative to iC (t_{195} = -32.48, 95% CI [-0.08, -0.07], p < 0.001) trials.



Supplemental Figure A.III.S3.

Supplemental Figure A.III.S3. *Demographics and self-reported psychopathology spread.*

(A) *Age, IQ and psychiatric symptoms score distributions across participants.*

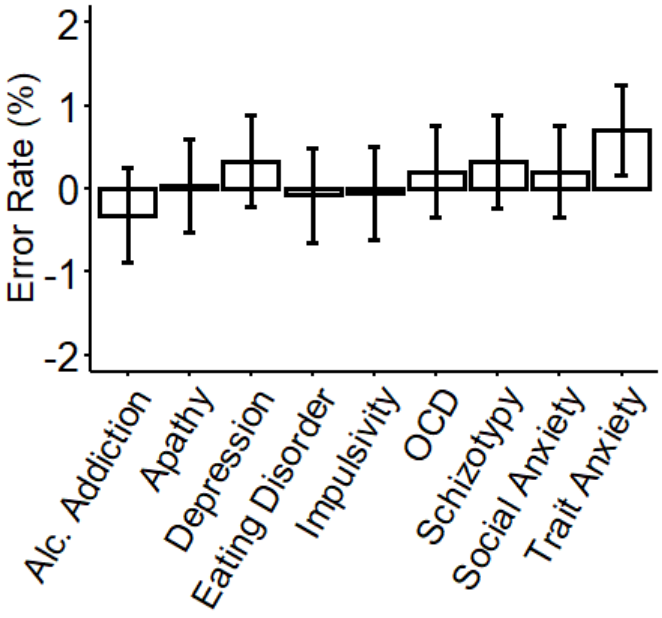
(B) *Correlation matrix of mean scores of the nine pyschiatric questionnaires or transformed dimension scores (AD: anxious-depression, CIT: compulsive behaviour and intrusive thought, SW: social withdrawal), including age and IQ. Colour scale indicates correlation coefficient, size of colour patch indicates significance.*

	Alcohol Addiction	Apathy	Depression	Eating Disorder	Impulsivity	OCD	Schizotypy	Social Anxiety	Trait Anxiety
Alcohol Addiction									
Apathy	0.14								
Depression	0.09	0.62							
Eating Disorder	0.10	0.06	0.22						
Impulsivity	0.26	0.39	0.27	0.14					
OCD	-0.05	0.23	0.39	0.25	0.18				
Schizotypy	0.08	0.46	0.60	0.27	0.51	0.49			
Social Anxiety	-0.03	0.30	0.42	0.20	0.16	0.43	0.48		
Trait Anxiety	0.05	0.51	0.75	0.16	0.27	0.45	0.61	0.46	
Reliability	0.87	0.86	0.85	0.86	0.81	0.89	0.85	0.92	0.93

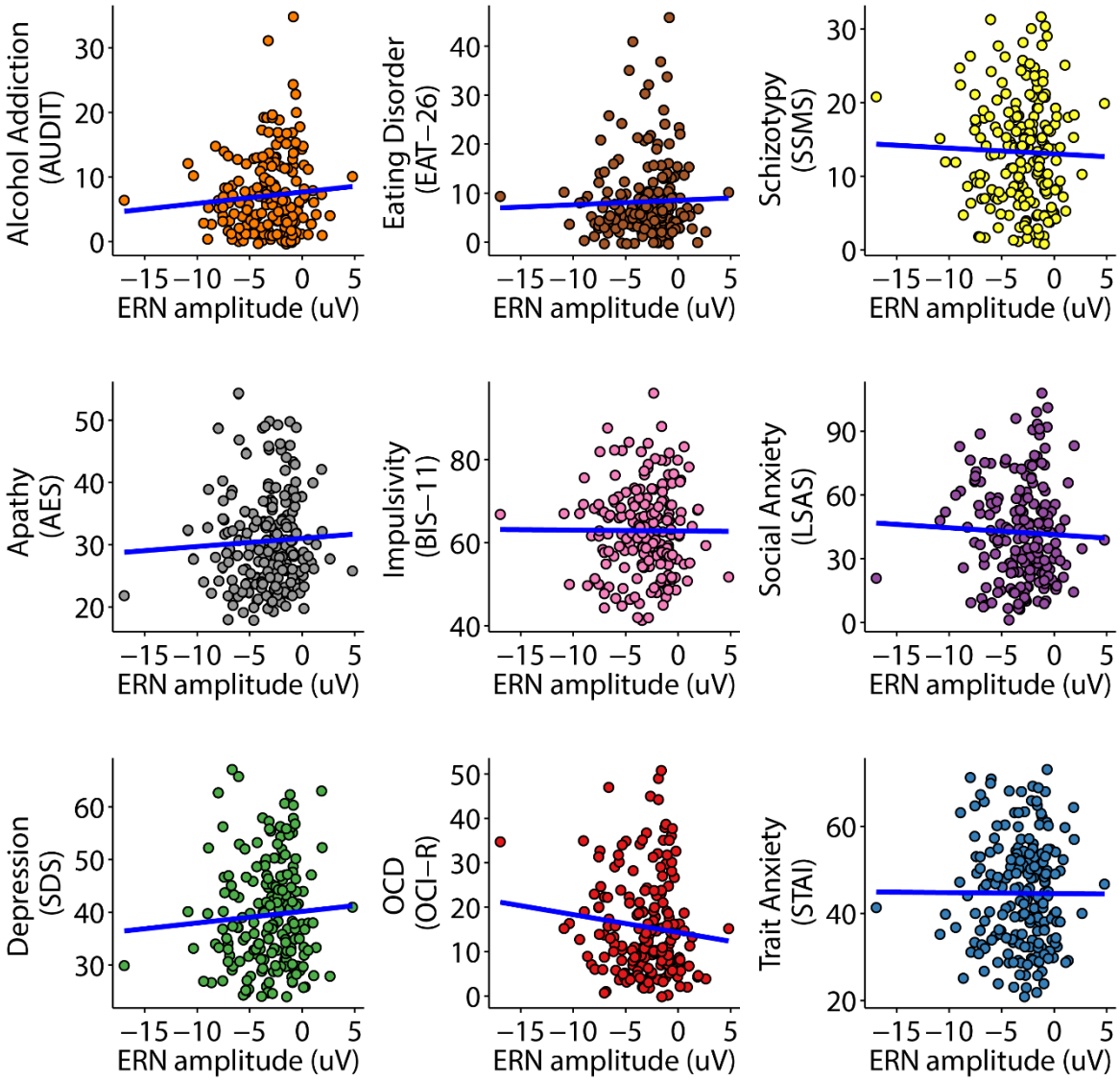
Supplemental Table A.III.S1. *Pearson's correlations between total questionnaire scores. Cronbach's Alpha was used to calculate reliability for questionnaires.*

	AD	CIT	SW
AD			
CIT	0.33		
SW	0.38	0.39	

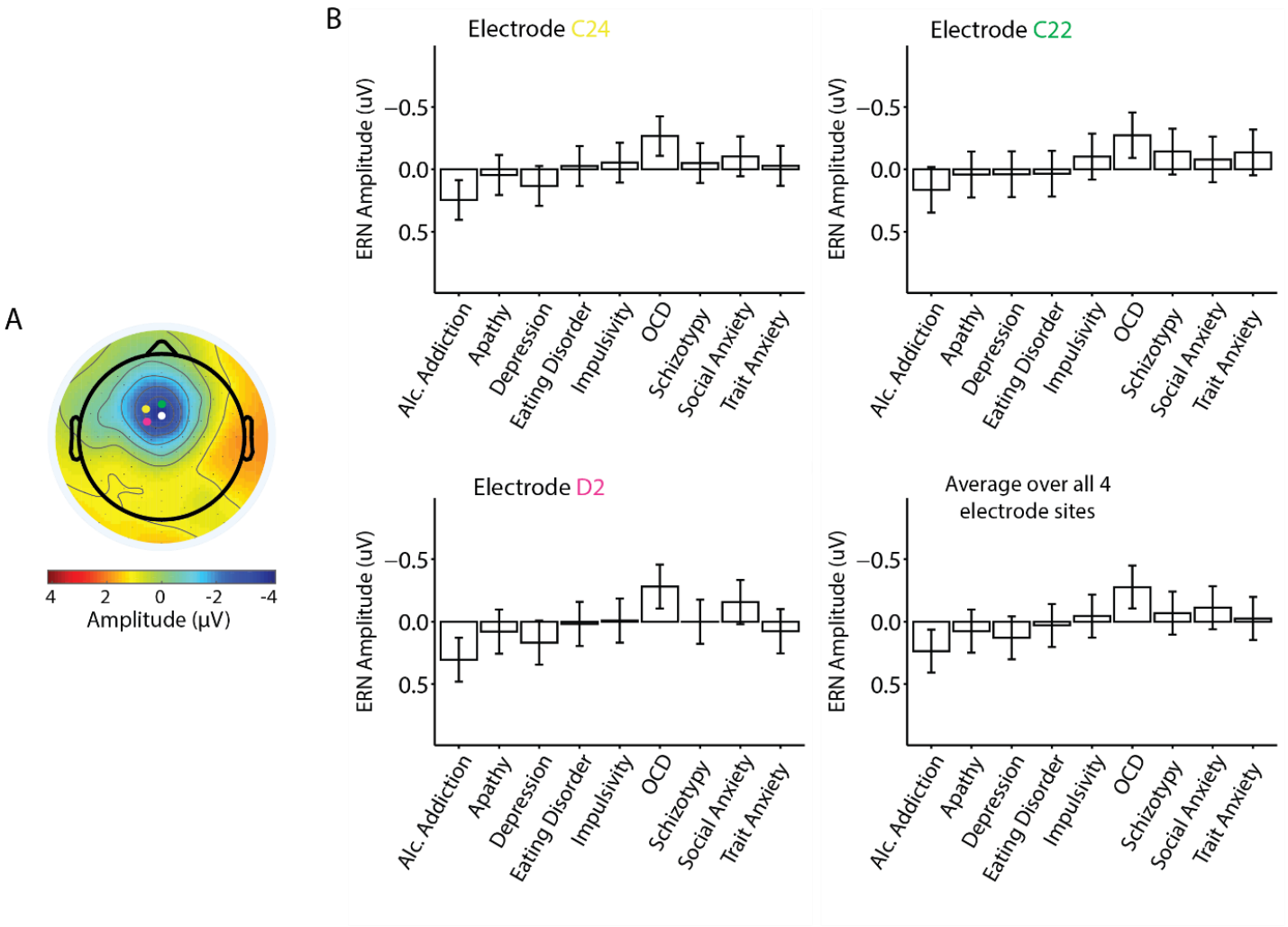
Supplemental Table A.III.S2. *Pearson's correlations between transdiagnostic dimensions scores (AD: 'anxious-depression', CIT: 'compulsive behaviour and intrusive thought', SW: 'social withdrawal').*



Supplemental Figure A.III.S4. *Associations between questionnaire scores with mean error rate (%). Error bars denote standard errors. The Y-axes indicate the change in error rate as a function of 1 standard deviation (SD) increase of questionnaire scores. No questionnaire score was significantly associated to changes in error rate.*



Supplemental Figure A.III.S5. *Scatterplots of ERN amplitude and total questionnaire scores. Coloured markers represent an individual's total score for the corresponding questionnaire. See Figure 2 and Table 1. We note that a possible outlier appears to exist (ERN amplitude > -15 uV), but when the data point is removed, the associations between questionnaire scores and the ERN do not statistically differ from the original correlations (William's test of correlation difference: all z < 0.40 , all p > 0.69).*

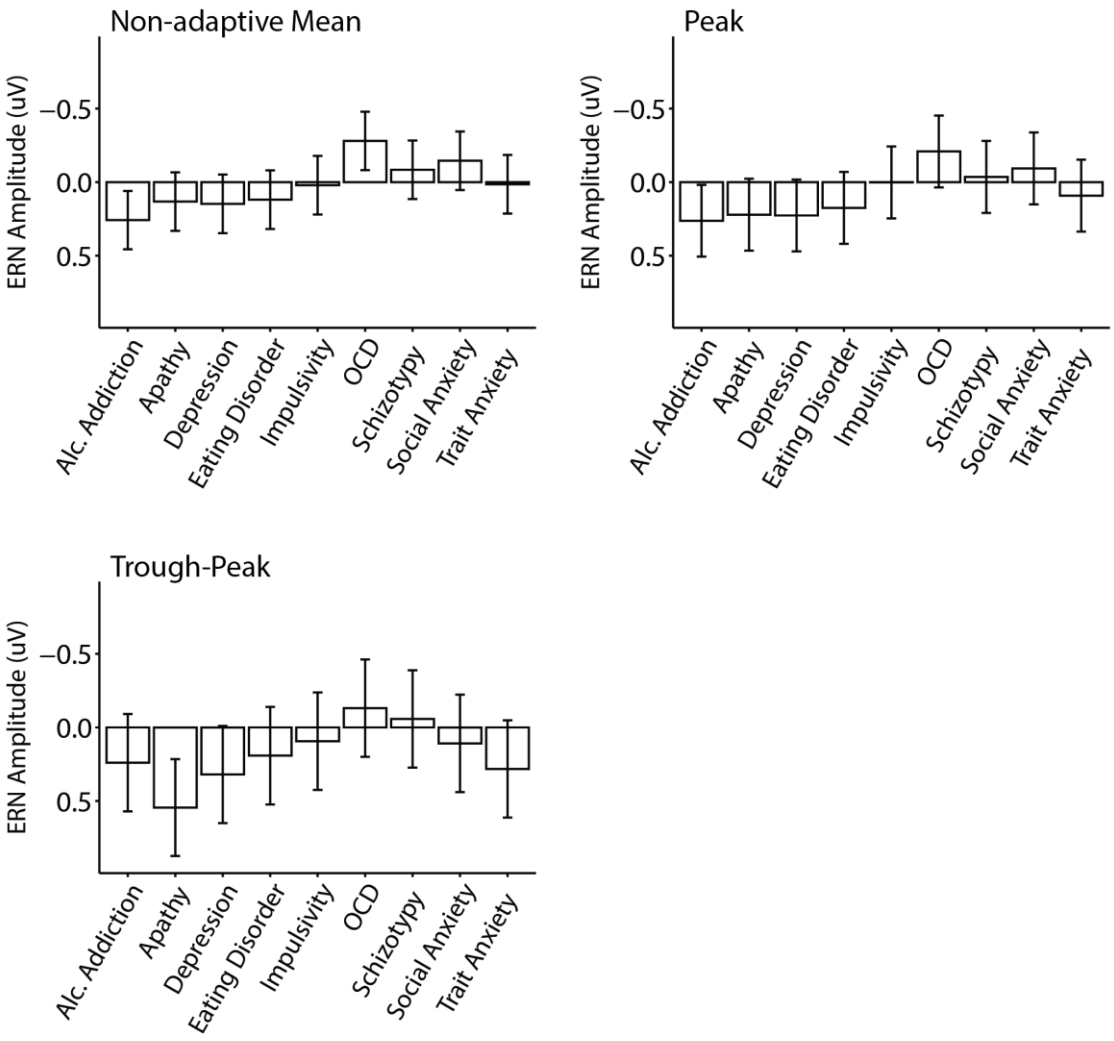


Supplemental Figure A.III.S6.

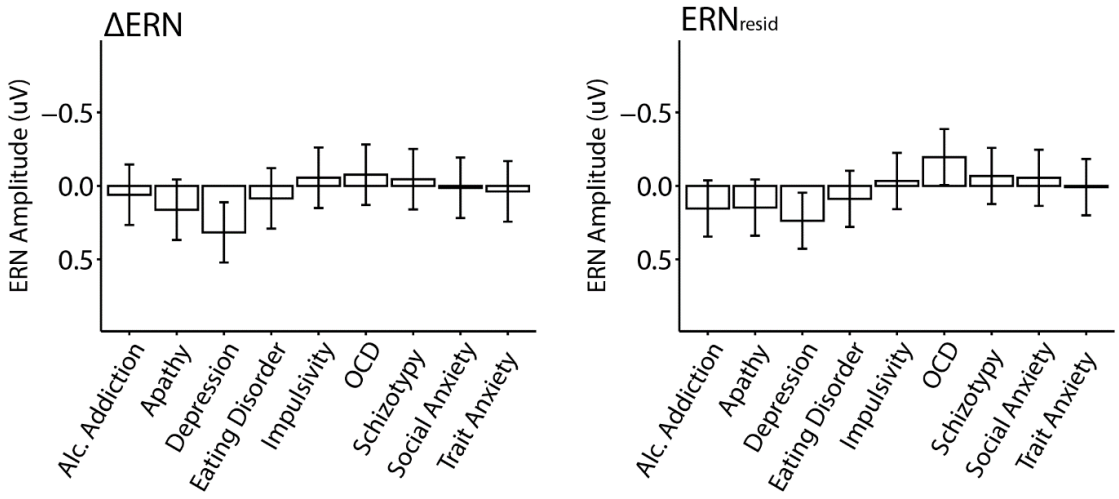
Supplemental Figure A.III.S6. *ERN quantification at various electrode sites with the adaptive mean method.*

(A) *Scalp map displays the voltage distribution at 37.61ms, the average latency of the most negative peak. Coloured dots indicate electrode positions around ERN peak; FCz: white, C22: green, C24: yellow, D2: pink.*

(B) *Associations between questionnaire total scores with ERN amplitude quantified at various electrode sites. Error bars denote standard errors. The Y-axes indicates the change in in ERN amplitude as a function of 1 SD increase of questionnaire scores.*



Supplemental Figure A.III.S7. *Associations between psychiatric symptoms with ERN amplitude quantified by various methods at electrode FCz. Error bars denote standard errors. The Y-axes indicate the change in ERN amplitude as a function of 1 SD increase of questionnaire scores.*



Supplemental Figure A.III.S8. *Associations between psychiatric symptoms with ΔERN and ERN$_{resid}$ amplitude. Error bars denote standard errors. The Y-axes indicate the change in ERN amplitude as a function of 1 SD increase of questionnaire scores. See* ***Table A.III.S3****.*

	ΔERN			***ERN$_{resid}$***		
Psychiatric Questionnaire	*β (SE)*	*z-value*	*p-value*	*β (SE)*	*z-value*	*p-value*
Alcohol Addiction	0.06 (0.20)	0.29	0.77	0.15	0.81	0.42
Apathy	0.16 (0.20)	0.79	0.43	0.15	0.77	0.44
Depression	0.32 (0.20)	1.55	0.12	0.24	1.24	0.21
Eating Disorder	0.09 (0.20)	0.42	0.68	0.09	0.46	0.64
Impulsivity	-0.06 (0.20)	-0.27	0.79	-0.03	-0.17	0.86
OCD	-0.08 (0.20)	-0.37	0.71	-0.20	-1.02	0.31
Schizotypy	-0.05 (0.20)	-0.22	0.83	-0.07	-0.35	0.73
Social Anxiety	0.01 (0.20)	0.07	0.95	-0.05	-0.29	0.77
Trait Anxiety	0.04 (0.20)	0.18	0.86	0.009	0.04	0.96
Transdiagnostic Dimension	*β (SE)*	*t-value*	*p-value*	*β (SE)*	*t-value*	*p-value*
Anxious-depression	0.11 (0.23)	0.49	0.62	0.16	0.76	0.45
Compulsive behaviour and intrusive thought	-0.01 (0.22)	-0.06	0.95	-0.06	-0.29	0.77
Social withdrawal	-0.02 (0.24)	-0.10	0.92	-0.11	-0.51	0.61

Supplemental Table A.III.S3.

Supplemental Table A.III.S3. *Associations between ΔERN or* ERN_{resid} *amplitude with total scores of self-report psychiatric questionnaires or transdiagnostic dimensions. SE = standard error. For psychiatric questionnaires, each row reflects the (uncorrected for multiple comparisons) results from an independent analysis where each psychiatric questionnaire score was regressed against ERN amplitude. For transdiagnostic dimensions, all three dimensions scores were included in the same regression model. See* ***Figure A.III.S8****.*

Disorder	*Diagnosis*
Mood disorders	
Major depressive disorder	18
Suicide behavior disorder	1
Bipolar Disorder	1
Anxiety disorders	
Panic Disorder	12
Agoraphobia	4
Generalised Anxiety Disorder	15
Social Anxiety Disorder	11
Obsessive-Compulsive Disorder	5
Posttraumatic Stress Disorder	0
Substance use disorders	
Alcohol Use Disorder	7
Substance Use Disorder (Non-alcohol)	9
Psychotic Disorders	
Psychotic Disorders	0
Eating Disorders	
Anorexia Nervosa	0
Bulimia Nervosa	1
Binge-Eating Disorder	4
Other disorders	
Antisocial Personality Disorder	1

Supplemental Table A.III.S4. *Mini International Neuropsychiatric interview (M.I.N.I.) diagnostic information summary for participants who presently met the criteria for at least one DSM-V disorder (N = 38).*