

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

Metacognition: A Transdiagnostic Mechanism of Psychopathology

Celine Fox

School of Psychology and Trinity College Institute of Neuroscience

Trinity College Dublin, the University of Dublin

A thesis submitted to

Trinity College Dublin, the University of Dublin

For the degree of

Doctor of Philosophy

2024

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.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).



Signed: Celine Fox
Date: 29/07/2024

Summary

There has been slow progress in understanding precise cognitive mechanisms underlying mental illness. For example, metacognition, which reflects one's confidence in decision-making, has been implicated in a broad range of psychiatric disorders. Yet, the overall evidence of confidence disturbances in psychiatric conditions has been ambiguous, inconsistent and non-specific (Hoven et al., 2019). Halted progress in mechanistic research can be partly attributed to an overreliance on small-sampled, cross-sectional and case-controlled studies that defined mental illness in accordance with traditional diagnostic taxonomies. The experiments in this thesis aimed to resolve these issues, by measuring metacognition through time and considering mental illness from a transdiagnostic, dimensional perspective. By factor analysing a wide variety of psychiatric symptoms, transdiagnostic dimensions that cut across traditional disorder boundaries can be derived (Gillan et al., 2016). The aims of this thesis were to 1) investigate the association between metacognition and transdiagnostic psychiatric dimensions among clinical patients, 2) test if metacognitive abilities change with improvements in transdiagnostic dimensions, 3) develop a gamified smartphone task that facilitates valid and reliable measurement of metacognition over time and, 4) derive transdiagnostic dimensions of psychopathology that maximally correspond to individual differences in metacognition. To achieve this, the experimental chapters of this thesis leveraged online and smartphone-based research methods, which accelerate data collection both within and across individuals at scale.

Chapter 2 investigated the association between metacognitive abilities and transdiagnostic psychiatric dimensions among individuals initiating internet-based cognitive behavioural therapy. Metacognitive bias (mean confidence across trials, controlling for accuracy), was associated with transdiagnostic dimensions among this large sample of treatment-seeking patients. Specifically, those with higher levels of 'Anxious-depression' had lower confidence, while those with higher levels of 'Compulsivity and Intrusive Thought' had elevated confidence. This

replicated prior evidence of specific and bi-directional associations between metacognitive bias and transdiagnostic dimensions observed in general population samples (Benwell et al., 2022; Hoven et al., 2023; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). However, this cross-sectional research does not reveal if metacognitive bias is a trait-like marker of vulnerability to psychopathology, or if it resolves when symptoms improve. Measuring metacognition and transdiagnostic dimensions in this patient sample again four weeks later, Anxious-depression reduced from baseline to follow-up, and this was accompanied by a significant increase in metacognitive bias. Importantly, these changes were correlated; those with the greatest reductions in Anxious-depression levels had the largest increase in confidence in performance. Examining smaller comparison groups of individuals who either initiated antidepressant medication treatment or underwent no treatment, results suggested that the state-dependence of confidence was unlikely to be treatment-specific. Rather, confidence increased in concert with a reduction in levels of Anxious-depression. These data suggest the negative biases that characterise Anxious-depression are therefore state-dependent and might be normalised through clinical improvements.

To facilitate dense sampling of metacognition across time, Chapter 3 aimed to assess the psychometric properties of Meta Mind, a gamified smartphone task designed for valid, reliable and remote measurement of confidence with as few trials as possible. The validity and reliability of Meta Mind were investigated across two studies. In the first study, paid participants played Meta Mind and completed a traditional metacognitive task. Confidence was positively correlated across tasks, which indicated that Meta Mind had acceptable convergent validity. Additionally, the split-half reliability of confidence was excellent in Meta Mind. In the second study, over 3000 citizen scientists played Meta Mind and reported levels of transdiagnostic dimensions. The bi-directional associations between metacognitive bias with Anxious-depression and Compulsivity and Intrusive Thought could be detected in as few as 40 trials. Additionally, the split-half reliability of Meta Mind stabilised after 40 trials. Across sessions, the test-retest reliability of metacognitive bias was very high and there were no significant practice effects with Meta Mind. These results indicated that Meta Mind is valid

and reliable, allowing for precise and rapid measurement of metacognitive bias. Therefore, Meta Mind is positioned as a viable tool for future research aiming to investigate the temporal dependence of metacognitive bias and psychopathology.

While Chapters 2 and 3 provide further support that metacognitive bias is bi-directionally linked to Anxious-depression (underconfidence) and Compulsivity and Intrusive Thought (overconfidence), this factor structure of psychopathology was derived to explain covariance in psychiatric symptoms, not cognition. Factors based on symptom correlations are purely descriptive and do not necessarily have mechanistic grounding. This issue is further compounded by extensive analytic freedom when conducting factor analysis to derive psychiatric dimensions. It is therefore plausible that an alternative factor solution potentially maps more strongly onto individual differences in metacognition. The final experiment, Chapter 4, tested this idea empirically. In Chapter 4, large-scale multiverse analysis was conducted to test the mechanistic validity of thousands of competing factor solutions, operationalised as the strength of their association with different aspects of cognition, including metacognitive bias. Using nine datasets comprising N=7565 individuals, we found that the latent structure underlying the covariation of psychiatric symptoms (i.e., the factor structure that included Anxious-depression and Compulsivity and Intrusive Thought) performed better than thousands of engineered alternatives, including performing twice as well as traditional questionnaire total scores. Moreover, we showed that factor analysis at the descriptive level performed equivalently to a supervised approach to derive factors based on cognitive performance. These results suggest that factor solutions generated based on the correlation of experienced symptoms are clinically sensible and cognitively meaningful. Unsupervised factor analysis of psychiatric symptoms represents a potentially powerful framework for determining specific cognitive mechanisms of mental illness.

List of Publications and Presentations

This thesis incorporates material already published or currently under review in the following manuscripts:

- **Fox, C. A., Teckentrup, V., Donegan, K. R., Seow, T. X. F., Benwell, C., Tervo-Clemmens, B., & Gillan, C. (2024).** Mechanistic arbitration between candidate dimensions of psychopathology. *PsyArXiv.* https://doi.org/10.31234/osf.io/7waey
- **Fox, C. A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R. J., Hanlon, A. K., Gallagher, E., Rouault, M., & Gillan, C. M. (2024).** Reliable, rapid, and remote measurement of metacognitive bias. *Scientific Reports*, *14*(1), 14941. https://doi.org/10.1038/s41598-024-64900-0
- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X. F., Lynch, K., Harty, S., Richards, D., Palacios, J., O'Keane, V., Stephan, K. E., & Gillan, C. M. (2023).** An observational treatment study of metacognition in Anxious-depression. *eLife*, *12*, RP87193. https://doi.org/10.7554/eLife.87193

The following are presentations arising from this work:

- **Fox, C. A., Teckentrup, V., Donegan, K. R., Seow, T. X. F., Benwell, C., Tervo-Clemmens, B., & Gillan, C. M.** Mechanistic arbitration between candidate dimensions of psychopathology. Poster presentation at the Computational Psychiatry Conference, University of Minnesota, USA, 2024.
- **Fox, C. A., Teckentrup, V., Donegan K. R., & Gillan, C. M.** Data-driven computational factor modelling of psychopathology: Optimising the explanation of individual differences. Poster presentation at the School of Psychology Research Symposium, Trinity College Dublin, Ireland, 2024.
- **Fox, C. A., Teckentrup, V., Donegan K. R., & Gillan, C. M.** Mapping transdiagnostic psychiatric dimensions onto individual differences in cognition: A data-driven factor modelling approach. Invited oral presentation for the Translational Science Methods Club at the Alan Turing Institute, virtual, 2024.

- **Fox, C. A. & Senta, J.** The X Factor: Identifying neurocomputational mechanisms underlying transdiagnostic psychiatric dimensions using factor modelling. Invited oral presentation for the Predictive Brain Health Modelling Group, Trinity College Dublin, Ireland, 2023.
- **Fox, C. A., McDonogh, A., Donegan, K. R., Tighe D., Hanlon, A. K., & Gillan, C. M.** Validation of a gamified smartphone assessment of metacognition: Associations with transdiagnostic psychiatric dimensions. Poster presentation at the Society of Biological Psychiatry Annual Meeting, Hilton San Diego, USA, 2023.
- **Fox, C. A., McDonogh, A., Donegan, K. R., Tighe D., Hanlon, A. K., & Gillan, C. M.** Validation of a gamified smartphone assessment of metacognition: Associations with transdiagnostic psychiatric dimensions. Poster presentation at the Neuroscience Society of Ireland Conference, Trinity College Dublin, Ireland, 2022.
- **Fox, C. A & Gillan, C. M.** A dynamic network approach to understanding compulsivity. Invited oral presentation at the Icahn School of Medicine at Mount Sinai, New York, USA, 2022.
- **Fox, C. A., McDonogh, A., Donegan, K. R., Tighe D., Hanlon, A. K., & Gillan, C. M.** Meta Mind: A gamified smartphone assessment of metacognition. Invited oral presentation at the Computations of Metacognition in Mental Health Workshop, University College London, UK, 2022.
- **Fox, C. A., McDonogh, A., Donegan, K. R., Tighe D., Hanlon, A. K., & Gillan, C. M.** Validation of a gamified smartphone assessment of metacognition: Associations with transdiagnostic psychiatric dimensions. Poster presentation at the British Association of Psychopharmacology, Imperial College London, UK, 2022.
- **Fox, C. A. & Gillan, C. M.** Citizen Neuroscience: A digital revolution with smartphone-based research. Invited oral presentation at the ADAPT Research Centre, Trinity College Dublin, Ireland, 2022.
- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., ... & Gillan, C. M.** Metacognitive biases in Anxious-depression improve with therapeutic gains: evidence from a 4-week observational study of iCBT and antidepressant treatment. Poster presentation at the MQ Mental Health Summit, virtual, 2021.

- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., ... & Gillan, C. M.** Do metacognitive biases in Anxious-depression improve with therapeutic gains? Evidence from a 4-week observational study of iCBT and antidepressant treatment. Invited oral presentation at the Transdiagnostic Approaches to Mental Health International Online Conference, virtual, 2021.
- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., ... & Gillan, C. M.** Do metacognitive biases in Anxious-depression improve with therapeutic gains? Evidence from a 4-week observational study of iCBT and antidepressant treatment. Invited oral presentation for the Neuroscience Ireland 2021: Winners' Spotlight by Ulysses Ireland, virtual, 2021.
- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., ... & Gillan, C. M.** Do metacognitive biases in Anxious-depression improve with therapeutic gains? Evidence from a 4-week observational study of iCBT and antidepressant treatment. Poster presentation at the Neuroscience Society of Ireland Conference, virtual, 2021.
- **Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., ... & Gillan, C. M.** Are metacognitive biases in Anxious-depression ameliorated following successful depression treatment? A longitudinal, observational study. Poster presentation at the Society of Biological Psychiatry Annual Meeting, virtual, 2021.

Acknowledgements

No woman is an island. This thesis would not have been possible without all the wonderful people that have surrounded me over the past four years. Many, many thanks are in order:

First and foremost, thank you to my supervisor, Prof. Claire Gillan. Claire, I could not have asked for a better supervisor. Thank you for all your guidance, patience and encouragement. You have been extremely generous with your time, and I always knew you were in my corner. Your fierce passion for mental health research inspired me every single day. Thank you for allowing me space to explore my curiosity, while always keeping me on a clear path. Although my time as a PhD student has come to an end, I am grateful to have you as a mentor for years to come.

Thank you to the board of examiners, Prof. Redmond O'Connell and Prof. Micah Allen, for your expertise and appraisal of this thesis. I would like to thank my collaborators, Prof. Naomi Fineberg and Dr. Laura Berner, for providing an invaluable clinical perspective on this research. Thank you to Eoghan Gallagher and Ana Petre, for tirelessly working to bring Meta Mind and Pattern Wise to life. I am very grateful to the Irish Research Council for financially supporting me through the Government of Ireland Postgraduate Scholarship.

Beginning my PhD journey in 2020 during the COVID-19 pandemic was a challenging experience, marked by isolation and solitude. However, as the vaccine became available, I was finally able to return to the lab. Working on campus alongside my lab mates transformed my PhD experience in profound ways. Thank you so much to the members of the Gillan lab, past and present, for creating such a joyous working environment. It has been such a pleasure working with you all. Special gratitude goes to Anna Hanlon, Kelly Donegan, Parnian Rafei, Anna Rosická, Vanessa Teckentrup, Judith Milner, Alice Cox and Sarah Moran for your tremendous help, especially in the last few months. Thank you

also to all the brilliant and kind people at the Trinity College Institute of Neuroscience.

Thank you to my teammates, Aoife Costigan, Ellen Tuck and Francesca Matrá for the lovely chats off the pitch. We did it!

Thank you to my dearest friends, Michelle McLoughlin, Lauren McConway, Elaine Doyle, Amy Holdcroft, Zoë Renshaw and Richáel Farrell Courtney. Even though we haven't seen each other as much as I would like recently, it's great to know that I have friends who are always there for me.

Thank you to Conor Fox. I am very lucky to not only call you my brother, but also one of my closest friends. Your dry sense of humour always makes me laugh, even on the toughest days.

Grazie di tutto, Moreno Stellini. You have been an unwavering source of support, strength and stability. My gratitude to you is ineffable. Ti amo.

Last, but not least, thank you Mam and Dad. I am so incredibly grateful for everything you have done for me. Thank you for encouraging to do whatever makes me happy and for being there to catch me when I fall. You have always celebrated me and loved me unconditionally, and for that I am forever grateful. I dedicate this thesis to you both. I hope that it makes you proud.

Table of Contents

Declaration 2

Summary 3

List of Publications and Presentations 6

Acknowledgements 9

Table of Contents 11

Chapter 1 – General Introduction 13

1.1 Metacognition and confidence 14
1.2 The traditional, disorder-based classification system 17
1.3 Transdiagnostic dimensional approaches to psychopathology 23
1.4 Metacognition: A mechanism of transdiagnostic mental illness 28
1.5 The state-dependence of metacognition 30
1.6 Embracing online research in mental health 32
1.7 Optimising the transdiagnostic dimensional space 37
1.8 Experiments and hypotheses 47

Chapter 2 – An Observational Treatment Study of the State-dependence of Metacognition in Anxious-depression 50

2.1 Introduction 50
2.2 Methods 52
2.3 Results 58
2.4 Discussion 64

Chapter 3 – Reliable, Rapid, and Remote Measurement of Metacognitive Bias 68

3.1 Introduction 68
3.2 Methods 71
3.3 Results 83
3.4 Discussion 94

Chapter 4 – Mechanistic Arbitration Between Candidate Dimensions of Psychopathology 101

4.1 Introduction 101
4.2 Methods 104
4.3 Results 113
4.4 Discussion 125

Chapter 5 – General Discussion 130

5.1 Summary 130
5.2 Strengths, limitations and future directions 135
5.3 Clinical implications 151

5.4 Conclusions 159

6 – References 160

7 – Supplementary Material 186

7.1 Supplementary material for Chapter 2 186
7.2 Supplementary material for Chapter 3 197
7.3 Supplementary material for Chapter 4 202

Chapter 1 – General Introduction

Cognitive dysfunction is a pervasive, pathological feature of mental illness. Yet, precise cognitive mechanisms underlying disorders remain elusive (Doyle et al., 2018; East-Richard et al., 2020; Wright et al., 2014). Despite decades of research from thousands of independent studies, no disorder-specific cognitive markers have been identified with adequate specificity, sensitivity, or predictive validity, to have sufficient diagnostic or therapeutic utility (Abi-Dargham et al., 2023; Abramovitch et al., 2021; Tiego et al., 2023). This slow progress partly reflects an overreliance on antiquated research practices, including a strong allegiance to small sampled and case-controlled studies, which narrowly focus on cross-sectional evidence obtained from cognitive tasks without established reliability. By tackling shortcomings associated with the traditional research model, this thesis aims to uncover precise cognitive mechanisms underlying psychopathology.

This thesis will focus specifically on metacognition, a central facet of cognitive functioning that has been implicated across a broad range of psychiatric disorders (Hoven et al., 2019; Philipp et al., 2020; Sun et al., 2017). By employing complementary methods and large datasets in this thesis, metacognition is evaluated through time and through treatment. An enhanced understanding of metacognitive disturbances in mental illness is then leveraged to reshape how we define psychopathology entirely. To outline the general introduction, metacognition is firstly defined and described, with emphasis on the limited understanding of its role in mental illness. Following this, novel approaches in mental health science are outlined, including transdiagnostic psychiatry, online and smartphone-based research methods, the citizen science framework and longitudinal, within-person monitoring. The importance of task validity and reliability is also discussed, as critical prerequisites for individual differences research. This thesis employs a transdiagnostic, dimensional perspective on psychopathology throughout. As such, the final section of the general introduction reflects on how transdiagnostic dimensions are generated using factor analysis, and the implications of this for mechanistic research.

1.1 Metacognition and confidence

Imagine you are preparing for your driving test. Coming up to the test, you may take many lessons and practise driving repeatedly to ensure you are sufficiently skilled. When you believe that you are ready to take the test, you may feel confident that you do not need any additional lessons or practice. During the driving test itself, you need to make a variety of decisions. When you come to a red light, you confidently stop the car. When the traffic light turns green, you then confidently start driving again. In other instances, you might be much less confident. For example, you stop at a yield sign, but you are unsure whether it is safe to procede. You may also decide not to overtake a car because you lack confidence that the other side of the road is clear. Therefore, your decions-making and behaviour before and during the test is guided by the confidence ascribed to your choices. While waiting for the test results, you may think back over all the decisions that you made during test, feeling confident that you passed. This capacity to evaluate and reflect upon your beliefs, decisions and actions is known as metacognition.

Metacognition is considered a central facet of human consciousness (Heyes et al., 2020). In daily life, we are often not provided with any objective feedback on decision-making. This absence of objective feedback on decision-making leads to a heavy reliance on metacognition, through which we continuously monitor our moment-to-moment decisions (Fleming & Lau, 2014), abilities (Rouault et al., 2019) and self-worth (Seow, Rouault, et al., 2021). This capacity to introspect is highly adaptive, supporting successful learning (Guggenmos et al., 2016) and informing appropriate behaviour (Fleming & Daw, 2017) and communication in social interactions (Fleming et al., 2012). Therefore, impaired metacognition can give rise to aberrant decision-making and behaviour (Dar et al., 2022; Hoven et al., 2019). Returning to the driving test anecdote, underconfidence that it is safe to proceed at a yield sign could result in you failing the test for a lack of progression. However, overconfidence that the road is clear, despite poor visibility, could cause a deadly car collision.

As alluded to, metacognition can be interrogated with retrospective confidence judgements (Fleming, 2024). Confidence refers to the subjective belief that a choice or action is accurate, given the available evidence (Pouget et al., 2016). While metacognition operates across cognitive domains, perceptual decision-making tasks have become a staple behavioural paradigm for measuring confidence (Rahnev, 2021). In commonly employed metacognitive task, participants make an innocuous perceptual judgement (e.g., which of two stimuli contains more dots), and then subsequently rate their confidence in that judgement. Using repeated confidence ratings across multiple decisions, related facets of metacognition can be derived: bias and efficiency (Xue et al., 2021). Metacognitive bias refers to the tendency to give high or low confidence ratings on average, which is typically indexed as mean confidence across correct and incorrect trials (Fleming & Lau, 2014). Compared to metacognitive bias, calculating efficiency is relatively more complex. Metacognitive efficiency is the extent to which confidence judgements map onto decision accuracy (Fleming & Lau, 2014). While there are many alternative ways to measure metacognitive efficiency, the gold standard is M-Ratio (Rahnev, 2023). M-Ratio leverages signal detection theory to calculate metacognitive sensitivity (meta-d'; the extent to which confidence judgements distinguish between correct and incorrect responses) and task performance sensitivity (d') (Fleming & Lau, 2014). In signal detection theory, d' indexes an observer's hit rate against false alarm rate, quantifying how well an observer distinguishes between signal (a correct response) and noise (an incorrect response). Meta-d' extends this concept to confidence judgements, representing the expected d' given the observed hit and false alarm rates across different confidence ratings (Maniscalco & Lau, 2012). The ratio of meta-d' to d' provides an estimate of metacognitive efficiency, offering insights into how efficiently an observer uses performance accuracy to inform metacognitive judgements (Fleming & Lau, 2014).

Studies have linked metacognitive bias and efficiency to a variety of individual differences. Across different cognitive domains, females demonstrated lower confidence, but had greater metacognitive efficiency relative to males (Hoogervorst et al., 2024). With age, a decrease in metacognitive bias indicated that confidence is lower in later life (McWilliams et al., 2023). There was

conflicting evidence that metacognitive efficiency declines (Palmer et al., 2014) and remains stable (McWilliams et al., 2023) across the lifespan, which may reflect a reliance of studies on cross-sectional effects. When task accuracy is matched, individuals with higher intelligence tended to have lower levels of confidence (Rouault et al., 2018). Examining polarizing socio-political attitudes, individuals with more radical political beliefs had lower metacognitive sensitivity and were less likely to update their confidence with post-decision evidence (Rollwage et al., 2018). Similar patterns of reduced metacognitive sensitivity and confidence updating with evidence accumulation were observed among those who endorsed greater climate change scepticism (De Beukelaer et al., 2023). This demonstrates how metacognitive abilities on low-level perceptual tasks have implications for individual differences in age, gender, intelligence and political and scientific decision-making.

Examining individual differences in mental illness, confidence disturbances have been reported across a wide range of psychiatric disorders (Hoven et al., 2019). For example, obsessive compulsive disorder, classically described as a disorder of doubt, is characterised by underconfidence on memory and perception tasks (Dar et al., 2022). Pathological underconfidence may manifest as self-doubt and obsessions, fuelling compulsive behaviours (Hoven, Rouault, et al., 2023). As a simplified example, a patient with OCD may lack confidence that a door was properly closed, despite a clear memory of locking the door, which could lead to intrusive thoughts about home invasion and repeated door checking behaviour. However, underconfidence in decision-making is not a unique feature of OCD. Low confidence has been reported across multiple other disorders, including depression, anxiety, schizophrenia and addiction (Hoven et al., 2019). Within disorder categories, there is opposing evidence for the direction of confidence effects. For example, overconfidence was associated with schizophrenia in some studies, while others reported evidence of underconfidence or null confidence effects in schizophrenia (Hoven et al., 2019). Overall, the reported effects of metacognitive disturbances in psychiatric disorders are non-specific, ambiguous and inconsistent.

1.2 The traditional, disorder-based classification system

Most studies on metacognition and mental health adopted a case-control study design, in which diagnosed patients were compared to 'healthy' controls. In these studies, psychiatric disorders are typically defined in terms of discrete, diagnostic categories, in accordance with the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2022) and the International Statistical Classification of Diseases (ICD; World Health Organization, 2019). Defining psychopathology in terms of discrete disorders is not unique to the metacognitive literature, as this approach is adopted broadly in cognitive science and psychiatry. A key tenet of both the DSM and ICD frameworks is that disorders are discontinuous, and can be differentiated based on diagnostic signs and symptoms (Lilienfeld & Treadway, 2016). The presence or absence of a specific disorder (e.g., major depressive disorder, OCD or schizophrenia) is determined by a number of pre-defined criteria, and experiences of distress or impairment along those criteria (Stein et al., 2021). Additionally, the specific symptoms that arbitrate the presence or absence of a disorder are separate to the symptom criteria for other disorder categories (e.g., symptom criteria for schizophrenia are distinct from criteria for OCD). While there is much overlap in disorder categories between the DSM and ICD, the DSM is predominantly used in research and for highly trained mental health professionals, while the ICD is adopted more often in global clinical settings (Clark et al., 2017; Rief et al., 2023).

This diagnostic framework was developed with the primary goal of providing a shared language for communication between clinicians and their patients (First et al., 2015). Since the mid-twentieth century, the DSM in particular has dominated psychiatric research and public discourse (Lafrance & McKenzie-Mohr, 2013). The DSM not only informs core texts in psychiatry but majorly influences insurance, government and legal policies, pharmaceutical industry practices, art and literature (Dalgleish et al., 2020). Indeed, it is widely assumed that "major diagnostic categories are obvious and even 'natural'" (Kendler, 2009, p.1939). There are many factors that have contributed to this global reign. Firstly, socio-political forces have majorly influenced the development of the diagnostic approach. These influences include diverse interests external to the field of

psychiatry, such as insurance, pharmaceutical, legal and political powers, in addition to specific cultural events (e.g., deinstitutionalisation) (Strand, 2011). Additionally, the diagnostic system is steeped in specific historical contingencies. To illustrate this, key features of the criteria for major depression published in the DSM-III were heavily reliant on prior criteria developed by Cassidy and colleagues in 1957 (Cassidy et al., 1957). When Cassidy was asked how he decided on the original threshold of six out of 10 criteria for major depression, he replied, “It sounded about right” (Kendler et al., 2010, p.136). Indeed, clinical intuition and expert consensus have informed the established disorder criteria, with less reliance on systematic evidence-based approaches (Kendler & Solomon, 2016).

This diagnostic classification system has been passively accepted and reified, despite a lack of grounding in empirical scientific evidence (Hyman, 2010). In reality, the complex landscape of mental illness does not ascribe to this convenient framework. As outlined below, disorder categories are characterised by arbitrary thresholds for binarisation, poor reliability, high rates of comorbidity, shared symptomology across disorders and symptom heterogeneity within disorders.

1.2.1 Arbitrary thresholds for disorder diagnosis

Thresholds for binarisation are used to demarcate the presence or absence of disorders (e.g., requiring the presence of at least five of nine diagnostic criteria for 2 weeks or longer to receive a DSM-5 diagnosis of major depression) (American Psychiatric Association, 2022). If psychopathology is truly taxonomic in nature, then real disorder boundaries would exist. Among over 12,000 individuals with major depression, symptoms were continually distributed in the population, with no clear ‘zone of rarity’ to demark the presence of disorder (Fried et al., 2022). Reviewing 177 studies (with combined data from over half a million individuals) that employed established statistical techniques to investigate the categorical nature of psychopathology, there was little evidence of psychological taxa (Haslam et al., 2012). Specifically, mood, anxiety and eating disorders, and disorders in the externalising domain had poor fit within a discrete, categorical structure (Haslam et al., 2012). While there is some evidence to support potential

taxa within substance use disorders and schizotypy, this was not consistent across studies (Haslam et al., 2012).

A substantial proportion of individuals have ‘subthreshold’ presentations for disorders, but suffer from significant morbidity (Kotov et al., 2017). For example, subthreshold depression, in which the full DSM criteria are not fulfilled, has comparable functional impairment (Noyes et al., 2022) and is a major risk factor for major depression (Cuijpers & Smit, 2004; Zhang et al., 2023). This is similar for presentations of subthreshold generalised anxiety disorder, which is highly prevalent and associated with increased psychosocial impairment (Haller et al., 2014). Examining the DSM-5 criteria for bulimia nervosa, there were no meaningful differences in clinical severity or impairment between individuals with sub- and full-threshold presentations (Chapa et al., 2018). Across psychiatric disorders, there is a lack of consensus on definitions for subthreshold presentations (Cuijpers & Smit, 2004; Haller et al., 2014), which leads to consistent marginalisation of those suffering (Okasha, 2009). Additionally, arbitrary thresholds for diagnosis have direct implications for treatment pathways (Bakker, 2019) and clinical trial research (Von Glischinski et al., 2021). Across 1600 studies that conducted clinical trials for major depressive disorder, there is no clear consensus on the optimal cut-off for determining inclusion or evaluating trial efficacy (Von Glischinski et al., 2021). Only a minority of clinical trials provide any justification for arbitrary threshold choices (Courtney et al., 2021). In short, arbitrary binarisation does not reflect clinical reality and neglects subthreshold presentations that are clinically meaningful.

1.2.2 Poor inter-rater reliability

Diagnostic reliability measures how likely it is that one clinician's diagnosis of a mental disorder will be independently confirmed by a second clinician assessing the same patient (Freedman et al., 2013). Accurate diagnosis is critical precondition for determining prognosis and treatment planning (Markon et al., 2011). With each new publication of the DSM, a field trial is conducted to assess the reliability of psychiatric diagnosis in clinical practice (Freedman et al., 2013). In the DSM-5 field trials, interviewers from multiple clinical sites assessed the diagnosis of 33 separate disorders among a large representative patient sample

(Regier et al., 2013). Across disorders, diagnostic agreement ranged from very good to unacceptable (Regier et al., 2013). While post-traumatic stress disorder was one of the most reliable (kappa = 0.67), more commonly diagnosed disorders had questionable reliability, including major depressive disorder (kappa = 0.28) and generalised anxiety disorder (kappa = 0.20) (Regier et al., 2013). When structured interviews were conducted by general practitioners specifically, the reliability for major depressive disorder was unacceptable (kappa = 0.16) (Fernández et al., 2010). These reliability rates are a major concern, with direct implications for the validity of major depression and generalised anxiety in particular (Freedman et al., 2013). Based on simulated data, the low inter-rater reliability of major depressive disorder reported in the DSM-5 field trials severely impacted true diagnostic rates, by producing false positive diagnoses (i.e., diagnosing a nondepressed person as depressed) and false negatives (i.e., diagnosing a depressed person as nondepressed) (Fried et al., 2022).

1.2.3 High comorbidity

Psychiatric co-morbidity refers to the simultaneous co-occurrence of multiple psychiatric disorders during the clinical course of a patient (Maj, 2005). Although the DSM and ICD categorise psychiatry disorders as distinct entities, psychiatric co-morbidity is highly prevalent, resulting in complicated clinical presentations (Clark et al., 2017). In the United States, a epidemiological study using the National Comorbidity Survey indicated that 45% of individuals with a psychiatric disorder had at least 2 diagnoses, and 23% had 3 or more diagnoses (Kessler et al., 2005). Rampant rates of co-morbidity were also reported in epidemiological studies with adolescents (Kessler et al., 2012) and across many different countries (McGrath et al., 2020). Examining the prevalence rates of major depression and generalised anxiety, co-morbidity was 400% higher than the rate predicted if these disorders co-occurred by chance alone (i.e., if there was no correlation between the disorders) (Eaton et al., 2010). Some argue that the reported rates of psychiatric comorbidity are artificially inflated, representing an artefact of narrow disorder categorisation (First, 2005; Hyman, 2021; Maj, 2005; Von Glischinski et al., 2021). Additionally, disorder co-occurrence may be an artefact of DSM diagnostic thresholds becoming less stringent with subsequent editions. Diagnostic criteria relaxed from the DSM-III to the DSM-5 for eating

disorders, attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder, resulting in an increased diagnostic rate for each disorder (Fabiano & Haslam, 2020). Finally, co-morbidity rates may be artificially inflated by shared symptomatology across disorders.

1.2.4 Shared symptomatology across disorders

Similar experiences can be endorsed across multiple diagnostic categories. For example, hallucinations are recognised as criteria for schizophrenia, schizoaffective disorder and bipolar disorder (Allsopp et al., 2019). Across the 628 distinct symptoms that comprise the 202 disorders in the DSM-5, 231 symptoms (36.8%) were repeated across multiple diagnoses (Forbes et al., 2024). More than a third of disorders have every symptom repeating in at least one other diagnosis, with major depression dominating the list of symptoms that repeat (e.g., insomnia, difficulty concentrating, low mood) (Forbes et al., 2024). This lack of symptom specificity across disorders may inflate comorbidity rates, particularly for disorders that share many symptoms. While co-occurrence was pervasive across disorder pairings, bivariate comorbidity was highest among more closely related disorders (e.g., major depression and generalised anxiety) (Kessler et al., 2005). Examining the diagnostic overlap of major depression and generalised anxiety in a large primary care sample, comorbidity was attributed to the overlapping symptoms (e.g., fatigue, poor centration and sleep disturbances), rather than the nonoverlapping symptoms (e.g., low mood in major depression, irritability in generalised anxiety) across disorders (Zbozinek et al., 2012). This provides further support that psychiatric comorbidity is inflated by symptom overlap. Shared symptomatology across disorders is also relevant to diagnostic accuracy, which may be hampered following misattribution of symptoms to alternative disorders. This was demonstrated in a study that used language analysis of online mental disorder support groups for depression, anxiety, bipolar, borderline personality and schizophrenia (Horvath & Todd, 2023). Forum posts related to anxiety, bipolar, borderline personality and schizophrenia were often misclassified as depression support groups, which was attributed to the shared depressive symptomatology across these disorders (Horvath & Todd, 2023).

1.2.5 Symptom heterogeneity within disorder categories

The DSM and ICD approach emphasises the importance of sum scores and thresholds, rather than the specific symptoms that constitute each disorder (Fried & Nesse, 2015). There are a total 227 possible ways to meet diagnostic criteria when considering the nine DSM-5 symptoms of major depression (Zimmerman et al., 2015). The majority (170/227) of these potential symptom profiles were endorsed by patients, demonstrating considerable diagnostic heterogeneity (Zimmerman et al., 2015). Considering that DSM-5 symptoms for major depression have sub-symptoms and extreme poles (e.g., increased/decreased weights), the estimated potential profiles that qualify as a major depression diagnosis increases from 227 to 16,400 (Fried & Nesse, 2015). Considering all potential profiles, 1030 unique symptom profiles were identified among 3703 patients seeking depression treatment (Fried & Nesse, 2015). Almost half (48.6%) of symptom profiles were unique to an individual patient, and most profiles (83.9%) were endorsed by five or less patients (Fried & Nesse, 2015). This illustrates that there is considerable symptom variation, despite all patients being classified as having the same 'major depressive disorder'. Substantial phenotypic heterogeneity is evident within other disorder categories. Using DSM-5 symptom criteria, there are 23,442 ways to have panic disorder (Galatzer-Levy & Bryant, 2013). As an extreme example, there are 636,120 unique symptom profiles for posttraumatic stress disorder (Galatzer-Levy & Bryant, 2013). In many disorder categories, two individuals can share a diagnosis, and have no symptoms in common (Olbert et al., 2014). Recognising comorbidity, there are over 270 million ways to meet the criteria for posttraumatic stress disorder with major depression, and this skyrockets to over one quintillion (10^{18}) ways when five other common comorbidities for posttraumatic stress are also considered, which is more symptom profiles than stars in the Milky Way (Young et al., 2014). This astronomical amount of unique symptom profiles renders disorders amorphous, with no consistent form across individuals (Young et al., 2014).

In summary, the traditional, disorder-based framework is plagued with profound shortcomings. Case-control studies of disorder categories obscure the reality of clinical complexity, which hampers mechanistic research in psychiatry. The disorder-based framework places an upper limit on the size and specificity of our scientific observations, which cannot be addressed by recent initiatives to gather

larger and more diverse datasets (e.g., The Brainstorm Consortium), or indeed advances in neuroimaging techniques (Etkin, 2019), computational modeling (Huys et al., 2021) and artificial intelligence (Bzdok & Meyer-Lindenberg, 2018). The current nosology is suffering from a ‘crisis of confidence’, being widely perceived as a failing scientific paradigm (Zachar & Kendler, 2017). To overcome the limitations inherent in the disorder-based framework, recent research efforts have focused on developing alternative classification methods (Kotov et al., 2017). While multiple promising alternatives are in development, the 'transdiagnostic dimensional' approaches are leading a distinct paradigm shift.

1.3 Transdiagnostic dimensional approaches to psychopathology

As an alternative to the disorder-based framework, the symptom space of mental health can be considering multidimensional. Dimensions (or factors) of psychopathology are proposed to be unbroken, existing on a continuum of severity levels from absent to very high (Kotov et al., 2017). Under this assumption, dimensions cut across the boundaries set by traditional diagnostic categories, making dimensions inherently ‘transdiagnostic’ (Dalgleish et al., 2020). A dimension of mental illness can be considered as a latent construct – a factor that explains patterns in observed data, but cannot be directly observed itself (Conway et al., 2019). Therefore, observable symptoms co-occur because they share influence from the same underlying dimension. Factor analysis, a family of statistical techniques, can be used to identify the underlying latent factor structure (e.g., the psychiatric dimensions) in a set of observed, interrelated variables (e.g., the psychiatric symptoms) (Yong & Pearce, 2013). The main aim of factor analysis is to explain the maximum amount of covariance in a correlation matrix of observed data, using the smallest number of explanatory factors (Field, 2013). In this way, factor analysis permits dimensionality reduction: a large number of psychiatric symptoms can be reduced into a smaller number of dimensions, which capture the essence of related symptoms.

Psychiatric symptoms load onto the same dimensions based on the degree of interrelation between those symptoms. Unrelated symptoms load onto different

dimensions, resolving the issue of symptom heterogeneity within disorder categories. This enables parsimony – complex psychiatric profiles comprising of many different symptoms are simplified into (partially) distinct dimensions. Therefore, the dimensional approach offers a more parsimonious description of psychopathology, embracing comorbidity and clinical complexity. Traditional diagnostic systems struggle to account for rampant comorbidity, homogeneity across disorders and symptom heterogeneity within disorders. In contrast, transdiagnostic approaches offer a more coherent explanation for these anomalies, by establishing overlapping dimensions of dysfunction that contribute to the manifestation of diverse symptoms (Eaton et al., 2023).

1.3.1 Comparing the disorder-based and transdiagnostic perspectives

Transdiagnostic approaches overcome many of the issues inherent to traditional nosology. Firstly, the transdiagnostic dimension approach disregards arbitrary thresholds of disorder binarisation. This aligns with well-replicated evidence that mental health problems are best conceptualised continuously, extending to subthreshold manifestations. Additionally, the transdiagnostic dimensional approach improves validity and reliability. Across two meta-analyses of 58 studies with nearly 60,000 participants, there was a 15% increase in reliability and a 37% increase in validity when mental health problems were conceptualised as continuous rather than dichotomous constructs (Markon et al., 2011). Improved validity and reliability with continuous measurements was consistent across samples and forms of psychopathology (Markon et al., 2011).

The transdiagnostic approach is more consistent with genetic evidence that disorders represent the extremes of a continuous trait, with dimensional variation in the population. Psychiatric disorders are highly polygenic, meaning that many thousands of alleles, distributed across the human genome, each exert a small effect to cumulatively increase the risk of psychiatric phenotypes (Hyman, 2021). A population-based twin study found that heritable risk factors for psychopathology transcend diagnostic boundaries and were largely shared across disorders (Kendler et al., 2011). Genome-wide association studies have also shown that genetic risk factors are shared across disorders. For example, genome-wide genotype data for the Psychiatric Genomics Consortium calculated

the genetic correlation across schizophrenia, bipolar disorder, major depression, autism spectrum disorder and ADHD (Lee et al., 2013). Genetic aetiology was shared across these major disorders, including those with little phenotypic overlap (e.g. schizophrenia and ADHD) (Lee et al., 2013). Schizophrenia and bipolar disorder, two disorders with a longstanding history of differentiation (Smoller et al., 2019), have a genetic correlation of nearly 0.7 (Lee et al., 2013). The Brainstorm Consortium separately reported a high degree of genetic correlation between psychiatric disorders using data from 200,000 patients (The Brainstorm Consortium et al., 2018). Widespread cross-disorder heritability indicates that psychiatric disorder categories do not reflect any specific genetic aetiology (Smoller et al., 2019).

Neurobiological disturbances reported in psychopathology align more closely with dimensions of psychopathology. A shared pattern of grey matter loss was evident across psychiatric disorders, which may explain executive dysfunction as a transdiagnostic feature of psychopathology (Goodkind et al., 2015). Relatedly, aberrant activation in brain regions implicated in cognitive control was transdiagnostic, rather than disorder specific (McTeague et al., 2017). This was further supported in another meta-analysis, which reported that transdiagnostic cognitive deficits reflect shared disruptions in functional connectivity architecture across psychiatric disorders (Sha et al., 2019). In a case-control study of 1809 individuals, neuroimaging markers could not distinguish ‘healthy’ individuals from those with major depression, as neural signatures were highly similar across groups (Winter et al., 2022). Capturing specific aspects of psychosis neurobiology required cutting across conventional DSM disorder categories with psychotic elements (i.e., schizophrenia, schizoaffective and bipolar disorders) (Koen et al., 2023; Tamminga et al., 2014). More broadly, common and dissociable patterns of network connectivity were more consistently associated with psychiatric dimensions than disorder categories (Xia et al., 2018). Therefore, a transdiagnostic perspective may reveal the shared neural architecture underlying cognitive deficits shared across disorders.

Treatments for mental illness are inherently transdiagnostic, targeting shared mechanisms between disorders that account for clinical complexity and

comorbidity. The publication of the DSM-III led to the establishment of effective, evidence-based treatments for specific diagnoses, including pharmacological interventions (Mayes & Horwitz, 2005) and psychotherapy (e.g., Cognitive Behavioural Therapy (CBT)) (Schaeuffele et al., 2021). These evidence-based treatments were initially established by focusing on single disorder categories specifically (Schaeuffele et al., 2021). However, clinical interventions have generalised efficacy across a broad range of disorders, leading to transdiagnostic prescription. For example, selective serotonin reuptake inhibitors are prescribed as the first-line treatment for depression, but also for anxiety, obsessive compulsive disorder and posttraumatic stress disorder (Bostic & Rho, 2006). Antipsychotic drugs are effective for psychosis but also for anxiety disorders (Garakani et al., 2020). Rather than being disorder specific, CBT treatments offer a ‘one size fits all’ approach to a broad range of psychiatric disorders (Schaeuffele et al., 2021). Importantly, mental health professionals do not consider diagnostic classification useful for determining treatment pathways (First et al., 2018). Accordingly, treatments are often prescribed based on symptoms, rather than diagnoses (Taylor, 2016). As treatments do not map onto specific DSM disorders, medications prescribed to patients are better predicted by transdiagnostic dimensions than disorder categories (Waszczuk et al., 2017). In sum, the overwhelming weight of evidence suggests that the transdiagnostic dimensional perspective better represents the clinical reality of psychopathology.

1.3.2 Computational factor modeling

To identify transdiagnostic dimensions with precise neurocognitive mechanisms, research has employed computational factor modeling (Gillan & Seow, 2020; Wise et al., 2023). Rather than focusing on the upper range of symptom severity endorsed by patients, computational factor modeling involves identifying candidate dimensions in large, unselected samples. The use of patient samples to derive dimensions permits mapping of only the most severe clinical experiences. Therefore, unselected samples are recruited to gauge the normal variation of psychiatric symptoms within the general population to access the full spectrum of symptomatology. General population sampling also permits a shift away from reliance on unreliable DSM diagnostic interviews (Regier et al., 2013), to self-asessment of psychiatric experiences. Computational factor modeling

involves two steps: 1) data-driven dimensionality reduction of self-report symptoms and 2) theory-driven modeling of objective behaviour (Wise et al., 2023). First, candidate dimensions are extracted using factor analysis. Once candidate dimensions have been defined, they are externally validated against well-defined computational parameters of behaviour. The performance of candidate dimensions is then compared to a benchmark (e.g., questionnaire total scores or diagnosis). By examining the association between participants' transdiagnostic diemension scores and cognition, it has been suggested that more precise mechanisms of underlying psychopathology can be uncovered (Wise et al., 2023).

One of the first applications of computational factor modeling was to investigate model-based planning as a transdiagnostic mechanism of compulsive disorders. Model-based planning is the cognitive process of using a mental model informed by action-outcome learning to strategically achieve goals (Daw et al., 2011). Dysfunction in model-based planning is a common feature of DSM disorders characterised by compulsivity, including obsessive compulsive disorder, binge eating disorder and alcohol and substance use disorders (Voon et al., 2017). To resolve non-specific findings between model-based planning and multiple psychiatric disorders, Gillan and colleagues (2016) recruited a large, unselected sample (N=1413) to complete a wide battery of psychiatric self-report scales and then applied factor analysis to the correlation matrix of questionnaire responses. From this, three related transdiagnostic dimensions were produced: 'Anxious-depression', 'Compulsivity and Intrusive Thought' and 'Social Withdrawal' (Gillan et al., 2016). Examining the association between these dimensions and model-based planning task performance, the overlapping links between deficits in model-based planning and certain disorders could be explained by the Compulsivity and Intrusive Thought dimension. This was specific to Compulsivity and Intrusive Thought, as Anxious-depression and Social Withdrawal were not significantly associated with model-based planning abilities. The association between deficits in model-based planning and Compulsivity and Intrusive Thought was subsequently replicated across multiple independent datasets with large samples (Donegan et al., 2023; Patzelt et al., 2019; Seow, Benoit, et al., 2021). This demonstrates how computational factor modeling can explain

individual differences in model-based planning, above and beyond disorder-specific questionnaires.

Using alternative questionnaire sets and behavioural tasks, computational factor modeling has successfully identified transdiagnostic dimensions with specific correlates across various cognitive domains, including uncertainty processing (Gagne et al., 2020), affective biases in reward processing (Daniel-Watanabe et al., 2022), aversive learning (Wise & Dolan, 2020) and temporal discounting (Keidel et al., 2024). These examples demonstrate how computational factor modeling represents a viable method to identify precise neurocognitive mechanisms underlying transdiagnostic symptoms. The questionnaire set used by Gillan and colleagues (2016) was selected to discern theoretically-relevant compulsive and noncompulsive dimensions (Gillan & Seow, 2020). This set includes 209 questionnaire items that measured aspects of mental illness related to alcohol misuse, anxiety, apathy, depression, eating disorders, impulsivity, obsessive compulsive disorder and schizophrenia. The three-factor structure derived by Gillan and colleagues (2016) has been replicated within multiple independent samples that conducted factor analysis on the same questionnaire set (Rouault et al., 2018; Seow et al., 2024). This indicated that Anxious-depression, Compulsivity and Intrusive Thought and Social Withdrawal are robust dimensions of mental illness, capturing the natural co-variance of psychiatric experiences in the general population. In addition to model-based planning abilities, this three-factor structure has informed our understanding of disturbances in metacognition.

1.4 Metacognition: A mechanism of transdiagnostic mental illness

Conceptualising psychopathology along data-driven transdiagnostic dimensions reveals specific and consistent associations with metacognitive abilities. Using the same three-factor structure of psychopathology generated by Gillan and colleagues (2016), transdiagnostic psychiatric dimensions explained more variance in metacognitive bias than disorder-specific diagnostic categories (Rouault et al., 2018). Tightly controlling for task accuracy, the transdiagnostic

dimension Anxious-depression was associated with significantly lower levels of confidence, while Compulsivity and Intrusive Thought was related to elevated confidence, in a large sample (N=496) from the general population (Rouault et al., 2018). Importantly, most of the classic questionnaires did not have a significant association with metacognitive bias (Rouault et al., 2018). This speaks to the fundamental issues with the disorder definitions upon which these questionnaires are based. This might be attributable to overlapping compulsive and anxiety components within questionnaires, which have opposing effects on confidence and thus averaged out. These findings were specific to confidence bias, as metacognitive efficiency was not significantly associated with transdiagnostic dimensions. Evidence of underconfidence in Anxious-depression and overconfidence in Compulsivity and Intrusive Thought has been replicated in multiple large-scale independent studies (Benwell et al., 2022; Hoven, Denys, et al., 2023; Seow et al., 2024; Seow & Gillan, 2020). Consistently replicated bi-directional effects of transdiagnostic dimensions on metacognitive bias demonstrate the utility of computational factor modeling to tease apart opposing mechanisms underlying psychopathology, which may be impossible to study using a case-control framework.

As transdiagnostic dimensions are not orthogonal, it is possible for individuals to exhibit high scores on Anxious-depression and Compulsivity and Intrusive Thought, simultaneously. Yet, these dimensions are associated with opposing directions in metacognitive bias (underconfidence for Anxious-depression, overconfidence for Compulsivity and Intrusive Thought). This curious finding suggests that differential processes are driving these opposing effects. For Anxious-depression, underconfidence may be explained by dysfunction across the metacognitive hierarchy (Seow, Rouault, et al., 2021). Across the metacognitive hierarchy, confidence can manifest across many levels of abstraction (Hoven, Denys, et al., 2023). Bias and efficiency are metacognitive facets at the lowest level of the hierarchy, as they reflect confidence judgements at the ‘local’ trial-level of decision-making (Seow, Rouault, et al., 2021). Ascending the metacognition hierarchy, ‘local’ confidence in isolated trial-level decisions reciprocally interacts with ‘global’ self-performance estimated at the task-level and core beliefs integral to our sense of self, such as self-esteem

(Seow, Rouault, et al., 2021). In support of this model, isolated local confidence evaluations aggravated to form global self-performance evaluations at the task-level (Rouault et al., 2019).

In a large general population sample (N=489), Anxious-depression was associated with widespread underconfidence across the metacognitive hierarchy, from local confidence to global self-beliefs (Hoven, Denys, et al., 2023). The negative association between local confidence and Anxious-depression was explained by lower self-beliefs, indicating that negative self-beliefs may permeate down through the hierarchy to impact a relatively low-stakes decision, such as trials on a perceptual task (Hoven, Denys, et al., 2023). This was not the case for Compulsivity and Intrusive Thought, as there was a marked decoupling of local and global confidence as Compulsivity and Intrusive Thought severity increased (Hoven, Denys, et al., 2023). Across the various levels of metacognition, Compulsivity and Intrusive Thought was negatively associated with global confidence and self-beliefs, but positively associated with local confidence (Hoven, Denys, et al., 2023). This suggests that the bidirectional associations with local metacognition in Anxious-depression and Compulsivity and Intrusive Thought may have distinct origins. While metacognitive bias in Anxious-depression can be explained by widespread disturbance across the metacognitive hierarchy (from global self-beliefs to trial-level underconfidence), overconfidence in Compulsivity and Intrusive Thought may be differentially explained by general difficulties in constructing an accurate mental model of the world (Seow & Gillan, 2020). This further demonstrates how the computational factor modeling can inform our understanding of metacognition, across hierarchical levels of abstraction.

1.5 The state-dependence of metacognition

Prior studies examined metacognition and transdiagnostic dimensions at a single time-point only (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). This correlational evidence does not indicate whether the normalisation of metacognitive impairments translates into improved psychopathology, and vice versa. It remains unknown if

metacognitive dysfunction is a distinct trait underlying specific psychiatric dimensions, or a state-dependent epiphenomenon, only present during or resulting from an acute psychiatric episode. There is evidence that metacognition is malleable and has the capacity to change with metacognitive training. Provision of feedback on how closely confidence judgements tracked accuracy was associated with enhanced metacognitive bias and efficiency (Carpenter et al., 2019). Similar effects of metacognitive training were reported by Engeler and Gillbert (2020), as feedback on metacognitive judgements improved metacognitive accuracy and removed underconfidence bias. While both these studies included general population samples (Carpenter et al., 2019; Engeler & Gilbert, 2020), there is also preliminary evidence that metacognitive training can ameliorate confidence biases among psychiatric patients (Moritz et al., 2014). Yet, it remains unknown if metacognitive abilities change independently of any improvements in psychiatric symptoms.

In clinical studies, research has identified confidence abnormalities in at-risk populations (Eisenacher et al., 2015; Gawęda et al., 2018), suggesting that metacognition is trait-dependent (i.e., stable across clinical states). However, it has also been shown that those with former substance use have better metacognition than current stimulant users, suggesting state-dependence (Moeller et al., 2016). So far, the evidence of the state- or trait-dependence of metacognition has been confined to cross-sectional, case-control studies (Eisenacher et al., 2015; Gawęda et al., 2018; Moeller et al., 2016), which provide no insight into the dynamic nature of metacognition. Longitudinal, within-subject designs are needed to build upon this work, to understand if abnormalities in metacognition can improve, and if changes in metacognition directly coincide with any real-world improvement in psychiatric symptoms. As such, the state-dependence of metacognition was evaluated in **Chapter 2** of this thesis, by examining whether metacognitive bias changed following treatments with established efficacy for improving transdiagnostic psychiatric symptoms: internet-based Cognitive Behavioural Therapy (iCBT) (Schaeuffele et al., 2021) and antidepressant medications (Bostic & Rho, 2006). Additionally, **Chapter 2** of this thesis aimed to replicate and extend the original cross-sectional findings observed in general population samples, by demonstrating the association

between metacognition bias and transdiagnostic psychiatric dimensions in a large, treatment-seeking sample of clinical patients. To capture metacognitive and clinical changes through treatment in a large sample, **Chapter 2** leverages novel, online-based technology.

1.6 Embracing online research in mental health

Relative to standard laboratory procedures, online research facilitates rapid acquisition of inter- and intra-individual data at scale. Importantly, the reported effects of transdiagnostic dimensions on metacognition (local and global) are significant, but small (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). This means sample sizes comprising hundreds of individuals are required to detect these effects. Additionally, the large samples required for computational factor modeling, to permit dimension generation and external validation, cannot be easily acquired using standard laboratory recruitment. To achieve this, online crowdsourcing has been considered a viable sampling methodology to speed up and scale up recruiment within and between participants (Berinsky et al., 2012).

1.6.1 Online crowdsourcing

Crowdsourcing refers to use of the internet to outsource labour openly to large samples of individuals (Strickland & Stoops, 2019). Within cognitive science, nearly half of research articles source participants from crowdsourcing platforms, such as Amazon's Mechanical Turk (AMT) and Prolific (Stewart et al., 2017). Indeed, AMT enabled recruitment of adequately sized samples for studies that identified cognitively-relevant dimensions of psychopathology (e.g., Gillan et al., 2016). Crowdsourcing platforms permit rapid and cost-effective sampling, providing easier access to difficult-to-research populations (Strickland & Stoops, 2019). As the 'crowd' from online labour markets are unselected, they are more diverse than participants obtained from other convenient sampling methods (e.g., college student samples) (Buhrmester et al., 2011). Although more diverse, crowdsource samples are not more representative of national populations, especially for demographic and clinical characteristics. For example, AMT samples were younger, more educated and endorsed higher rates of social

anxiety and autistic traits relative to the general population (Chandler & Shapiro, 2016).

Given the dramatic rise in popularity of crowdsourcing platforms, significant concerns have been raised regarding the quality of data obtained. Although initial reports indicated that AMT data had acceptable quality (Buhrmester et al., 2011), recent years have seen hightened scrutiny of crowdsourcing platforms for generating poor quality data. There has been a surge in low-quality data provided by AMT samples, to the detriment of study results (Chmielewski & Kucker, 2020). To counter this, it was recommended that validity indicators to check data quality be integrated into the study design (Chmielewski & Kucker, 2020). Validity indicators that flag careless or inattentive responses include response inconsistency, statistically improbably responses and suspious response times (Chmielewski & Kucker, 2020; Zorowitz et al., 2023). Comprehension tests at the beginning of an experiment also help mitigate the risk of careless responding (Donegan & Gillan, 2022). However, invalid responses have become so ubiquitous on AMT that recommended protections fail to address the problem of low quality data (Burnette et al., 2022; Donegan & Gillan, 2022). In a study by Burnette and colleague (2022), 90% of responders failed data quality checks, leaving the data unsuitable for statistical analyses. Low-quality data obtained from crowdsourcing platforms has been linked to a misalignment in motivations between researchers and participants - AMT samples are often incentivised by monetary compensation, rather than by a shared interest in advancing scientific knowledge (Donegan & Gillan, 2022). Moving away from reliance on crowdsourced platforms and financially motivated participants, Donegan and Gillan (2022) encourage a post-AMT era that adopts a ‘citizen science’ approach to online research.

1.6.2 The citizen science framework

Citizen science refers to the practice of actively engaging the general public in scientific research (Vohland et al., 2021). According to the citizen science framework, some individuals voluntarily participate in research, for motivational factors unrelated to financial gain (Gillan & Rutledge, 2021). Relative to paid laboratory and AMT samples, uncompensated online samples were more

representative of the general population, with a broader spread of demographic and clinical data (Gillan & Rutledge, 2021). Poor quality data is less of a concern with unpaid participant, demonstrating equivalent data quality to lab-tested samples (Germine et al., 2012). While paid participants are sourced through open labour markets, citizen engagement can be acquired through smartphone applications (Gillan & Rutledge, 2021). Multiple studies have shown that high quality data collection among citizen scientists can be achieved by releasing cognitive tasks through free smartphone apps (Brown et al., 2014; Coutrot et al., 2018; Donegan et al., 2023; Rutledge et al., 2016). For example, Coutrot and colleagues (2018) created a smartphone app to study spatial abilities in 2.5 million people worldwide, revealing an age-related decline in navigation skills. The substantially larger samples acquired through smartphone-based studies outweighed the added noise from collecting data in uncontrolled environments (Brown et al., 2014; Gillan & Rutledge, 2021). Without the lure of payment, researchers often gamify cognitive tasks in smartphone apps to entice unpaid participants (Donegan & Gillan, 2022).

1.6.3 Gamification of cognitive tasks

Standard experimental paradigms are often simple and carefully controlled, allowing for measurement of unconfounded cognitive abilities. To complement this, games can be used to determine if behaviours generalise to more complex, ecologically valid environments (Allen et al., 2024). Gamification of cognitive tasks functions to maximise participant engagement and enjoyment (Gillan & Rutledge, 2021). Often, classic cognitive tasks are not intuitive or enjoyable (Allen et al., 2024). Qualitative feedback indicated that most participants disliked standard cognitive experiments, describing them as challenging, long and repetitive (Lee et al., 2023). This negative feedback was specifically targeted at tasks that measure cognitive capacities implicated in mental illness, including metacognition (Lee et al., 2023). Lack of task enjoyment can discourage intrinsic motivation (Allen et al., 2024) and promote careless or inattentive responding, a critical threat to online data quality (Zorowitz et al., 2023). Zorowitz and colleagues (2023) found that careless and inattentive responders induce spurious correlations between online cognition tasks and self-reported psychopathology, which increased with sample size. While protective checks can be applied to

screen out this systematic bias (Zorowitz et al., 2023), the risk of poor data quality would ideally be minimised in the first instance, by designing tasks that are more positively received (Allen et al., 2024). This has been achieved by modifying established experimental paradigms to be more gamified, shorter and less difficult (Brown et al., 2014). After developing a gamified model-based planning in the smartphone app 'Neureka', Donegan and colleagues (2023) replicated the association between Compulsivity and Intrusive Thought and decreased model-based planning in a large sample of citizen scientists (N=5005). This demonstrates that large samples of citizen scientists can be recruited via smartphone apps to play gamified cognitive tasks, revealing specific cognitive mechanisms underlying transdiagnostic psychopathology.

1.6.4 The validity and reliability of gamified cognitive tasks

To understand mechanisms of mental illness, it is not sufficient to merely describe when and under what circumstance metacognitive impairments are observed. Rather, understanding how metacognition dynamically relates to psychiatric symptoms can reveal if changes in metacognition *causally* influence how psychiatric symptoms develop and improve, or vice versa. To assess casual relations between dynamic processes, multiple observations over time are required (Molenaar, 2019). This kind of dense, repeated sampling can be achieved by using smartphone-based technology, making it relatively convenient and minimally burdensome for participants (Gillan & Rutledge, 2021). Additionally, individuals are more inclined to revisit tasks repeatedly when tasks have been gamified (Allen et al., 2024). Adapting lab-based cognitive task for gamified smartphone-based research involves modifying core design features of the task. Specifically, integrating superfluous game features into a task can potentially confound the cognitive capacity of interest (Allen et al., 2024). As such, design changes necessary for gamification are potentially detrimental to the task's psychometric properties, including reliability and validity. While reliability refers to the stability or replicability of a measure over time, construct validity is the extent to which a measurement reflects the intended domain of content (Drost, 2011). Establishing task reliability and validity is an important prerequisite for examining individual differences in cognition (Karvelis et al., 2023). Poor reliability can induce imprecise correlations by either underestimating or inflating

effects (Kadlec et al., 2024). Low construct validity is also problematic, negating the generalisability of findings (Yarkoni, 2022). Recently, concerns have been raised regarding the validity and reliability of cognitive tasks commonly employed to study inter-individual differences (Karvelis et al., 2023). A variety of cognitive tasks were shown to have overall poor test-retest reliability (Hedge et al., 2018; Xu et al., 2022) and lower reliability relative to self-report questionnaires (Enkavi et al., 2019; Vrizzi et al., 2023). Additionally, a lack of construct validity has been reported among cognitive tasks that claim to measure the same underlying construct (e.g., Eisenberg et al., 2019; Noreen & MacLeod, 2015), which directly undermines interpretability.

There are a multitude of reasons why a cognitive tasks may have poor reliability for individual differences research. The most widely used index to measure task reliability is the intra-class correlation coefficient (ICC). The ICC represents the ratio of between-individual variance to total variance (Koo & Li, 2016). Curiously, many cognitive tasks with robust group-level effects have poor reliability. Considering ICC explains this seemingly paradoxical effect – replicable group-level effects require low between-participant variance, which inherently produces a poor ICC index and low reliability for individual differences research (Hedge et al., 2018). Therefore, a cognitive task may demonstrate robust group-level effects, but this does not mean it is necessarily suitable for individual differences research. To detect differences between individuals, higher between-participant variability is required. Between-participants variability can be increased by minimising ceiling and floor effects and by recruiting larger and more diverse samples (Zorowitz & Niv, 2023). Additionally, task reliability may be improved by incorporating gamification elements, which reduces measurement error by increasing participant engagement (Zorowitz & Niv, 2023). While poor reliability may be a consequence of low between-participant variability or measurement error, other factors may also skew reliability estimates. Multiple layers of influence, at the state- and trait-level, can cause cognition to dynamically fluctuate over time on a momentary timescale (Palminteri & Chevallier, 2018). In particular, practice and affective state effects dynamically impact cognitive estimates (Schurr et al., 2024). With practice, the confounding effect on cognition was largest during initial assessments and subsequently diminished with time

(Schurr et al., 2024). Therefore, concluding that a task is ‘unreliable’ may be misleading when sources of temporal variation (i.e., practice and affective state) have not been accounted for.

To summarise, determining the causal role of metacognition in psychiatric experiences requires dense, longitudinal assessment. Leveraging smartphone technology makes repeatedly sampling considerably easier. High quality data can be acquired from unpaid citizens when tasks are gamified and made freely available through smartphone apps. An important precondition to this is evaluating the psychometric properties of gamified tasks, as poor reliability and validity hinders task utility for individual differences research. Considering this, **Chapter 3** of this thesis aimed to investigate the psychometric properties of ‘Meta Mind’, a gamified smartphone task designed to measure local and global metacognition validly and reliably over time.

1.7 Optimising the transdiagnostic dimensional space

There is substantial evidence for specific associations between the three-factor structure of psychopathology derived by Gillan and colleagues (2016) and individual differences in model-based planning (e.g., Gillan et al., 2016) and metacognitive bias (e.g., Rouault et al., 2018). Importantly, the factor analytic pipeline used to generate this specific factor structure involved multiple steps and specification decisions.

1.7.1 Steps for conducting factor analysis

There are two major classes of factor analysis – exploratory and confirmatory (Williams et al., 2010). While exploratory factor analysis does not impose any preconceived structure on the dimensional space, confirmatory factor analysis requires a priori specifications to test hypotheses with strict constraints (Williams et al., 2010). This means that exploratory factor analysis is more flexible, making it more appropriate for the initial stages of research when the dimensional space is being explored. The dimensional structure generated by Gillan and colleagues (2016) was exploratory. Exploratory factor analysis refers to two different models: principal components analysis (PCA) and common factor analysis (Watkins,

2018). While PCA aims to maximise the variance explained by uncorrelated principal components, common factor analysis can identify dimensions that explain correlations in the observed data (Watkins, 2018). Common factor analysis may be more appropriate than PCA for psychiatric research (Costello & Osborne, 2019), as it aligns with the goals of understanding how symptoms cluster together in the real-world and are represented by interpretable latent factors.

When conducting exploratory factor analysis, there are a number of crucial steps (Figure 1.1). As a first step, data suitability should be assessed. While there is no definitive rule of thumb, a sample size greater than 1000 is ideal (Williams et al., 2010). As factor analysis extracts dimensions from the correlation matrix of observed data, any deviations from linearity and normality should be considered in the context of the correlation matrix. Measured variables chosen for factor analysis should be theoretically-informed, and variability in interrelations above and below r=0.3 should be evident in the correlation matrix (Williams et al., 2010) (Figure 1.1A). Once the suitability of the data is determined, the number of factors to be extracted should be specified. Given that the goal of factor analysis is to reduce the number of items into a lower-dimensional space, the smallest plausible number of dimensions should be extracted. When determining the number of factors to retain, many extraction approaches are reliant on the scree plot - a graphic that shows the explained variance of each newly defined component in the data (Cattell, 1966) (Figure 1.1B). The percentage or absolute value of the explained variance for each component is known as the eigenvalue (Field, 2013). Different methods to interpret eigenvalues are used to indicate the most suitable number of factors to keep:

1. *The Elbow Method (Cattell, 1966):* Using the Elbow Method, the number of factor retained is indicated by the ‘point of inflection’ - where eigenvalues begin to plateau on the scree plot (i.e.,the point where the plot forms an elbow-like shape), as this indicates where variance explained by subsequent components flattens (Figure 1.1B). However, interpreting the point of inflexion is highly subjective and requires researcher judgement.
2. *Kaiser’s criteria (Kaiser, 1960):* According to Kaiser’s criteria, any factor with an eigenvalue ≥1 should be retained. This method tends to

overestimate the number of factors, as absolute eigenvalues are contingent on the number of variables included: an eigenvalue of 1 with 100 variables indicates a factor explains 1% of the variance, but with 10 variables this factor explains 10% of the variance (Field, 2013).

3. *Parallel analysis (Horn, 1965):* When conducting parallel analysis, eigenvalues obtained from the actual data are compared to the eigenvalues obtained from random data (with no underlying factors). Factors with eigenvalues in the real data greater than those in the random data are retained. While parallel analysis is more objective than the elbow method or Kaiser's rule, it is sensitive to sample size and reliant on the random data adequately representing the structure of the true data.
4. *Cattell-Nelson-Gorsuch (CNG) indices (Gorsuch & Nelson, 1981):* Finally, the CNG scree test calculates the slopes linking three adjacent eigenvalues and selects the number of factors were the slope difference is largest. While this method involves less subjective interpretation, a minimum of three components is required. Gillan and colleagues (2016) retained three transdiagnostic dimensions based on CNG indices.

As these four methods not necessarily produce identical solutions, subjective assessment of each method is required to determine the most appropriate number of factors to retain. Another important consideration when conducting factor analysis is deciding if a variable relates to more than one factor (Williams et al., 2010). The initial factor structure may result in some variables having strong correlations with multiple factors, and other variables having very weak correlations with all factors. This can hinder interpretation of the factor structure, as variables do not have clear loadings onto any specific factors. Given that factors represent linear combinations of the original variables, rotation alters the orientation of factors to align with an underlying theoretical framework (Field, 2013) (Figure 1.1C). By rotating the location of the x- and y- axes to fit the actual data better, the factor structure is more easily interpretable. Rotation reduces cross-loadings, ensuring that variables are associated more strongly with one factor, and weakly with others. Additionally, rotation minimises the influence of sample-specific characteristics on the factor structure, enhancing generalisability and reproducibility across datasets. There are different types of rotation, with the

main distinction being whether or not factors are assumed to be correlated or uncorrelated with each other (Yong & Pearce, 2013). ‘Oblique’ rotation types allow for factors to be correlated, while ‘orthogonal’ rotations assume factors are not correlated (i.e., a 90-degree rotation matrix) (Figure 1.1C).

The use of oblique or orthogonal rotation types is informed by theoretical judgement by the researcher (Field, 2013). Arguably, factors that emerge from psychiatric phenomena will always be related to some degree, making orthogonal rotation less theoretically plausible (Costello & Osborne, 2019; Field, 2013; Watkins, 2018; Yong & Pearce, 2013). As such, Gillan and colleagues (2016) applied an oblique rotation to generate the three-factor structure. The interpretability of factor analysis is also influenced by the structural model applied to the data. In a simple (first-order) factor model of psychopathology, all latent dimensions are on the same level of a hierarchy (e.g., Gillan et al., 2016; Kotov et al., 2011) (Figure 1.1D). For a hierarchical (second-order) factor model, shared features among lower-level factors are represented by higher-order factors (e.g., Forbes et al., 2017; Kotov et al., 2017) (Figure 1.1D). Finally, the bi-factor model adopts a factor structure where observed variables are influenced by multiple specific factors, as well as a single general factor (e.g., Caspi et al., 2014; Caspi et al., 2024) (Figure 1.1D).

For researchers, there are a myriad of specifications required for factor analysis (i.e., analysis type, number of factors, rotation type, the structural model). Given this, a common concern with factor analysis is determining the ‘optimal’ model (Sellbom & Tellegen, 2019). Factor model superiority can be established with model fit indices (e.g., Caspi et al., 2024), which indicate how well the estimated model can reproduce the observed data. However, fit indices can be biased by models that overfit the data (Luciano et al., 2020), and can show goodness of fit even for uninterpretable or theoretically implausible factor structures (Bornovalova et al., 2020). Rather than overreliance on fit indices, theoretical consideration is critical when evaluating alternative factor models (Sellbom & Tellegen, 2019). The final stage of factor analysis is interpretation of the optimal factor structure, which involves examining how variables are associated with each factor and labelling the factors (Williams et al., 2010). The strength of the

association between each variable and the factor is indicated by the factor loadings – higher (absolute) loadings indicate a variable is more strongly correlated with that factor (Yong & Pearce, 2013). For example, social anxiety has strong positive loadings onto the Social Withdrawal dimension, while impulsive behaviour is less related and has weaker negative loadings (correlations closer to zero) onto that dimension (Gillan et al., 2016). While factor loadings represent the correlation between each variable and a given factor, factor weights indicate how much each variable contributes to the estimated latent factors (Field, 2013). Factor scores are calculated by multiplying factor weights by the observable data (Field, 2013). An individual's factor score represents their relative position on a dimension. Factor labels, and their inherent meaning, are determined by the researcher (Williams et al., 2010; Yong & Pearce, 2013). As factor labelling is highly subjective, factor names do not imply reification (i.e., factors correspond to 'real' and natural entities). This subjectivity has given rise to the so called 'jingle jangle fallacies', whereby factors with the same name do not necessarily reflect the same construct and factors with different names may not be different constructs (Watkins, 2018). Labelling merely eases communication of the factor structure.

In summary, factor analysis is used to identify the number and nature of latent dimensions that map onto psychiatric experiences. Factor analysis is a complex statistical approach, involving multiple steps and specification decisions. Many of these decisions are reliant on theoretical reasoning and subjective judgements. Choosing alternative specifications (e.g., exploratory vs confirmatory analysis, oblique vs orthogonal rotation type) can majorly impact the factor structure, which determines any inferences drawn about the underlying dimensional landscape of psychopathology. An ambitious and comprehensive initiative that has applied factor analysis to develop a dimensional classification system of mental disorders is the Hierarchical Taxonomy of Psychopathology (Kotov et al., 2017).

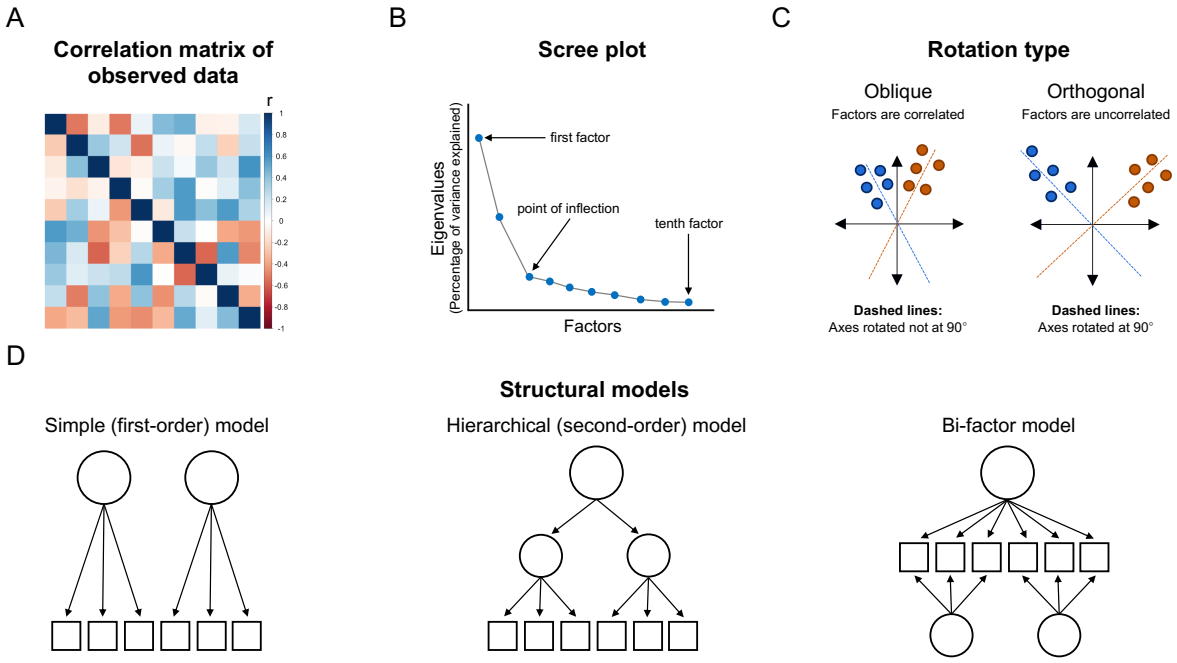


Figure 1.1. Schematic diagram illustrating steps for conducting factor analysis. (A) Generate a correlation matrix of observed data. (B) Graph scree plot to determine number of factors to retain. (C) Choose an oblique or orthogonal rotation type. (D) Choose the most appropriate structural model: a simple (first-order), hierarchical (second-order) or bi-factor model.

1.7.2 The Hierarchical Taxonomy of Psychopathology

The Hierarchical Taxonomy of Psychopathology (HiTOP) is a dimensional framework of psychopathology developed by a consortium of clinical researchers. HiTOP is composed of inter-related dimensions across a hierarchical structure (Kotov et al., 2017). The HiTOP consortium aim to overcome the shortcomings of DSM and ICD frameworks, such as arbitrary disorder thresholds, poor reliability, symptom heterogeneity and comorbidity (Kotov et al., 2018). HiTOP dimensions, derived based on the patterns of natural covariance among symptoms and disorders, are arranged across a hierarchy from the most narrow to broad psychiatric phenomena (Kotov et al., 2017, Kotov et al., 2022). This contrasts the factor structure developed by Gillan and colleagues (2016), which is a simple (first-order) model without a hierarchical structure (Figure 1.1D). At the top of the HiTOP hierarchy is an extremely broad 'general' dimension of psychopathology, which explains six lower level 'spectra' (the internalising, externalising (disinhibited), externalising (antagonistic), thought disorder, detachment and somatoform dimensions) (Kotov et al., 2017, Kotov et al., 2022). The spectra-level dimensions explain co-variance among lower levels 'sub-factors' (e.g., the

internalising dimension explains the co-occurrence of sub-factors for fear, distress and eating problems), which in turn cascade down to the most narrow units of psychiatric phenomena (individual signs, symptoms and maladaptive behaviours) (Kotov et al., 2017; Kotov et al., 2022). Within the HiTOP framework, a patient's profile for diagnosis compromises spectra dimension scores (e.g., a patient may be high-scoring on internalising and thought disorder dimensions, but low-scoring on the somatoform dimension) (Kotov et al., 2018).

HiTOP aims to be a fully empirical classification system. The DSM and ICD adopt an 'authoritative' approach to classification, as the structure of disorders is more reliant on general consensus from appointed experts (Kendler & Solomon, 2016; Krueger et al., 2018). In contrast, HiTOP ascribes to an 'empirical' approach, in which the classification rubric is informed by data from scientific research (Krueger et al., 2018). The architecture of HiTOP has been informed by many independent studies comprising of data on more than 100,000 individuals (Dalgleish et al., 2020). In an early study that investigated the structue of psychopathology among 8098 adults, patterns of comorbidity across 10 DSM disorders (including major depression, anxiety and panic disorders) could be represented by three sub-factors, all of which mapped onto a single 'internalising' dimension (Krueger, 1999). Using multiple representative studies (N=23,557), the best-fitting meta-analytic model that explained comorbidity among 11 DSM disorders included higher-order internalising and externalising factors (Krueger & Markon, 2006). Expanding to a larger number of DSM categories, co-occurrence of 25 disorders among 2900 outpatients seeking treatment were best explained by the five 'spectra' dimensions, labelled 'somatoform', 'internalising', 'thought disorder', 'externalising' and ''antagonism (Kotov et al., 2011). In this study, major depression, anxiety and eating disorders all mapped onto the internalising dimension, while substance and substance use disorders were explained by the externalising dimension (Kotov et al., 2011). The existence of an internalising dimension of psychoapthology was also supported in a large-scale nationally representative study comprising of 16,000 individuals (Forbush & Watson, 2013). In this study, patterns of co-morbidity among common DSM disorders (such as mood, anxiety and eating disorders) were explained by latent sub-factors (related to distress, fear and eating pathology), and those sub-factors mapped onto an

internalising dimension (Forbush & Watson, 2013). Similar to adult populations, covariance across DSM disorders among children and adolescents could be explained by higher-order ‘internalising’ and ‘externalising’ dimensions (Lahey et al., 2008).

Hierarchical structural analysis of DSM clinical and personality disorders provided support for a three-tier model of psychopathology: co-morbidity among disorder categories was explained by spectra at the bottom of the tier, which were nested under superspectra that all mapped onto a general factor of psychopathology at the top of the tier (Forbes et al., 2017). Hierarchical organisation is a distingushing feature of HiTOP, and the broad general factor of psychopathology at the apex of the hierarchy has garnered significant scientific interest. The pervasive overlap of psychiatric disorders led to the notion that there is a general propropensity for developing all psychiatric disorders (Lahey et al., 2012). This propensity can be measured as a single dimension of general psychopathology, termed the ‘p’ factor (Caspi et al., 2014). The 'p' factor conceptually aligns with the general 'g' factor of intelligence, exerting an overarching influence on the various lower-level features of psychopathology (Caspi et al., 2014). Unifying all disorder, the ‘p’ factor is suggested to reflect general liability, comorbidity, persistence, and severity of psychopathology across the lifespan (Caspi & Moffitt, 2018). The ‘p’ factor has been associated with various life outcomes, including suicidality, psychiatric hospitalisation and criminal convinction (Caspi et al., 2014). The ‘p’ factor can be extracted using either hierarchical or bi-factor modelling (Figure 1.1D). As evidence for the existence of the ‘p’ factor, incorporating a general factor into the structure of DSM disorders improves model fit, relative to simple (first-order) models (Caspi et al., 2014; Constantinou et al., 2019; Lahey et al., 2012). However, a simple one-factor model with ‘p’ alone is insufficient to describe psychopathology, and inclusion of lower-level dimensions (e.g., internalising and externalising dimensions) ensured goodness of fit (Caspi et al., 2014). This suggested that psychiatric experiences share some commonalities (i.e., the 'p' factor), while also having unique aspects specific to different subsets of experiences, (i.e., spectra dimensions).

HiTOP aligns with the higher-order genetic structure of psychopathology – pleiotropic genes increase risk of general psychopathology, while other genes confer a specific influence on individual spectra dimensions (Waszczuk et al., 2020). In addition to genetic risk, HiTOP dimensions have been correlated with various neurocognitive capacities implicated in mental illness, including working memory, rewarding learning and cognitive control (Michelini et al., 2021). Based on this, HiTOP dimensions were suggested to have 'clear', 'consistent' and 'robust' links with neurocognitive processes (Michelini et al., 2021). However, the mechanistic validity of HiTOP dimensions might be overstated, as these claims are based on independent studies that measured the association between cognition and DSM disorders, not scores on specific HiTOP spectra. Overall, HiTOP has been heavily reliant on DSM categorisation to establish its dimensional framework. Despite the HiTOP rubric not explicitly including DSM disorders as a hierarchical level, psychiatric disorder categories are often used to generate dimensions (Kotov et al., 2021). Influential studies that shaped HiTOP used patterns of comorbidity at the DSM disorder-level to model psychiatric dimensions (e.g., Forbush & Watson, 2013; Kotov et al., 2011; Krueger, 1999; Lahey et al., 2008; Forbes et al., 2017). Given this reliance on DSM classification, the 'p' factor and HiTOP spectra (e.g., externalising dimension) were shown to be nearly identical to the sum of diagnoses that contributed to them (Fried et al., 2021). In this regard, one key critique is that HiTOP may represent a re-expression of the DSM, rather than a novel classification system.

HiTOP can be referred to as a 'soft' transdiagnostic approach, as the established diagnostic framework (i.e., the DSM) is retained, with comparisons across disorders used to determine commonalities (Dalgleish et al., 2020). While HiTOP captures commonalities across disorder categories, it neglects diversity within. In contrast to HiTOP, computational factor modeling represents a 'hard' transdiagnotic method, as it is unconstrained by DSM definitions, which allows for new formulations of mental illness to be developed. Unlike HiTOP, the aim of computational factor modeling is to establish novel definitions of mental health that map more closely onto genetic, neural, cognitive and behavioural mechansims from the outset (Gillan & Seow, 2020). HiTOP dimensions are phenotypic constructs, based on subjective information only, without integration

of any objective task performance (Kotov et al., 2021). This constrains the mechanistic validity of HiTOP, as dimensions are defined prior to any consideration of putative mechanisms. In contrast, computational factor modeling interrogates psychiatric dimensions and neurocognitive mechanisms in tandem (Wise et al., 2023).

1.7.3 Mechanistic arbitration of transdiagnostic dimensions

Transdiagnostic dimensions are derived from the covariance matrix of clinical symptomatology, irrespective of whether dimensions align with the HiTOP framework or computational factor modeling (e.g., Gillan et al., 2016). However, symptoms that co-occur do not necessarily share a common mechanism. For example, a transdiagnostic dimension of negative symptoms, which cuts across DSM disorder boundaries, can be reached via multiple different mechanistic pathways (Strauss & Cohen, 2017). Rather than focusing on the factor models that explain symptom covariance, alternative factor solutions may have better mechanistic validity, providing more explanatory power for individual differences in cognition.

These alternative models can be derived by considering the multiverse of factor analysis specifications, including the number of factors derived, the questionnaire items considered, the rotation type and the structural model. Rather than biasing factor models with researcher subjectivity (Haeffel et al., 2022), a data-driven approach involves comparing the mechanistic validity of many competing factor solutions. Factors that do not represent a significant axis of variation in a questionnaire set may not be selected for retention; but they may nonetheless constitute an important and mechanistically homogenous dimension of psychopathology. With this, the ‘winning’ solution would produce transdiagnostic dimensions that explain the highest proportion of variance in the cognitive domain of interest. The three-factor solution consistently associated with model-based planning and metacognition was derived from a first-order factor analysis with oblique rotation, and applied to transdiagnostic sets of the maximal number of self-report symptoms available to the researcher (Gillan et al., 2016). However, it is possible that a solution with alternative specifications (e.g., a 4-factor solution with orthogonal rotation) produces dimensions that are more predictive of

individual differences in cognition. Currently, it remains unknown if the covariance-based framework generates clinical phenotypes with optimal mechanistic utility. Rather than only relying on the covariance of symptoms, future efforts should prioritise identifying data-driven factor solutions that maximally correspond to cognitive mechanisms. Accordingly, **Chapter 4** of this thesis compared thousands of competing factor solutions to identify dimensions of psychopathology with the strongest neurocognitive mapping, operationalised as the strength of the association between candidate dimensions and cognitive abilities with transdiagnostic clinical correlates (metacognition and model-based planning).

1.8 Experiments and hypotheses

1.8.1 Experiment 1 - An observational treatment study of the state-dependence of metacognition in Anxious-depression

Experiment 1 (Chapter 2) examined how metacognitive bias changes following clinical intervention. Data were collected as part of a large-scale transdiagnostic, four-week observational study of individuals initiating iCBT or antidepressant medication. Self-reported clinical questionnaires and perceptual task performance were gathered to assess transdiagnostic dimensions and metacognitive bias at baseline and 4-week follow-up. Primary analyses were conducted for individuals who received internet-based CBT (n=649), with comparisons between smaller samples that received antidepressant medication (n=82) and a control group receiving no intervention (n=88). A trait-view of metacognition suggests that metacognitive abilities will not change following treatment, despite significant improvement in psychiatric symptoms and transdiagnostic dimensions. In contrast, the state-view proposes that metacognition will significantly improve with response to treatment. Experiment 1 represents an exploratory analysis of the state- and trait-view of metacognition across psychiatric symptom dimensions, given the lack of previous longitudinal research. This experiment also aimed to replicate and extend the original cross-sectional findings (Rouault et al., 2018), by demonstrating bi-directional associations between metacognitive bias and transdiagnostic psychiatric dimensions in a clinical cohort.

1.8.2 Experiment 2 - Reliable, rapid, and remote measurement of metacognitive bias

Experiment 2 (Chapter 3) evaluated a brief gamified smartphone task called 'Meta Mind', designed to measure metacognitive bias rapidly and remotely. The psychometric properties of Meta Mind, including reliability (split-half and test-retest), and convergent validity, were assessed across two studies. In the first study, paid participants (N=52) completed Meta Mind and a traditional perceptual decision-making task, from which Meta Mind was adapted. Convergent validity was indicated as the correlation between confidence across the gamified and traditional versions of the task, while split-half reliability was measured as the correlation of within-task confidence ratings. In the second study, a large sample of over 3000 citizen scientists completed Meta Mind and mental health questionnaires within the smartphone app. The aim of the second study was to externally validate Meta Mind, by replicating the relations between the various levels of the confidence hierarchy and psychiatric dimensions (Hoven, Denys, et al., 2023). Individuals often describe traditional metacognitive tasks as lengthy and unengaging (Lee et al., 2023), which hinders the ability to gather large samples and to sample the same individual repeatedly over time. Given this, Experiment 2 aimed to establish the minimum number of trials required to measure metacognitive bias, while retaining adequate reliability and well-established clinical correlates. To permit repeated assessments, the test-retest reliability of confidence was examined in sub-samples of citizen scientists that played Meta Mind twice within 30 days (N=100) or 15 times over eight weeks (N=120).

1.8.3 Experiment 3 - Mechanistic arbitration between candidate dimensions of psychopathology

In Experiment 3 (Chapter 4), a large-scale comparison of competing factor solutions was conducted to determine the factor-structure with the strongest specific cognitive correlates. This included data from over 7565 individuals across 9 independent datasets (Benwell et al., 2022; Donegan et al., 2023; Fox et al., 2023; Gillan et al., 2016; Patzelt et al., 2019; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020), which included multiple variations of cognitive tasks

measuring model-based planning and metacognition, various age ranges, different data-collection mediums, and representing a spread of general, crowd-sourced and clinical populations. Thousands of competing factor solutions were generated, spanning variations in the number of factors retained, combinations of questionnaires, rotation setting and structural model type, and tested for the maximal effect size for the cognitive capacity of interest. Performance was benchmarked against a baseline of established transdiagnostic dimensions (i.e., Anxious-depression and Compulsivity and Intrusive Thought), which were derived from a three-factor solution, an oblique rotation, and based on the full set of items (Gillan et al., 2016). Finally, we compared the mechanistic validity and identity of factors derived from the covariance of symptoms alone (i.e., exploratory factor analysis) versus a fully supervised method (partial least squares regression) that uses cognitive information to inform the derivation of factors.

Chapter 2 – An Observational Treatment Study of the State-dependence of Metacognition in Anxious-depression

---

2.1 Introduction

Metacognition refers to the ability to accurately monitor and appraise one's own cognitive experience (Fleming & Lau, 2014). Metacognition is crucial for adaptive behaviour: it allows for flexible adjustment of behavioural strategies in order to improve performance, signals when to engage or withdraw from an activity, and guides the engagement in social interactions (Fleming et al., 2012; Fleming & Daw, 2017). Metacognitive abilities vary across individuals; one can be under- or over-confident, and these biases can be associated with maladaptive thoughts, feelings and behaviours (Philipp et al., 2020). There is a growing interest in these confidence abnormalities in psychiatry, with studies implicating alterations of metacognition in depression, obsessive-compulsive disorder, and psychosis (Hoven et al., 2019). While case-control studies have mainly found patterns of reduced confidence across several disorders, newer methods that can separate transdiagnostic dimensions of mental health using large online samples have revealed specific and bi-directional effects of confidence (Wise et al., 2023). Using these methods, studies have shown that the transdiagnostic dimension Anxious-depression is linked to under-confidence in one's own performance, while a separate dimension Compulsivity and Intrusive Thought is related to elevated confidence (Rouault et al., 2018; Seow & Gillan, 2020; Hoven et al., 2022; Seow et al., 2021; Benwell et al., 2022; Seow et al., 2024).

A major gap in this area is that studies to-date only measure metacognition and transdiagnostic psychopathology at a single time point. Therefore, it is unclear if metacognitive biases are stable, fixed traits, or if they might change with treatment response. Preliminary evidence suggests metacognition may indeed be malleable; metacognitive abilities can be improved with metacognitive interventions, such as training, in unselected online samples (Carpenter et al., 2019; Engeler & Gilbert, 2020) and in clinical populations (Jelinek et al., 2017;

Lysaker et al., 2018). However, it remains unknown if metacognitive changes generalise beyond its specific training context and are associated with any real-world improvement in psychiatric symptoms. In clinical studies, research has identified confidence abnormalities in at-risk populations (Eisenacher et al., 2015; Gawęda et al., 2018), suggestive of a trait-dependence. In contrast, stimulant use disorders remitters have better metacognition than active users, suggesting state-dependence (Moeller et al., 2016). Within-subject designs are needed to extend this work and understand if metacognition can improve in parallel to symptom alleviation, or if those with greater metacognitive deficits are simply the most vulnerable to illness onset and persistence.

The present study aimed to address this by examining metacognition in a large cohort of individuals before and after iCBT. iCBT has emerged as an important intervention for reducing the treatment-gap in mental healthcare provision globally; it is low-cost, scalable, geographically unconstrained and flexible (Mogoașe et al., 2017; Webb et al., 2017). iCBT offers patients standardised content and records objective metrics of treatment engagement, making it particularly well-suited to treatment-oriented research in psychiatry (Lee et al., 2023). Additionally, iCBT has demonstrated clinical effectiveness in terms of symptom improvement (Andersson et al., 2019; Eilert et al., 2021; Karyotaki et al., 2021). While one study found that iCBT modified self-reported metacognitive beliefs (Newby et al., 2014), it remains unknown if metacognitive confidence in decision-making improves following successful iCBT. In the current study, we used an objective task measure of metacognition (Fleming et al., 2014), which allowed us to test if successful treatment is linked to within-person improvements in metacognition. We also tested if any changes in metacognition were iCBT-specific, by comparing data gathered from smaller samples of individuals receiving antidepressant medication and a control group receiving no intervention. Similar to iCBT, antidepressants have established transdiagnostic efficacy (Cipriani et al., 2018; Gøtzsche & Dinnage, 2020; Skapinakis et al., 2016). However, studies examining the impact of antidepressants on cognition have typically focused on cognitive capacities other than metacognition (Prado et al., 2018; Rosenblat et al., 2015; Salagre et al., 2017; Vernon et al., 2014). Accordingly, a secondary aim of this study was to compare metacognitive

changes across the different intervention arms, which may shed light on differential therapeutic mechanisms and potentially augment therapeutic decision-making in the future.

2.2 Methods

2.2.1 Participants

Participants were recruited as part of the Precision in Psychiatry (PIP) study (Lee et al., 2023), an observational, longitudinal study in which participants underwent a four-week course of iCBT or antidepressant medication. Further details of the PIP study procedures that are not specific to this study can be found in a prior publication (Lee et al., 2023). Ethical approval for the PIP study was obtained from the Research Ethics Committee of School of Psychology, Trinity College Dublin and the Northwest-Greater Manchester West Research Ethics Committee of the National Health Service, Health Research Authority and Health and Care Research Wales. A power analysis was carried out using effect sizes from a previous study examining cross-sectional associations between metacognition and Anxious-depression, and Compulsivity and Intrusive Thought (Rouault et al., 2018). Sample sizes of N=454 and N=332 respectively were required to detect these associations with 80% power. The sample sizes of the antidepressant arm and control group were smaller and used for secondary and more exploratory analyses.

iCBT arm. Individuals initiating iCBT provided by SilverCloud Health were recruited from two sites: 1) the National Health Service Berkshire Foundation in the UK and 2) Aware mental health charity in Ireland. Participants included in the study either started their iCBT intervention ≤2 days prior to signing up, or provided a treatment start date in the near future, and scored ≥10 on the Work and Social Adjustment Scale (WSAS) at baseline, which indicated significant functional impairment due to clinical symptoms (Mundt et al., 2002). Figure 2.1A shows the disposition of participants throughout the study. N=2404 were screened, of whom N=1496 were eligible, N=836 completed baseline assessments and met inclusion criteria (detailed in Supplementary Material 7.1.1). A final N=649 completed and met inclusion criteria for follow-up

assessments. While study follow-up data was collected after four weeks of treatment, iCBT could last up to 12 weeks (Lee et al., 2023). The final sample was, on average, 32.2 years old (SD=11.0), mostly female (n=501, 77.4%), living in the United Kingdom (n=546, 84.4%), and had some or completed undergraduate level education (n=342, 52.9%) (Table 2.1).

Antidepressant arm. Individuals were recruited globally using advertisements placed on Google search, in addition to social media platforms, mental health websites, local pharmacies and General Practitioner waiting rooms. Participants were included if they started or planned to start treatment ≤2 days of study sign-up, scored ≥10 on the WSAS at baseline, and provided a valid photograph of an antidepressant medication prescription. N=270 individuals were screened, of whom N=174 were eligible, N=102 completed and met inclusion criteria at baseline and a final N=82 had follow-up data (Figure 2.1A). Participants were mostly female (n=60, 73.2%), mean age=30.5 (10.5), were living in Ireland or the United Kingdom (n=66, 80.5%) and had some or completed undergraduate level education (n=49, 59.8%) (Table 2.1).

Control group. Participants in the no treatment control group were recruited through university mailings lists and advertisements posted online and around Trinity College Dublin. Participants included in this arm scored <10 on an adapted version of the WSAS (where they rated functional impairment from their general problems rather than mental health problems) and self-reported that they had no current mental health problems and were not undergoing treatment for any mental health problems at the time of screening. N=444 individuals were screened, of whom N=245 were eligible, N=113 had baseline data and a final N=88 completed follow-up assessments and met inclusion criteria for the study (Figure 2.1A). Participants in the control group were matched for sociodemographic characteristics in the antidepressant arm, with no significant differences across the two groups in gender (Fisher's exact test p=0.498), or age (Welch's t(166.66)=-0.81, p=0.420), or levels of educational attainment ($\chi^2(2) = 0.66$, p=0.718). Participants in the control group came from the United Kingdom and Ireland (>93%), whereas the antidepressant arm was more international, with 20% coming from other countries ($\chi^2(2)=13.59$, p=0.001).

When comparing sociodemographic characteristics across the three study groups, there were no significant differences between groups in gender proportions (Fisher's exact test p=0.414), or levels of educational attainment ($\chi^2(4)=7.11$, p=0.130), as reported in Table 2.1. Age differed across the three arms (F(2, 813)=3.68, p=0.026), with post hoc Tukey tests indicating that mean age was higher in the iCBT arm (M=32.2, SD=11.0) compared to the controls (M=29.1, SD=12.0) (p_{adj}=0.034), but not when compared to the antidepressant arm (M=30.5, SD=10.5) (p_{adj}=0.376) (Table 2.1). The countries participants were living in varied across the study arms ($\chi^2(4)=211.73$, p<0.001), as there was a higher proportion of individuals in the iCBT arm living in the UK (n=546, 84.4%) when compared to the antidepressant arm (n=34, 41.5%) and the control group (n=24, 27.3%) (Table 2.1).

Table 2.1. Baseline Sociodemographic Characteristics of Participants.

Characteristic	iCBT (*n* = 649)[a]	Antidep. (*n* = 82)	Control (*n* = 88)	F/X^2 (df)	*p*
Gender, No. (%)					0.414[b]
Male	141 (21.8)	20 (24.4)	22 (25.0)		
Female	501 (77.4)	60 (73.2)	66 (75.0)		
Other	5 (0.8)	2 (2.4)	0 (0.0)		
Age, *M (SD)*	32.2 (11.0)	30.5 (10.5)	29.1 (12.0)	3.68 (2, 813)	0.026
Country of residence, No. (%)				211.73 (4)	<0.001
Ireland	89 (13.8)	32 (39.0)	58 (65.9)		
United Kingdom	546 (84.4)	34 (41.5)	24 (27.3)		
Other	12 (1.9)	16 (19.5)	6 (6.8)		
Highest level of education, No. (%)				7.11 (4)	0.130
Below undergraduate	147 (22.7)	12 (14.6)	13 (14.8)		
Some/completed undergraduate	342 (52.9)	49 (59.8)	57 (64.8)		
Above undergraduate	158 (24.4)	21 (25.6)	18 (20.5)		

[a] Two participants in the iCBT arm were missing data for age, gender, country of residence and highest level of education.
[b]Gender proportions were compared using Fisher's exact due to cell count < 5 in the 'Other' group.

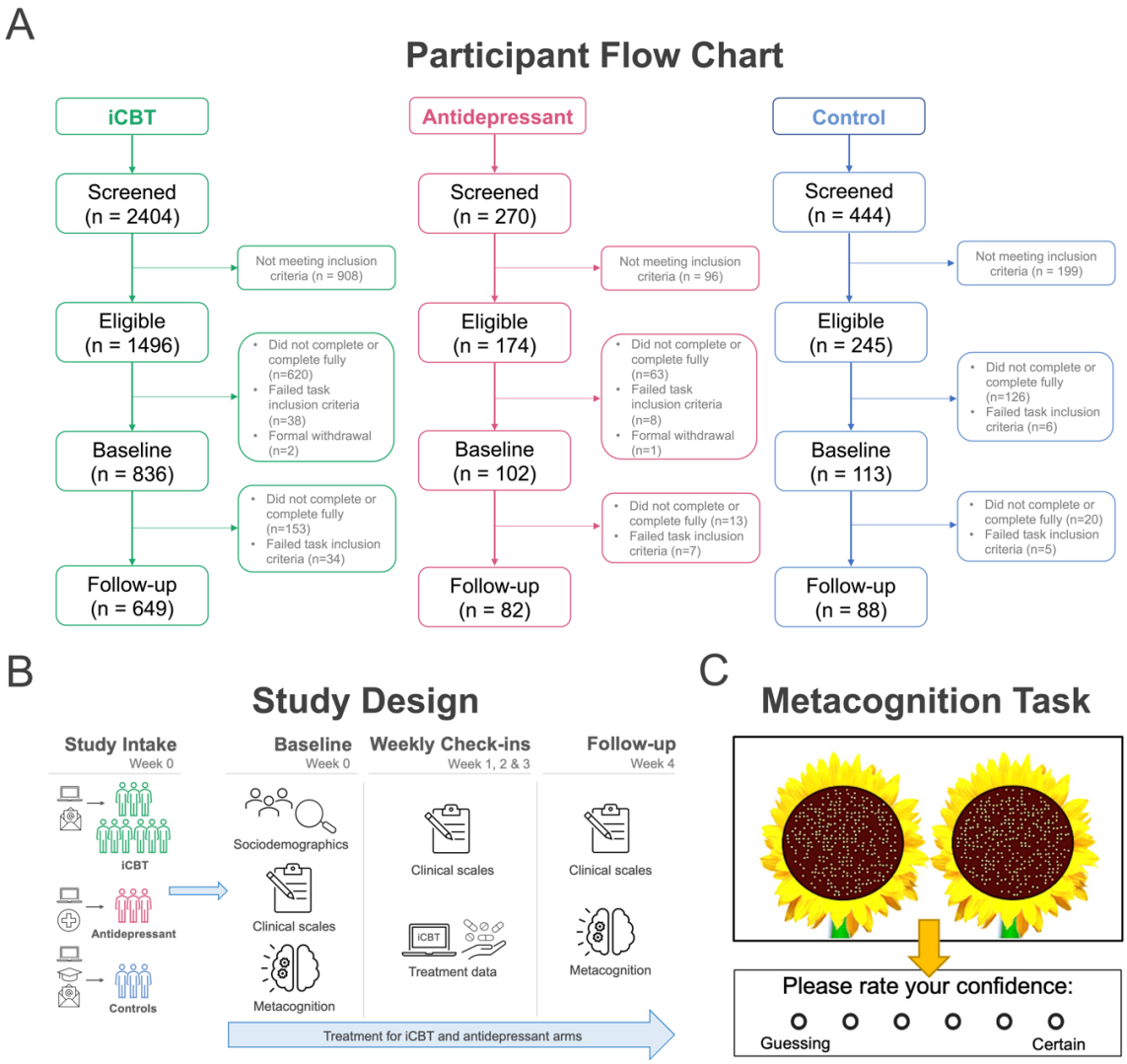


Figure 2.1. Study methods. (A) Participant flow chart (CONSORT chart). Participants were considered 'completers' if they had metacognitive and transdiagnostic psychiatric dimension data at baseline and follow-up and met task inclusion criteria. (B) Overview of study design from study intake (week 0) to follow-up (week 4) assessments across groups. (C) Metacognitive (visuo-perceptual decision-making) task design (N = 210 trials). On each trial, participants were asked to judge and choose the sunflower that contained more seeds (i.e., higher number of dots) and then provide a confidence rating on their decision.

2.2.2 Procedure

Figure 2.1B shows an overview of the study design, including the assessments involved at each timepoint. For the purposes of this study, we focused on a select set of sociodemographic characteristics (gender, age, country of residence, level of educational attainment), self-reported psychiatric questionnaires, metacognitive task performance and treatment data from the PIP study (Lee et al., 2023).

Self-reported psychiatric questionnaires. Participants completed nine standard self-report questionnaires (209 items in total) at baseline and four-week follow-up that assess a variety of psychiatric symptoms, including depression (Zung Self-Rating Depression Scale) (Zung, 1965), trait anxiety (State Trait Anxiety Inventory) (Spielberger et al., 1983), schizotypy (Short Scales for Measuring Schizotypy) (Mason et al., 2005), impulsivity (Barratt Impulsiveness Scale 11) (Patton et al., 1995), OCD (Obsessive-Compulsive Inventory-Revised) (Foa et al., 2002), social anxiety (Liebowitz Social Anxiety Scale) (Liebowitz, 1987), eating disorders (Eating Attitudes Test) (Garner et al., 1982), apathy (Apathy Evaluation Scale) (Marin et al., 1991), and alcohol misuse (Alcohol Use Disorders Identification Test) (Saunders et al., 1993) (see Supplementary Material 7.1.1 for further questionnaire details).

Metacognitive task. Participants completed a visuo-perceptual decision-making task (Fleming et al., 2014) to assess metacognition (Figure 2.1C). On each of the 210 trials, participants were asked to judge and choose the sunflower that contained more seeds (i.e., higher number of dots) and then provide a confidence rating on their decision. Mean accuracy was tightly controlled in this task using a 'two-down one-up' staircase procedure, in which equal changes in dot difference occurred after each incorrect response and after two consecutive correct responses, detailed further in Supplementary Material 7.1.1.

Treatment data. Objective indicators of treatment engagement were provided by SilverCloud for 640 participants in the iCBT arm, which comprised of percentage of program viewed, time (minutes) spent in the program, and program type (listed in Supplementary Material 7.1.1). Information on concurrent treatment and treatment adherence across each group are reported in Supplementary Material 7.1.1.

2.2.3 Data preparation and analysis

Self-reported psychiatric questionnaires. Individual scores on dimensions of Anxious-depression, Compulsivity and Intrusive Thought, and Social Withdrawal were calculated by multiplying each of the 209 item scores on the nine self-report clinical scales by the 209 corresponding item weights from a previously published

factor analysis on these scales (Gillan et al., 2016). Dimension scores were scaled to centre on zero, with higher scores indicating higher levels of transdiagnostic psychopathology. We demonstrated the consistency of this factor structure among clinical patients by carrying out factor analysis on these 209 items in this dataset and testing their correlation with the prior work in a general population sample (reported in Supplementary Material 7.1.1). Careless or inattentive responding to questionnaires was detected with embedded 'catch' items across questionnaires, detailed in Supplementary Material 7.1.1.

Metacognition task. The perceptual decision-making task performance was used to quantify our primary cognitive outcome measure, metacognitive bias, the mean confidence rating across trials. Task difficulty was measured as the mean dot difference across trials, where more difficult trials had a lower dot difference between stimuli. Other task measures included mean accuracy (proportion of correct selections) and mean reaction time (in seconds), which are described further in Supplementary Material 7.1.1. As metacognitive efficiency was not previously associated with transdiagnostic dimensions cross-sectionally (Benwell et al., 2022; Hoven et al., 2022; Rouault et al., 2018), results pertaining to efficiency are reported separately in Supplementary Material 7.1.1.

Statistical analysis. We tested for relationships between baseline task measures and the psychiatric symptom dimensions using linear regression analysis, controlling for age, gender, and education. To examine pre-post changes, we carried out linear mixed-effects models with measures of metacognition or psychopathology as the dependent variable, time (baseline=0; follow-up=1) as the independent variable and participants as random effects. To determine the association between change in confidence and change in Anxious-depression, we used (1) Pearson correlation analysis to correlate change indices for both measures and, (2) regression-based repeated-measures analysis: mean confidence ~ time*Anxious-depression score change, where mean confidence is entered with two datapoints (one for pre- and one for post-treatment i.e., within-person) and Anxious-depression change is a single value per person (between-person change score). Exploratory linear regression analyses tested the specificity of the effects, replacing Anxious-depression with each of the measures

of psychopathology in turn as follows: Mean confidence~ time*dimension/psychiatric scale score change. We additionally ran regression analyses to test if concurrent treatment or the degree of objective engagement in iCBT interacted with the effect of time on mean confidence. Exploratory ANOVA analyses were also conducted to compare changes in Anxious-depression, task difficulty and confidence across the three arms directly. For all tests, statistical significance was defined as $p<0.05$, with two-tailed p-values used. All regressors were scaled as Z scores to compare the regression coefficients of independent variables within each model. Adjustments for multiple comparisons were not conducted for analyses of replicated effects, or exploratory analyses (Althouse, 2016). The code to reproduce statistical analyses are available at https://osf.io/89xzq/. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. All data from this study will be made publicly available once the primary treatment predict analysis of the PIP study are published on a pre-print server.

2.3 Results

2.3.1 Cross-sectional findings at baseline: iCBT

At baseline, participants with higher levels of Anxious-depression had lower levels of mean confidence (β=-0.09, SE=0.03, p=0.008; Figure 2.2A & 2.2B), while those with higher levels of Compulsivity and Intrusive Thought had elevated mean confidence (β=0.11, SE=0.03, p=0.002; Figure 2.2A & 2.2C), controlling for age, gender, and education. Levels of Social Withdrawal were not associated with mean confidence (β=-0.05, SE=0.03, p=0.168; Figure 2.2A).

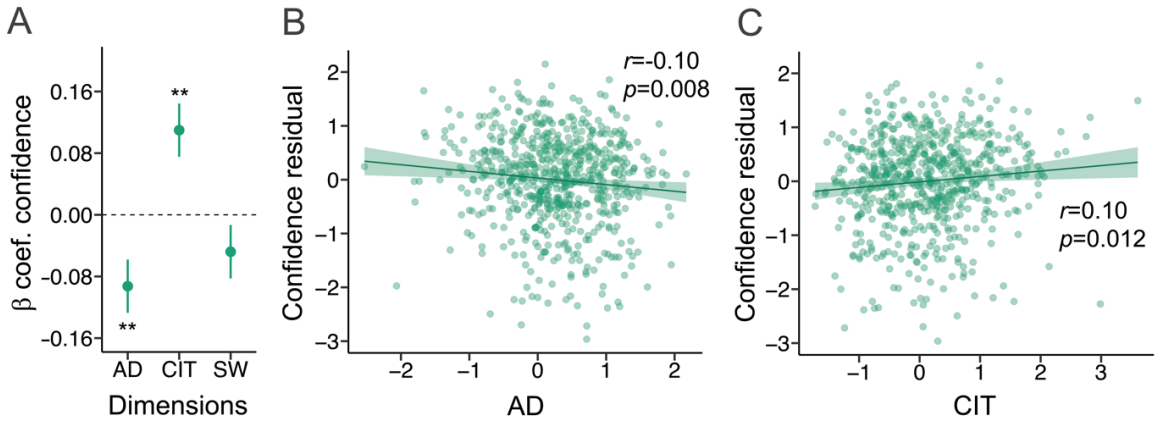


Figure 2.2. Baseline associations between confidence and transdiagnostic dimensions in the iCBT arm. β = standardised beta coefficient, *r* = correlation coefficient, *p* = p-value, AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal. The error bars represent the standard error around the standardised beta coefficient. (A) AD and CIT were associated with metacognitive bias, while SW was not. (B) The residual values for confidence (controlling for age, gender and education) were negatively associated with AD. (C) The residual values for confidence (controlling for age, gender and education) were positively associated with CIT.

2.3.2 Treatment findings: iCBT

The transdiagnostic dimensions and psychiatric scale scores all significantly improved from baseline to four-week follow-up, except for impulsivity (Figure 2.3A and Supplementary Table S7.1.1). In tandem with these clinical changes, there was a small but significant increase in mean confidence from baseline (M=3.78, SD=0.85) to follow-up (M=3.95, SD=0.89), (β=0.17, SE=0.02, p<0.001, r^2=0.01) (Figure 2.3B). Although overall accuracy remained stable due to the staircasing procedure, participants' ability to detect differences between the visual stimuli improved. This was reflected as the overall increase in task difficulty to maintain the accuracy rates from baseline (dot difference: M=41.82, SD=11.61) to follow-up (dot difference: M=39.80, SD=12.62), (β=-2.02, SE=0.44, p<0.001, r^2=0.01) (Figure 2.3C). Additionally, the effect of time on confidence was not dependent on how much participants engaged in iCBT, as indexed by time spent in the program (β<0.01, SE<0.01, p=0.756) and percentage of the iCBT program viewed (β=0.09, SE=0.21, p=0.650). Change in confidence was not different among those receiving concurrent treatment versus not (β=-0.03, SE=0.06, p=0.566), with 175 participants (27.0%) in the iCBT group receiving another treatment during the study (further detailed in the Supplementary Material 7.1.2).

To test if changes in confidence from baseline to follow-up scaled with changes in

Anxious-depression, we ran a repeated measure regression analyses with per-person changes in Anxious-depression as an additional independent variable. We found this was the case, evidenced by a significant interaction effect of time and change in Anxious-depression on confidence (β=-0.12, SE=0.04, p=0.002). Those with the largest decrease in Anxious-depression had the greatest increase in confidence. This was similarly evident in a simple correlation between change in confidence and change in Anxious-depression (r(647)=-0.12, p=0.002) (Figure 2.3D). This effect was specific to Anxious-depression; the interaction effect of time and change in Compulsivity and Intrusive Thought on mean confidence was not significant (β=-0.06, SE=0.05, p=0.221). Similarly, the significant interaction effect of time and Anxious-depression on mean confidence held when including change in the other transdiagnostic dimensions (Compulsivity and Intrusive Thought and change in Social Withdrawal) as covariates in the model (β=-0.07, SE=0.03, p=0.005). The interaction effect of time and change in Anxious-depression on task difficulty was not significant (β=0.14, SE=0.69, p=0.835). To test the extent to which baseline differences in mean confidence or anxious depression might drive the results, we re-ran these regression analyses with baseline measures instead of change indices. There was no significant interaction effect of time and baseline confidence (β=0.06, SE=0.04, p=0.144) or an interaction effect of time and baseline Anxious-depression (β=-0.03, SE=0.05, p=0.611) on mean confidence. Exploratory analyses further tested the specificity of these effects to Anxious-depression by examining the interaction effect of time and change in each psychiatric score on mean confidence. Changes in trait anxiety (β=-0.08, SE=0.02, p=0.002), depression (β=-0.06, SE=0.02 p=0.011) and alcohol misuse (β=-0.05, SE=0.02, p=0.037) also showed an association with changes in confidence (Figure 2.3E and Supplementary Table S7.1.2).

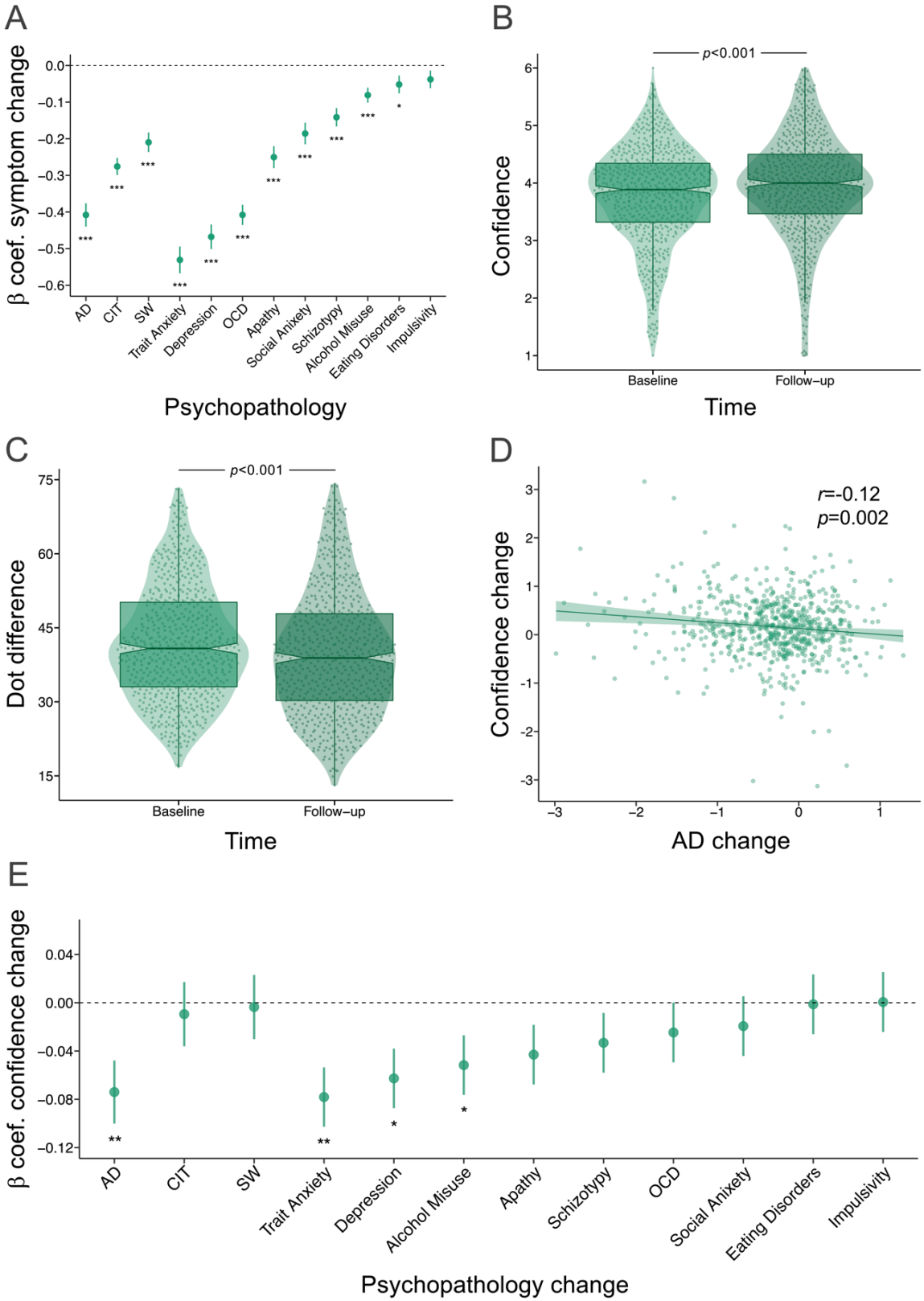


Figure 2.3. Treatment findings in the iCBT arm. β = standardised beta coefficient, AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, OCD = Obsessive compulsive disorder, *r* = correlation coefficient, *p* = p-value (unadjusted), *** = *p*<0.001, ** = *p*<0.01, * = *p*<0.05. (A) Psychopathology symptoms improved with four weeks of iCBT. (B) Confidence was significantly higher and, (C) the task was more difficult at four-week follow-up. (D) Those with the largest improvements in AD had the greater increases in

confidence. (E) Change in confidence also scaled with improvements in trait anxiety, depression and alcohol misuse.

2.3.3 Comparing iCBT, antidepressant and control groups

When comparing the three groups directly, ANOVA analysis predicting Anxious-depression scores with group and time as independent variables revealed a main effect of time ($F(1, 1632)=62.99$, $p<0.001$), a main effect of group ($F(2, 1632)=249.74$, $p<0.001$), and an interaction effect of group and time ($F(2, 1632)=9.23$, $p<0.001$). Examining simple effects in the antidepressant arm, there was a significant reduction in Anxious-depression from baseline to follow-up ($\beta=-0.61$, $SE=0.09$, $p<0.001$). Among controls, levels of Anxious-depression did not significantly change ($\beta=0.10$, $SE=0.06$, $p=0.096$). Further details of transdiagnostic clinical changes for the antidepressant and controls groups are presented in Figure 2.4A and Supplementary Table S7.1.3.

Predicting confidence scores using ANOVA analysis with group and time as independent variables revealed a main effect of time ($F(1, 1632)=16.26$, $p<0.001$), and no significant main effect of group ($F(2, 1632)=2.35$, $p=0.096$). The interaction effect of group and time on mean confidence was not significant ($F(2, 1632)=0.60$, $p=0.550$), suggesting that change in confidence did not differ across the three groups. Tests of simple effects revealed that mean confidence significantly increased from baseline ($M=3.77$, $SD=0.88$) to follow-up ($M=4.07$, $SD=0.79$) in the antidepressant arm ($\beta=0.31$, $SE=0.08$, $p<0.001$) (Figure 2.4B). Among controls, there was no significant change in confidence from baseline ($M=3.68$, $SD=0.86$) to follow-up ($M=3.79$, $SD=0.92$) ($\beta=0.11$, $SE=0.07$, $p=0.103$) (Figure 2.4B).

With respect to task performance, there was a significant main effect of time ($F(1, 1632)=15.17$, $p=0.001$) and group ($F(2, 1632)=4.56$, $p=0.011$) on mean dot difference when the three groups were included in the model. The interaction effect of time and group on mean dot difference was not significant ($F(2, 1632)=1.91$, $p=0.148$), suggesting no differences across the groups in task difficulty changes. In the antidepressant arm, mean dot difference decreased from baseline ($M=41.2$, $SD=13.3$) to follow-up ($M=35.3$, $SD=13.1$) ($\beta=-5.91$,

SE=1.25, p<0.001), indicating increased task difficulty. There was no significant change in task difficulty among controls from baseline (M=43.0, SD=11.8) to follow-up (M=41.4, SD=13.6) (β=-1.64, SE=1.30, p=0.210) (Figure 2.4C).

While our sample was underpowered to examine individual differences, we conducted an exploratory analysis examining the connection between changes in Anxious-depression symptoms and changes in confidence in the antidepressant and controls groups. When examining the effects of time, group and Anxious-depression change on mean confidence, there was a significant interaction effect of time and Anxious-depression change on mean confidence ($F(1, 1626)=4.04$, $p=0.045$), suggesting change in confidence is associated with change in Anxious-depression. There was no significant three-way interaction of Anxious-depression change, time and group on mean confidence when comparing the three groups ($F(2, 1626)=0.08$, $p=0.928$), indicating that the significant association between confidence change and Anxious-depression change was not specific to any group. Although not significant, the association between change in confidence and change in Anxious-depression was in the expected negative direction in the antidepressant arm ($r(80)=-0.10$, $p=0.381$), and among controls ($r(86)=-0.17$, $p=0.111$) (Figure 2.4D).

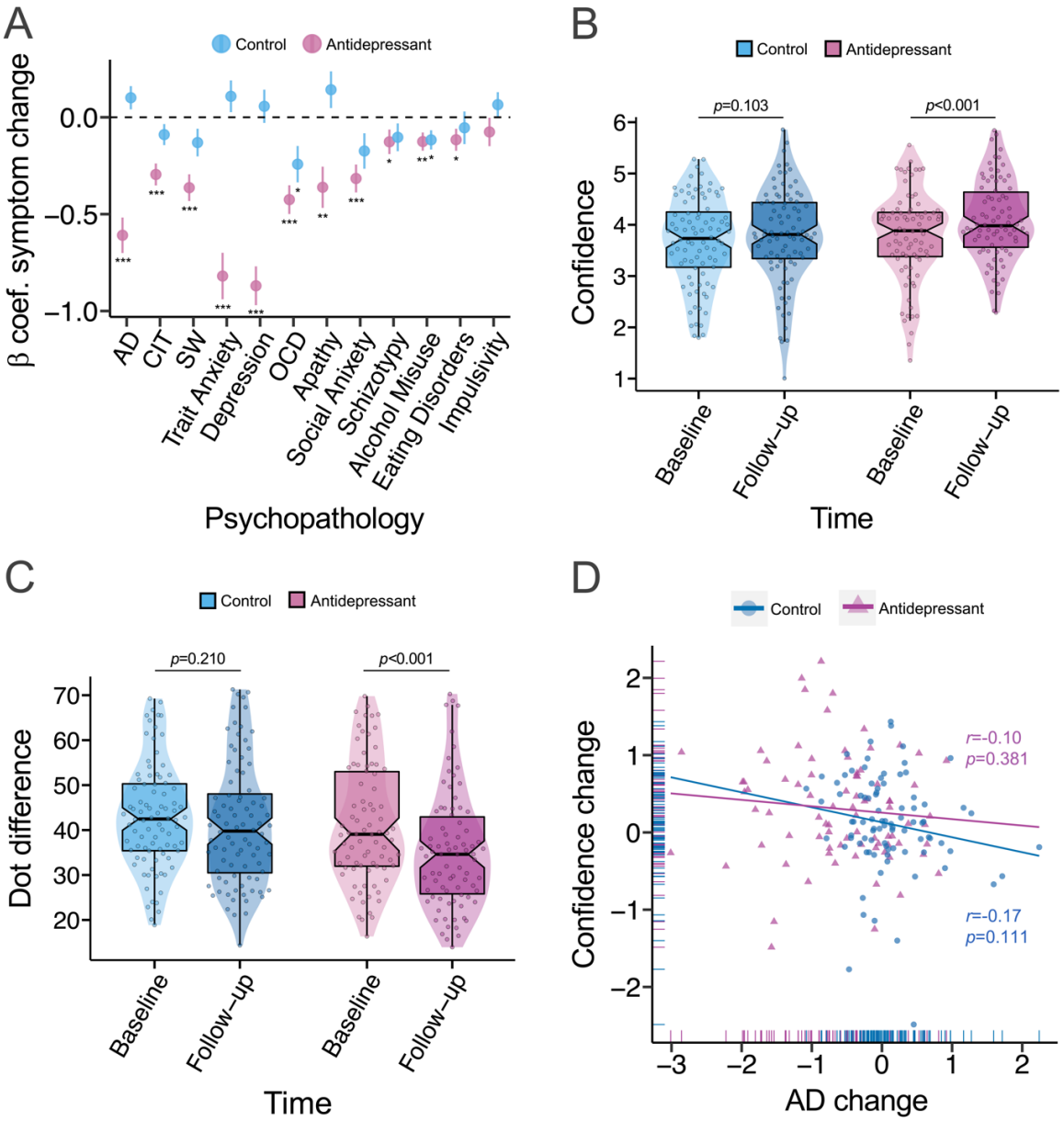


Figure 2.4. Comparing the iCBT, antidepressant and control groups. b = standardised beta coefficient, AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, OCD = Obsessive compulsive disorder, *r* = correlation coefficient, *p* = p-value, *** = p<0.001, ** = p<0.01, * = p<0.05. (A) The majority of psychiatric scales improved in the antidepressant arm after 4 weeks of treatment, while the controls only had significant reductions in OCD symptoms and alcohol misuse at follow-up. (B) The larger increase in confidence in the antidepressant arm compared to controls was trended towards significant. (C) The antidepressant arm had a greater increase in task difficulty (a reduction in dot difference across stimuli) from baseline to follow-up, relative to controls. (D) Although not significant, the association between change in confidence and change in Anxious-depression was in the expected negative direction in the antidepressant arm and among controls.

2.4 Discussion

Metacognitive biases are linked to transdiagnostic dimensions of mental health, but it is presently unclear if these biases are stable traits, or if they fluctuate

alongside symptoms and change during the course of treatment (Seow et al., 2021). To answer these questions, we administered a previously validated adaptive task of metacognitive ability that controls for objective performance differences (Fleming et al., 2014) in a large sample of individuals before and after four weeks of iCBT or antidepressant medications (Lee et al., 2023). As expected, a four-week course of iCBT or antidepressant medication led to transdiagnostic improvements in mental health (Cipriani et al., 2018; Gøtzsche & Dinnage, 2020; Hofmann et al., 2012; Skapinakis et al., 2016). Alongside this, there was a significant increase in metacognitive confidence following four weeks of iCBT or antidepressant medication. Not simply a practice effect, we found that individuals in the iCBT arm with the greatest improvements in Anxious-depression had the largest increase in confidence at follow-up. This association with clinical improvements was specific to metacognitive changes, and not changes in task performance, suggesting that changes in confidence do not merely reflect greater task familiarity at follow-up. These findings suggest that metacognitive biases in Anxious-depression are state-dependent. This builds on previous findings in small samples that have shown iCBT improves self-reported metacognitive self-beliefs (Newby et al., 2014) and that metacognition can be altered through adaptive training (Carpenter et al., 2019; Engeler & Gilbert, 2020; Jelinek et al., 2017; Lysaker et al., 2018).

At baseline, we replicated the previously observed bi-directional associations between metacognitive bias and Anxious-depression and Compulsivity and Intrusive Thought (Benwell et al., 2022; Hoven et al., 2022; Rouault et al., 2018; Seow & Gillan, 2020). While higher levels of Anxious-depression is associated with lower confidence, those with higher levels of Compulsivity and Intrusive Thought have elevated confidence. This is a somewhat surprising dissociation, as Compulsivity and Intrusive Thought and Anxious-depression are themselves positively correlated in the population. One way this can be reconciled is if the mechanisms underlying these opposing confidence biases are distinct. In Anxious-depression, there appears to be more pervasive metacognitive biases that affect confidence in many domains and levels of a metacognitive hierarchy (spanning confidence in low level perceptual decisions to ideas of self-worth) (Hoven et al., 2022; Seow et al., 2021). In contrast, inflated confidence in

compulsivity may be based on more specific biases in learning and inference (Seow & Gillan, 2020).

The present study was observational and therefore did not randomly assign participants to a different treatments. To partially remediate that limitation, we included two smaller groups receiving antidepressant medication and a control group. Levels of transdiagnostic psychiatric dimensions remained stable across time among controls, while they significantly improved in the antidepressant arm. Similarly to iCBT, we found that confidence improved in the antidepressant group, but not among controls. The interaction, however, was not significant, meaning that we cannot reject the null hypothesis that confidence improved to the same degree across the three groups. As increased task difficulty among clinical groups was not significantly greater relative to controls, changes in task difficulty may simply reflect greater task familiarity at follow-up across groups, as opposed to gains in general perceptual performance among clinical arms. Examining the three groups together, the data suggests that confidence changes are unlikely to be treatment specific, rather, confidence scales with improvements in Anxious-depression. This was evident in an overall association between change in Anxious-depression and change in confidence that was not modified by treatment arm. Additionally, levels of iCBT engagement and concurrent treatments did not bolster changes in confidence. Overall, the results indicated that metacognition fluctuates with Anxious-depression state, regardless of treatment type or exposure. Future research with larger samples across treatment arms are required to address this definitively.

Confidence change and Anxious-depression change were significantly but weakly associated. Similarly, the relative change in confidence across treatment arms was small. Therefore, while tests of metacognitive confidence can inform theoretical models, like most cognitive tests, they are likely of limited utility in clinical practice, at least when used in isolation (Mogoașe et al., 2017). Given the complexity of mental health causes and presentations, multivariable models are needed to see practical value from such tests. We did not assess confidence or Anxious-depression to treatment cessation and so the causal path and temporal dependence, if they exist, cannot be derived from these data. Future research

should consider assessing metacognition and Anxious-depression continuously through treatment, in order to elucidate the causal relationship between Anxious-depression and metacognition with mediation analysis (Newby et al., 2014). More intensive, repeating testing in future studies may also reveal the temporal window at which metacognition has the propensity to change, which could be more momentary in nature. While this study examined changes in metacognition with iCBT generally, future research should examine if the strength of the association between confidence change and Anxious-depression change is greater following iCBT modules targeting metacognition or following metacognitive intervention (Philipp et al., 2020). The iCBT programs in this study primarily targeted depression and anxiety, which may explain why changes in confidence did not scale with improvements in compulsivity. Future research is required to assess if treatments aimed at compulsive disorders decrease the over-confidence commonly observed in those high in Compulsivity and Intrusive Thought. As the antidepressant and control groups were much smaller than the iCBT arm, we were unable to compare changes in confidence across the types of antidepressant medications individuals received and we were underpowered more generally for individual differences analyses and multi-arm comparisons. Exploratory analyses were nonetheless presented and can form the basis for future investigations.

In conclusion, our findings replicated the cross-sectional evidence that higher levels of Anxious-depression are associated with under-confidence. We demonstrate that metacognitive confidence increases following four weeks of iCBT or antidepressant treatment. Overall, we observed that the greater the improvement in Anxious-depression, the more confident participants became, which did not appear to be dependent on treatment type. This suggests that metacognitive biases in Anxious-depression are state-dependent and might be normalised through clinical improvements.

Chapter 3 – Reliable, Rapid, and Remote Measurement of Metacognitive Bias

---

3.1 Introduction

Metacognition, the ability to reflect upon and evaluate cognitive experiences, is a central facet of human consciousness (Heyes et al., 2020). As feedback is often absent in daily life, metacognition facilitates the continuous monitoring of our moment-to-moment decisions, informing – for better or worse - our beliefs about our skills, abilities (Rouault et al., 2019) and even self-worth (Seow et al., 2021). Aside from informing self-concepts, at a more granular level, metacognition facilitates learning and guides behaviours in the absence of direct feedback (Fleming & Daw, 2017; Guggenmos et al., 2016), and allows us to communicate uncertainty in decision-making to others (Fleming et al., 2012). The prototypical example of metacognition is the confidence we hold in our own decisions (Fleming, 2024). By gathering repeated confidence judgements from individuals as they make choices, we can measure important facets of metacognition: bias and efficiency (Xue et al., 2021). Bias refers to the tendency to give high or low confidence ratings on average, while efficiency is the extent to which our confidence levels reliably discern correct from incorrect decisions (Fleming & Lau, 2014).

Crucially, individuals vary in their metacognitive abilities and these differences correspond to where an individual sits along a spectrum of transdiagnostic mental health symptoms (Hoven et al., 2019; Wise et al., 2023). Specifically, a transdiagnostic dimension of ‘Anxious-depression’ is linked to underconfidence in one’s own performance, while a separate dimension ‘Compulsivity and Intrusive Thought’ is related to elevated confidence (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2022; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). Metacognitive bias in these studies was originally calculated from confidence judgements at the ‘local’ trial-level, but recent work has shown that metacognition manifests across a hierarchy (Seow et al., 2021). Within this

hierarchy, isolated local confidence evaluations in single decisions are aggregated slowly over time to form more 'global' beliefs about one's ability in a given domain, and may generalise into broader self-beliefs (Rouault et al., 2019). Indeed, recent work has shown that disturbances in local confidence across transdiagnostic psychiatric dimensions are reflected in similar patterns of biased global self-performance evaluations spanning longer timescales (Hoven et al., 2022). If generalised outside of a single domain, these biases could conceivably contribute over time to generalised negative schemata central to cognitive models of depression (Beck, 2002).

To date, investigations have been largely cross-sectional and between-person (Benwell et al., 2022; Hoven et al., 2022; Rouault et al., 2018; Seow & Gillan, 2020). This limits what we can learn about temporal dynamics, or cause and effect. For example, while it is possible that metacognitive biases in depression at the local level play a role in shaping global self-beliefs, it is equally plausible that changes in self-esteem reciprocally impact local confidence (Seow et al., 2021). To address this gap, recent work has begun to adopt within-person designs, measuring metacognition within the same person over time. These studies have provided evidence to suggest that metacognitive biases fluctuate over time in healthy individuals (Da Fonseca et al., 2023) and negative confidence bias reduces following cognitive behavioural therapy and antidepressant medication (Fox et al., 2023). This suggests metacognitive bias is not a fixed or final trait, but instead may be malleable, and potentially a target for intervention. However, it remains poorly understood how these biases temporally relate to changes in psychopathology. One way to sample metacognition densely over time, and spanning periods of significant clinical change within an individual, is through online, remote data collection.

Prior studies have achieved remote testing by using web-based metacognitive tasks and recruiting well-powered samples through crowd-sourcing platforms, such as Amazon Mechanical Turk and Prolific (Benwell et al., 2022; Hoven et al., 2022; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). However, paid crowd-sourced samples have come under scrutiny for generating poor quality data (Chmielewski & Kucker, 2020), which is not necessarily resolved with

the established protective quality measures (Burnette et al., 2022). As an alternative recruitment avenue, ‘Citizen Science’ is a valuable paradigm with improved data quality (Donegan & Gillan, 2022), in which individuals participate in research voluntarily, due to motivational factors unrelated to financial gain (Gillan & Rutledge, 2021). Uncompensated, self-selected online samples provide comparable data quality to lab-based perceptual experiments, with the advantage of speeding up and scaling up data collection (Germine et al., 2012), and being more representative (Gillan & Rutledge, 2021). Employing cognitive tasks through smartphone applications specifically has proven particularly beneficial in ensuring high-quality data collection among citizen scientists (Brown et al., 2014; Coutrot et al., 2018; Donegan et al., 2023; Rutledge et al., 2016).

One barrier to this, however, is growing concerns that many of the most commonly used cognitive tests in psychiatry suffer from poor reliability (Hedge et al., 2018; Xu et al., 2022). This is in contrast to self-report clinical questionnaires, which typically demonstrate good to excellent reliability (Enkavi et al., 2019; Vrizzi et al., 2023). Metacognitive bias differs from standard objective task outcomes, as it is typically measured using a hybrid approach that incorporates elements of cognitive assessment and self-report. Much like a typical cognitive test (and unlike a self-report questionnaire), some metacognitive tasks tightly control actual “Type I" performance (e.g., by titrating the task difficulty to each person’s ability), thereby preventing actual performance differences from confounding metacognitive judgements (Fleming et al., 2010). A key metric of metacognitive abilities, however, is not behavioural, but subjective – the estimate of confidence in one’s decisions (Fleming & Lau, 2014). For these reasons, metacognitive bias might enjoy a higher level of test-retest reliability that is more similar to a questionnaire than classic cognitive tests.

Poor reliability is an issue for developments in the field of computational psychiatry, as prior findings on inter-individual differences may be imprecise and invalid (Karvelis et al., 2023). Alternatively, a less pessimistic view of poor reliability among cognitive outcomes is that behavioural tasks provide legitimate estimates of momentary cognitive capacities when tested, but these are liable to fluctuations over time (Palminteri & Chevallier, 2018). In line with this, cognitive

capacities tend to temporally covary with affect and practice factors, which are often not accounted for in reliability assessments (Schurr et al., 2024). This further illustrates the need for within-subject longitudinal assessments in computational psychiatry, to infer individual clinical and cognitive phenotypes.

The present study aimed to test this using a brief gamified smartphone task called ‘Meta Mind’, designed to measure metacognitive bias reliably, remotely, and in as few trials as possible. To this end, we evaluated the psychometric properties of Meta Mind, including reliability (split-half and test-retest), and convergent validity, across two experiments. In the first experiment, paid participants completed Meta Mind and a traditional perceptual-decision making task, from which Meta Mind was adapted (Fox et al., 2023). In the second experiment, a large sample of over 3000 citizen scientists completed Meta Mind and mental health questionnaires within the smartphone app. Visuo-perceptual decision-making tasks have become a staple method for measuring confidence (Rahnev, 2021), providing estimates of domain general metacognition (Benwell et al., 2022), and so this was the task type upon which Meta Mind was based (Fox et al., 2023). However, prior work has shown that this type of task is criticised by participants and described as tedious, lengthy and difficult (Lee et al., 2022). This is a threat to research quality, as a lack of task engagement can increase rates of careless or inattentive responding, in some cases leading to spurious associations between cognition and self-reported psychopathology (Zorowitz et al., 2023). To address this, we focused on not just gamification, but determining the minimum number of trials required to measure metacognitive bias, while retaining adequate reliability and well-established clinical correlates.

3.2 Methods

3.2.1 Experiment 1: Paid participant sample

3.2.1.1 Participants

Individuals in Experiment 1 were recruited by convenience sampling; through word of mouth, social media, online forums and university mailing lists. Participants were included if they were over the age of 18 and had access to a computer (desktop/laptop) and smartphone (Apple/Android device). Of the N=116

participants that consented to participate, N=52 met inclusion criteria and fully completed Meta Mind on their personal smartphone device and the traditional metacognitive task on in a web-browser. The sample had a mean age of 23.62 (SD=7.96), was mostly female (n=39, 75.00%), living in Ireland (n=49, 94.20%), and had obtained at least secondary school level education (n=50, 96.10%) (Table 3.1). Participants were paid €10 for taking part. Power analysis was based on a prior study that gathered data on the same traditional browser-based metacognitive task at baseline and 4 weeks later. This study reported an ICC of 0.73 for metacognitive bias, with 95% confident interval=0.67-0.77 (Fox et al., 2023). We anticipated that the convergent validity (Pearson correlation) estimates when comparing tasks in this study might be smaller, given differences across the tasks in length, instructions, interface, graphics and confidence rating scales. Conservatively, we powered our study to detect a medium effect size (r=0.40) with 0.80 power, requiring 46 participants.

Table 3.1. Baseline Sociodemographic Characteristics of Paid Participants in Experiment 1.

Characteristic	Paid Participants (N = 52)
Gender, No. (%)	
Male	13 (25.0)
Female	39 (75.0)
Age, M (SD)	23.62 (7.96)
Country of residence, No. (%)	
Ireland	49 (94.2)
United Kingdom	3 (5.8)
Education, No. (%)	
Primary level	1 (1.9)
Secondary level	32 (61.5)
Undergraduate degree	18 (34.6)
Above undergraduate degree	1 (1.9)
Ethnicity, No. (%)	
White or Caucasian	45 (86.5)
Asian or Pacific Islander	4 (7.7)
Multiracial or biracial	3 (5.8)
Employment status, No. (%)	
Unemployed	27 (51.9)
Part-time employed	23 (44.2)
Full-time employed	2 (3.8)

3.2.1.2 Procedure

Participants accessed the study information and an electronic consent form via Qualtrics, the link for which was embedded in the study advertisement. After providing consent, participants provided their email address and the following sociodemographic information: age, gender, level of education, country of residence, ethnicity and employment status. Participants that met inclusion criteria received an email from the research team containing their unique Study ID, with instructions on how to download the app and complete Meta Mind and a hyperlink to access the traditional metacognitive task in their browser. The email specified the sequence that each participant should complete the tasks, which was counterbalanced across the sample.

Traditional metacognitive task. The traditional metacognitive task was a visuo-perceptual decision-making task, which has been previously described (Fox et al., 2023). The task could be completed by participants via a web-browser on their personal computer (Figure 3.1A). On each trial, participants were shown a fixation cross for 1000 milliseconds (ms), followed by two sunflowers, positioned on the left and right of the screen for 300 ms. After the sunflowers disappeared from the screen, participants had unlimited time to make a judgement about which contained more seeds. The chosen sunflower was highlighted for 500ms, but no feedback on accuracy was provided. Participants then rated their confidence in each judgement, on a scale from 'Guessing' to 'Certain'. There was a total of 210 trials, divided equally into five blocks. Accuracy was controlled using a 'two-down one-up' staircase procedure, in which the task became easier (i.e. a larger seed difference between sunflowers) after each incorrect response and more difficult (i.e. a smaller seed difference between sunflowers) after two consecutive correct responses. This maintained objective performance across all participants between 60% and 85% correct, which ensured that estimates of confidence were not confounded by performance differences, and confidence biases can be accessed when accuracy does not vary across individuals. The first 25 trials participants experienced were in tutorial format and used as burn-in (i.e., to stabilise accuracy), and thus not used in the calculation of behavioural metrics like mean confidence. One sunflower was always half-filled (313 dots out of 625 positions), while the other box contained an increment of +6 to +81 dots compared to the standard. Changes in difficulty (seed differences between stimuli) were calculated in log-space, with a starting log difference of 4.2 (+70 dots). Differences in step size changed by ± 0.4 for the first five trials, ± 0.2 for the next five trials and ± 0.1 for the remainder of the task. Seed differences on each trial could range from as few as six dots (1.79 in log-space – the hardest to discriminate) to as many as 81 dots (4.39 in log-space – the easiest).

Meta Mind. Participants downloaded the smartphone app *Neureka* and entered their unique Study ID to their app profile, which differentiated their data from that of unpaid citizen scientists. In Meta Mind, players travel in a spaceship through the brain and make a series of choices based on stimuli they meet along the way (Figure 3.1B). At the start of each round, players are presented with a fact about

how brain health may be impacted by factors like sleep, diet and spending time in nature. Players are then instructed to navigate their ship to the stimuli containing more dots. On each trial, players presented trial-by-trial with pairs of moving icons representing that brain health fact (e.g. brains), that differ in the number of dots contained within them. As icons descend down the screen, players must select the icon with more dots, touching the screen to navigate their ship left or right to collide with the icon. After players make a choice, they then rate their confidence in the accuracy of their choice. After 20 trials, players evaluate their overall accuracy on that round, forming a 'global' self-performance evaluation based on round performance. Although no direct feedback on accuracy is provided, for gameplay reasons, participants are informed that the task will get harder when they get better at it and at the end of each round, the difficulty level for the next round is indicated to participants.

Meta Mind has three instructed tutorial trials, which are followed by 100 game play trials, divided into five rounds of 20 trials. The first 20 trials of game play are used to burn-in the 'two-down one-up' staircase, which maintains the average accuracy within a range of 0.60 to 0.85, as per the traditional task. The first 20 trials are not used for the calculation of task outcomes, leaving the 80 subsequent trials for analyses. On each trial, there is always one stimulus with 100 dots, randomly presented on left- or right-hand side of the screen. The minimum number of dots in the comparison icon is 101 and the maximum number of dot positions in the comparison stimuli is 149, which still left areas of unfilled space within the stimuli. As per the traditional task, change in dot differences is calculated in a log-space. The difficulty ranged from level 1 (easiest level, dot difference of 49 and a corresponding log value of 3.9) to level 26 (most difficult level, dot difference of 1 and a corresponding log value of 0.4). The tutorial trials all have a difficulty level of 1 and the first game play trial has a difficulty level of 13 (dot difference of 15 and a corresponding log value of 2.7), which changes based on trial-level accuracy over the subsequent 100 trials.

Ethical approval for both experiments was obtained from the Research Ethics Committee of School of Psychology, Trinity College Dublin (Approval ID: SPREC072019-01). All methods were performed in accordance with the relevant

guidelines and regulations. Both experiments only included adults over the age of 18.

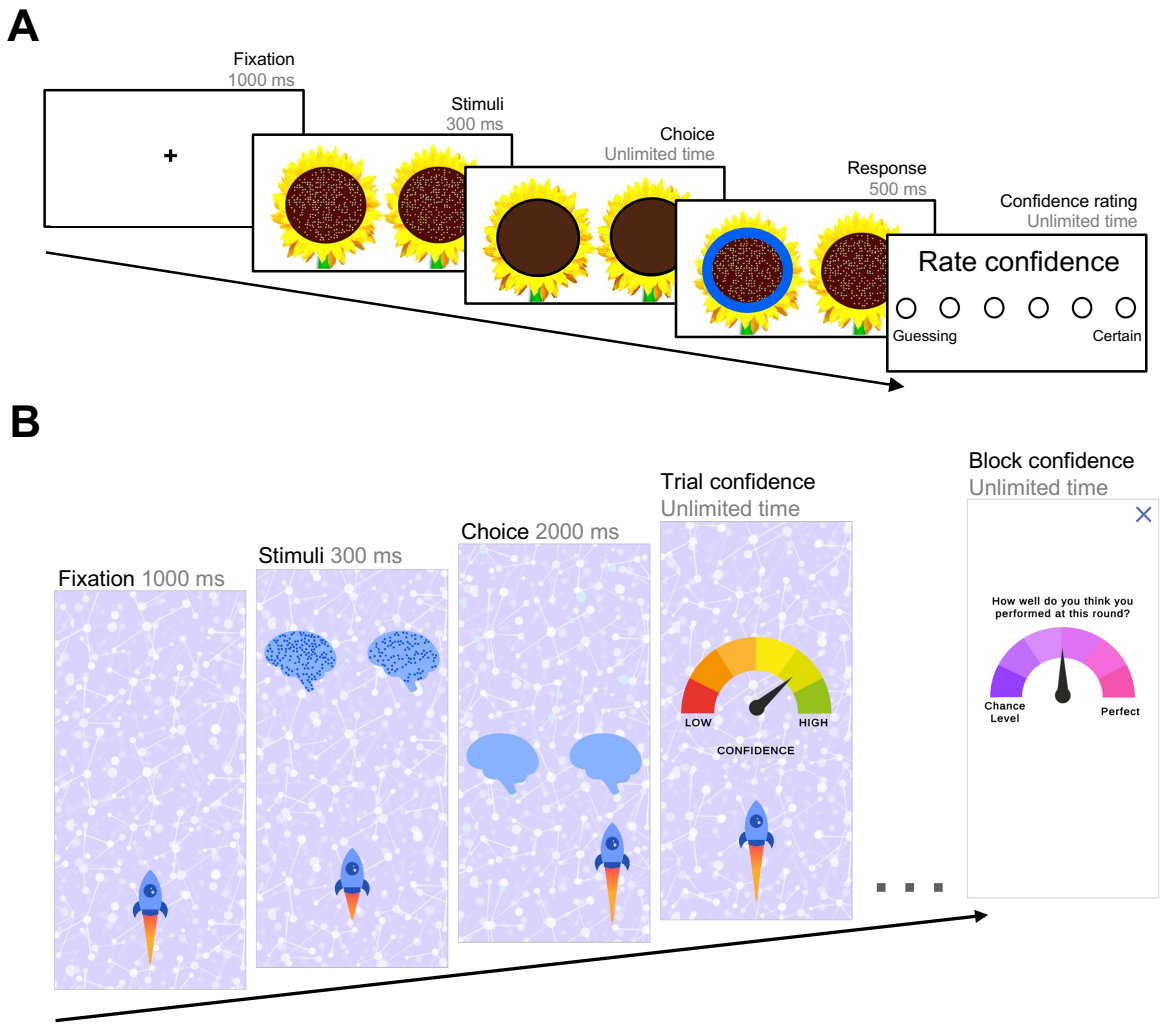


Figure 3.1. Task designs. **(A)** Perceptual decision-making task design (N=210 trials). On each trial, participants are shown a fixation cross for 1000 milliseconds (ms) before being shown two sunflowers with a different number of seeds for 300 ms. Participants are asked to judge and choose the sunflower contained more seeds (i.e., higher number of dots). This chosen sunflower is highlighted blue for 500 ms. Participants then have unlimited time to provide a confidence rating on their decision. **(B)** In Meta Mind, players are instructed to navigate their ship to the stimuli with more dots. At the start of each Meta Mind round (N=20 trials), players are shown a screen with no icons for 1000 ms. Icons then appear at the top of the screen and drift towards the bottom. Icons contain dots for 300 ms. After the dots disappear, blank icons appear on screen for up to 2000 ms, until a choice is made. Players tap the left or right side of the screen to choose the correct stimuli. Missed trials are recorded and repeated with icons that have the same dot difference. After Meta Mind players choose a stimuli, they then rate their local confidence in the accuracy of that choice on the trial, with unlimited time. After 20 Meta Mind trials, players evaluate their overall accuracy on that block (round-level, global confidence).

3.2.1.3 Data preparation and analysis

Behavioural outcomes and exclusions. Explicit confidence judgements are the conventional measure of metacognition in experimental tasks, required to evaluate meta-representations of the self[7]. For the traditional metacognitive task and Meta Mind, our primary outcome measure was metacognitive bias, calculated as mean confidence across trials. For both tasks, confidence on each trial was rated on a 6-point numeric scale, where 'Guessing' (traditional task)/'Low' (Meta Mind) = 1, and 'Certain' (traditional task)/'High' (Meta Mind) = 6. As shown in Figure 3.1, the text on the scales, but not the corresponding numbers, were presented to participants. Metacognitive efficiency (M-Ratio) was also calculated, which is the ratio of metacognitive sensitivity to mean accuracy, where sensitivity is the extent to which confidence ratings discriminate between correct and incorrect trials. M-Ratio, the gold-standard measure of metacognitive efficiency (Rahnev, 2023), was calculated in a hierarchical Bayesian framework (single-subject estimations) using the freely available HMeta toolbox (Fleming, 2017), http://github.com/smfleming/HMM, accessed June 2022. An M-Ratio value of 1 indicates that confidence was fully informed by accessing the total perceptual information available. 'Global' self-performance evaluations was calculated as the mean of round-level accuracy ratings across the four game rounds (every 20 trials). Self-performance evaluations were rated on a scale at the end of each block, from 'Change Level (1)' to 'Perfect (6)'. Task difficulty was measured as the mean seed/dot difference across trials, where more difficult trials had a smaller difference between stimuli.

Mean reaction time to stimulus choice across trials was measured in seconds and task accuracy was calculated as the mean proportion of correct responses across trials. Given the remote, online study design, participants had unlimited time to complete the tasks, and could do so across multiple days. Therefore, to estimate game completion time, we only considered completion times under 60 minutes (removing n=5 for Meta Mind and n=2 for the traditional task from this specific metric). Of N=116 participants recruited, N=62 completed the traditional task and the N=59 participants that played Meta Mind, with N=52 completing both. We employed a number of established exclusion criteria to ensure high data quality from the metacognitive task (Fox et al., 2023). Due to a software bug,

data for single trials on both tasks were missing choice response time for a small proportion of total trials (mean percentage of trials missing data was 0.02% for each task). These trials with missing data were discarded when calculating behavioural outcomes. For both tasks, participants who selected the right or left stimuli on more than 95% of trials or who had a mean accuracy <0.60 or >0.85 were to be excluded, but no participants met these criteria.

Statistical Analysis. Linear regression analyses were conducted with task type (Meta Mind or traditional task) as the independent variable to determine the effect of task type on the following task performance characteristics as separate dependent variables: time to complete, mean confidence, M-Ratio, mean accuracy and mean difficulty. Convergent validity, the extent to which two tasks measure the same underlying construct, was assessed using Pearson product moment correlation analyses for 1) mean confidence and, 2) M-Ratio. Split-half reliability, the consistency across each half of a measure, was assessed through Pearson product moment correlation analyses between odd and even trials on 1) Meta Mind and, 2) the traditional task, consistent with prior publications (Guggenmos, 2021; Rahnev, 2023). To determine the effect of task order on performance characteristics, linear regression were conducted with the order of task (Meta Mind being completed first or second) as the independent variable and the following separate dependent variables: mean confidence, M-Ratio, mean accuracy and mean difficulty (for Meta Mind/the traditional task). To evaluate the association between levels of the metacognitive hierarchy, Pearson correlation analysis were used to correlate local (trial-level) mean confidence with global (round-level) mean self-performance evaluations.

3.2.2 Experiment 2: Unpaid ‘citizen scientist’ sample

3.2.2.1 Participants

Of the 5997 general users of the Neureka app that completed the Meta Mind tutorial, 3776 (62.96%) played the full 100 trials between December 2021 and September 2023. General users of the Neureka app were unpaid ‘citizen scientists’. After applying exclusion criteria (detailed below), N=3410 individuals were retained for analyses. Participants were primarily female (n=2401, 70.41%), with a mean age of 50.77 (SD=13.88), were living in the United Kingdom

(n=2392, 70.15%) and had completed at least undergraduate level education (n=2082, 61.06%) (Table 3.2). A power analysis was carried out using effect sizes from a previous study examining cross-sectional associations between metacognition with Anxious-depression and Compulsivity and Intrusive Thought (Rouault et al., 2018). Sample sizes of N=454 and N=332 respectively were required to detect these associations with 80% power (linear regression analyses, two-tailed test). Therefore, the sample was well-powered to detect the association between metacognition and transdiagnostic psychiatric dimensions.

Table 3.2. Baseline Sociodemographic Characteristics of Unpaid Participants in Experiment 2.

Characteristic	Unpaid participants (N = 3410)
Gender, No. (%)	
Male	936 (27.45)
Female	2401 (70.41)
Other gender identity	73 (2.14)
Age, M (SD)	50.77 (13.88)
Country of residence, No. (%)	
United Kingdom	2392 (70.15)
United States	474 (13.90)
Ireland	299 (8.77)
Other	245 (7.18)
Education, No. (%)	
Below undergraduate level	1328 (38.94)
Completed at least undergraduate level	2082 (61.06)

3.2.2.2 Procedure

Neureka. Neureka, a nonprofit smartphone application developed and managed by the Gillan Lab at Trinity College Dublin, is available to download for free on Apple or Android phones through the Apple Store or Google Play Store. Since being launched in 2020, Neureka has over 23,000 registered users across 139 countries as of September 2023. The Neureka project aims to enrol members of the general public in scientific research, by voluntarily playing games that tap into

distinct cognitive processes underlying brain health (Donegan et al., 2023). After downloading Neureka, users provide informed consent and provide the following sociodemographic information upon registration: age, gender, levels of education and country of residence.

Meta Mind. Citizen scientists were able to play Meta Mind in Neureka by either completing 1) Meta Mind (n=1130, 33.14%), a single session science challenge consisting of the game plus mental health questionnaires, or 2) another science challenge called 'Brain Changer' (n=2280, 66.86%), a repeated session science challenge which includes a full version of Meta Mind (100 trials) and the same questionnaire to be completed on day 1, and is then followed by an abbreviated version of Meta Mind (reduced to 40 trials) bi-daily for 8 weeks alongside another bidaily cognitive test and daily self-report measures, which are not the topic of the present study. The design of Meta Mind was identical for all participants, regardless of which challenge was used to access the game. Of the 3410 participants that completed Meta Mind and the questionnaires described below, 110 played the 100-trial version Meta Mind twice within 30 days, once in Brain Changer and once in the stand-alone challenge. An additional 120 played an abbreviated, 40-trial version of Meta Mind 15 times across 8 weeks. These sub-samples of participants were used to examine the test-retest reliability of Meta Mind.

*Self-report psychiatric questionnaires***.** Participants completed 49 items taken from six self-report questionnaires that assess a variety of psychiatric symptoms, including depression (Zung Self-Rating Depression Scale) (Zung, 1965), trait anxiety (State Trait Anxiety Inventory) (Spielberger et al., 1983), impulsivity (Barratt Impulsiveness Scale 11) (Patton et al., 1995), obsessive-compulsive disorder (Obsessive- Compulsive Inventory-Revised) (Foa et al., 2002), eating disorders (Eating Attitudes Test) (Garner et al., 1982), apathy (Apathy Evaluation Scale) (Marin et al., 1991) (see Supplementary Material 7.2.1 for the full list of items). These items were chosen based on a previous study that demonstrated the original set of 209 items used to generate the Anxious-depression and Compulsivity and Intrusive Thought dimensions could be reduced to 49 (Wise & Dolan, 2020). In Brain Changer, the questionnaire items were presented to

participants before the Meta Mind game. In the stand-alone Meta Mind challenge, items were completed by participants after playing the game.

3.2.2.3 Data preparation and analysis

Behavioural outcomes and exclusions. All behavioural measures were the same as in Experiment 1, except we did not examine metacognitive efficiency because it exhibited poor psychometric properties in Experiment 1 and was not previously associated with transdiagnostic dimensions cross-sectionally (Benwell et al., 2022; Hoven et al., 2022; Rouault et al., 2018). In total, 3776 citizen scientists completed all 100 trials of Meta Mind. Data for single trials were excluded when choice response time were missing, due to a software bug (mean percentage of trials removed was 0.01%). No participants selected the right or left stimuli on greater than 95% of trials, but 12 (0.32%) participants had mean accuracy below 0.60 or above 0.85. N=3764 participants therefore progressed to the next step.

Transdiagnostic psychiatric dimensions and exclusions. Individual scores on dimensions of Anxious-depression and Compulsivity and Intrusive Thought were calculated by multiplying each of the 49 individual questionnaire responses by the corresponding weights from a previously published regularised regression model (Wise & Dolan, 2020) trained to predict the original factors (Gillan et al., 2016). Dimension scores were scaled to centre on zero, with higher scores indicating higher levels of transdiagnostic psychopathology. Of the 3764 participants with adequate Meta Mind data, 3442 (91.45%) completed all the questionnaire items. To determine the proportion of careless/inattentive responders on the self-report clinical questionnaires, we included a ‘catch’ question that was embedded in the impulsivity scale (*I competed in the 1917 Summer Olympics Game*) (Zorowitz et al., 2023). Thirty-two (0.93%) participants did not respond ‘Never’ to this question, and were subsequently excluded from analyses, leaving 3410 participants that were retained for statistical analyses.

Statistical analysis. Test-retest reliability of local metacognitive bias was evaluated using the ICC (two-way mixed-effects model, absolute agreement, single rater (Koo & Li, 2016)) of mean local confidence among the sub-sample of

individuals that played Meta Mind twice. ICC values between 0.50 and 0.75 indicate moderate reliability, values between 0.75 to 0.90 have good reliability, and values greater than 0.90 have excellent reliability (Koo & Li, 2016). To determine the association between levels of the metacognitive hierarchy, we used Pearson correlation analysis to correlate local (trial-level) mean confidence with global (round-level) mean self-performance evaluations. To examine simple effects, we used correlation (Pearson and biserial) analyses to test for relationships between metacognition (local/global) with task outcomes (mean accuracy, difficulty and reaction time), sociodemographic factors (age, gender and education) and device type (Apple/Android). To examine the relationship between metacognition (local/global) and transdiagnostic psychiatric dimensions, we included Anxious-depression and Compulsivity and Intrusive Thought as independent variables within the same model, with age, gender and levels of education as covariates.

To examine the effect of increasing trial number on mean confidence estimates, we firstly calculated mean confidence step-wise, cumulatively across the first 20 trials (the burn-in period). That is, we calculated confidence on trial 1, then the average of trials 1 and 2, and so on until all 20 trials were averaged over. We then repeated this for the main game trials after the burn-in period, calculating mean confidence from the 21st trial and then increasing data with every subsequent trial, until the 100th trial. Linear mixed-model regression analyses were conducted to examine the effect of increasing trial number, with participants as random effects, on mean local confidence, mean global evaluations and mean accuracy for the burn-in trials and main game play trials separately. We also ran linear mixed-model analyses to examine the fixed effect of longitudinal assessment timepoint (15 in total) on mean local confidence, mean global evaluations, mean accuracy and mean difficulty, with participants as a random factor.

The split-half reliability for mean local confidence was calculated using Pearson correlation analyses of odd and even trials, from the first 2 trials in bins of 2 trials until 100 trials (for the burn-in period and main game trial separately). We fit an exponential model to the area under the curve (AUC) sequence of split-half

reliability values across main game play trials and calculated the minimal number of trials required to reach the asymptotic limit of split-half reliability for local confidence (95% of its asymptote across trial bins) (Buyalskaya et al., 2023). To determine the minimal number of cumulative trials required to detect the association between metacognition and psychiatric dimensions, we examining the interaction effect of psychiatric dimension and trial number on mean local confidence/global evaluations, controlling for age, gender and level of education (across the burn-in trials and main game trials separately). An interaction effect with $p>0.05$ for this analysis would indicate that the association between metacognition and psychiatric dimensions was not dependent on the trial number (i.e., the association was stable across trials).

For all tests in Experiment 1 and 2, statistical significance was defined as $p<0.05$, with two-tailed p-values used. Adjustments for multiple comparisons were not conducted. For regression analyses, all the dependent variables and continuous independent variables were z-scored before entering the models as to obtain standardised (i.e. comparable) regression coefficients. Gender was coded numerically (male=1, female=-1, other=0).

3.3 Results

3.3.1 Experiment 1: Paid participant sample

Comparing the tasks, Meta Mind was relatively shorter, taking on average 7.86 minutes (SD=1.86) to complete, while the traditional task took 21.50 minutes (SD=6.87) (β=1.60, SE=0.12, t=13.15, $p<0.001$). Figure 3.2 shows performance characteristics across the two metacognitive tasks. Metacognitive bias in Meta Mind, operationalised as mean local confidence, was significantly higher (M=4.33, SD=0.67) than in the traditional task (M=3.80, SD=0.74) (β=-0.70, SE=0.18, t=-3.81, $p<0.001$) (Figure 3.2A). In addition to metacognitive bias, we also quantified metacognitive efficiency via M-Ratio (i.e., the ratio of metacognitive sensitivity to mean accuracy, where sensitivity is the extent to which confidence ratings discriminate between correct and incorrect trials (Rahnev, 2023)). M-Ratio was higher for Meta Mind (M=0.97, SD=0.54) compared to the traditional task (M=0.77, SD=0.38) (β=-0.41, SE=0.19, t=-2.13,

p=0.036) (Figure 3.2B). Despite the use of a staircase procedure, task accuracy was higher in Meta Mind (M=74%, SD=3) than the traditional task (M=70%, SD=3) (β=-1.03, SE=0.17, t=-6.08, *p*<0.001) (Figure 3.2C). As expected, the traditional task had a significantly higher mean dot difference (M=41.89, SD=15.38) than Meta Mind (M=7.91, SD=2.81) (β=1.83, SE=0.08, t=23.31, *p*<0.001), given the constraints with increasing the dot number on relatively smaller phone screens when designing Meta Mind. There was modest evidence for learning effects for task difficulty only. Those who completed Meta Mind after first completing the traditional task achieved a higher level of objective difficulty (i.e., a lower dot difference) on the Meta Mind game (M=8.92, SD=3.13 vs. M=6.91, SD=2.05, β=-0.36, SE=0.13, t=-2.74, *p*=0.008). The analogous effect was not significant for the traditional task (β=-0.18, SE=0.14, t=-1.29, *p*=0.202). There were no effects of task order on mean confidence, M-Ratio or accuracy in either task (all *p*>0.19).

Mean local confidence was moderately correlated across tasks (*r*(50)=0.64, *p*<0.001), indicating adequate convergent validity (Figure 3.2D). Within Meta Mind, local confidence and global self-performance evaluations were correlated (*r*(50)=0.75, *p*<0.001) (Figure 3.2F). The split-half reliability of local mean confidence was excellent for both Meta Mind (*r*(50)=0.91, *p*<0.001) and the traditional task (*r*(50)=0.98, *p*<0.001) (Figure 3.2G and 3.2H). In contrast, there was no significant association between M-Ratio across task versions (*r*(50)=0.03, *p*=0.853) (Figure 3.2E) and split-half reliability for M-Ratio in Meta Mind neared 0 (*r*(50)=0.04, *p*=0.794) and was poor for the traditional task (*r*(50)=0.25, *p*=0.075) (Figure 3.2I and 3.2J). This follows recent work demonstrating that 100 should be considered the lowest boundary for the sufficient number of trials when estimating M-Ratio (Rahnev, 2023), as Meta Mind and the traditional task only had 40 and 105 trials in each split (odd vs. even), respectively.

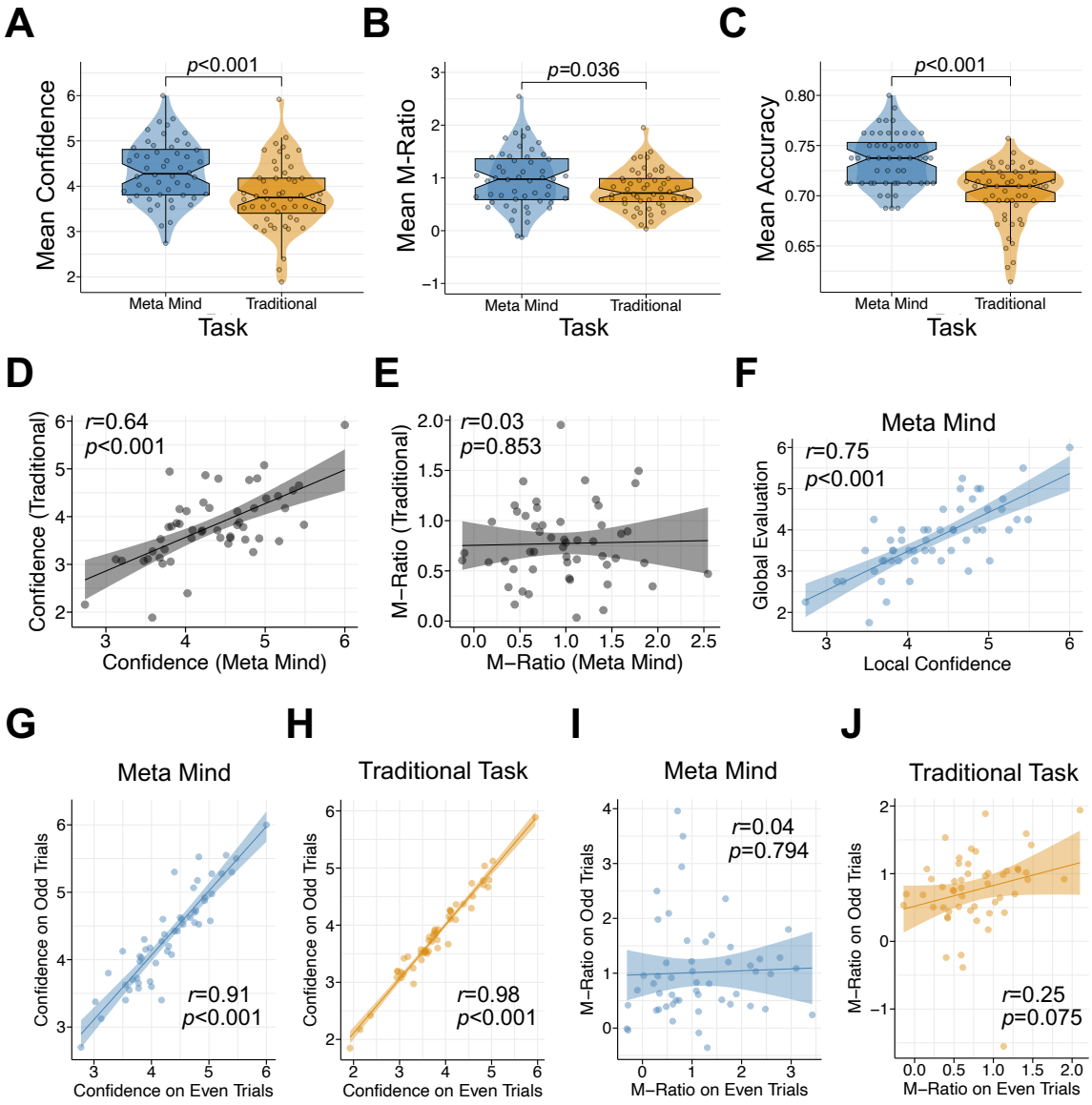


Figure 3.2. Comparing Meta Mind against the traditional metacognitive task (N=52). Mean confidence **(A)**, M-Ratio **(B)** and mean accuracy **(C)** were all significantly higher with Meta Mind (blue), compared to the traditional task (orange). While mean local confidence was significantly associated across tasks **(D),** there was no association between M-Ratio measure through Meta Mind and the traditional task **(E)**. In Meta Mind, those with higher mean local confidence had higher mean global evaluations of performance **(F)**. There was a high correlation between odd and even trials for mean confidence on both metacognitive tasks, indicating sufficient split-half reliability **(G, H)**. Split-half reliability was poor for M-Ratio across tasks **(I, J)**.

3.3.2 Experiment 2: Unpaid 'citizen scientist' sample

3.3.2.1 Test-retest reliability with the 100-trial version of Meta Mind

Among the 110 unpaid, citizen scientists that played the 100-trial version of Meta Mind twice, the median time interval between test and retest game completion was 2 days (SD=7.74) (Figure 3.3A). An intra-class correlation (ICC; two-way mixed-effects model with absolute agreement, single rater) of 0.86, with 95% confident interval = 0.80-0.90, $p<0.001$), was calculated for local confidence bias among repeated Meta Mind players, indicating good test-retest reliability (Koo & Li, 2016) (Figure 3.3B). A Pearson correlation coefficient of $r(108)=0.87$, $p<0.001$ also indicted a strong association between test and retest mean confidence.

3.3.2.2 Individual differences in local, trial-level confidence

Examining correlations across task outcomes in the full sample (N=3410), local confidence was slightly higher in individuals with greater task accuracy ($r(3408)=0.05$, $p=0.008$), in those with an increased mean dot difference (easier difficulty level on average) ($r(3408)=0.14$, $p<0.001$), and in those with faster reaction times ($r(3408)=-0.08$, $p<0.001$). Of note, the large sample size in this study (N=3410) may contribute observations of statistical significance, even for very weak associations (e.g., $r(3408)=0.05$, $p=0.008$ for the correlation between mean local confidence and mean accuracy). There was a strong correlation between local trial-level confidence and global round-level self-performance evaluations ($r(3408)=0.79$, $p<0.001$) (Figure 3.3C), replicating the association reported in Experiment 1. When examining the effects of device type (Apple/Android), participants with Apple devices (n=506) had higher mean confidence ($r_{pb}(3408)=0.07$, $p<0.001$), higher mean global evaluations ($r_{pb}(3408)=0.05$, $p<0.001$) and slower reaction times ($r_{pb}(3408)=0.11$, $p<0.001$) compared to Android users (n=2904). There was no association between device type and task accuracy ($r_{pb}(3408)=0.01$, $p=0.725$), difficulty ($r_{pb}(3408)=-0.02$, $p=0.352$), or levels of educational attainment ($\chi^2(1)=0.08$, p=0.771). Android devices were more common among female ($\chi^2(2)=8.59$, $p=0.003$) and older participants ($r_{pb}(3408)=-0.24$, $p<0.001$).

Including sociodemographic factors in separate models, older adults

(r(3408)=0.06, p<0.001), males (r(3408)=0.09, p<0.001) and those with lower levels of educational attainment (r_{pb}(3408)=0.10, p<0.001) had significantly higher mean local confidence. Controlling for age, gender and levels of education in the model, participants with higher levels of Anxious-depression had lower levels of mean confidence (β=-0.27, SE=0.02, t=-13.41, p<0.001, r^2=0.05) while those with higher levels of Compulsivity and Intrusive Thought had elevated mean confidence (β=0.08, SE=0.02, t=3.97, p<0.001, r^2=0.004) (Figure 3.3D). Although mean accuracy and device type were associated with mean confidence, including these as additional covariates in the model did not affect the significant association between confidence bias with Anxious-depression (β=-0.26, SE=0.02, t=-12.98, p<0.001) or Compulsivity and Intrusive Thought (β=0.08, SE=0.02, t=3.94, p<0.001).

After controlling for mental health dimensions, the effect of age on local confidence was no longer significant (β=0.02, SE=0.02, t=1.04, p=0.299), while the effects of gender (β=0.09, SE=0.02, t=4.84, p<0.001) and educational attainment (β=0.28, SE=0.03, t=8.14, p<0.001) held within this model. As the controlled effect of age on local confidence was not significant in the model, we examined which variables accounted for the significant correlation between age and local confidence. Specifically, accounting for Anxious-depression in the model removed the significant effect of age on local confidence (β<0.001, SE=0.02, t=0.03, p=0.979).

3.3.2.3 Individual differences in global, round-level self-performance evaluations

Unlike local confidence, mean global self-performance evaluations were not significantly correlated with task accuracy (r(3408)=0.03, p=0.071), but like local confidence, global evaluations were higher in those who had an increased mean dot difference (easier difficulty level) (r(3408)=0.12, p<0.001), those who had faster response times (r(3408)=-0.08, p<0.001), and those with Apple devices (r_{pb}(3408)=0.05, p=0.001).

Similar to local confidence, older adults (*r*(3408)=0.06, *p*<0.001), males (*r*(3408)=0.15, *p*<0.001) and those with lower levels of educational attainment (*r*(3408)=0.05, *p*<0.001) had significantly higher global self-performance evaluations, when included in separate models. Consistent with the local confidence results, those with higher levels of Anxious-depression had lower global self-performance evaluations (β=-0.26, SE=0.02, t=-13.28, *p*<0.001, r^2=0.05), and those with higher levels of Compulsivity and Intrusive Thought had increase global evaluations (β=0.04, SE=0.02, t=2.27, *p*=0.023, r^2=0.001), controlling for age, gender and education in the model (Figure 3.3D). Similar to local confidence, the uncontrolled effect of age on global evaluations became non-significant when the transdiagnostic dimensions were included in the model specifically (β=0.004, SE=0.02, t=0.23, *p*=0.816).

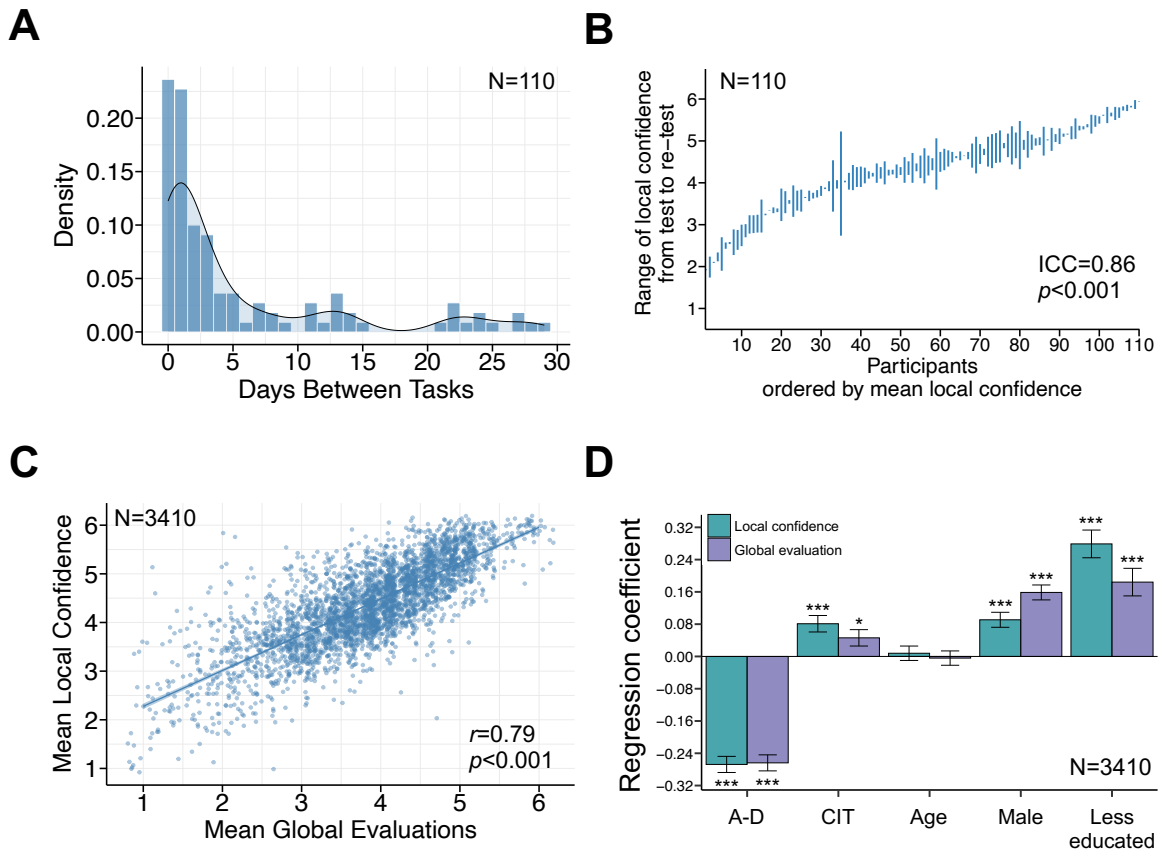


Figure 3.3. Validation of Meta Mind in a large sample of unpaid citizen scientists. N=sample size, ICC=intra-class correlation coefficient, *r*=correlation coefficient, *p*=p-value, A-D=Anxious-depression, CIT=Compulsivity and Intrusive Thought, ***=*p*<0.001, **=*p*<0.01, *=*p*<0.05. **(A)** Density plots of days between Meta Mind test and retest completion (median=2, SD=7.74) (N=110). **(B)** There was good test-retest reliability for mean confidence among sub-samples of citizen scientists that played 100 trials of Meta Mind twice within 30 days (N=110). **(C)** Those with higher local, trial-level confidence had elevated, round-level global evaluations of task performance (N=3410). **(D)** Among N=3410, mean local

confidence and global evaluations were higher among males and those that were less educated. Those with higher levels of Anxious-depression (A-D) had lower local confidence and global evaluations, while those with higher levels of Compulsivity and Intrusive Thought (CIT) have elevated local confidence. The positive association between CIT and mean global evaluations was marginally significant at p=0.023. The error bars represent the standard error around the standardised beta coefficient.

3.3.2.4 Measuring metacognition in few trials

To examine the impact of trial number on estimates, we removed 22 participants that were missing data for at least one trial due to a software bug, leaving 3388 participants with 100 trials of confidence and task performance data. First we examined split half-reliability for local confidence (Figure 3.4A). The exponential function fitted to the sensitivity curve of split-half reliability for main game trials accounted for a significant proportion of variance explained in local confidence ratings (r^2=0.98). Internal consistency for mean confidence with the main game trials reached the 95% asymptote with the exponential model fitted to the split-half reliability values at just 40 trials (*r*=0.90, 95% CI [0.89-0.90]) (Figure 3.4A, vertical dashed line). Split-half reliability then plateaued and reached *r*=.97 (95% CI [0.97-0.97]) when based on the full 100 trials (Figure 3.4A).

Figure 3.4B shows how the association between mean confidence with Anxious-depression and Compulsivity and Intrusive Thought changes with increasing trial number, controlling for age, gender, levels of education in the model. As 40 trials were required to optimise reliable estimates of local confidence, we examined the stability of clinical correlates with local confidence from 40 to 100 trials. The negative association between mean local confidence and Anxious-depression was significant with 40 trials (β=-0.24, SE=0.02, t=-12.13, *p*<0.001) and with 100 trials (β=-0.26, SE=0.02, t=-13.11, *p*<0.001). The association between local confidence and Anxious-depression remained stable from 40 to 100 main game trials, with no significant interaction effect of Anxious-depression and trial number on mean confidence (β=-0.004, SE=0.003, t=-1.41, *p*=0.158) (Figure 3.4B). Similarly, the positive association between local confidence and Compulsivity and Intrusive Thought was evident at 40 trials (β=0.06, SE=0.02, t=2.84, *p*=0.005) and at 100 trials (β=0.08, SE=0.02, t=3.90 *p*<0.001), and remained stable from 40 to 100 main game trials (β=0.004, SE=0.003, t=1.42, *p*=0.155) (Figure 3.4B). Therefore, clinical correlates with local confidence could be detected with only 40

trials. Excluding the first block's rating, there was no significant interaction effect of block number with Anxious-depression (β=-0.004, SE=0.01, t=-0.41, *p*=0.680) or Compulsivity and Intrusive Thought (β=0.008, SE=0.01, t=0.82, *p*=0.414) on mean global evaluations, indicating clinical correlates with global evaluations were also stable across blocks (Figure 3.4C). Clinical correlates with mean local confidence and global evaluations were also stable across the binned trials (see Supplementary Material 7.2.2).

Finally and for completion, we examined changes in the mean level estimates from the task over trials. In the first block, there was a significant decrease in mean confidence with the accumulation of trials (β=-0.07, SE=0.002, t=-49.16, *p*<0.001) (Figure 3.4D, trials 1-20). For the subsequent 80 main game trials, mean confidence slightly decreased further, but this effect was small (β=-0.002, SE<0.001, t=-4.05, *p*<0.001) (Figure 3.4D, trials 21-100). Excluding the first block rating, mean global evaluations, in contrast, slightly increased with the accumulation of blocks (β=0.03, SE=0.003, t=10.09, *p*<0.001) (Figure 3.4E). As expected, the mean dot difference reduced throughout the first block (β=-0.27, SE=0.002, t=-149.30, *p*<0.001) and main game trials (β=-0.16, SE=0.001, t=-273.6, *p*<0.001), reflecting an increased in task difficulty with continued game play (Figure 3.4F). Finally, accuracy significantly declined during the burn-in trials (β=-0.07, SE=0.003, t=-24.79, *p*<0.001) and continued to decline throughout the game (β=-0.15, SE=0.002, t=-90.91, *p*<0.001) (Figure 3.4G). Despite this, accuracy remained firmly within the bounds of the upper and lower limits of the acceptable range (60%–85%).

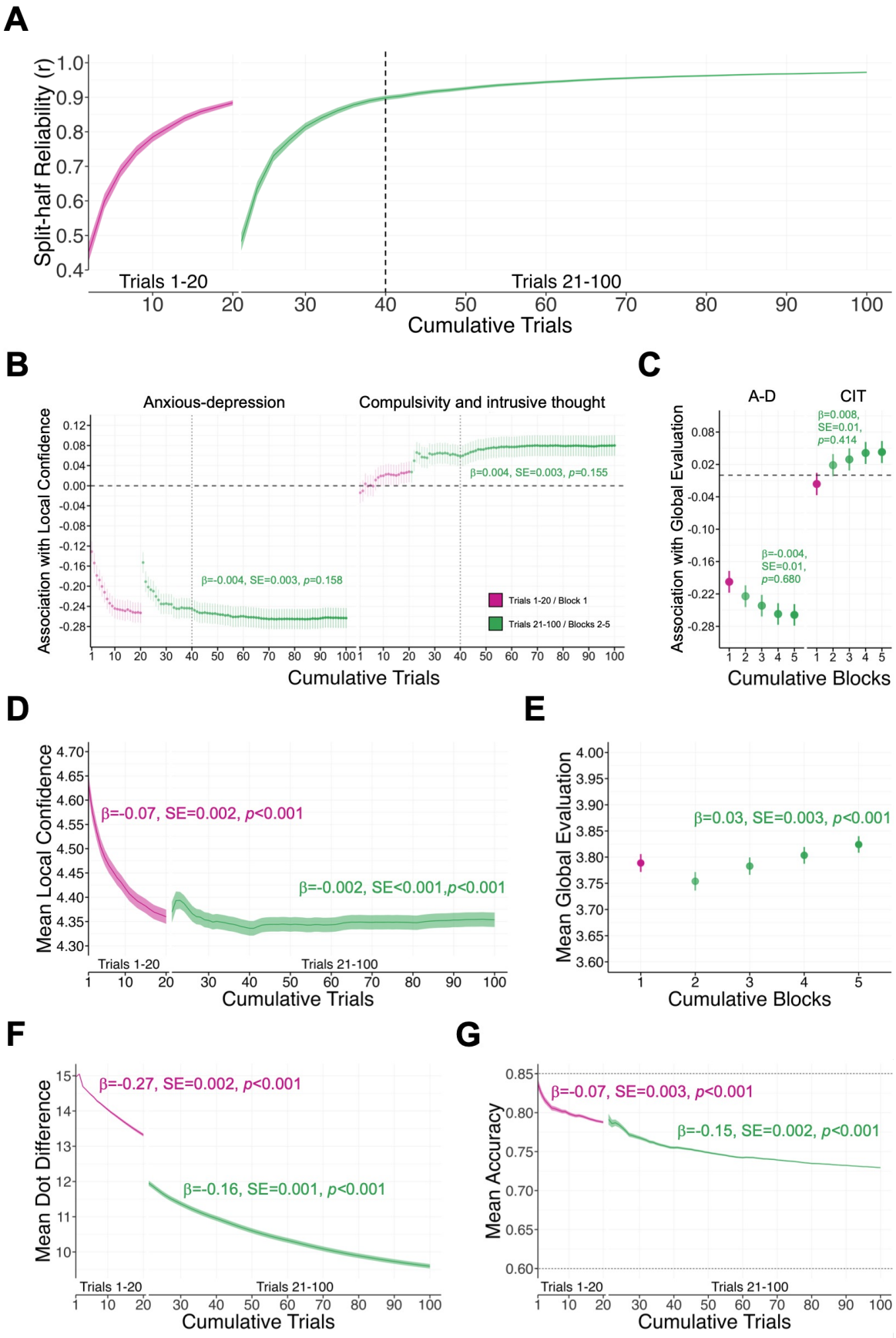


Figure 3.4. Impact of increasing trial number on mean confidence (N=3388).
β=standardised beta coefficient, SE=standard error, *p*=p-value, ***=*p*<0.001, **=*p*<0.01, *=*p*<0.05
A-D = Anxious-depression, CIT = Compulsivity and Intrusive Thought. The error bars represent

the standard error around the standardised beta coefficient. **(A)** Split-half reliability of mean local confidence was calculated separately for the initial 20 burn-in trials (pink) and the 80 main game play trials (green). Only 40 trials were required to estimate the split-half reliability of local confidence at r=0.90. **(B)** The significant associations between mean local confidence with Anxious-depression and Compulsivity and Intrusive Thought was evident at 40 trials (vertical lines) and remained stable to 100 trials. **(C)** Following the burn-in block (pink error bar, block 1), there was no significant interaction effects of additional block number and transdiagnostic dimensions on mean global evaluation, indicating that associations with global evaluations remained stable with cumulative round-level ratings (green error bars, blocks 2-5). **(D)** While mean confidence significantly reduced during the burn-in period for the staircase (pink line, trials 1-20), confidence estimates were more stable across the remaining 80 main game trials (green line, trials 21-100). **(E)** Following the burn-in block (pink bar, block 1), mean global evaluations increased with cumulative blocks (green error bars, blocks 2-5). **(F)** Mean dot difference and **(G)** mean accuracy declined across the burn-in and then continued to decline across main game trials.

3.3.2.5 Measuring metacognition repeatedly with the 40-trial version of Meta Mind

A sub-sample of 120 citizen scientists played an abbreviated, 40-trial version of Meta Mind 15 times over an 8-week period, with a median time interval of 2 days (SD=1.06) between games played. Among this subsample, test-retest reliability for local confidence was good across the 15 assessment points (ICC (A,1) [CI] = 0.86 [0.83, 0.89], p<.001) (Figure 3.5A). Similarly, global self-performance estimates had good reliability across the 8-week period (ICC (A,1) [CI] = 0.71 [0.65 0.76], p<0.001) (Figure 3.5B). To characterise any potential practice effects, we ran linear mixed-model analyses to examine the fixed effect of assessment timepoint on mean local confidence, mean global evaluations, mean accuracy and mean difficulty, with participants as a random factor. There was a marginally significant increase across assessment points in mean confidence (β=0.02, SE=0.008, t=1.97, p=0.050) (Figure 3.5C), but no significant change in global self-performance estimates (β=0.004, SE=0.01, t=0.35, p=0.727) (Figure 3.5D). Mean dot difference significantly decreased across assessment points (β=-0.04, SE=0.01, t=-2.51, p=0.012), reflecting an increase in task difficulty across time as participants became better at the perceptual discrimination task (Figure 3.5E). Accuracy was staircased, and as a result the values were highly stable across days (Figure 3.5F). Indeed, due to low variance, the mixed-model across the 15 assessment points would not converge for the analysis of mean accuracy. As we could not fit this model, we instead compared first vs. 15th assessment, and

found there was no significant change in accuracy when comparing the assessment timepoints (β=0.19, SE=0.12, t=1.60, *p*=0.110).

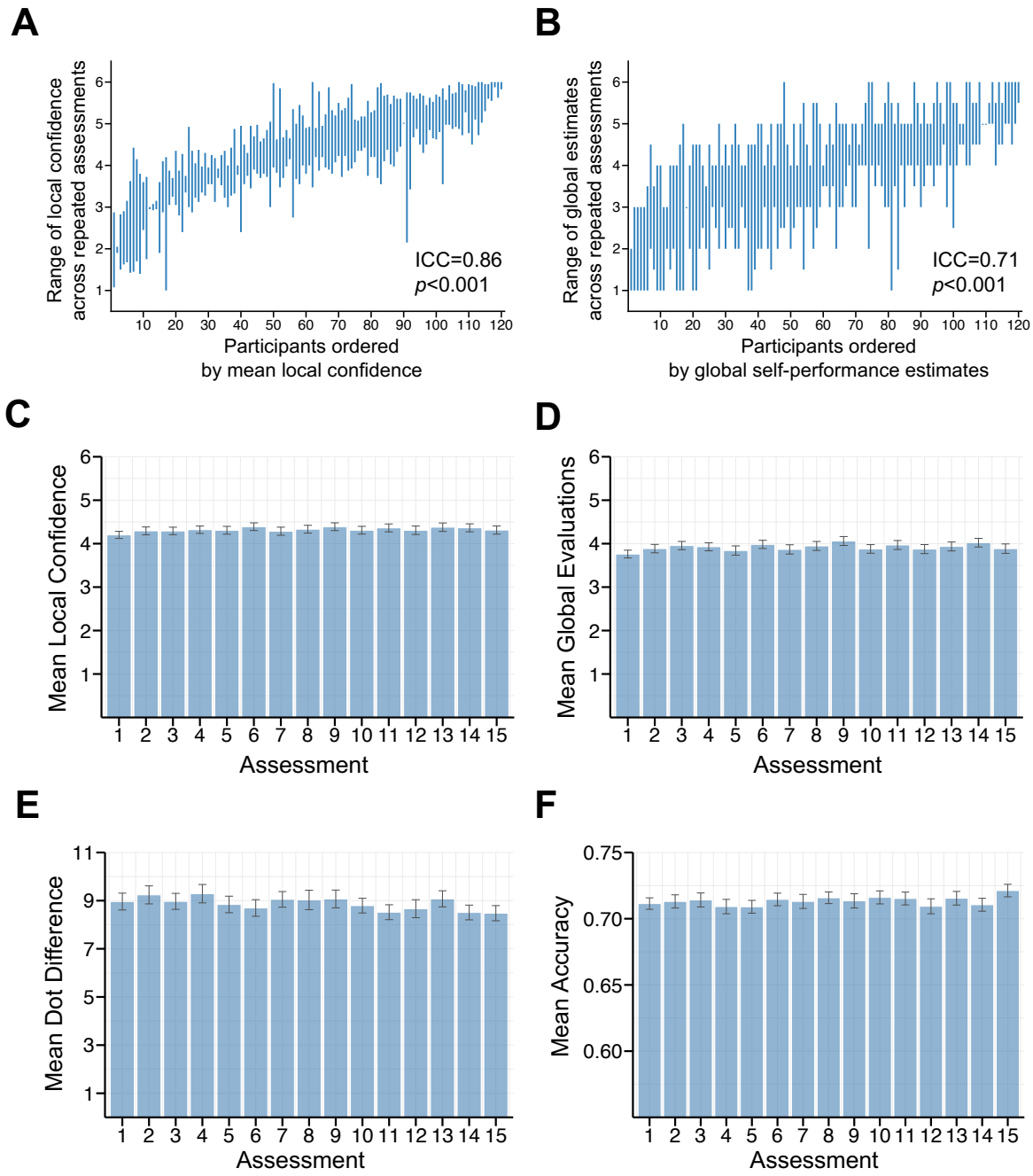


Figure 3.5. Measuring metacognition longitudinally (N=120). ICC=intra-class correlation coefficient, SE=standard error, *p*=p-value. Good test-retest reliability was evident for **(A)** mean local confidence and **(B)** global self-performance estimates among the N=120 that played the abbreviated version of Meta Mind 15 times over 8 weeks. **(C)** Mean local confidence marginally increased, but **(D)** global self-performance evaluations did not significantly change when examining the impact of timepoint with binned assessments. **(E)** Task difficulty increased across the 15 assessment points, as indexed by a decrease in mean dot difference. **(F)** Due to low variance, the mixed-model across the 15 assessment points would not converge for mean accuracy. In a simpler comparison of 1st vs. 15th session, there was no significant change in accuracy.

3.4 Discussion

There is growing interest in the study of metacognition in psychiatric populations (Hoven et al., 2019; Seow et al., 2021), with well-replicated observations of reduced confidence in those with higher levels of Anxious-depression and elevated confidence in those endorsing symptoms of Compulsivity and Intrusive Thought (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2022; Rouault et al., 2018; Seow & Gillan, 2020). But progress in understanding the mechanisms underlying these biases in transdiagnostic psychiatric dimensions has been slower, in part due to overreliance on cross-sectional study designs, paid participants from crowdsourced platforms and unknown psychometric properties of mainstay tests. In this study, we aimed to support a move towards repeated within-person and remote assessment by developing a brief and reliable task to measure metacognition via a smartphone application among citizen scientists.

In a considerably abbreviated smartphone task, metacognitive bias had acceptable convergent validity, as mean confidence was moderately correlated across traditional and smartphone tasks. Split-half reliability for metacognitive bias was excellent and empirically stabilised after 20 game trials, which included the burn-in period, corresponded to requiring 40 trials in total. With both the 100- and 40-trial versions of Meta Mind, metacognitive bias had good test-retest reliability, indicating that metacognitive estimates for each individual were highly stable across time. This is consistent with previous findings of strong test-retest reliability for confidence ratings in a traditional visual metacognitive task, with as few as 50 trials (Rahnev, 2023). This high level of reliability in our study contrasts with recent reports that other cognitive tasks under study in the field of computational psychiatry suffer from poor reliability (Karvelis et al., 2023). In contrast, prior work suggests that self-report measures are considerably more reliable than behavioural readouts alone (Enkavi et al., 2019; Vrizzi et al., 2023). This facet of our task, coupling self-report with the experimental control of a behavioural test (Fleming et al., 2010), may explain the high reliability that we observed. This sort of ‘hybrid self-report cognition’ task may be one way to bridge the recently discussed reliability gap in computational psychiatry. This approach may be particularly suited for studying the mechanisms of biases in thinking and

feeling specifically, which may only noisily lead to downstream changes in behaviour on a task. We showed that task accuracy was highly stable across multiple sessions, indicating there was no significant practice effects with Meta Mind. This was in line with the task design, as stability in task performance is maintained by adjusting in task difficulty, which increased across sessions. In addition to practice factors, affective state can be a source of significant temporal variability in cognitive task performance (Schurr et al., 2024). However, we did not measure any changes in affective state or psychiatric symptoms across time. Given that confidence bias is state-dependent (Fox et al., 2023) and changes over a period of days (Da Fonseca et al., 2023), we would expect stable metacognitive abilities in our sample to coincide with stability in psychiatric states across time. An interesting target of future research may be to uncover the dynamic nature of the relationship between metacognition and mental health and as it fluctuates with mental state. This could be achieved by adopting Meta Mind as a clinical tool to monitor these fluctuations across various timescales (e.g., over hours, days, weeks). Assessing these dynamic interactions repeatedly over time is the next step to test causal models of metacognition and psychopathology.

As a further demonstration of validity, we replicated the established patterns of local metacognitive biases across transdiagnostic psychiatric dimensions (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2022; Rouault et al., 2018; Seow & Gillan, 2020). As the associations between metacognitive bias and psychiatric dimensions are replicable across different task types, including a learning task (Seow & Gillan, 2020) and for general knowledge (Benwell et al., 2022), this would suggest that our findings on metacognition are not task-dependent, and would translate to other decision-making tasks, regardless of the cognitive facet. Specifically, those with higher levels of Anxious-depression had lower confidence, while those with higher levels of Compulsivity and Intrusive Thought had elevated confidence in their performance. We additionally found associations between psychiatric dimensions and global evaluations, which were analogous to the local confidence level results. Anxious-depression was associated with reduced global evaluations of performance, which is consistent with prior work showing widespread biases across the metacognitive hierarchy,

from low-level perceptual decisions up to notions of self-worth (Hoven et al., 2022; Seow et al., 2021). Conversely, greater levels of Compulsivity and Intrusive Thought were associated with higher global estimates (Hoven et al., 2022), but were only marginally significant with a negligible effect size. This could suggest that overconfidence in Compulsivity and Intrusive Thought manifests more at the trial-level, due to specific learning biases for local decision-making. Indeed, this was observed in a prior study using a reinforcement learning task, where compulsivity, but not Anxious-depression, was linked to a failure to use trial-level feedback to update metacognitive bias (Seow & Gillan, 2020).

While the effect of Compulsivity and Intrusive Thought on local confidence bias was significant, this was weaker than the effects reported in prior studies that used dot discrimination metacognitive tasks (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2022; Rouault et al., 2018). This may be because we used a reduced set of questions that a prior study demonstrated were highly predictive of scores (Wise & Dolan, 2020), based on the full 209 questionnaires items from the original paper (Gillan et al., 2016). It is therefore possible that this decreased the sensitivity to detecting overconfidence in Compulsivity and Intrusive Thought. The reduced set omitted items from the schizotypy scale for example, which might be important for capturing positive symptoms that are linked to overconfidence bias (Schmack et al., 2021). Given this, future studies with smaller sample sizes should consider using the full questionnaire items to have sufficient power to detect the associations between local confidence and psychopathology (Benwell et al., 2022; Fox et al., 2023; Hoven et al., 2022; Rouault et al., 2018). Although we provide evidence for Meta Mind's convergent validity, this is specifically within the visuo-perceptual domain. While there is evidence that visuo-perceptual metacognition generalises across domains (e.g., convergence with confidence in general knowledge [15]), the validity of Meta Mind as a measure of domain-general metacognition should be further investigated by comparing Meta Mind to metacognitive tasks in alternative domains (e.g., general knowledge, memory, or even other perceptual modalities).

It must be noted that the distinction between global and local assessments was subtle in Meta Mind. Across both experiments, there was overall a strong cross-

sectional correlation between local confidence and global evaluations. While this is in line with other studies (Friedemann et al., 2023; Hoven et al., 2022), this strong association could mean that our measure of global evaluations (task-level rating) might be too proximal to the local level to reflect its aggregation over time. Future research should consider using alternative measures for global metacognition, such as task-level choices, which may be less proximal to local confidence ratings (Hoven et al., 2022; Rouault et al., 2019). Although we included measures of local confidence and global performance evaluations within the task, a limitation of the current study is that we did not collect a measure of global self-beliefs. Global self-beliefs, such as self-esteem, are higher-order stable traits that are thought to be partially informed by local and global confidence (Seow et al., 2021). Recent work using this hierarchical framework demonstrated that self-esteem, a form of global self-belief, is lower in Anxious-depression and to a lesser extent, Compulsivity and Intrusive Thought (Hoven et al., 2022). This differs markedly from observations at the bottom of the hierarchy – i.e., local confidence – where opposing patterns of association are consistently observed for these transdiagnostic traits. Further, discrepancies in how local confidence and global self-beliefs interact were recently observed in a parallel paper using the same smartphone task from the present study (Meta Mind) (Friedemann et al., 2023). In this study, self-esteem in problem gamblers was found to be correlated with local confidence, but no such correlation was observed in comparison subjects who did not gamble (Friedemann et al., 2023). To unpack this, future studies with large samples could consider including measures of high level self-beliefs, to investigate the contemporaneous and temporal relationships between the various hierarchical levels of metacognition and psychiatric dimensions, with particular focus on how they develop and interact over time.

The association between local confidence and transdiagnostic dimensions could be detected rapidly, with 40 trials required to detect stable estimates of underconfidence in Anxious-depression and overconfidence in Compulsivity and Intrusive Thought, respectively. Considering this with split-half reliability findings, valid and precise estimates of local metacognitive biases can be obtained with as few as 40 trials. Although global evaluations of self-performance increased with

cumulative ratings, the association between global evaluations of self-performance and Anxious-depression could be detected with a single rating and did not increase with more ratings. The weaker association between global evaluations and Compulsivity and Intrusive Thought was also stable across block ratings.

In contrast to metacognitive bias, we found that metacognitive efficiency was neither valid nor reliable in Meta Mind. While M-Ratio is the dominant standard measure of metacognitive efficiency, estimates are dependent on trial number (Guggenmos, 2021; Rahnev, 2023). Unacceptable internal consistency for M-Ratio with Meta Mind is consistent with previous findings of poor split-half reliability for M-Ratio with trial numbers below 100 (Guggenmos, 2021; Rahnev, 2023). This raises questions over the continued use of this metric in individual differences research, despite the widespread use of lower trial numbers than recommended. Even with 400 trials, test-retest reliability of M-Ratio was found to be very low (Rahnev, 2023), indicating that M-Ratio estimates may lack stability, as opposed to being 'unreliable' per se. To account for any instability in M-Ratio estimates, future studies should account for potential sources of temporal variability, such as state-like changes[33,35]. Overall, our data lend support to the idea that researchers should carefully choose a reliable measure, suitable for the task design and appropriate to the inferences about individual differences in metacognitive efficiency they aim to make.

Examining sociodemographic factors, males and individuals with lower levels of educational attainment had higher local confidence consistent with previous findings (Fox et al., 2023). With respect to global evaluations of performance, the direction of effects was the same but interestingly, effect sizes for gender were much larger for global compared to local and the opposite was observed for educational attainment. While uncontrolled analyses suggested that older adults had higher local confidence and global evaluations, this effect was also no longer present when we accounted for psychiatric dimensions in the model. While higher local confidence and global evaluations have been previously reported among younger adults (McWilliams et al., 2023), the true direction of these effects may have been obscured, as mental health related factors were not accounted for.

Indeed, prior inconsistent findings as to whether metacognitive abilities vary according to age (Rouault et al., 2018; Fox et al., 2023; McWilliams et al., 2023; Weil et al., 2013) or gender (Fox et al., 2023; Rivers et al., 2021; Rouault et al., 2018; Xue et al., 2023) may be due to a lack of consideration of psychiatric dimensions as potential covariates in the model. Local confidence and global evaluations varied according to device type, indicating software and hardware variability should be considered when utilising smartphone-based research methods (Germine et al., 2019). This raises new challenges for research, as device types, far from being randomly allocated in the population, are associated with important socio-demographic characteristics and cognitive capacities we wish to study (Passell et al., 2021).

In terms of basic game mechanics, relative to the traditional task, Meta Mind was considerably shorter at 7.86 vs 21.50 minutes. Mean accuracy and confidence were also higher on average for Meta Mind. Trial-by-trial analysis revealed that accuracy steadily declines with additional trials, suggesting that although the staircase procedure maintained mean accuracy within a narrow and desired range (74% correct on average), more trials would decrease it somewhat further before reaching an asymptote. While task accuracy was weakly associated with local confidence, controlling for accuracy in our model did not affect the association between metacognition and transdiagnostic dimensions. The staircase functioned optimally across assessments, as task accuracy was stable with repeated game play. Relative to a metacognitive questionnaire, the main benefit of Meta Mind as an experimental task is the assessment of confidence while controlling for performance accuracy across participants. For example, an individual could report low confidence in their visual abilities on a self-report scale, but this may be an accurate appraisal of their abilities if they indeed have poor vision. With Meta Mind, we can determine true bias towards lower or higher confidence, controlling for visual task performance. Similar to accuracy, dot difference steadily declines throughout the task. This may reflect the design of Meta Mind, a relatively easier task, especially during the initial trials. Readjusting the dot difference of the starting trials to be more difficult may stabilise task difficulty and accuracy over fewer trials.

Overall, the study provided further support for utilising a combination of smartphone-based methods and a citizen scientist framework to conducting large-scale mental health research. The sample size of this study is the largest to date that examined how metacognitive abilities vary with psychopathology. With that, data quality was excellent. We employed a battery of checks for our task and questionnaire data (Zorowitz et al., 2023), and less than 2% of participants were excluded for careless or inattentive responses. This is in stark contrast to recent reports with paid crowdsourced participants (Burnette et al., 2022). As citizen scientists were not financially incentivised to complete Meta Mind, their intrinsic motivations may translate to conscientious participation. Although Meta Mind was designed to be more enjoyable and engaging than the traditional task, we did not collect any qualitative feedback on playing Meta Mind from participants. While the majority of citizen scientists that completed the tutorial trials went on to play the game in full, we did not have a direct indicator of task enjoyment.

Relative to the traditional task, Meta Mind was much shorter, but was able to replicate previous finding of disruption to metacognition at different levels of the hierarchy across psychiatric dimensions. This brief gamified task demonstrated validity and high reliability even with as little as 40 trials, allowing for the precise and rapid measurement of metacognitive bias. This study also provided more general support for utilising smartphone-based methods and a citizen science framework, to scale up and speed up cognitive science. The next frontier is to uncover the dynamic interactions between metacognition and psychopathology, which can be achieved by using Meta Mind as a tool to monitor within-person disruptions to metacognitive stability over time.

Chapter 4 – Mechanistic Arbitration Between Candidate Dimensions of Psychopathology

4.1 Introduction

Dimensional and transdiagnostic approaches to psychiatric classification have gained traction over the last decade (Dalgleish et al., 2020; Kotov et al., 2018), positioned as an alternative to the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2022). Their advantages are that they incorporate the lack of clear boundary between healthy and unhealthy states (Fried et al., 2022; Haslam et al., 2012) and their existence accounts for the heterogeneity of symptom presentation within diagnostic categories (Fried & Nesse, 2015; Zimmerman et al., 2015), the explicit overlap in symptoms (Forbes et al., 2024) and high rates of co-morbidity (Kessler et al., 2005; McGrath et al., 2020) across disorders. These challenges in defining valid and reliable clinical phenotypes have translated into a lack of progress in developing a mechanistic understanding of mental health (Abi-Dargham et al., 2023; Strawbridge et al., 2017; Tiego et al., 2023). Differences in cognitive functioning between healthy individuals and patients are small and disorder non-specific (Abramovitch et al., 2021; East-Richard et al., 2020), as are genetic (Smoller et al., 2019) and brain-behaviour associations (Marek et al., 2022; Sha et al., 2019; Winter et al., 2022). To achieve progress in our mechanistic understanding of psychiatric conditions, there is a growing need for an empirically derived, valid and reliable, set of transdiagnostic dimensional clinical phenotypes for research. Without this, a ceiling is placed on the size and specificity of clinical correlates, which cannot be addressed by recent initiatives to gather larger and more diverse datasets, or indeed advances in neuroimaging techniques (Etkin, 2019), computational modelling (Huys et al., 2021) and artificial intelligence (Bzdok & Meyer-Lindenberg, 2018).

Although transdiagnostic, dimensional approaches to psychopathology have many proponents (Dalgleish et al., 2020; Eaton et al., 2023) and are central to

major funding initiatives (Insel et al., 2010), there remains a lack of clarity around how candidate dimensions should be identified and validated. One increasingly popular approach is to derive dimensions by examining the natural covariation of symptoms between-subjects in the population using factor analysis (Kotov et al., 2017; Wise et al., 2023). A key conceptual critique of this method, however, is that it operates at the symptom level only (Wittchen & Beesdo-Baum, 2018). Descriptive taxonomies based on symptom covariance are limited, as symptoms that hang together do not necessarily have a shared mechanism or aetiology (Haeffel, Jeronimus, Kaiser, et al., 2022). Rather, a single dimension of related symptoms across DSM disorders can be reached via multiple mechanistic pathways (Strauss & Cohen, 2017). Symptoms can cause one another (Borsboom, 2017), the same symptoms can be arrived at from different causes and conversely, a single underlying cause can produce different symptoms in different individuals for example depending on early childhood experiences (Caspi & Moffitt, 2018). For these reasons, it is unclear if a covariation-based framework can lead us to aetiologically or mechanistically useful clinical phenotypes.

Emerging evidence suggests that, despite this potential limitation, psychiatric dimensions derived from factor analysis of symptoms nonetheless outperform classic frameworks in their association with cognitive-computational processes. For example, model-based planning, the tendency to use prospective mental maps to guide behaviour (Daw et al., 2011), has a stronger and more specific association to a transdiagnostic dimension Compulsivity and Intrusive Thought than traditional disorder categories (Donegan et al., 2023; Gillan et al., 2016; Patzelt et al., 2019; Seow et al., 2021). Metacognitive bias - the confidence one has in their own performance (Fleming & Lau, 2014) – tends to be decreased in a non-specific way across many DSM disorder categories and questionnaires (Hoven et al., 2019), but in fact has a highly specific and bi-directional association with two different transdiagnostic dimensions. Individuals high in Anxious-depression are underconfident, whereas those with Compulsivity and Intrusive Thought express overconfidence (Benwell et al., 2022; Fox, Lee, et al., 2023; Fox, McDonogh, et al., 2023; Rouault et al., 2018; Seow et al., 2024; Seow & Gillan, 2020).

These examples illustrate that covariance-based clinical phenotypes may have more mechanistic validity compared with traditional diagnoses (Wise et al., 2023). However, while conceptually symptom-based latent factors are hypothesised to give rise to co-occurrence among disorders, factor solutions are fundamentally constrained by the specific dataset at hand. As such, they change depending on the inclusion or omission of symptoms or diagnoses (Wittchen et al., 2009). It is possible that clinical dimensions that map even more closely to cognition exist, but do not correspond to the most substantial axes of variation in a given questionnaire set. A related issue is that factor modelling involves multiple degrees of analytic freedom (Haeffel, Jeronimus, Kaiser, et al., 2022), for example regarding the number of factors retained (Krueger, 1999), the inclusion of a higher- order general factor (Caspi et al., 2014), and whether dimensions of psychopathology should be orthogonal or partially correlated (Caspi et al., 2024). There is currently no consensus on these issues and problematically, specification decisions can produce a large, possibly infinite, set of alternative and well-fitting models (Simonsohn et al., 2020). Finally, and as articulated above, these models are designed to find the best solution for the covariance matrix of symptoms provided, these need not necessarily be optimal for or differentially correspond to distinct aetiologies or mechanisms.

This paper aimed to resolve these issues by using data from over 7000 individuals from nine independent datasets (Benwell et al., 2022; Donegan et al., 2023; Fox, Lee, et al., 2023; Gillan et al., 2016; Lee et al., 2023; Patzelt et al., 2019; Rouault et al., 2018; Seow et al., 2021; Seow & Gillan, 2020), with multiple variations of cognitive tasks measuring model-based planning and metacognition, various age ranges, data-collection methods, and representing a spread of general, crowd-sourced and clinical populations. To generate an analytic consensus on the structure of psychopathology, we generated thousands of competing factor solutions, spanning variations in the number of factors retained, a multiverse of expanded and contracted questionnaire sets, and examining bifactor, orthogonal and oblique rotations and tested for the maximal association with the cognitive-computational capacity of interest. We benchmarked performance against a baseline of established transdiagnostic dimensions (i.e.,

Anxious-depression and Compulsivity and Intrusive Thought), which were derived from a 3-factor solution, an oblique rotation, and based on the full set of items. Finally, we compared the mechanistic validity and identity of factors derived from the covariance of symptoms alone – as is the case in prominent dimensional frameworks such as the Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2017), versus a fully supervised method (partial least squares 'PLS' regression) that uses cognitive information to inform the derivation of factors.

4.2 Methods

4.2.1 Participants

In total, nine datasets were included in this study (Figure 4.1A). Five of these datasets included measures of model-based planning (Donegan et al., 2023; Gillan et al., 2016; Lee et al., 2023; Patzelt et al., 2019; Seow et al., 2021), and four included measures of metacognitive bias (Benwell et al., 2022; Fox, Lee, et al., 2023; Rouault et al., 2018; Seow & Gillan, 2020). N=1413 individuals were included in the discovery sample, previously described in Gillan et al. (2016) (labelled AMT 1, to distinguish it from other AMT samples). We chose this dataset as the discovery sample, as this study was the first to derive the established transdiagnostic dimensions and is the sample from which the other studies have applied the factor weights to generate independent dimension scores. Individuals in this samples were recruited remotely, online through AMT (Gillan et al., 2016). The sample had a mean age of 32.97 (SD=10.81), was mostly female (n=823, 58.20%), with a mean IQ of 98.00 (SD=9.55) (Table 4.1).

Table 4.1. Sociodemographic Characteristics

	Model-based planning					Metacognitive bias			
Socio-demographic characteristic	AMT 1 (N = 1413)	Smartphone App (N = 1684)[a]	AMT 2 (N = 820)	Student (N = 192)	Patient 1 (N = 862)	AMT 3 (N = 496)	AMT 4 (N = 437)	Patient 2 (N=825)	Prolific (N = 817)
Gender, *N (%)*									
Male	590 (41.80)	483 (28.68)	421 (51.34)	80 (41.67)	178 (20.65)	256 (51.61)	241 (55.15)	173 (20.97)	331 (40.51)
Female	823 (58.20)	1163 (69.06)	399 (48.66)	112 (58.33)	676 (78.42)	240 (48.39)	196 (44.85)	644 (78.06)	486 (59.49)
Other		36 (2.14)			8 (0.93)			8 (0.97)	
Age, *M* (*SD*)	32.97 (10.81)	46.14 (14.70)	34.89 (10.05)	31.89 (12.10)	32.07 (11.06)	35.59 (10.57)	37.54 (10.39)	32.00 (11.02)	25.58 (9.79)
IQ, *M* (*SD*)	98.00 (9.55)		99.00 (9.74)	7.95 (3.29)		7.98 (3.47)	-0.27 (0.79)		
Education, *N (%)*									
Below undergraduate		609 (36.16)			203 (23.55)			196 (23.76)	
Completed undergraduate		678 (40.26)			453 (52.55)			433 (52.48)	
Above undergraduate		395 (23.46)			206 (23.90)			196 (23.76)	

[a] = missing data for 2 participants. AMT = Amazon's Mechanical Turk, IQ = Intelligence quotient, M = mean, SD = standard deviation, N = count.

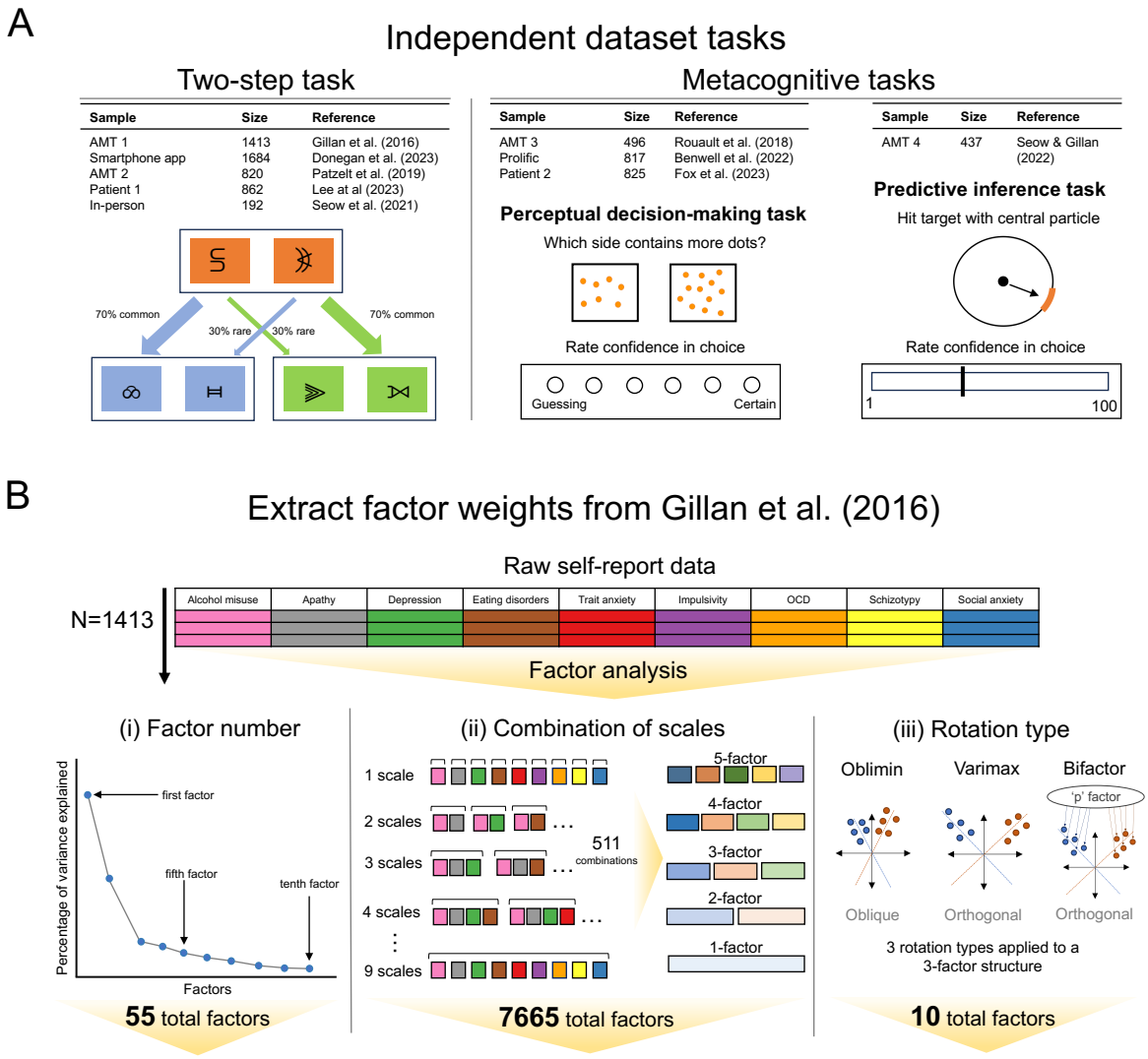


Figure 4.1. General Methodology. AMT = Amazon's Mechanical Turk. The four AMT and two patient datasets are numerically labelled to help distinguish samples (e.g., AMT 1 = Gillan et al. (2016) dataset (N=1413)). **(A)** Five datasets included alternative versions of the two-step task to assess model-based planning abilities. Four datasets included metacognitive tasks: Three included alternative versions of a visuo-perceptual decision-making task, and one dataset used a predictive inference task. **(B)** Factor weights were extracted from a discovery dataset (Gillan et al., 2016), following separate manipulations of (i) the number of factors retained (ranging from a single- to 10-factor structure, resulting in 55 total factors), (ii) the combinations of scales (ranging from a single questionnaire to all nine questionnaires considered, with each combination

subsequently analysed as a single- to 5-factor structure, resulting in 7665 total factors), and (iii) the rotation type (oblique 'oblimin', orthogonal 'varimax' or 'bifactor' rotations, each as a three-factor structure (in addition to a hierarchical factor for the bifactor rotation), resulting in 10 total factor solutions). **(C)** (i) The performance of each dimension was assessed by examining how within dataset factor scores (ii) were associated with model-based planning/metacognitive bias, controlling for age, gender and IQ/education, using linear regression analyses to extract the effect size (Cohen's f^2 value of each factor). (iii) The weighted average Cohen's f^2 values across datasets was used to determine which factor had the largest 'winning' effect size for predicting individual differences in model-based planning (lowest Cohen's f^2 value), overconfidence (highest Cohen's f^2 value), and underconfidence (lowest Cohen's f^2 value) respectively.

The other samples with measures of model-based planning included three previously published datasets (Donegan et al., 2023; Patzelt et al., 2019; Seow et al., 2021) and unpublished data from the Precision in Psychiatry (PIP) study (Lee et al., 2023). Participants in the Patzelt et al. (2019) dataset were recruited online using (N=820) (labelled AMT 2). We had a slightly lower sample than the sample published in the Patzelt et al. (2019) paper, as we only included those with full model-based planning, questionnaire and sociodemographic information. Participants in the Seow et al. (2021) dataset were from the general population and recruited in-person (N=192). The Donegan et al. (2023) dataset included participants from the Neureka Project, which enrols members from the general public who voluntarily download a smartphone application to contribute to scientific research. We had a slightly larger sample (N=1785) than the sample published in Donegan et al. (2023) paper, as we included additional Neureka users who have completed the model-based planning task since the study publication. The PIP study recruited participants from clinical sites that made referrals for internet-based cognitive behavioural therapy (Lee et al., 2023). Among PIP study completers, N=862 completed a model-based planning task (Lee et al., 2023) (labelled Patient 1), and N=825 completed the metacognitive task (Fox, Lee, et al., 2023) (labelled Patient 2). The sample size in this paper (N=825) differs from the size published in Fox, Lee, et al. (2023), as we included the updated sample of individuals that completed the metacognitive task, relevant self-report questionnaires and sociodemographic information. Three of the other datasets with measures of metacognitive bias were pre-existing, published datasets, that recruited participants online through crowdsourcing AMT (Rouault et al., 2018, N=496 (labelled AMT 3) and Seow & Gillan, 2020, N=437 (labelled AMT 4)), or Prolific (Benwell et al., 2022, N=817). The sociodemographic descriptives for all datasets are included in Table 4.1.

4.2.2 Procedures

Self-report clinical questionnaires. In each study, participants completed 209 items from nine self-report clinical questionnaires that assess a variety of psychiatric symptoms, including depression (Zung Self-Rating Depression Scale) (Zung, 1965), trait anxiety (State Trait Anxiety Inventory) (Spielberger et al., 1983), schizotypy (Short Scales for Measuring Schizotypy) (Mason et al., 2005), impulsivity (Barratt Impulsiveness Scale 11) (Patton et al., 1995), OCD (Obsessive-Compulsive Inventory-Revised, OCI-R) (Foa et al., 2002), social anxiety (Liebowitz Social Anxiety Scale) (Liebowitz, 1987), eating disorders (Eating Attitudes Test) (Garner et al., 1982), apathy (Apathy Evaluation Scale) (Marin et al., 1991), and alcohol misuse (Alcohol Use Disorders Identification Test) (Saunders et al., 1993). These questionnaires were chosen based on prior factor analysis in the discovery sample study (Gillan et al., 2016), which demonstrated that these items could be used to generate the established Anxious-depression and Compulsivity and Intrusive Thought dimensions previously associated with model-based planning and metacognitive bias.

Model-based planning tasks. Alternative versions of the reinforcement-learning 'two-step' task (Daw et al., 2011; Kool et al., 2017) were used in the five samples that measured model-based planning. In the two-step task used to quantify model-based planning, participants are presented with a series of choices between two stimuli (often represented as abstract symbols or images). Each choice leads to another set of options, creating a two-step decision-making process (Figure 4.1A). The key feature of the two-step task is that the outcomes of the initial choices are probabilistically associated with different outcomes in the subsequent steps. Participants must learn these associations through trial and error. Model-based learning refers to the ability to learn and utilize an explicit model of the task structure to make decisions. In the context of the two-step task, this involves understanding the probabilistic relationships between choices and outcomes and using this knowledge to plan and select actions that are expected to yield favourable outcomes in the long run. More detailed descriptions of the adapted two-step task versions are included in the original publications for each

study (Donegan et al., 2023; Gillan et al., 2016; Lee et al., 2023; Patzelt et al., 2019; Seow et al., 2021).

Metacognitive tasks. Metacognitive bias was measured through adapted versions of a visuo-perceptual decision-making task (Benwell et al., 2022; Fox, Lee, et al., 2023; Rouault et al., 2018) or a predictive inference task (Seow & Gillan, 2020). In the visuo-perceptual decision-making tasks, participants make a choice as to which of two stimuli contains more dots, and then subsequently rate their confidence in the accuracy of their choice, across multiple trials (Figure 4.1A). In contrast, the predictive inference task involves participants aiming a particle from the centre of a large circle to hit a target. Participants then rate their confidence that the particle would hit the target (Figure 4.1A). Detailed descriptions of the tasks can be found in the prior publications from which the datasets were taken (Benwell et al., 2022; Fox, Lee, et al., 2023; Rouault et al., 2018; Seow & Gillan, 2020).

4.2.3 Data preparation and analysis

Quantifying model-based planning. The 'two-step' task can be used to assess an individual's model-based planning by using logistic regression analyses with mixed-effect models to predict their choice on the subsequent trial. Specifically, model-based planning was indexed as the interaction effect of reward and transition on their choice stochasticity (Donegan et al., 2023; Gillan et al., 2016; Lee et al., 2023; Seow et al., 2021). Alternatively, model-based planning was calculated as the fit of a computational learning model, in which choices were reflected as the weighted combination of model-free and model-based planning (Patzelt et al., 2019). For all datasets, deficits in model-based planning were indicated by lower model-based learning values.

Quantifying metacognitive bias. Explicit post-decisional confidence judgements were used to measure metacognition across the four datasets (Benwell et al., 2022; Fox, Lee, et al., 2023; Rouault et al., 2018; Seow & Gillan, 2020). Metacognitive bias was calculated as the mean confidence reported by participants across experimental trials. Higher mean confidence, relative to

within-sample estimates, was used to index ‘overconfidence’, while ‘underconfidence’ was indicated as relatively lower mean confidence.

Extracting factor weights from discovery dataset. Using the self-reported clinical questionnaire data from the Gillan et al. (2016) dataset only, we aimed to determine which transdiagnostic dimension had the strongest association with individual differences in model-based planning and metacognitive bias, following three separate manipulations (Figure 4.1B):

1. *Variation of the number of factors selected for retention.* First, we generated a heterogenous correlation matrix of the 209 items from the nine questionnaires using the hector function in the polycor package in R. We conducted maximum likelihood estimation (MLE) factor analysis using the fa function from the psych package in R. We specified that each iteration would run from a single to 10-factor structure, using regression with an oblique rotation. The oblique rotational ‘oblimin’ was used, consistent with prior factor analysis of this questionnaire set (Gillan et al., 2016). Factor analysing each factor number from a single to 10-factor structure generated factor weights and loadings for 55 candidate dimensions in total (Figure 4.1B(i)).

2. *Variation in the selection of clinical symptoms.* Candidate dimensions for multiverse analysis were identified in the Gillan et al. (2016) sample by factor analysing every possible combination of the nine clinical questionnaires. Combinations ranged from each questionnaire on its own, to all nine questionnaires considered together, giving a total of 511 possible combinations (Figure 4.1B(ii)). We then took each of the standalone combinations of questionnaires and generated a heterogenous correlation matrix for the items included in that combination. The correlation matrices were then factor analysed (MLE in the fa package in R), with the oblique rotational ‘oblimin’. Rather than using the scree plot to determine factor number, we set each of the 511 dimensions as having a single factor structure up to a 5-factor structure. Factor analysing each combination with up to 5 potential factor structures generated factor

weights and loadings for 7665 candidate dimensions in total (Figure 4.1B(ii)).

3. *Higher- versus first-order factor rotation.* The heterogenous correlation matrix of all 209 item responses was factor analysed (MLE in the fa package in R), specifying a three-factor structure for analyses. A three-factor structure was chosen following the results from the factor number manipulation analyses. To manipulate factor rotation type, we considered 3 commonly employed rotation types: the oblique rotation, 'oblimin' (used in the prior sections), and orthogonal rotations, 'varimax', and a 'bifactor' solution (Figure 4.1B(iii)). While the no rotation, oblimin and varimax rotations were generated using the fa package in R, the bifactor solution was calculated using the 'omega' function from the Psych package in R, as per prior research on bifactor modelling in psychiatry (Gagne et al., 2020). The within-solution correlations between factor scores verified the oblique and orthogonal nature across rotation types (Supplementary Figure S7.3.6). Factor analysing each rotation type with a three-factor structure (with an additional hierarchical general 'p' factor for the bifactor solution) produced weights and loadings for 10 candidate dimensions in total (Figure 4.1B(iii)).

Testing candidate dimensions across datasets. Considering each manipulation type separately (factor number, combinations of scales and rotation type), we calculated the factor scores for each participant within each of the eight datasets separately. Participants' factor scores were calculated as the sum of the item weight by participants' response to that item, for each factor (i.e., dimension score = sum(item response*weight)) (Figure 4.1C(i)). This only differed slightly for our bifactor model, as the 'omega' output does not provide factor weights. To account for this, factor scores for our bifactor model were calculated from the factor loadings using the Anderson-Rubin method, in which a least-squares formula is applied to maintain the orthogonality of the general and specific factor scores (Anderson & Rubin, 1956).

We then ran linear regression analysis to determine the association between neurocognitive abilities (model-based planning/metacognitive bias) and factors

scores, controlling for age, gender and IQ/education. Age, gender and IQ/education were included as covariates to account for their potential effects on model-based planning and confidence. For each regression analysis, we calculated the effect size of each factor as Cohen's f^2 value (Selya et al., 2012) (Figure 4.1C(ii)). For the factor number and rotation type manipulations, our regression models included all the dimensions generated within that factor structure. For example, the regression model with a three-factor structure would be: model-based planning/metacognition ~ factor 1 scores + factor 2 scores + factor 3 scores + age + gender + IQ/education (Figure 4.1C(ii)). We included all factors within the structure in the same model, as the dimensions across the factor structure are obliquely rotated, meaning that factor scores are related across the structure. Including all factors means the model accounts for the effects of the related factors. This is more important for metacognition, as it has two poles of clinical association. For the interested reader, see Supplementary Figure S7.3.2 for sensitivity analyses showing individual model results without these controls, where overall effect sizes are smaller. When we tested the impact of the inclusion/exclusion of questionnaires, we only included individual factor scores in each of the 7665 models predicting model-based planning. For metacognitive bias, we included Anxious-depression as an additional covariate to determine which factors explained the most variance in overconfidence. When we were interested in factors that predicted underconfidence, we reran the models and included Compulsivity and Intrusive Thought instead of Anxious-depression as a covariate.

Following this, we took each dataset with a measure of model-based planning and calculated the weighted (by sample size) average Cohen's f^2 for each factor across the samples, consistent with a meta-analytic approach. We then repeated this separately for each dataset that had a measure of metacognitive bias to get the weighted average Cohen's f^2 when mean confidence was the dependent variable. The weighted average Cohen's f^2 were used to determine which factor performed best when predicting individual differences in model-based planning and metacognitive bias (i.e., which factors were the winning solutions). Therefore, mechanistic validity was defined as the magnitude of the effect size of a given dimension in predicting a given cognitive capacity, controlling for age, gender and

IQ/education. For model-based planning, we were specifically interested in deficits (lowest Cohen's f^2 values). For metacognitive bias, we were interested in both directions (highest and lowest Cohen's f^2 values, measuring over- and under-confidence respectively) (Figure 4.1C(iii)).

Partial least squares 'PLS' regression. We used partial least squared regression to identify transdiagnostic latent dimensions, comprised of the 209 questionnaire items, which are linked to individual differences in cognitive outcomes. The AMT 1 sample (N=1413) (Gillan et al., 2016) was used as discovery datasets to generate the PLS component for model-based planning, as this sample was also used to discover the factor structures from unsupervised factor analysis. For metacognitive bias, the AMT 3 sample (N=496) (Rouault et al., 2018) were used as discovery datasets to generate the PLS component. The Rouault et al. (2018) sample was chosen because it is the largest dataset available with a general population sample and data on metacognition and all covariates (age, gender and IQ). To avoid model overfitting, we split each discovery dataset into training and test sets, comprising of 75% and 25% of the data, respectively (Wise & Dolan, 2020). To identify the optimal number of components and equivalent item loadings for components within the training set, we used a 10-fold cross-validation procedure, fitting the model on 90% and testing performance on the left-out 10% of the data. The mean squared error of the model's predictions was then averaged across test folds to provide an index of the model's predictive accuracy with different numbers of components. Within the training set, the optimal models for model-based planning (RMSE=0.99), overconfidence (RMSE=0.97) and underconfidence (RMSE=0.98) all contained a single component. The item loadings for components were used to generate participants' dimension scores (dimension score = sum(item response x loading)). We then evaluated if component scores were significantly associated with cognitive outcomes within the training and test sets separately, before testing model performance out of sample.

4.3 Results

4.3.1 Comparing transdiagnostic dimensions to questionnaire total scores

As a preliminary step, we calculated the weighted average effect sizes for total scores across the nine clinical questionnaires and the established transdiagnostic dimensions, for associations with model-based planning and metacognitive bias (Figure 4.2). To compare the performance of total scores on the nine clinical scales to the three established transdiagnostic dimensions, we conducted separate regression analyses per cognitive task within each dataset, with the questionnaires/dimensions as an independent variable, along with age, gender and education/IQ. Regression analyses with clinical questionnaires included each questionnaire in a separate model (e.g., metacognitive bias ~ depression scale score + age + gender + IQ), while regression analyses for dimensions included the three dimensions in the same model (e.g., metacognitive bias ~ Anxious-depression + Compulsivity and Intrusive Thought + Social Withdrawal + age + gender + IQ). Examining the weighted average effect sizes (Cohen's f^2) across datasets for each questionnaire and dimension, transdiagnostic dimensions outperformed clinical questionnaires across all cognitive facets (Figure 4.2). For deficits in model-based planning, Compulsivity and Intrusive Thought was the top performing factor (Cohen's f^2 = -0.014), performing better than all the clinical questionnaires (Figure 4.2A). Similar for metacognitive bias, Compulsivity and Intrusive Thought had the largest positive effect size for individual differences in overconfidence (Cohen's f^2 = 0.038), and Anxious-depression had the largest negative effect size for individual differences in underconfidence (Cohen's f^2 = -0.030) (Figure 4.2B). The effect sizes of Compulsivity and Intrusive Thought and Anxious-depression for metacognitive bias were twice as large as the effect sizes of the next best performing questionnaires (the Obsessive-Compulsive Inventory-Revised for overconfidence with a Cohen's f^2=0.016, and the Apathy Evaluation Scale with a Cohen's f^2 = -0.015 for underconfidence) (Figure 4.2B).

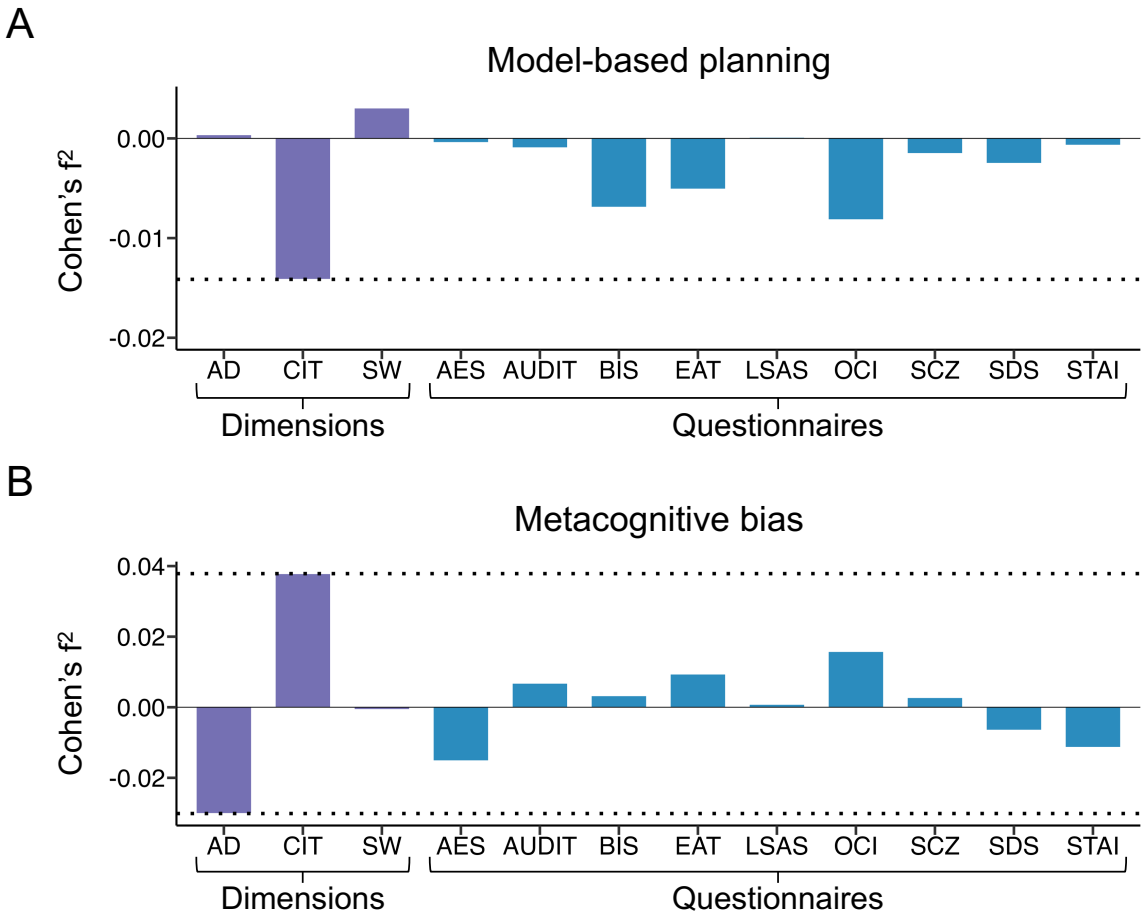


Figure 4.2. Weighted average effect sizes of total scores on clinical questionnaires and transdiagnostic dimensions. AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, AES = Apathy Evaluation Scale, AUDIT = Alcohol Use Disorders Identification Test, BIS = Barratt Impulsiveness Scale, EAT = Eating Attitudes Test, LSAS = Liebowitz Social Anxiety Scale, OCI = Obsessive-Compulsive Inventory-Revised, SCZ = Short Scales for Measuring Schizotypy, SDS = Zung Self-Rating Depression Scale, STAI = State Trait Anxiety Inventory. (A) Weighted effect sizes of dimensions and questionnaires from linear regression analyses predicting model-based planning, averaged across 5 datasets (N=4990). (B) Weighted effect sizes of dimensions and questionnaires from linear regression analyses predicting metacognitive bias, averaged across 4 datasets (N=2575).

4.3.2 Variation of the number of factors selected for retention

Factors that do not represent a significant axis of variation (based on our selection criteria, see below) in a questionnaire set may not be selected for retention; but they may nonetheless constitute an important and mechanistically homogenous dimension of psychopathology. Examining how variations in the number of factors retained per model (Figure 4.1B(i)), we generated a total of 55 factors, with the first from a 1-factor solution, the following 2 from a 2-factor solution, all the way to a 10-factor model (Figure 4.3A). For model-based planning, none of these factors outperformed the benchmark of Compulsivity and

Intrusive Thought, which was the 2nd factor of a three-factor solution (Figure 4.3A), as suggested by the Cattell-Nelson-Gorsuch (CNG) indices (Gillan et al., 2016), a mathematical formalisation of the scree plot method (Gorsuch & Nelson, 1981). We repeated this analysis for metacognitive bias, which has two poles of clinical association – Compulsivity and Intrusive Thought is associated with higher confidence and Anxious-depression with lower confidence. We found that once again no factor outperformed the benchmark of Compulsivity and Intrusive Thought in their positive association with metacognitive bias (i.e. higher confidence) (Figure 4.3A). However, the first factor from a 2-factor solution had a nominally stronger negative association with metacognitive bias than Anxious-depression (Figure 4.3A). The top factors had statistically significant effects on model-based planning and metacognitive bias (all $p<0.05$) within each individual dataset (Figure 4.3B). Comparing the top performing factor with Anxious-depression, they were effectively identical; scores from participants in the discovery dataset (N=1413) were correlated at $r(207)=.99$, $p<0.001$ (Figure 4.3C). Likewise, the loadings were correlated at $r(207)=.94$, $p<0.001$ (Figure 4.3D). Additionally, scores and loadings from the second factor from the 2-factor solution were highly correlated ($r_{both}>0.80$) with Compulsivity and Intrusive Thought (Supplementary Figure S7.3.2).

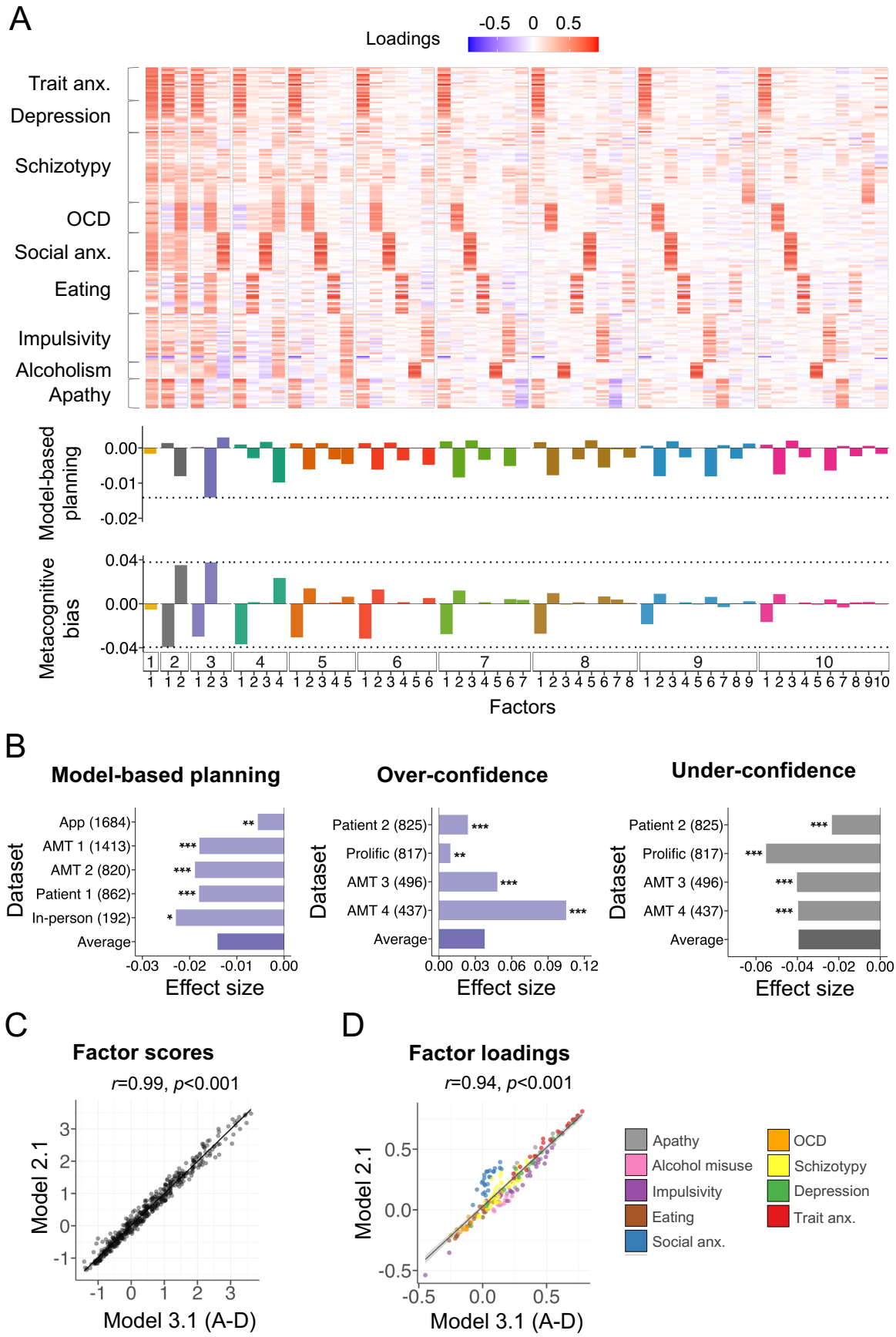


Figure 4.3. Variation of the number of factors selected for retention. AMT = Amazon's Mechanical Turk. A-D = Anxious-depression, OCD = Obsessive compulsive disorder, Social anx = Social anxiety, Trait anx = Trait anxiety, r= Pearson correlation coefficient, p = p-

value. **(A)** Fifty-five factors were generated from solutions retaining 1-factor, 2-factors, …, 10-factors. Heatmap indicates the loading of individual items onto each resulting factor (top panel). Weighted effect sizes from 10 linear regression analyses predicting model-based planning, averaged across 5 datasets (N=4990) (middle panel) and metacognition in 4 datasets (N=2575) (bottom panel). Factor loadings and effect sizes are aligned along the same x-axis. **(B)** Top-performing factors' effect sizes within each dataset and the weighted averaged effects across datasets. The top-performing factors had a statistically significant effects on model-based planning overconfidence and underconfidence within each dataset (all $p<0.05$). *=p<0.05, **=p<0.01, ***=p<0.001. The top performing factor for model-based planning and overconfidence was extracted from the 3-factor model (in purple) and the top performing factor for underconfidence came from the 2-factor model (in grey). **(C)** Correlation between scores on the first factor in a 3-factor solution (model 3.1 i.e., Anxious-depression) and the first factor from a 2-factor solution (model 2.1) in the discovery dataset. **(D)** Correlation between factor loadings of Anxious-depression and model 2.1 in the discovery dataset.

4.3.3 Variation in the selection of clinical symptoms

We conducted a multiverse analysis that systematically varied not just the number of factors retained, but the symptom feature set itself. Specifically, we examined all combinations of 1-9 questionnaires, considering factor solutions 1-5 for each, resulting in 7665 candidate dimensions. For their association with model-based planning, there was a large positive spike in effect sizes around 0, but the majority of factors had negative Cohen's f^2 effect sizes (81.03%, 6211/7665) (Figure 4.4A). Compulsivity and Intrusive Thought was the 9th best factor for explaining individual differences in model-based planning deficits (Cohen's f^2=-0.011), performing better than 99.88% of all factors generated (Figure 4.4A). The top performing factor (Cohen's f^2=-0.012) and Compulsivity and Intrusive Thought both had the highest loadings for OCD items and the lowest loadings for apathy items, with the top factor being generated from a set of questionnaires that omitted the schizotypy and alcohol addiction questionnaires (Supplementary Figure S7.3.3A). Scores across the factors were highly correlated ($r(1411)=0.85$, $p<0.001$) in the discovery dataset (N=1413) (Figure 4.4B). Within datasets, the top performing factor outperformed Compulsivity and Intrusive Thought for the association with model-based planning in 2/5 cases (Figure 4.4C). The top performing factor (Cohen's f^2=-0.012) and Compulsivity and Intrusive Thought (Cohen's f^2=-0.011) both had very small weighted average effect sizes (Cohen's f^2<0.02)(Selya et al., 2012) (Figure 4.4C). The mean difference between the weighted average effect size of the top performing factor and Compulsivity and Intrusive Thought was not substantial in magnitude (difference in Cohen's f^2=0.001) (Figure 4.4C). A heatmap of the top 1500

dimensions illustrated the relative importance of each questionnaire for explaining individual differences in model-based planning (Supplementary Figure S7.3.4A). Averaging item-level loadings across the top 100 dimensions Supplementary (Figure S7.3.5A), items related to OCD (M=0.51, SD=0.01), followed by eating disorders (M=0.20, SD=0.07), had the highest average loadings (Supplementary Figure S7.3.5D).

Compulsivity and Intrusive Thought had a stronger association with overconfidence than 99.70% of factors (Cohen's f^2=0.039), making it the 23rd top factor across analyses controlling for age, gender, IQ/levels of education and levels of Anxious-depression (Figure 4.4D). Relative to Compulsivity and Intrusive Thought, the top performing factor (Cohen's f^2=0.041) omitted the trait anxiety and eating disorder questionnaires (Supplementary Figure S7.3.3B), and scores on both factors were positively correlated in the discovery dataset at $r(1441)=0.62$, $p<0.001$ (Figure 4.4E). While the effect size was nominally greater for the top performing factor across samples, this difference was minuscule (Cohen's f^2 difference = 0.002), and the top factor did not outperform Compulsivity and Intrusive Thought within half (2/4) of the datasets (Figure 4.4F). Items related to OCD, schizotypy and impulsivity had strong positive loadings across the top 1500 performing factors (Supplementary Figure S7.3.4B) and specifically, items from the OCD questionnaire contributed most predominantly to the top 100 dimensions associated with overconfidence (Supplementary Figure S7.3.5B), and had the highest questionnaire-level average loadings (M=0.36, SD=0.02) (Supplementary Figure S7.3.5D).

When examining negative associations with confidence bias, Anxious-depression was the 731st best performing factor (better than 90.46% of factors) in terms of its association with confidence (Cohen's f^2=-0.036) controlling for age, gender, IQ/levels of education and levels of Compulsivity and Intrusive Thought (Figure 4.4G).Anxious-depression scores were highly correlated at ($r(1411)=0.95$, $p<0.001$), with the top performing factor (Cohen's f^2=-0.041) (Figure 4.4H), which had high loadings for the trait anxiety items (Supplementary Figure S7.3.3C). Despite having a larger negative effect size across samples, the top generated factor did not consistently outperform Anxious-depression, which had a larger

effect size in the Prolific dataset (N=817) (Figure 4.4I). Overall, the differences in effect size between the top performing dimensions versus Anxious-depression were negligible, reflecting a weighted effect size difference of r=0.005 for underconfidence (Figure 4.4I). Visualising the heatmap of item loadings across the top 1500 factors, trait anxiety contributed highly to explaining individual differences in underconfidence (Supplementary Figure S7.3.4C), and had the highest average item- and questionnaire-level loadings across the top 100 dimensions (M=0.59, SD=0.02) (Supplementary Figures S7.3.5C-D).

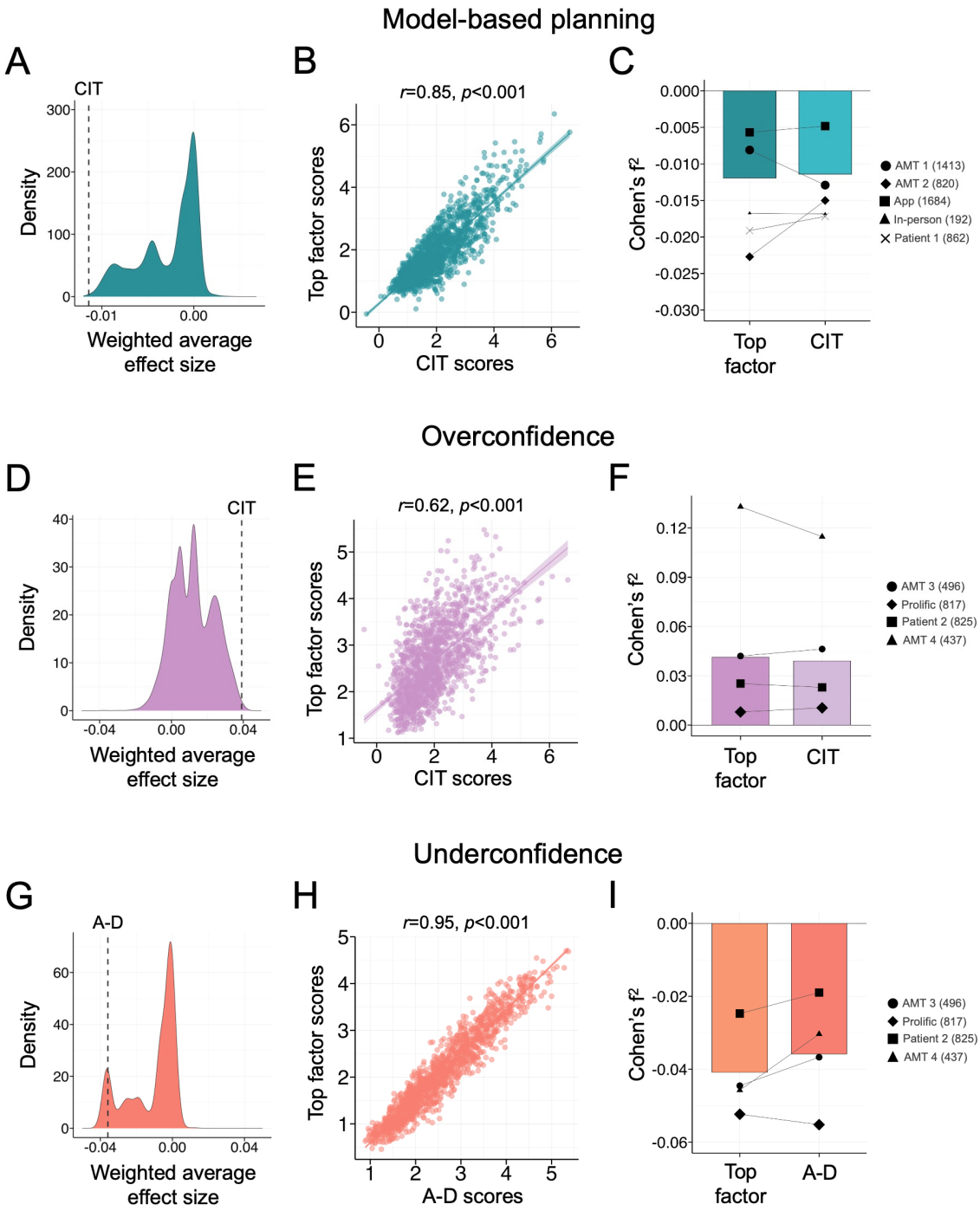


Figure 4.4. Variation in the selection of clinical symptoms. CIT = Compulsivity and Intrusive Thought, A-D = Anxious-depression, r= Pearson correlation coefficient, p = p-value. **Model-based planning. (A)** Density plot of effect sizes from 7665 candidate dimensions derived from all combinations of 9 questionnaires, retaining 1-5 factor solutions, predicting model-based planning. CIT was the 9th best performing of 7665. **(B)** Final scores for the top factor and CIT correlated at r=0.85, p<0.001 within the discovery dataset (N=1413). **(C)** Weighted average effect sizes across datasets (bars) and individual dataset effect sizes (dots) for the top factor and CIT associations with model-based planning. Overall effect sizes gains were negligible and inconsistent across datasets. **Overconfidence. (D)** Density plot of effect sizes predicting overconfidence. CIT was the 80th best performing of 7665. **(E)** Final scores between the top factor and CIT were correlated at r=.62, p<0.001 in the discovery dataset. **(F)** Weighted average effect sizes across datasets (bars) and individual dataset effect sizes (dots) for the top factor and CIT associations with overconfidence. Overall effect sizes gains were negligible and inconsistent across datasets. **Underconfidence. (G)** Density plot of effect sizes predicting underconfidence.

A-D was the 731st best performing of 7665. **(H)** Final scores for the top factor and A-D were correlated at r=0.95, p<0.001 in the discovery dataset. **(I)** Weighted average effect sizes (bars) and individual dataset effect size (dots) for the top factor and A-D associations with underconfidence. Overall effect sizes gains were negligible, and the top factor inconsistently outperformed A-D.

4.3.4 Higher- versus first-order factor rotation

Next we examined how factor models might best be rotated, and tested for evidence of utility for a higher order solution, which includes a general hierarchical ('p') factor. The cross-solution correlations (Figure 4.5A) of scores were very high across oblique (oblimin) and orthogonal (varimax) first-order solutions, but the oblique rotation (2nd factor – Compulsivity and Intrusive Thought) produced the nominally largest association with deficits in model-based planning (Cohen's f2 = -0.014) across our 5 datasets (Figure 4.5B). The same was true for both under- and over-confidence in metacognitive biases. Compulsivity and Intrusive Thought showed the largest positive association (Cohen's f2 = 0.038) and Anxious-depression the largest negative association (Cohen's f2 = -0.030) (Figure 4.5B). Notably, the general ('p') factor from the bifactor model showed the relatively smaller associations with all three cognitive domains (Figure 4.5B). Zooming in on this, associations with the 'p' factor were not statistically significant within any of the datasets for model-based planning (all p>0.05). The 'p' factor was associated with metacognitive bias in two datasets, but these were in opposing directions (negative association in the Prolific, N=817 and positive association in the AMT 4, N=437 datasets) (Figure 4.5C).

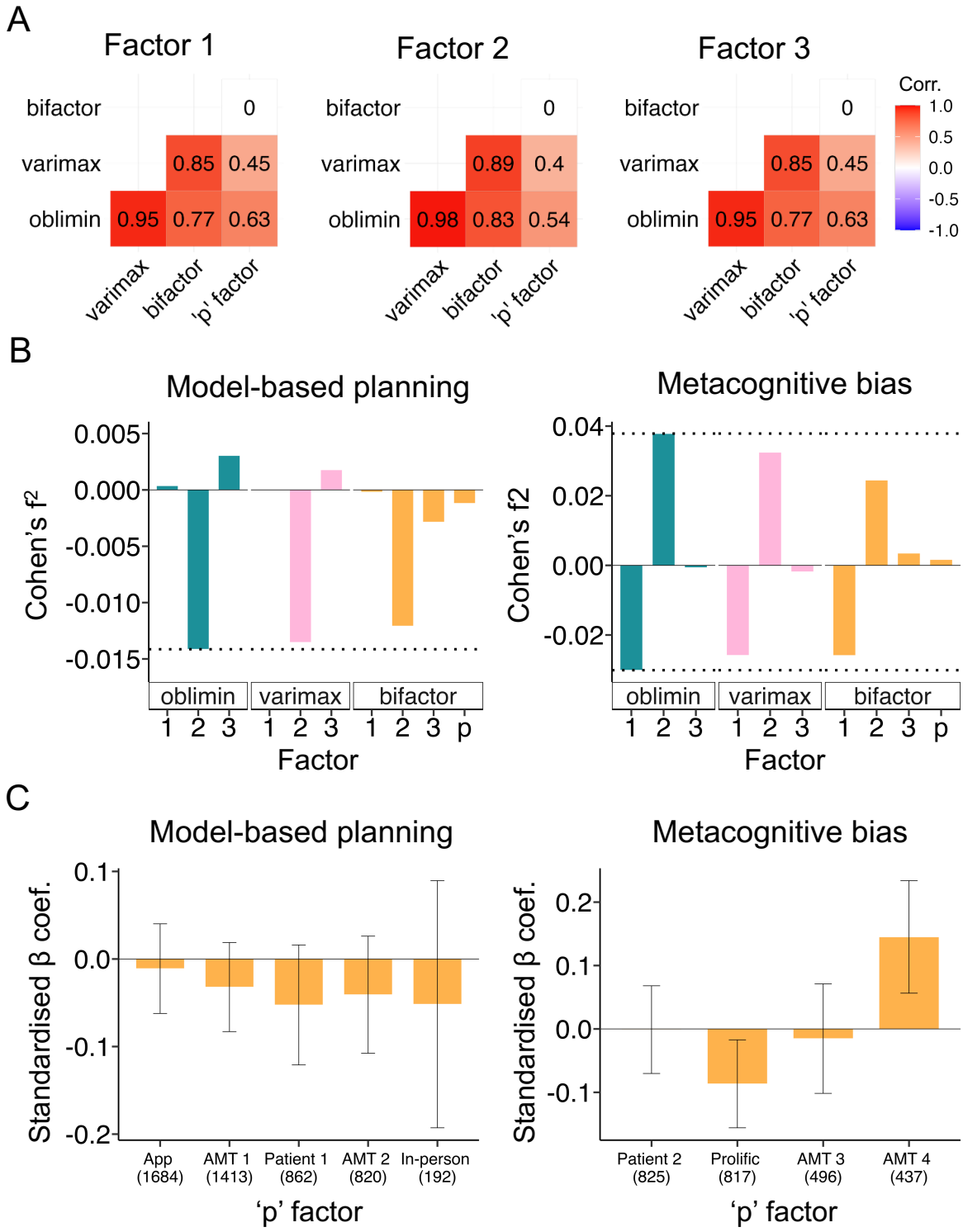


Figure 4.5. Higher- versus first-order factor rotation. AMT = Amazon's Mechanical Turk. p = general factor extracted from bifactor model. **(A)** Correlation of scores on factor 1, 2 and 3 from the 3 derivations – oblique, orthogonal and bifactor models in the discovery dataset (N=1413). Each correlation matrix also includes the correlation to the bifactor models general ('p') factor. **(B)** Weighted effect size for the association between derived factors and model-based planning and metacognitive bias. An oblique solution produces the winning model in all cases. **(C)** Individual results from component datasets for the 'p' factor reveals no consistent pattern of association with the cognitive constructs in question.

4.3.5 Comparison to a supervised method to derive factors

The preceding analyses focused on the extent to which natural patterns of symptom covariation possess mechanistic validity. In a final analysis, we compared the resulting factors to those derived using a fully supervised approach, where cognitive performance (i.e., model-based planning and metacognitive bias) informs the selection of the factor itself. Specifically, we used partial least squares regression with 10-fold cross-validation within 75% of the discovery dataset (Gillan et al., 2016) to first identify a transdiagnostic component linked to individual differences in model-based planning (Supplementary Figure S7.3.7A). Across the test folds of the cross-validation procedure run on the discovery set (N=1061), higher component scores were associated with model-based planning (residualised for age, gender and IQ) with an average r^2=0.02 in the test folds of the cross-validation. Evaluating the model in the held out 25% of the data from the discovery sample (N=352), the PLS component score was significantly negatively associated with the residual of model-based planning abilities, r(350)=-0.15, p=0.004 (Supplementary Figure S7.3.8A). Testing model performance out of sample, the PLS component performance was higher than Compulsivity and Intrusive Thought in 3 of the 4 independent datasets (Supplementary Figure S7.3.8B). While the weighted average effect size for the PLS component (Cohen's f^2 = -0.015) was nominally larger than the effect size for Compulsivity and Intrusive Thought (Cohen's f^2 = -0.011), the average gain in effect size was not substantial (difference in Cohen's f^2=0.004) (Supplementary Figure S7.3.8B). In the full discovery dataset (N=1413), scores on the PLS factor were strongly correlated at r(1411)=0.86, p<0.001 with Compulsivity and Intrusive Thought (Supplementary Figure S7.3.8C).

A PLS component was generated to predict individual differences in overconfidence within the AMT 3 sample (N=496) (Rouault et al., 2018) (Supplementary Figure S7.3.7B). The PLS component was associated with the residual values of confidence (after controlling for age, gender, IQ and levels of Anxious-depression) at r^2=0.07, averaged across the cross-validation test folds in the training 75% of the data (N=372). The PLS component was significantly positively associated with mean confidence in the held out 25% of the data (N=124), r(122)=0.21, p=0.017 (Supplementary Figure S7.3.8D). With out-of-

sample testing, the weighted average effect size of the PLS component (Cohen's f^2=0.037) was not nominally larger than the effect of Compulsivity and Intrusive Thought (Cohen's f^2=0.037) when predicting individual differences in overconfidence across all 3 independent datasets (Supplementary Figure S7.3.8E). Within the full discovery dataset (N=496), scores on the PLS component were highly correlated (r(494)=0.91, p<0.001) with Compulsivity and Intrusive Thought scores (Supplementary Figure S7.3.8F). We repeated this analysis residualising Compulsivity and Intrusive Thought, instead of Anxious-depression, from the confidence measure to identify a PLS component significantly negatively associated with confidence (Supplementary Figure S7.3.7C). This component had an average r^2=0.06 in the training set (N=372) and was negatively associated with the residual of confidence in the held out 25% of the data (N=124), r(122)=-0.17, p=0.055 (Supplementary Figure S7.3.8G). The PLS component did not outperform Anxious-depression in 2 of the 3 out-of-sample tests, and had a lower weighted average effect size (Cohen's f^2=-0.035) compared to Anxious-depression (Cohen's f^2=0.036) (Supplementary Figure S7.3.8H). Scores for the PLS component factor were strongly correlated (r(494)=0.95, p<0.001) with Anxious-depression scores (Supplementary Figure S7.3.8I).

4.4 Discussion

There is broad agreement that current diagnostic categories present challenges for mechanistic research in psychiatry (Kapur et al., 2012; Stephan et al., 2016). There is less agreement, however, about how alternative classification frameworks should be generated. Some have suggested that correlation at the level of symptoms can reveal an important latent structure of mental health, which is statistically robust and reproducible (Caspi et al., 2014; Kotov et al., 2017; Wise et al., 2023). Others have argued that this approach may be fraught because there is no reason to think that a descriptive construct, albeit one that is statistically valid, has any specific mechanistic influence (Haeffel, Jeronimus, Fisher, et al., 2022; Wittchen & Beesdo-Baum, 2018). This issue is compounded by the analytic degrees of freedom inherent in such correlation-based

approaches, wherein researcher choices at the time of data collection and analysis can produce many alternative factor solutions.

In this paper, we conducted an exhaustive multiverse analysis of potential factor solutions and tested their mechanistic validity, operationalised as their association with three aspects of cognition. In nine datasets, comprising N=7565 individuals, we provide evidence that those factors with the greatest mechanistic validity converge on the factors derived from a first-order factor analysis with oblique rotation, applied to a transdiagnostic sets of the maximal number of theoretically-informed self-report symptoms available to the researcher. These canonical factors (Compulsivity and Intrusive Thought and Anxious-depression) performed better than 99% and 90% (respectively) of alternatives in their association with cognition. These factors consistently out-performed total scores on constituent questionnaires and, in some cases, dimension effects were twice as large as classic clinical questionnaire effects. In cases where they performed nominally worse than the top-performing factor from multiverse analyses, the differences in effect size were trivial, inconsistent across datasets and correlations between competing factors approached unity. Perhaps most surprisingly, we showed that dimensions derived from factor analysis at the symptom-level had as strong an association to cognition as dimensions derived from a supervised approach that expressly maximised that same association to cognitive performance. This suggests that novel dimensions of psychopathology can be identified using symptoms alone, opening the door to leveraging other large existing datasets, without cognitive test data, to identify novel transdiagnostic dimensions. Our results generalised across nine independent samples with markedly different characteristics, including patients, the general population, paid and unpaid participants, of different ages, collected in-person, online or in-app, and using several variants of the cognitive tasks in question. This remarkable stability across heterogeneous datasets suggests that factor solutions generated based on the correlation of experienced symptoms are a valid and powerful approach for defining novel transdiagnostic dimensions for research in psychiatry.

As part of this multiverse analysis, we showed that a first-order correlated factors model outperformed a bifactor model. The ‘p’ factor has garnered significant interest in psychiatric research since its initial identification, and has been suggested by some to reflect the liability, comorbidity, persistence, and severity of psychopathology (Caspi & Moffitt, 2018). It features prominently in the HiTOP framework as a ‘superspectrum’, and although it is without doubt a statistical feature of symptom features sets, concerns have been raised about the interpretation of the ‘p’ factor (Aristodemou & Fried, 2020; Fried et al., 2021), and more broadly the evaluation and interpretation of bifactor models, including model overfitting (Bornovalova et al., 2020; Luciano et al., 2020), and poor replication across samples (Watts et al., 2024). Across the nine datasets and their aggregates, we found no association between the ‘p’ factor and any of the three cognitive phenotypes under study. These results suggest two things. First, it suggests that the ‘p’ factor, if it constitutes an important level of clinical description, may reflect nonspecific psychopathology (Caspi & Moffitt, 2018). Second, the ‘p’ factor is near indistinguishable from the sum of the feature matrix that generates it (Fried et al., 2021), and so the finding that it is unrelated to cognitive test performance further underscores that the associations we report between cognition, Anxious-depression and Compulsivity and Intrusive Thought are not generalised effects.

Mechanistic validity is just one way in which we might evaluate a candidate classification system in psychiatry. Alternative objective functions, for example maximising discriminative treatment-prediction or future risk, could produce different dimensions that are of more immediate clinical value (Haeffel, Jeronimus, Fisher, et al., 2022), even if they provide little-to-no insights with respect to aetiology or brain-behaviour correspondence. We do not see these as competing goals, *per se,* but rather different levels of analysis. The current focus on improved mapping of symptom dimensions to neurocomputational mechanisms is designed to provide a new, theory-based path to developing precision therapeutics that target cognitive processes, or using these symptom dimensions to guide selection of existing treatments (Strauss & Cohen, 2017). Future research should test if treatment-oriented vs. mechanistic approaches to symptom validation are distinct and complementary or converge on the same or

similar factors. The present study cannot speak directly to this potential. However, there is some preliminary evidence that transdiagnostic dimensions derived from computational factor modelling may help us understand how treatments may work. For example, those with the largest increases in metacognitive confidence following internet-based cognitive behaviour therapy had the greatest reductions in Anxious-depression (Fox, Lee, et al., 2023). To further elucidate the causal computational mechanisms of this effect, more intensive repeated testing would be a promising approach to account for temporal and contextual variation (Karvelis et al., 2023).

While the top performing factors had considerably stronger cognitive correlates than individual questionnaire total scores, the average effect sizes remained small in magnitude overall. This underscores how individual cognitive tests can at best provide a highly simplistic mechanistic account of psychopathology (Huys et al., 2021). For a comprehensive explanatory model, multivariable models are required, incorporating micro (e.g., neural circuits) and macro (e.g., societal factors) levels of information (Kendler, 2008). An additional limitation of the current study is that we only considered nine specific clinical questionnaires, which yielded three factors, only two of which had robust cognitive correlates. Future work aiming to capture the full spectrum of psychopathology must consider additional facets of mental health and illness (e.g., autism spectrum disorder or attention deficit hyperactive disorder (Keidel et al., 2024)), which are theoretically relevant to the cognitive facet on interest. Finally, cognitive test performance is but one layer of mechanistic precision – future work aiming to link brain-based measures to these dimensions may reveal stronger associations than the existing literature has been able to (Seow et al., 2021; Xia et al., 2018), providing a more compelling path towards novel pharmacological or stimulation-based interventions.

In conclusion, we used data-driven multiverse analysis of large-scale datasets to demonstrate the factor structure of psychopathology with optimal mechanistic validity. The natural covariance patterns of symptoms corresponded to specific cognitive mechanisms, regardless of manipulations to factor number retained, combinations of questionnaires or rotation type. Specifically, dimensions that

maximally corresponded to distinct cognitive capacities were generated through first-order factor analysis, with oblique rotation, a factor number indexed by the scree plot and a broad set of theoretically-informed psychiatric questionnaires. This demonstrated that computational factor modelling can derive dimensions of psychopathology that are clinically sensible and cognitively meaningful.

Chapter 5 – General Discussion

5.1 Summary

Historically, research on cognitive mechanisms underlying mental illness has been limited by strong allegiance to small, case-control studies with a cross-sectional design. As outlined in the general introduction, case-control studies suffer from a plethora of issues, as the traditional nosology is characterised by high inter-individual variability within disorder categories (Fried & Nesse, 2015; Zimmerman et al., 2015), shared symptomatology across diagnoses (Forbes et al., 2024), high rates of comorbidity (Kessler et al., 2005; McGrath et al., 2020) and arbitrary thresholds for indicating the presence of disorders, which all contribute to questionable diagnostic reliability (Regier et al., 2013). It is therefore unsurprising that researchers struggle to identify precise mechanisms that directly correspond to psychiatric symptoms and disorders. Not only an issue for individual differences in cognitive abilities (Doyle et al., 2018; East-Richard et al., 2020; Wright et al., 2014), but also differences in genes (Smoller et al., 2019) and brain structure and function (Sha et al., 2019), which also do not abide by the rigid boundaries of disorder categories.

Motivated by the constraints inherent in the traditional categorical nosology, psychiatric classification has seen a paradigm shift towards transdiagnostic dimensional approaches (Kotov et al., 2018). Transdiagnostic approaches conceptualise psychopathology along continuous dimensions, cutting across boundaries set by disorder categories (Eaton et al., 2023). These clinical dimensions can be uncovered with factor analysis, an intuitive and parsimonious method that aims to capture complexity and comorbidity across psychiatric disorders. The dimensional approach delineates mechanistic explanations of specific psychiatric phenomena. Compared to classic disorder categories, individual differences in model-based planning and were better explained by a specific transdiagnostic dimension Compulsivity and Intrusive Thought (Donegan et al., 2023; Gillan et al., 2016; Patzelt et al., 2019; Seow, Benoit, et al., 2021).

Metacognitive bias has inconsistent associations with various DSM disorder categories (Hoven et al., 2019), but shows specific, bi-directional associations with two transdiagnostic dimensions: Individuals with higher levels of Anxious-depression were underconfident, while those with higher levels of Compulsivity and Intrusive Thought were overconfident (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault, Seow, et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). This illustrates how covariance-based clinical phenotypes can outperform traditional disorder categories in their association with candidate cognitive mechanisms.

While highly suggestive, the evidence so far has been restricted to general population samples, rather than patient cohorts, which limits clinical applicability. Furthermore, these cross-sectional findings fail to capture the temporal dependence of transdiagnostic psychopathology on cognition. Additionally, the resulting factor structure is descriptive and constrained by subjective researcher biases, and thus may not be suitable to find the most meaningful axes of clinical variation. This thesis aimed to address these shortcomings, focusing specifically on metacognition as the cognitive mechanism of interest. The central objectives of this thesis were to 1) replicate the association between metacognitive bias and transdiagnostic dimensions in a clinical, treatment-seeking sample 2) elucidate the state-dependence of metacognitive bias in transdiagnostic psychopathology, 3) develop a gamified smartphone task that facilitates dense sampling of metacognitive bias over time and, 4) determining transdiagnostic dimensions that maximally correspond to individual differences in metacognitive bias and model-based planning. Each chapter of this thesis leveraged online and smartphone-based technology to speed and scale up data collection across and within participants.

One aim of the first experimental chapter, Chapter 2, was to test if the association between metacognitive bias and transdiagnostic psychiatric dimensions was reproducible in a clinical cohort. We found this was the case, replicating the bi-directional associations between metacognitive bias and transdiagnostic dimensions in a large sample of individuals initiating iCBT. Specifically, those with the higher levels of Anxious-depression had the lower confidence, while those

with higher levels of Compulsivity and Intrusive Thought had elevated confidence. The transdiagnostic dimension Social Withdrawal was not significantly associated with metacognitive bias, consistent with prior research (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault, Seow, et al., 2018; Seow & Gillan, 2020). While this three-factor structure of psychopathology has been previously reproduced, it was until now restricted to general population samples (Rouault, Seow, et al., 2018; Seow et al., 2024). Factors derived from the covariance of psychiatric symptoms in the general population do not necessarily reflect how symptoms co-occur among those with significant clinical impairment. To test this in Chapter 2, we factor analysed the correlation matrix of questionnaire responses among those in the iCBT group. This produced a near-identical factor structure to that generated from general population samples, as the factor weights derived from the clinical sample were highly correlated with factor weights reported by Gillan and colleagues (2016). This replication of the three-factor structure of psychopathology in a clinical population is an important validation for research aiming to leverage general population samples for psychiatric research.

The second aim of Chapter 2 was to examine changes in metacognition following transdiagnostic intervention. If confidence bias improved with successful treatment, it would suggest that metacognition is malleable and state-dependent. However, if confidence biases were resistant to change despite therapeutic gains, it would indicate that metacognition is more trait-like. Following four weeks of iCBT, there were widespread improvements in transdiagnostic symptomatology. In concert with clinical changes, there was a smaller, but significant, increase in confidence from baseline to four-week follow-up. Importantly, change in confidence scaled with changes in Anxious-depression – those with the greatest reductions in Anxious-depression had the largest increase in confidence. While other major aspect of task performance improved, such as task difficulty, this was in a manner unrelated to clinical change. This suggested that improved Anxious-depression coincided specifically with confidence changes, rather than merely reflecting a task practice effect. To partially remediate the lack of random treatment allocation in this observation study, we included two smaller groups: individuals starting antidepressant medications and individuals not undergoing

treatment. Similar to iCBT, confidence significantly increased in the antidepressant group, but not among controls. Across groups, changes in confidence scaled with improvements in Anxious-depression, indicating that this is not a treatment-specific effect. Nor was the increase in confidence treatment dose-dependent, as confidence changes were not greater among those who had higher iCBT engagement or those who received concurrent treatments. Overall, the results indicated that metacognitive biases in anxious-depression are state-dependent; when symptoms improve with treatment, so does confidence in performance.

Examining metacognition pre- and post- clinical treatment in Chapter 2 indicated that changes in metacognitive abilities scale with symptom improvement. However, investigation at only two timepoints is not sufficient to infer how later improvements in psychopathology *depend on* earlier changes in metacognition (and vice versa). To address this effectively, more dense assessments of metacognition are required (e.g., hourly, daily or weekly evaluations). This necessitates the development of a metacognitive task that is conveniently accessible to the general public and encourages sustained engagement over time (Gillan & Rutledge, 2021). Unfortunately, traditional metacognitive tasks do not lend themselves to repeated testing, as they are negatively perceived by the public, being described as ‘repetitive’ and ‘boring’ (Lee et al., 2023). Given this, the second experiment (Chapter 3) of this thesis describes ‘Meta Mind’, a gamified smartphone task developed to measure metacognition repeatedly and remotely. Task validity and reliability are important prerequisites for conducting any research on individual differences. As such, the aim of Chapter 3 was to establish the validity and reliability of Meta Mind across two studies. The first study in Chapter 3 included paid participants that completed Meta Mind and traditional metacognitive task. Despite being considerably abbreviated, Meta Mind had acceptable convergent validity, as confidence was moderately correlated across tasks. Additionally, the split-half reliability for metacognitive bias was excellent in Meta Mind.

In the second study of Chapter 3, which included over 3000 citizen scientists that played Meta Mind, we replicated the established patterns of local metacognitive

biases across transdiagnostic psychiatric dimensions, which were evident in as few as 40 trials. The split-half reliability of Meta Mind also stabilised after 40 trials. Across sessions, metacognitive biases in decision-making were stable, exhibiting very high test-retest reliability for the 100-trial and 40-trial versions of Meta Mind. Overall, Chapter 3 demonstrated that Meta Mind is valid and reliable, permitting precise and rapid assessment of metacognitive bias. Additionally, Chapter 3 provides more general support for utilising a combination of smartphone-based methods and a citizen scientist framework to conduct longitudinal cognitive research at scale. The findings in Chapter 3 position Meta Mind as a viable tool for investigating the temporal dependence of metacognitive bias and psychopathology.

Transdiagnostic psychiatric dimensions assessed in Chapters 2 and 3 were based on the three-factor structure originally generated by Gillan and colleagues (2016). Within Chapters 2 and 3, we provide further support that metacognitive bias is bi-directionally linked to Anxious-depression (underconfidence) and Compulsivity and Intrusive Thought (overconfidence). While compelling, this three-factor structure of psychopathology was derived to explain covariance in psychiatric symptoms, not cognition. Factors based on symptom correlations are purely descriptive and do not necessarily have mechanistic grounding. It is therefore plausible that an alternative factor solution potentially maps more strongly onto individual differences in metacognition. The final experiment, Chapter 4, tested this idea empirically. In Chapter 4, we conducted a large-scale comparison of competing factor solutions that allowed us to determine if the latent structure underlying the covariation of psychiatric symptoms has robust and specific cognitive correlates. Focusing on metacognition and model-based planning as the cognitive constructs of interest, using nine datasets comprising N=7565 individuals, we found strong evidence that this was the case. The factors with the best fit to cognition were those derived from a first-order factor analysis on the maximal number of theoretically informed symptoms available. These factors (Compulsivity and Intrusive Thought and Anxious-depression) performed better than thousands of engineered alternatives, including performing twice as good as traditional questionnaire total scores. Crucially, this unsupervised approach based on symptom correlation performed on-par with a partial least

squares analysis, a supervised approach to deriving factors based on cognition. These results were consistent across several datasets, including patients, healthy individuals, a broad set of age ranges and task variants. These results suggest that unsupervised factor analysis of symptoms is a practical method for rethinking how we define mental health and illness and affords clear opportunities for enhancing our understanding of specific underlying cognitive processes.

5.2 Strengths, limitations and future directions

The following section of the thesis outlines the strengths and limitations of the experimental chapters, in addition to potential avenues for future research.

5.2.1. Robust evidence of metacognition bias effects

In every experimental chapter of this thesis, the bi-directional association between metacognitive bias and transdiagnostic dimensions was demonstrated. We consistently found that those with higher levels of Anxious-depression had lower mean confidence, while those with higher levels of Compulsivity and Intrusive Thought had elevated confidence. These robust and replicable effects has been found across multiple independent studies that included paid, crowdsourced samples (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault, Seow, et al., 2018; Seow et al., 2024; Seow & Gillan, 2020). Replicated findings of the specific patterns of metacognitive bias in transdiagnostic dimensions dispel any concerns that the patterns observed in the original study by Rouault et al. (2018) were just noise (Nosek et al., 2022). For the first time, this thesis replicated this effect among clinical patients seeking treatment (Chapter 2) and unpaid citizen scientists (Chapter 3). Replication of metacognitive bias effects in these novel populations cements the generalisability of these results. This is major strength of this thesis, as replicability is the cornerstone of science (Simons, 2014). Replicability is facilitated by transparent research practices, which include sharing analytic decisions and outcomes (Nosek et al., 2022). To promote transparent research practices and further replicability of the experimental chapters in this thesis, the analytic code and data have been published on the Open Science Framework (Foster & Deardorff, 2017). We further extended these findings by demonstrating that the association

between metacognitive bias and transdiagnostic dimensions can be established in as few as 40 trials (Chapter 3).

Standard metacognitive tasks, which typically include hundreds of trials, have been criticised for being too lengthy (Lee et al., 2023). In this thesis, we present evidence that task durations need not be this long, as stable estimates of underconfidence in Anxious-depression and overconfidence in Compulsivity and Intrusive Thought can be detected rapidly in 40 trials. In addition to external validity, highly reliable estimates of metacognitive bias were evident with as little as 40 trials (Chapter 3). Overall, this thesis provides strong evidence that valid and reliable estimates of metacognitive bias can be acquired rapidly and remotely, to explain individual differences in transdiagnostic psychopathology.

5.2.2. Measuring metacognitive efficiency

In contrast to metacognitive bias, valid and reliable estimates of metacognitive efficiency could not be acquired with 100 trials (Chapter 3) or even 210 trials (Chapter 2 and 3). As a reminder, metacognitive efficiency refers to the extent to which confidence judgements discriminate between correct and incorrect responses, ratioed against objective performance (Fleming & Lau, 2014). In this thesis, we used the gold standard measure of metacognitive efficiency, known as ‘M-Ratio’ (Rahnev, 2023). In Chapters 2 and 3, we reported the unacceptable split-half reliability (internal consistency) for M-Ratio. This was likely a consequence of low trial number, as M-Ratio has poor split-half reliability when tasks have less than 400 trials (Guggenmos, 2021). Poor internal consistency has direct consequences for research on individuals differences in metacognitive efficiency. In Chapter 2, we reported that there was no significant association between metacognitive efficiency and any of the transdiagnostic dimensions (i.e., Anxious-depression, Compulsivity and Intrusive Thought and Social Withdrawal). This was consistent with other studies that did not find a significant effect of metacognitive efficiency on transdiagnostic dimensions (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault, Seow, et al., 2018; Seow et al., 2024). However, these studies also included tasks with an insufficient number of trials to reliably estimate metacognitive efficiency (i.e., <400 trials). Therefore, it is possible that metacognitive efficiency does indeed vary according to

transdiagnostic psychiatric dimensions, but future studies require tasks with a sufficiently large trial number (>400) to reliably index M-Ratio. Although this large trial number is acceptable for single time-point analyses, these findings suggest that the M-Ratio, and other existing measures of efficiency, are unsuitable for more rapid, repeated assessments over time. Therefore, Meta Mind is incompatible with lengthy tasks required to measure metacognitive efficiency.

In addition to poor reliability, another concern with M-Ratio is dependence on task accuracy (Guggenmos, 2021). Many popular methods used to measure metacognitive efficiency (e.g., area under the type 2 ROC function) are biased by primary task performance (Fleming & Lau, 2014). M-Ratio was specifically designed to be independent on task performance (Fleming & Lau, 2014). However, both empirical and simulated data have indicated that M-ratio is dependent on performance accuracy, regardless of trial number (Guggenmos, 2021). We found further evidence of this in Chapter 2, as M-Ratio values were confounded by task accuracy, despite inclusion of a staircase procedure to tightly control for accuracy. Currently, no metacognitive performance measures are completely independent on task accuracy (Guggenmos, 2021; Rahnev, 2023). As such, future studies with a sufficient trial number to estimate metacognitive efficiency should take care to include accuracy as a covariate in model, to account for its confounding effects.

5.2.3 The domain generality of metacognition

The original evidence of an association between metacognitive bias and transdiagnostic dimensions was demonstrated using a perceptual decision-making task (Rouault, Seow, et al., 2018), which informed the metacognitive task design in Chapters 2 and 3. While visual metacognition is the dominant method for modelling metacognition (Rahnev, 2023), metacognition can be evaluation across other domains, such as memory, knowledge and motor tasks (Rahnev et al., 2020). Therefore, a critical consideration is whether findings in Chapter 2 and 3 generalise to other cognitive domains. Whether metacognitive abilities are domain-general or domain-specific is an ongoing debate (Rouault, McWilliams, et al., 2018). In a recent meta-analysis of cross-domain correlations, metacognitive abilities were positively correlated within the perceptual domain (e.g., visual,

auditory, tactile), but there was evidence of domain generality across cognitive domains (e.g., visual and memory) (Rouault, McWilliams, et al., 2018). However, this meta-analysis averaged across metacognitive facets (e.g., bias and efficiency), which were not calculated using a consistent metric across studies (Rouault, McWilliams, et al., 2018). Examining metacognitive bias specifically, there is evidence of within-domain generality, as confidence bias was highly correlated across perceptual and auditory domains (Ais et al., 2016). Across cognitive domains, mean confidence level also were correlated within individuals across perceptual and memory tasks (Baird et al., 2013).

In Chapter 4, we averaged across multiple datasets that measured metacognitive bias using adaptions of a perceptual decision-making task (Benwell et al., 2022; Fox et al., 2023; Rouault, McWilliams, et al., 2018) and a perceptual learning task (Seow & Gillan, 2020). Across the perceptual task types in Chapter 4, metacognitive bias was consistently linked with Anxious-depression (underconfidence) and Compulsivity and Intrusive Thought (overconfidence). Chapter 4 also included a dataset that measured metacognitive bias in a perceptual and general knowledge task (Benwell et al., 2022). In this study, bi-directional associations between metacognitive bias and transdiagnostic dimensions were evident across task types, suggesting cross-domain generality (Benwell et al., 2022). This effect has more recently been replicated in a memory task, providing further support for the mapping of metacognitive bias onto transdiagnostic dimensions is domain-general (Seow et al., 2024). Therefore, the reported associations between metacognitive bias and transdiagnostic dimensions in Chapters 2 and 3 do not appear task-dependent, and may translate to other decision-making tasks, regardless of the cognitive facet. However, domain-generality could be further explored by examining confidence biases across psychiatric dimensions with other task types, such as interoceptive tasks (Khalsa et al., 2018). This may indicate whether metacognitive abilities in cognitive domains (perception, memory) generalise to bodily states.

5.2.4 Explicit vs. implicit measures of metacognition

Chapters 2, 3 and 4 all measured metacognitive bias through explicit (self-report) confidence judgements. Explicit, self-reported confidence judgements are the

conventional measure of metacognition in experimental tasks (Fleming, 2024). As an alternative to explicit confidence, we considered measuring metacognition through implicit (non-verbal) task performance when designing Meta Mind (Chapter 2). With explicit ratings, game play is constantly broken when participants are 'removed' from the game to rate confidence (Chapter 3). As such, implicit metacognition initially seemed more appropriate in the realm of gamification. Implicit measures include non-verbal indicators of confidence. For example, the decision to make a task more difficulty may implicitly indicate confidence in performance. Implicit measures are useful in circumstances when explicit estimates can be accessed, especially in studies with pre-verbal infants (Goupil & Kouider, 2016) or non-human animals (Kornell, 2009). However, explicit and implicit metacognitive tasks are uncorrelated, suggesting that they do necessarily measure the same underlying construct (Nicholson et al., 2021). Importantly, implicit estimates do not necessarily tap into meta-representations of the self, but could rather reflect first-order risk-evaluations and affective appraisals (Nicholson et al., 2021). For example, performance on gambling tasks are often used to infer implicit metacognition in primates, in which primates evaluate bets as high-risk or low-risk, with the assumption that primates will choose higher risk bets when they have higher implicit confidence in choices (Kornell, 2009). However, this pattern of behaviour may reflect positive affective states or sensation seeking, rather than a metacognitive evaluation of choices per se (Nicholson et al., 2021). This represents a problematic source of new confounds, particularly when designing tasks to examine clinical correlates. While explicit metacognition can also be confounded (e.g., by a self-consistency bias (Caziot & Mamassian, 2021)), it appears that on balance, explicit measures of metacognition are more likely to ensure that estimates are 'metacognitive' in nature. For these reasons, confidence was measured explicitly in Meta Mind (Chapter 3).

5.2.5 Considering alternative questionnaire sets

In Chapters 2 and 4, transdiagnostic psychiatric dimensions were measured using the same nine self-report questionnaires with 209 items that measured aspects of alcohol misuse, apathy, impulsivity, eating disorders, depression, trait anxiety, social anxiety, obsessive compulsive disorder (OCD) and schizotypy.

These questionnaires were chosen specifically to reproduce previously generated transdiagnostic dimensions (Gillan et al., 2016) with established links to metacognitive bias (e.g., Rouault, Seow, et al., 2018). However, this dimensional framework is not considered fixed or final (Gillan & Seow, 2020). Rather, additional aspects of psychopathology, which were not considered in this thesis, may yield dimensions that are even more predictive of individual differences in metacognition. For example, the dimensional framework did not consider neurodevelopmental disorders, such as autism spectrum disorder. While factor modelling in this thesis incorporated an impulsivity questionnaire, we did not evaluate inattention or hyperactivity, which are cardinal features of ADHD. Similar to psychiatric disorders, autism and ADHD have phenotypic characteristics that do not respect conventional disorder boundaries and are better classified across transdiagnostic dimensions (Astle et al., 2022). Disturbances in metacognitive abilities have been reported in both autism (Nicholson et al., 2021) and ADHD (Butzbach et al., 2021). However, meta-analytic evidence of metacognitive impairments among individuals with autism is inconsistent (Carpenter & Williams, 2023), which may be explained by heterogenous symptom clusters within autism (Mottron & Bzdok, 2020). Therefore, confidence disturbances may be implicated in transdiagnostic features of neurodevelopmental disorders not considered in this thesis.

Integration of features in autism and ADHD into computational factor modelling may reveal novel dimensions of psychopathology that map onto individual differences in metacognitive biases. As an example of an alternative dimensional framework, Keidel and colleagues (2024) generated transdiagnostic factors from a wide battery of items that included autism and ADHD questionnaires. This produced an ‘Inattention-impulsivity-overactivity’ dimension, which explained individual differences in temporal discounting (devaluation of a future reward as a function of delayed receipt). In addition to neurodevelopmental disorders, pathological personality traits (e.g., borderline personality disorder) have been previously integrated into dimensional models of psychopathology (Forbes et al., 2017) and may explain additional variance in confidence abnormalities (Thome et al., 2016). Therefore, the alignment between symptom dimensions and metacognition may be systematically improved by considering factor solutions

that include additional aspects of psychopathology. These aspects of psychopathology should possess theoretical relevance to metacognitive abilities, as evidenced with autism, ADHD and pathological personality traits.

5.2.6 The Research Domain Criteria (RDoC) framework

In this thesis, computational factor modelling was used to generate dimensions with mechanistic validity for two specific cognitive facets – metacognitive bias (Chapters 2-4) and model-based planning (Chapter 4). Although Social Withdrawal is produced alongside Anxious-depression and Compulsivity and Intrusive Thought, this transdiagnostic dimension was not significantly associated with individual differences in metacognition bias (Chapters 2 and 4) or model-based planning (Chapter 4), consistent with prior publications (Benwell et al., 2022; Donegan et al., 2023; Gillan et al., 2016; Hoven, Denys, et al., 2023; Rouault, McWilliams, et al., 2018; Seow, Benoit, et al., 2021; Seow & Gillan, 2020). The replicated null effect of Social Withdrawal on metacognitive bias motivated the decision to omit Social Withdrawal related items from Meta Mind in Chapter 3, to further reduce participant burden. Despite receiving relatively less attention throughout this thesis, other aspects of cognition and behaviour may have greater mechanistic validity for Social Withdrawal. For example, higher sensitivity to new information affecting actions was associated specifically with Social Withdrawal, but not the other transdiagnostic dimensions (Seow & Gillan, 2020). This is consistent with electrocortical evidence of enhanced performance monitoring among individuals with social anxiety (Judah et al., 2016). Therefore, future research might aim to determine the distinct neurocomputational mechanisms underlying Social Withdrawal specifically. As there are a myriad of neurocognitive processes that could be investigated for associations with Social Withdrawal, those incorporated into the Research Domain Criteria (RDoC) framework may be a viable consideration.

The RDoC framework by the National Institute of Mental Health was initiated with the goal of guiding research on the etiology of psychopathology, moving away from categorical DSM constructs and instead organising research along precise biobehavioural dimensions (Cuthbert & Insel, 2013; Insel et al., 2010). Rather than using symptom-based definitions, RDoC focuses on deriving dimensions of

disorder-relevant causal mechanisms based on observable behavioural and neurobiological measures (Insel et al., 2010). RDoC focuses specifically on identifying separable biobehavioural mechanisms, rather than describing symptom covariance. This translational research perspective views mental illness as brain disorders, which reflect dysfunction in neural circuitry (Insel et al., 2010). Neurological processes, in addition to genetics and behaviour, are used as the starting points to derive dimensions linked to psychological constructs (Cuthbert & Insel, 2013). The RDoC matrix includes dimensions across six higher-level domains: positive valence systems, negative valence systems, cognitive systems, systems for social process, arousal/modulatory systems, and sensorimotor systems (Cuthbert & Insel, 2013). Each domain contains multiple related components, which can be interrogated across different units of analysis, spanning from genes to behaviour, to evaluate the full range of functioning (Kozak & Cuthbert, 2016). Psychopathology is examined as extreme disruptions to normative functioning across these domain components (Morris et al., 2022).

Importantly, RDoC is designed for research, not for practical use in clinical settings (Cuthbert & Insel, 2013). As RDoC is more concerned with basic neurological processes, it lacks a comprehensive description of psychiatric phenotypes (Kotov et al., 2017). Therefore, it is possible that RDoC could be integrated into the dimensions derived from computational factor modelling. For example, temporal discounting, the tendency to devalue future rewards, is conceptually consistent with the RDoC Reward (probability) subconstruct (Michelini et al., 2021) and has been associated with transdiagnostic dimension derived using computational factor modelling (Keidel et al., 2024). Other RDoC subconstructs may specifically map onto individual differences in Social Withdrawal. In theory, Social Withdrawal may have associations with RDoC domains related to negative valence (e.g., potential threat) and systems of social processes (e.g., perception/understand of self/others) (Kozak & Cuthbert, 2016). Using RDoC to guide transdiagnostic research has the potential to expand the mechanistic basis of computational factor modelling. Additionally, integrating mechanisms from RDoC with dimensions from computational factor modelling offers a path towards developing a unified system, in which dimensions of psychopathology are grounded in neurobiology. Despite the significant potential

of RDoC, it has been criticised for representing arbitrary domains that have not been externally validated (Ross & Margolis, 2019). RDoC was formulated based on consensus from internal workshops at NIHM (Kozak & Cuthbert, 2016). In this regard, RDoC is similar to the DSM, as the relevant domains and constructs were determined based on expert consensus on biobehavioral systems (informed by prior research), not data-driven evidence. As it stands, RDoC is a preliminary template for the foundation of research into the genetic and neurobiological processes underlying psychopathology, requiring major testing, evolution and refinement.

5.2.7 The ‘p’ factor of psychopathology

It has been purported that variation across all psychiatric symptoms can be explained by a general ‘p’ factor of psychopathology (Caspi et al., 2014; Caspi & Moffitt, 2018; Lahey et al., 2012). Despite being a relatively novel theoretical construct, hundreds of studies have been conducted in the last decade to extract and validate the ‘p’ factor of psychopathology (Watts et al., 2024). The mechanistic validity of ‘p’ factor was investigated in Chapter 4, to determine whether ‘p’ explained individual differences in metacognition or model-based planning. Compared to first-order dimensions (i.e., Anxious-depression and Compulsivity and Intrusive Thought), the 'p' factor did not demonstrate mechanistic utility in Chapter 4. We found no association between the ‘p’ factor and metacognitive bias or model-based planning, across nine independent datasets and their aggregates. This is consistent with prior observations that the ‘p’ factor represents an index of the data, but does not necessarily have any mechanistic utility (Bonifay et al., 2017; Fried et al., 2021). It has been argued that the ‘p’ factor emerges due to shared genetic influence on mental illness specifically, rather than environmental factors that are nonshared and differentiate psychiatric experiences (Lahey et al., 2012). However, this has not been supported by empirical research (Kendler et al., 2011). A single dimension of underlying genetic risk of psychopathology was not revealed in a large study that used multivariate twin analysis (Kendler et al., 2011). Instead, the structure of genetic risk factors reflected partially separable internalising and externalising factors (Kendler et al., 2011). This demonstrates that genetic and environmental

risk factors have specificity across dimensions of psychopathology, further limiting the mechanistic importance of the ‘p’ factor.

While there is clear statistical evidence of the 'p' factor, other concerns have been raised regarding its interpretability, model fit and validity (Watts et al., 2024). The ‘p’ factor literature has been reliant on evidence of superior fit demonstrated by bi-factor models (e.g., Caspi et al., 2014; Constantinou et al., 2019; Lahey et al., 2012). However, fit indices alone are insufficient for model selection (Sellbom & Tellegen, 2019). In particular, bi-factor models are liable to overfitting data (Bonifay et al., 2017; Watts et al., 2024). Bi-factor models are so adept at fitting data that superior fit has been demonstrated for data with random and nonsensical response patterns (Bonifay et al., 2017). Across studies, the ‘p’ factor has poor replicability attributable to model complexity, with inconsistent loading patterns across datasets (Watts et al., 2024). Overall, there is no direct evidence the ‘p’ factor reflects a causal entity (Aristodemou & Fried, 2020). Rather, the ‘p’ factor may be a statistical consequence of positive manifold, where all the observed data is positively correlated (Bonifay et al., 2017). Using data from a large, nationally representative sample, the ‘p’ factor was indistinguishable from the sum of diagnoses that contributed to ‘p’ (Fried et al., 2021). In other words, the ‘p’ factor merely reflects the summation of diagnoses experienced by an individual (Fried et al., 2021). Overall, evidence from this thesis and prior publicatons suggested that the ‘p’ factor may index severity of psychiatric symptom profiles at best, but it is not useful for determinig specific mechanisms underlying psychopathology.

In Chapter 4 the ‘p’ factor was extracted using bi-factor modelling, consistent with prior research on general psychopathology (Caspi et al., 2014; Lahey et al., 2012). However, the ‘p’ factor can also be extracted using hierarchical (second-order) modelling (Caspi et al., 2014) (Figure 1.1). General ‘p’ factors generated from either second-order and bi-factor models were very highly correlated (Caspi et al., 2024). As such, it is unlikely that the inclusion of second-order modelling in Chapter 4 would have changed our conclusions on the mechanistic validity of the ‘p’ factor. However, a second-order factor solution may have produced other higher-level dimensions with sensitivity to individual differences in metacognition

and model-based planning, which were not assessed in this thesis. This is the main structural difference between HiTOP and computational factor modeling – while HiTOP typically uses confirmatory factor analysis of *categorical disorder co-occurrence* to produce hierarchical, second-order dimensions (Kotov et al., 2017, 2022), this thesis used exploratory factor analysis of psychiatric *symptoms* to generate simple, first-order dimensions of psychopathology. Therefore, the results in Chapter 4 may have been further strengthened by comparing the performance of first-order dimensions (i.e., Anxious-depression and Compulsivity and Intrusive Thought) to higher, second-order dimensions (e.g. broader 'externalising' or 'internalising' dimensions). However, there is recent evidence that first-order factors have greater mechanistic validity than second-order dimensions. Following hierarchical dimensionality reduction of psychiatric symptoms, individual differences model-based decision-making were best explained by lower-level (first-order) dimensions, rather than higher order internalising/externalising dimensions or the 'p' factor (Wise et al., 2024). These results suggest that certain aspects of cognition are better explained by more fine-grained, first-order symptom dimensions (Wise et al., 2024). This represents a potential avenue for future research, as it is yet to be determined if variance in metacognition is better explained by second-order factor solutions, consistent with the HiTOP framework.

5.2.8 Small effects and multivariate models

In each experimental chapter of this thesis, the associations between metacognitive bias and transdiagnostic dimension were statistically significant, but weak. Despite using the dimensional structure with the strongest mechanistic validity (Chapter 4), only a small proportion of variance in psychopathology was explained by metacognitive bias. The effect sizes reported in this thesis are typical for individual differences research, where $r \leq 0.02$ is expected for cognitive (Gignac & Szodorai, 2016) and brain-wide (Marek et al., 2022) associations with mental health. A single neurobiological mechanism, such as metacognition, is unlikely to be the root of complex mental health problems. Rather, mental illness involves a complicated set of interrelated, context-dependent processes (Hitchcock et al., 2022). As such, a mechanistic understanding of mental illness requires multivariate models that incorporates micro- and macro- levels of

information (Kendler, 2008). For example, a mechanistic sketch of Compulsivity and Intrusive Thought would consider multilevel influences from genetics, neurocomputational processes (including model-based planning and metacognitive abilities) and socio-cultural factors (Kendler, 2008). In particular, macro-level factors such as poverty and social inequality play a critical role in the emergence and maintenance of mental illness (Dean, 2019). Integration of multiple explanatory perspectives, beyond single neurocomputational processes, will deepen our mechanistic understanding of transdiagnostic psychopathology.

5.2.9 Cultural effects

The three-factor structure of psychopathology originally generated by Gillan and colleagues has since been replicated in general (Rouault, Seow, et al., 2018; Seow et al., 2024) and clinical (Chapter 2 of this thesis) populations. However, this factor structure may not be universally representative, as neither ethnicity nor racial differences were considered by previous studies (Gillan et al., 2016; Rouault, Seow, et al., 2018; Seow et al., 2024) or this thesis. Response patterns to psychiatric scales vary across different racial and ethnic groups. For example, Asian and Black individuals have elevated endorsement of specific OCD symptoms (e.g. contamination items) compared to white individuals (Wu & Wyman, 2016). Higher presence of certain psychiatric symptoms (such as magical thinking or superstition) may reflect specific religious affiliations or cultural beliefs and norms, rather than any pathological disturbances per se (Katz et al., 2020). As such, cultural differences in symptom interpretation may inflate the reported severity of psychiatric symptoms (Daniel-Watanabe et al., 2022). Future research may choose to construct psychiatric dimensions by omitting questionnaire items that are deemed too culturally specific (e.g., Hopkins et al., 2022). Alternatively, future psychiatric research can promote diversity and inclusion by ensuring minoritised groups are sufficiently represented.

5.2.10 Group-level effects

This thesis investigated the mechanistic validity of transdiagnostic dimensions using cross-sectional, group-level effects. However, inferences based on group-level data do not necessarily generalise to the individual (Fisher et al., 2018). Lack of group-to-individual generalisability may occur if data sampled within an

individual at any given time is not stationary i.e., dynamic processes are not captured by single time-point estimation (Fisher et al., 2018). Determining consistency between group and individual correlations requires collecting many measurements within individuals over time (Fisher et al., 2018). This kind of dense and protracted sampling can be achieved by leveraging digital phenotyping techniques and citizen science research (Gillan & Rutledge, 2021). In Chapter 3, unpaid citizen scientists were motivated to play the smartphone game Meta Mind repeatedly for months. Across time, both local and global metacognition were reliably estimated with Meta Mind. These results support Meta Mind as an appropriate instrument for inter- and intra-individual levels of analysis. While Chapter 3 included repeated measurement of metacognition, psychiatric symptoms were not sampled concurrently over time. As such, the longitudinal dynamic between metacognition and psychopathology could not be investigated in Chapter 3. Chapter 2 indicated that metacognition is state-dependent, as confidence increased proportionally within improvements in Anxious-depression. Changes in confidence scaled with improvements in Anxious-depression regardless of an individual's baseline clinical severity of metacognitive abilities. These results, based on assessments at two timepoints only, provide initial evidence that metacognition and psychiatric symptoms are correlated across time. In another study that used more dense sampling, confidence fluctuated over a period of days, but this was temporally unrelated to changes in mood (Da Fonseca et al., 2023). However, this study measured confidence implicitly in a small sample of healthy individuals (Da Fonseca et al., 2023). Adopting Meta Mind to evaluate within-person contemporaneous and time-lagged associations between explicit confidence and psychiatric symptoms would inform whether group-level effects reported in this thesis generalise to the individual.

Even if transdiagnostic dimensions were incorporated into longitudinal research, the dimensions themselves would still be estimated at a single time-point. For example, an individual may play Meta Mind and report levels of Anxious-depression twice daily, but the dimension score is derived based on the original, cross-sectional factor analysis. The dimensional approaches to psychopathology, including HiTOP (Kotov et al., 2017) and computational factor modelling (Wise et

al., 2023), all model the dimensional space using between-persons, cross-sectional data. However, transdiagnostic dimensions estimated in this way are not necessarily time invariant (i.e., the factor structure holds irrespective of measurement time). An exclusively between-person approach fails to consider contextualised, within-persons processes (Rief et al., 2023). Unaccounted for within-person variability can confound between-person estimates from a single time-point (Rush & Hofer, 2014). Therefore, both between- and within-person elements should be incorporated into the model to fully understand the dimensional structure of psychopathology. Using a repeated measures design, the within- and between-person factor structure can be examined simultaneously using multilevel factor analysis (Rush & Hofer, 2014). In multilevel factor analysis, the within-person factor structure reflects the covariance of psychiatric symptoms that varies over time within the same person, while the between-person factor reflects the covariance of symptoms pooled across occasions and individuals (Wright et al., 2023). Multilevel factor analysis of daily fluctuation in psychiatric symptoms provided some initial evidence that the HiTOP structure is highly comparable across within- and between-person levels (Wright et al., 2023). In this study by Wright and colleagues (2023), multilevel models of daily psychiatric symptoms at the between- and within-person levels produced factor structures with the same core dimensions of psychopathology (Wright et al., 2023). Given this, the validity of the three-factor structure of psychopathology included in this thesis may be substantially reinforced if it is shown to manifest at the within-person level. To assess within-person processes in the future, the comprehensive set of psychiatric questionnaires used to measure Anxious-depression, Compulsivity and Intrusive Thought, and Social Withdrawal will need to be administered repeatedly over time.

5.2.11 Using machine learning to reduce the questionnaire set

In Chapter 4, transdiagnostic dimensions with the best fit to cognition were those derived from factor analysis on the maximal number of self-report symptoms available (i.e., all 209 questionnaire items). However, lengthy questionnaires are an obstacle for clinical research, increasing perceived burden (Rolstad et al., 2011) and compromising data quality (Galesic & Bosnjak, 2009), especially with repeated assessments (Eisele et al., 2022). In light of this, recent efforts have

been made to optimise the three-factor structure generated by Gillan and colleagues (2016), by reducing the set of questions (Hopkins et al., 2022; Wise & Dolan, 2020). Using machine learning analysis, the smallest set of questionnaire items that strongly approximated transdiagnostic factor scores was identified (Wise & Dolan, 2020). This reduced the total items from 209 to 63, minimising participant burden while retaining accurate measurement of the three transdiagnostic dimensions (Wise & Dolan, 2020). In Chapter 3 of this thesis, the reduced set of questions developed by Wise and Dolan (2020) were used to expedite measurement of Anxious-depression and Compulsivity and Intrusive Thought among the 3410 citizen scientists that played Meta Mind. While the associations between metacognitive bias and transdiagnostic dimensions were statistically significant in Chapter 3, the effect of Compulsivity and Intrusive Thought on local confidence bias was notably smaller than the standardised effects reported in prior studies (Benwell et al., 2022; Hoven, Denys, et al., 2023; Rouault, Seow, et al., 2018). To explain this, omission of certain items from the reduced set may have decreased sensitivity to detecting overconfidence in Compulsivity and Intrusive Thought. Specifically, schizotypy items were completely omitted from the reduced questionnaire set (Wise & Dolan, 2020). The schizotypy scale captures positive symptoms in schizophrenia, such as hallucinations, which are characterised by an overconfidence bias (Schmack et al., 2021).

The work by Wise and Dolan (2022) has since been extended in a larger and more varied dataset, to develop another reduce questionnaire set of 71 items (Hopkins et al., 2022). This reduced set of 71 items also omitted schizotypy items, and subsequently produced Compulsivity and Intrusive Thought scores that explained less variance in overconfidence than original scores generated with schizotypy items (Hopkins et al., 2022). This fits with the results in Chapter 4 of this thesis, as factors from the top 100 and 1500 dimensions that predicted individual difference in overconfidence included higher loadings for positive schizotypy items. In contrast, the reduced item sets produced Anxious-depression scores that were highly sensitive to underconfidence, which was evident in Chapter 3 of this thesis and in the study by Hopkins and colleagues (2022). The reduced item sets included trait anxiety and apathy items to measure

Anxious-depression (Hopkins et al., 2022; Wise & Dolan, 2020), which we showed were highly relevant for strongly predicting underconfidence in Chapter 4. Overall, reduced item sets reduce participant burden, but are not necessarily more sensitive to individual differences in metacognition. Relative to the full questionnaire set, reduced item sets produced scores for Anxious-depression that appeared equally (if not more) sensitive to underconfidence, while Compulsivity and Intrusive Thought scores that omitted schizotypy items were less sensitive to overconfidence. Given smaller standardised effects, future studies that adopt the reduced questionnaire sets may require larger samples that are sufficiently powered to detect the association between local confidence and Compulsivity and Intrusive Thought.

5.2.12 The network perspective

A broad aim of this thesis was to investigate the dimensional approach to psychopathology. The experimental chapters did this by demonstrating the mechanistic validity of transdiagnostic psychiatric dimensions, above and beyond established disorder categories. This was illustrated most strongly in Chapter 4, where we compared dimension effects on metacognition and model-based planning to stand-alone questionnaire effects, aggregated across nine independent datasets. Transdiagnostic dimensions consistently outperformed total scores on constituent questionnaires, and dimension effects were twice as large as questionnaire effects for associations with over- and under-confidence. The cumulative evidence from this thesis supports a shift towards a dimensional approach to elucidate cognitive mechanisms of psychopathology. However, the dimensional approach is just one alternative to the traditional categorical framework. Another alternative is the network approach, which views psychopathology as emerging from a complex system of interacting symptoms (Borsboom, 2017; McNally, 2021).

The dimensional and network approaches have developed in parallel, with divergent theoretical assumptions about the structure of psychopathology (Eaton et al., 2023). In particular, the dimensional approaches affirm that symptoms map onto a common latent construct (Kotov et al., 2017; Wise et al., 2023), while the network perspective does not assume any underlying latent structure (Rief et al.,

2023). As such, the dimension perspective focuses on variances shared across symptoms, while the network perspective considers unique variance between symptom pairs (Epskamp et al., 2017). Despite considering alternative perspectives, there is scope to integrate dimensional and network approaches. Statistically, there is no clear distinction between network and latent variable (factor) models (Bringmann & Eronen, 2018). Latent variable models can be successfully transformed into equivalent network models (Molenaar, 2010). When applied within the same dataset, factor and network analyses generated highly congruent structural information about psychopathology (Eadeh et al., 2021). There has been a push towards embracing the complexity of both approaches (Fried, 2020), by generating statistical hybrid models that feature latent variables and interacting symptoms (Epskamp et al., 2017). Therefore, the dimension and network approaches can be viewed as complementary perspectives, capturing variance across and within psychiatric phenomena. Future research may adopt this hybrid approach, by incorporating network modeling to further refine transdiagnostic classification.

5.3 Clinical implications

The results of Chapter 2 provide support for the state-dependence of metacognitive bias; when transdiagnostic symptoms improve with treatment, so does confidence in performance. We found that treatment-seeking individuals with the greatest improvements in Anxious-depression had the largest increases in confidence. Changes in metacognitive bias scaled unconditionally with symptom improvements, regardless of intervention type, levels of treatment engagement or concurrent treatments. While the association between changes in confidence and Anxious-depression was statistically significant, the correlation was weak. The iCBT programs undertaken by patients in Chapter 2 comprised of various components (e.g., behavioural activation, cognitive restructuring and relaxation techniques), which did not target metacognition specifically (Lee et al., 2023). Therefore, the association between metacognition and symptom changes may be stronger following treatments that directly target metacognitive abilities. For example, metacognitive therapy aims to identify and reframe maladaptive beliefs about disorder-specific cognitions (Wells et al., 2012). Metacognitive

therapy is a transdiagnostic treatment, with evidence of efficacy for depression, anxiety, OCD and eating disorders (Normann & Morina, 2018; Philipp et al., 2019; Sharma et al., 2022). In addition to ameliorating psychiatric symptom, metacognitive therapy improved dysfunctional metacognitive beliefs, including low confidence in cognitive capacities (Normann & Morina, 2018; Sharma et al., 2022). While studies have reported that metacognitive therapy increased confidence in cognitive abilities (Sharma et al., 2022), this was based on self-reported confidence on a metacognitive questionnaire (the MCQ-30; Wells & Cartwright-Hatton, 2004), rather than confidence in objective task performance. Relative to a metacognitive questionnaire, assessment of confidence in an experimental task is advantageous, permitting control for objective performance. For example, an individual could report low confidence in memory abilities on a self-report questionnaire, but this may be an accurate appraisal of their abilities if they indeed have poor memory. In contrast, a metacognitive task determines true bias towards lower or higher confidence, controlling for memory accuracy. To date, no studies have evaluated how confidence bias changes on an objective task through metacognitive therapy.

In Chapter 2, metacognitive bias and levels of Anxious-depression were only assessed at baseline and then four weeks after treatment commencement. While two timepoints are sufficient for evaluating changes over time, the temporal dependence or directionality of metacognitive effects could not be evaluated in Chapter 2. Earlier changes in metacognition may precede and predict later symptom improvements, or vice versa. One way to elucidate the temporal dependence of symptom improvements on metacognitive changes is through more continuous assessments during treatment. A previous study found that a reduction in self-reported positive metacognitive beliefs between pre-treatment and mid-treatment mediated the impact of iCBT on post-treatment depression and anxiety symptoms (Newby et al., 2014). While this study included a small sample and self-report metacognitive beliefs, as opposed to confidence in objective performance, the results indicated that early shifts in metacognition mediate the effects of iCBT on transdiagnostic symptoms, with temporal precedence (Newby et al., 2014). However, evaluating metacognition and symptoms pre-, mid- and post-treatment may not be the optimal temporal

window, as transdiagnostic symptoms have the capacity to significantly improve much earlier in the iCBT course (Lee et al., 2023). Therefore, more dense sampling of metacognitive bias and symptoms would be required in the initial weeks of treatment. To achieve this, Meta Mind (Chapter 3) could be integrated into the study design in Chapter 2. Completion of Meta Mind by participants daily, in addition to self-reported transdiagnostic symptoms through iCBT and antidepressant treatment may indicate whether symptom improvements are temporally dependent on changes in local confidence. Additionally, employing machine learning analysis may determine if early changes in metacognitive bias predict subsequent treatment response, as demonstrated with other facets of cognition implicated in depression (Browning et al., 2019). Quantifying within-person changes in metacognition and transdiagnostic symptoms is an imperative step towards translating computational research to clinic (Yip et al., 2023). This may advance precision in treatment prescription, by helping identify patients who are most likely to respond to a given treatment.

Inclusion of Meta Mind in the study design in Chapter 2 would have the added benefit of incorporating a metacognitive task that is brief and gamified, rather than the traditional task that was negatively received by participants in that study (Lee et al., 2023). While Meta Mind includes measures of confidence bias at the trial- and task-level, a measure of higher-order global self-beliefs were omitted from the game design. Therefore, incorporating a self-beliefs questionnaire into Meta Mind may inform how levels of the metacognitive hierarchy reciprocally interact through treatment (Seow, Rouault, et al., 2021). From one perspective, a rapid increase in local confidence with treatment may propagate up the metacognitive hierarchy, to inform later changes in global self-beliefs and subsequent improvements in Anxious-depression (Seow, Rouault, et al., 2021). Alternatively, treatment may directly improve global self-beliefs, which then acts as a prior for local confidence evaluations (Hoven, Denys, et al., 2023). All previous research that reported a temporal influence of metacognitive changes on symptom improvements have only focused on one level of the metacognitive hierarchy (e.g., Goldin et al., 2013; Newby et al., 2014). Accounting for the various levels of the metacognitive hierarchy is necessary to fully evaluate how metacognition predicts changes in transdiagnostic symptom and treatment

response. Importantly, temporal precedence is not sufficient from inferring causality (Ohlsson & Kendler, 2020). Even with mediation analysis of longitudinal data, a reduction in confidence does not necessarily *cause* the emergence of transdiagnostic symptoms. Likewise, a positive shift in confidence bias may not be *crucial* for therapeutic response. The challenge of inferring causality is due to the myriad of known and unknown confounds that might be driving the temporal effect (Ohlsson & Kendler, 2020).

In Chapter 2, the iCBT and antidepressant treatments were prescribed primarily to treat symptoms of depression and anxiety specifically. The iCBT programs in Chapter 2 included a module that targeted negative self-beliefs (Lee et al., 2023). This may explain why we did not observe an association between changes in confidence and improvements in Compulsivity and Intrusive Thought, as treatment did not target the overconfidence bias specifically. Therefore, this thesis could not evaluate the state-dependence of overconfidence in Compulsivity and Intrusive Thought. To investigate this, future research may consider examining confidence changes through treatments that target symptoms characterised by an overconfidence bias. For example, we demonstrated in Chapter 4 that dimensions with the strongest links to overconfidence had high loadings for transdiagnostic symptoms related to delusions and hallucinations, often seen in schizophrenia. Delusions and hallucinations can be effectively treated with haloperidol, a commonly prescribed antipsychotic drug that blocks D2 dopamine receptors (Andreou et al., 2014). In mice, hallucination-like perceptions can be modelled as high-confidence in false percepts (Schmack et al., 2021). Although high-confidence in false detections can be induced by optogenetic stimulation of the striatum, this effect was blocked with haloperidol administration (Schmack et al., 2021). Among healthy individuals, haloperidol significantly reduced subjective confidence in decisions (Moritz et al., 2016) and overconfidence in errors (Andreou et al., 2014). This evidence from mice and healthy individuals suggests that the normalisation of an overconfidence bias may be the mechanism of action underlying haloperidol. To evaluate this in a clinical population, further research could incorporate Meta Mind into a randomised controlled trial that compares haloperidol to a placebo. Repeated evaluations of metacognitive bias through Meta Mind with haloperidol vs. placebo intervention may indicate if reductions in

overconfidence precede and predict improvements in delusions and hallucinations.

In addition to haloperidol, overconfidence biases among psychiatric patients might be targeted through metacognitive training. While metacognitive training is a general term, this particular form of psychotherapy aimed to reduce cognitive biases implicated in aberrant thinking patterns, such as overconfidence in errors (Moritz et al., 2010). Overconfidence bias in delusional thinking is targeted with metacognitive training by encouraging patients to lower their confidence when objective evidence is lacking (Moritz et al., 2014). Meta-analytic evidence indicated that metacognitive training is associated with reduced delusional thinking among psychosis patients (Eichner & Berna, 2016; Liu et al., 2018). Metacognitive training had long-lasting positive effects on delusions, with reduced symptoms evident up to six months post-treatment (Liu et al., 2018). While most studies evaluated end of treatment symptomology as the primary outcome, there was preliminary evidence that metacognitive training directly ameliorated overconfidence bias (Moritz et al., 2014). This metacognitive training intervention has been adapted to attenuate overconfidence and improve symptoms among patients with borderline personality disorder (Schilling et al., 2018). Therefore, metacognitive training represents an evidence-based intervention that targets transdiagnostic overconfidence. Beyond psychosis and borderline personality disorder, it remains unknown if metacognitive training improves transdiagnostic symptoms characterised by an overconfidence bias more broadly, such as those implicated in Compulsivity and Intrusive Thought. Future research may also evaluate how confidence and transdiagnostic symptoms dynamically change throughout metacognitive training, to determine if a reduction in overconfidence is the mechanism of action underlying symptom improvements. Furthermore, metacognitive training is not effective for ameliorating delusions in all patients (Eichner & Berna, 2016; Liu et al., 2018). Another avenue for future research is to evaluate who metacognitive training works for, and whether this includes patients with the strongest overconfidence bias at baseline.

Metacognition can be causally manipulated by directly stimulating brain region underlying metacognitive abilities (Dijkstra & De Bruin, 2016). Anatomical and

functional neuroimaging findings implicated the importance of the prefrontal cortex as the neural basis of metacognition (Rouault, McWilliams, et al., 2018). Varying patterns of connectivity between prefrontal regions and the precuneus contributed distinctly to metacognitive judgements of perception and memory (Baird et al., 2013). These brain regions can be modulated using non-invasive, brain stimulation techniques such as transcranial magnetic stimulation (TMS) and transcranial direct-current stimulation (tDCS). While TMS uses magnetic fields to induce electric currents in specific areas of the brain, tDCS delivers a constant, low electric current through electrodes on the scalp that modulate neuronal excitability by either depolarising (anodal stimulation) or hyperpolarising (cathodal stimulation) neuronal membranes (Demirtas-Tatlidede et al., 2013). Application of TMS over the precuneus impaired metacognitive efficiency for memory (Ye et al., 2018). In another study, anodal tDCS over the dorsolateral prefrontal cortex induced a reduction in memory confidence, but did not affect memory accuracy (Bona & Silvanto, 2014). While both of these studies included small samples of 'healthy' individuals, non-invasive neuromodulation techniques might have the potential to normalise confidence abnormalities in psychiatric patients. Additionally, TMS and tDCS might bolster clinical improvements when used as an adjunct to other interventions. For example, the overconfidence bias associated with Compulsivity and Intrusive Thought might be normalised to a greater extent by combining haloperidol or metacognitive training with tDCS to the dorsolateral prefrontal cortex. Given the dearth of research, the clinical utility for TMS and tDCS through modification of metacognition is highly speculative. Substantially more research is required to evaluate whether non-invasive neuromodulate techniques can modify metacognitive abilities among psychiatric patients and how this impacts symptom presentations.

Throughout this thesis, we demonstrated that Compulsivity and Intrusive Thought is linked to overconfidence in performance. Compulsivity and Intrusive Thought is characterised by higher positive loadings for OCD-related symptoms. In Chapter 4 of this thesis, relatively higher loadings for OCD related symptoms were evident across dimensions that captured the greatest variance in overconfidence, which included Compulsivity and Intrusive Thought. However, this appears at odds with prior research showing that patients with OCD suffer from underconfidence,

rather than overconfidence (Dar et al., 2022; Hoven et al., 2019). It has been suggested that any evidence of underconfidence in OCD is an artefact of case-controlled comparisons, as heterogenous symptom presentations (e.g., comorbid anxiety) obscure specific, transdiagnostic processes (Gillan et al., 2017; Wise et al., 2023). However, it has also been argued that OCD patients might be systematically different from those in the general population that endorse higher levels of OCD-related symptoms (Hoven, Rouault, et al., 2023; Seow et al., 2024). One study found that when matched for levels of depression and anxiety, OCD patients had significantly lower confidence relative to 'highly compulsive' individuals, who were those defined as scoring high on a measure of obsessive-compulsive disorder symptoms (Hoven, Rouault, et al., 2023). Some caveats to interpretation are that this study did not measure Compulsivity and Intrusive Thought, the sample size was relatively small, and there are systematic differences in how samples were recruited (Hoven, Rouault, et al., 2023). Nonetheless, the results are suggestive that patterns of metacognitive biases may differ between clinical and general populations. In Chapter 2, we demonstrated that higher Compulsivity and Intrusive Thought is associated with overconfidence in a well-characterised clinical sample. However, the vast majority were being treated for depression and anxiety disorders, and only a small minority had clinically significant OCD symptoms (Lee et al., 2023). Therefore, future research might investigate how metacognitive bias is associated with OCD severity and levels of Compulsivity and Intrusive Thought among OCD patients. Another avenue for future research is to examine whether metacognitive bias among clinical patients is better explained by OCD status or levels of Compulsivity and Intrusive Thought. Using a large patient sample, deficits in model-based planning were better explained by a transdiagnostic compulsivity dimension, rather than OCD status (Gillan et al., 2020). This demonstrated that transdiagnostic dimensions can be more informative than diagnostic status for individual differences in cognition among clinical patients.

From the outset, this thesis argued that a dimensional, transdiagnostic approach offers advantages over the traditional, categorical framework. The experimental chapters of this thesis showcased that the dimensional approach is more powerful for explaining individual differences in metacognition and model-based

planning. However, mechanistic validity is just one way to evaluate a candidate classification system in psychiatry. Alternatively, a classification system can be judged by its utility in clinical practice. A major concern with the DSM and ICD frameworks is that disorders do not map onto specific interventions, which limits its utility for guiding treatment decisions (First et al., 2018). As an alternative, some have argued that the HiTOP framework is poised for integration into clinic settings (Conway et al., 2021; Hopwood et al., 2020; Ruggero et al., 2019). Using HiTOP, a patient's profile reflects their relative position across higher-order dimensions (e.g., high score on 'Internalising' and low score on 'Externalising spectra) (Kotov et al., 2018). A patient's profile, compromising multiple dimensional scores across the HiTOP framework, subsequently guides the clinician's treatment planning (Ruggero et al., 2019). While the clinical utility of HiTOP has demonstrated through hypothetical cases (Hopwood et al., 2020; Ruggero et al., 2019), there is no empirical evidence to support its utility in clinic. Currently, there is no clear evidence that HiTOP has greater clinical utility than the DSM (Haeffel et al., 2022). This is also the case for the three-factor structure considered throughout this thesis (Gillan et al., 2016); established mechanistic validity does not necessarily translate to clinical utility.

To assess clinical utility, future research should determine if clinicians can reliably and validly measure and interpret psychiatric dimensional profiles (Haeffel et al., 2022). Additionally, a clinically useful framework should aid communication with patients and guide treatment decisions (Ruggero et al., 2019). There is some preliminary evidence to support the feasibility of dimensional frameworks in clinic. Comparing dimension and categorical models of personality, clinicians generally endorsed the utility of a dimensional framework, which facilitated better communication and treatment planning relative to categorical models (Hansen et al., 2019; Morey et al., 2014). However, this evidence is confined to models of pathological personality, as opposed to general mental health problems. A true test of clinical utility would be if treatment outcomes are better for patients when clinicians use a dimensional framework, rather than other taxonomies, such as the DSM. This has yet to be evaluated. As it stands, the dimensional approach to psychopathology is a nascent framework that requires substantially more research to support a transition to clinical settings.

5.4 Conclusions

In summary, this thesis sought to evaluate metacognition as a transdiagnostic mechanism of psychopathology. Disentangling inconsistent effects observed with the traditional disorder-based framework, the transdiagnostic approach delineated neurocognitive mechanisms underlying specific dimensions of psychopathology. Harnessing novel, online methods to acquire large-scale datasets, this thesis replicated cross-sectional associations between metacognitive bias and transdiagnostic dimensions in a clinical, treatment-seeking cohort (Chapter 2) and an unpaid, citizen scientist sample (Chapter 3) for the first time. Data-driven multiverse analysis of large, independent datasets revealed transdiagnostic dimensions that captured the natural covariance patterns of psychiatric symptoms were not purely descriptive and had optimal mechanistic validity (Chapter 4). The next frontier is to investigate time-lagged, within-person associations between metacognition and transdiagnostic dimensions. Dynamic interactions between metacognition and psychopathology can be revealed using Meta Mind, a digital tool that provides rapid, reliable and remote estimates of metacognition (Chapter 3). While changes in metacognitive bias scaled with clinical improvements (Chapter 2), the directionality of this relationship remains unknown. Uncovering the temporal dependence of clinical improvements on metacognitive changes represents an opportunity to advance treatment prediction and precision.

6 – References

Abi-Dargham, A., Moeller, S. J., Ali, F., DeLorenzo, C., Domschke, K., Horga, G., Jutla, A., Kotov, R., Paulus, M. P., Rubio, J. M., Sanacora, G., Veenstra-VanderWeele, J., & Krystal, J. H. (2023). Candidate biomarkers in psychiatric disorders: State of the field. *World Psychiatry*, *22*(2), 236–262. https://doi.org/10.1002/wps.21078

Abramovitch, A., Short, T., & Schweiger, A. (2021). The C Factor: Cognitive dysfunction as a transdiagnostic dimension in psychopathology. *Clinical Psychology Review*, *86*, 102007. https://doi.org/10.1016/j.cpr.2021.102007

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. https://doi.org/10.1016/j.cognition.2015.10.006

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., de Leeuw, J. R., Ma, W. J., Murayama, K., Nelson, J. D., van Opheusden, B., Pouncy, T., Rafner, J., Rahwan, I., Rutledge, R. B., … Schulz, E. (2024). Using games to understand the mind. *Nature Human Behaviour*, 1–9. https://doi.org/10.1038/s41562-024-01878-9

Allsopp, K., Read, J., Corcoran, R., & Kinderman, P. (2019). Heterogeneity in psychiatric diagnostic classification. *Psychiatry Research*, *279*, 15–22. https://doi.org/10.1016/j.psychres.2019.07.005

Althouse, A. D. (2016). Adjust for multiple comparisons? It's not that simple. *The Annals of Thoracic Surgery*, *101*(5), 1644–1645. https://doi.org/10.1016/j.athoracsur.2015.11.024

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders: DSM-5-TR*. American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425787

Anderson, T., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 111–150.

Andersson, G., Carlbring, P., & Rozental, A. (2019). Response and remission rates in internet-based cognitive behavior therapy: An individual patient data meta-analysis. *Frontiers in Psychiatry*, *10*, 749. https://doi.org/10.3389/fpsyt.2019.00749

Andreou, C., Moritz, S., Veith, K., Veckenstedt, R., & Naber, D. (2014). Dopaminergic modulation of probabilistic reasoning and overconfidence in errors: A double-blind study. *Schizophrenia Bulletin*, *40*(3), 558–565. https://doi.org/10.1093/schbul/sbt064

Aristodemou, M. E., & Fried, E. I. (2020). Common factors and interpretation of the p factor of psychopathology. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59*(4), 465–466. https://doi.org/10.1016/j.jaac.2019.07.953

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, *63*(4), 397–417. https://doi.org/10.1111/jcpp.13481

Bakker, G. M. (2019). A new conception and subsequent taxonomy of clinical psychological problems. *BMC Psychology*, *7*(1), 46. https://doi.org/10.1186/s40359-019-0318-8

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience*, *33*(42), 16657–16665. https://doi.org/10.1523/JNEUROSCI.0786-13.2013

Beck, A. T. (2002). Cognitive models of depression. *Clinical advances in cognitive psychotherapy: Theory and application*. *14*(1), 29–61.

Benwell, C. S. Y., Mohr, G., Wallberg, J., Kouadio, A., & Ince, R. A. A. (2022). Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *Npj Mental Health Research*, *1*(1), 10. https://doi.org/10.1038/s44184-022-00009-4

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bona, S., & Silvanto, J. (2014). Accuracy and confidence of visual short-term memory do not go hand-in-hand: Behavioral and neural dissociations. *PLoS ONE*, *9*(3), e90808. https://doi.org/10.1371/journal.pone.0090808

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Bostic, J. Q., & Rho, Y. (2006). Target-symptom psychopharmacology: Between the forest and the trees. *Child and Adolescent Psychiatric Clinics of North America*, *15*(1), 289–302. https://doi.org/10.1016/j.chc.2005.08.003

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606–615. https://doi.org/10.1037/rev0000108

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science – the utility of smartphones. *PLoS ONE*, *9*(7), e100662. https://doi.org/10.1371/journal.pone.0100662

Browning, M., Kingslake, J., Dourish, C. T., Goodwin, G. M., Harmer, C. J., & Dawson, G. R. (2019). Predicting treatment response to antidepressant medication using early changes in emotional processing. *European Neuropsychopharmacology*, *29*(1), 66–75. https://doi.org/10.1016/j.euroneuro.2018.11.1102

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Burnette, C. B., Luzier, J. L., Bennett, B. L., Weisenmuller, C. M., Kerr, P., Martin, S., Keener, J., & Calderwood, L. (2022). Concerns and recommendations for using Amazon MTurk for eating disorder research. *International*

Journal of Eating Disorders, *55*(2), 263–272. https://doi.org/10.1002/eat.23614
Butzbach, M., Fuermaier, A. B. M., Aschenbrenner, S., Weisbrod, M., Tucha, L., & Tucha, O. (2021). Metacognition, psychopathology and daily functioning in adult ADHD. *Journal of Clinical and Experimental Neuropsychology*, *43*(4), 384–398. https://doi.org/10.1080/13803395.2021.1940104
Buyalskaya, A., Ho, H., Milkman, K. L., Li, X., Duckworth, A. L., & Camerer, C. (2023). What can machine learning teach us about habit formation? Evidence from exercise and hygiene. *Proceedings of the National Academy of Sciences*, *120*(17), e2216115120. https://doi.org/10.1073/pnas.2216115120
Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007
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. https://doi.org/10.1037/xge0000505
Carpenter, K. L., & Williams, D. M. (2023). A meta-analysis and critical review of metacognitive accuracy in autism. *Autism*, *27*(2), 512–525. https://doi.org/10.1177/13623613221106004
Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p Factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473
Caspi, A., Houts, R. M., Fisher, H. L., Danese, A., & Moffitt, T. E. (2024). The general factor of psychopathology (p): Choosing among competing models and interpreting p. *Clinical Psychological Science*, *12*(1), 53–82. https://doi.org/10.1177/21677026221147872
Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: mental disorders in one dimension. *American Journal of Psychiatry*, *175*(9), 831–844. https://doi.org/10.1176/appi.ajp.2018.17121383
Cassidy, W. L., Flanagan, N. B., Spellman, M., & Cohen, M. E. (1957). Clinical observations in manic-depressive disease: A quantitative study of one hundred manic-depressive patients and fifty medically sick controls. *Journal of the American Medical Association*, *164*(14), 1535–1546.
Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*. https://doi.org/10.1207/s15327906mbr0102_10
Caziot, B., & Mamassian, P. (2021). Perceptual confidence judgments reflect self-consistency. *Journal of Vision*, *21*(12), 8. https://doi.org/10.1167/jov.21.12.8
Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*(1), 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623
Chapa, D. A. N., Bohrer, B. K., & Forbush, K. T. (2018). Is the diagnostic threshold for bulimia nervosa clinically meaningful? *Eating Behaviors*, *28*, 16–19. https://doi.org/10.1016/j.eatbeh.2017.12.002

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

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

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: *ICD-11, DSM-5* , and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, *18*(2), 72–145. https://doi.org/10.1177/1529100617727266

Constantinou, M. P., Goodyer, I. M., Eisler, I., Butler, S., Kraam, A., Scott, S., Pilling, S., Simes, E., Ellison, R., Allison, E., & Fonagy, P. (2019). Changes in general and specific psychopathology factors over a psychosocial intervention. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(8), 776–786. https://doi.org/10.1016/j.jaac.2018.11.011

Conway, C. C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., Mullins-Sweatt, S. N., Shackman, A. J., Skodol, A. E., South, S. C., Sunderland, M., Waszczuk, M. A., Zald, D. H., Afzali, M. H., Bornovalova, M. A., Carragher, N., Docherty, A. R., Jonas, K. G., Krueger, R. F., … Eaton, N. R. (2019). A hierarchical taxonomy of psychopathology can transform mental health research. *Perspectives on Psychological Science*, *14*(3), 419–436.

Conway, C. C., Krueger, R. F., HiTOP Consortium Executive Board, Cicero, D. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Kotov, R., Latzman, R. D., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., Watson, D., Widiger, T. A., & Wright, A. G. C. (2021). Rethinking the Diagnosis of Mental Disorders: Data-Driven Psychological Dimensions, Not Categories, as a Framework for Mental-Health Research, Treatment, and Training. *Current Directions in Psychological Science*, *30*(2), 151–158. https://doi.org/10.1177/0963721421990353

Costello, A. B., & Osborne, J. (2019). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(1), 7. https://doi.org/10.7275/JYJ1-4868

Courtney, D. B., Watson, P., Chan, B. W., Bennett, K., Krause, K. R., Offringa, M., Butcher, N. J., Monga, S., Neprily, K., Zentner, T., Rodak, T., & Szatmari, P. (2021). Forks in the road: Definitions of response, remission, recovery, and other dichotomized outcomes in randomized controlled trials for adolescent depression. A scoping review. *Depression and Anxiety*, *38*(11), 1152–1168. https://doi.org/10.1002/da.23200

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V. D., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2018). Global determinants of navigation ability. *Current Biology*, *28*(17), 2861-2866.e4. https://doi.org/10.1016/j.cub.2018.06.009

Cuijpers, P., & Smit, F. (2004). Subthreshold depression as a risk indicator for major depressive disorder: A systematic review of prospective studies. *Acta Psychiatrica Scandinavica*, *109*(5), 325–331. https://doi.org/10.1111/j.1600-0447.2004.00301.x

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

Da Fonseca, M., Maffei, G., Moreno-Bote, R., & Hyafil, A. (2023). Mood and implicit confidence independently fluctuate at different time scales. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(1), 142–161. https://doi.org/10.3758/s13415-022-01038-4

Dalgleish, T., Black, M., Johnston, D., & Bevan, A. (2020). Transdiagnostic approaches to mental health problems: Current status and future directions. *Journal of Consulting and Clinical Psychology*, *88*(3), 179–195. https://doi.org/10.1037/ccp0000482

Daniel-Watanabe, L., McLaughlin, M., Gormley, S., & Robinson, O. J. (2022). Association between a directly translated cognitive measure of negative bias and self-reported psychiatric symptoms. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(2), 201–209. https://doi.org/10.1016/j.bpsc.2020.02.010

Dar, R., Sarna, N., Yardeni, G., & Lazarov, A. (2022). Are people with obsessive-compulsive disorder under-confident in their memory and perception? A review and meta-analysis. *Psychological Medicine*, *52*(13), 2404–2412. https://doi.org/10.1017/S0033291722001908

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

De Beukelaer, S., Vehar, N., Rollwage, M., Fleming, S. M., & Tsakiris, M. (2023). Changing minds about climate change: A pervasive role for domain-general metacognition. *Humanities and Social Sciences Communications*, *10*(1), 46. https://doi.org/10.1057/s41599-023-01528-x

Demirtas-Tatlidede, A., Vahabzadeh-Hagh, A. M., & Pascual-Leone, A. (2013). Can noninvasive brain stimulation enhance cognition in neuropsychiatric disorders? *Neuropharmacology*, *64*, 566–578. https://doi.org/10.1016/j.neuropharm.2012.06.020

Dijkstra, N., & De Bruin, L. (2016). Cognitive neuroscience and causal inference: implications for psychiatry. *Frontiers in Psychiatry*, *7*. https://doi.org/10.3389/fpsyt.2016.00129

Donegan, K. R., Brown, Vanessa, Price, Rebecca B, Gallagher, Eoghan, Pringle, Andrew, Hanlon, Anna K, & Gillan, Claire. (2023). Using smartphones to optimise and scale-up the assessment of model-based planning. *Commun Psychol*, *1*, 31. https://doi.org/10.1038/s44271-023-00031-y

Donegan, K. R., & Gillan, C. M. (2022). New principles and new paths needed for online research in mental health: Commentary on Burnette et al. (2021). *International Journal of Eating Disorders*, *55*(2), 278–281. https://doi.org/10.1002/eat.23670

Doyle, A. E., Vuijk, P. J., Doty, N. D., McGrath, L. M., Willoughby, B. L., O'Donnell, E. H., Wilson, H. K., Colvin, M. K., Toner, D. C., Hudson, K. E., Blais, J. E., Ditmars, H. L., Faraone, S. V., Seidman, L. J., & Braaten, E. B. (2018). Cross-disorder cognitive impairments in youth referred for neuropsychiatric

evaluation. *Journal of the International Neuropsychological Society*, *24*(1), 91–103. https://doi.org/10.1017/S1355617717000601

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105–123.

Eadeh, H.-M., Markon, K. E., Nigg, J. T., & Nikolas, M. A. (2021). Evaluating the viability of neurocognition as a transdiagnostic construct using both latent variable models and network analysis. *Research on Child and Adolescent Psychopathology*, *49*(6), 697–710. https://doi.org/10.1007/s10802-021-00770-8

East-Richard, C., R-Mercier, A., Nadeau, D., & Cellard, C. (2020). Transdiagnostic neurocognitive deficits in psychiatry: A review of meta-analyses. *Canadian Psychology / Psychologie Canadienne*, *61*(3), 190–214. https://doi.org/10.1037/cap0000196

Eaton, N. R., Bringmann, L. F., Elmer, T., Fried, E. I., Forbes, M. K., Greene, A. L., Krueger, R. F., Kotov, R., McGorry, P. D., Mei, C., & Waszczuk, M. A. (2023). A review of approaches and models in psychopathology conceptualization research. *Nature Reviews Psychology*, *2*(10), 622–636. https://doi.org/10.1038/s44159-023-00218-4

Eaton, N. R., South, S. C., & Krueger, R. F. (2010). The meaning of comorbidity among common mental disorders. In T. Millon, R. Krueger, & E. Simonsen (Eds.), *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (2nd ed., pp.223-241). New York: Guilford Publications.

Eichner, C., & Berna, F. (2016). Acceptance and efficacy of metacognitive training (MCT) on positive symptoms and delusions in patients with schizophrenia: A meta-analysis taking into account important moderators. *Schizophrenia Bulletin*, *42*(4), 952–962. https://doi.org/10.1093/schbul/sbv225

Eilert, N., Enrique, A., Wogan, R., Mooney, O., Timulak, L., & Richards, D. (2021). The effectiveness of Internet-delivered treatment for generalized anxiety disorder: An updated systematic review and meta-analysis. *Depression and Anxiety*, *38*(2), 196–219. https://doi.org/10.1002/da.23115

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151.

Eisenacher, S., Rausch, F., Ainser, F., Mier, D., Veckenstedt, R., Schirmbeck, F., Lewien, A., Englisch, S., Andreou, C., Moritz, S., Meyer-Lindenberg, A., Kirsch, P., & Zink, M. (2015). Investigation of metamemory functioning in the at-risk mental state for psychosis. *Psychological Medicine*, *45*(15), 3329–3340. https://doi.org/10.1017/S0033291715001373

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Engeler, N. C., & Gilbert, S. J. (2020). The effect of metacognitive training on confidence and strategic reminder setting. *PloS One*, *15*(10), e0240858. https://doi.org/10.1371/journal.pone.0240858

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Etkin, A. (2019). A reckoning and research agenda for neuroimaging in psychiatry. *American Journal of Psychiatry*, *176*(7), 507–511. https://doi.org/10.1176/appi.ajp.2019.19050521

Fabiano, F., & Haslam, N. (2020). Diagnostic inflation in the DSM: A meta-analysis of changes in the stringency of psychiatric diagnosis from DSM-III to DSM-5. *Clinical Psychology Review*, *80*, 101889.

Fernández, A., Pinto-Meza, A., Bellón, J. A., Roura-Poch, P., Haro, J. M., Autonell, J., Palao, D. J., Peñarrubia, M. T., Fernández, R., Blanco, E., Luciano, J. V., & Serrano-Blanco, A. (2010). Is major depression adequately diagnosed and treated by general practitioners? Results from an epidemiological study. *General Hospital Psychiatry*, *32*(2), 201–209. https://doi.org/10.1016/j.genhosppsych.2009.11.015

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

First, M. B. (2005). Mutually exclusive versus co-occurring diagnostic categories: The challenge of diagnostic comorbidity. *Psychopathology*, *38*(4), 206–210.

First, M. B., Rebello, T. J., Keeley, J. W., Bhargava, R., Dai, Y., Kulygina, M., Matsumoto, C., Robles, R., Stona, A., & Reed, G. M. (2018). Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry*, *17*(2), 187–195. https://doi.org/10.1002/wps.20525

First, M. B., Reed, G. M., Hyman, S. E., & Saxena, S. (2015). The development of the ICD-11 clinical descriptions and diagnostic guidelines for mental and behavioural disorders. *World Psychiatry*, *14*(1), 82–90. https://doi.org/10.1002/wps.20189

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27). https://doi.org/10.1073/pnas.1711978115

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

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. https://doi.org/10.1093/brain/awu221

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

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.

Forbes, M. K., Kotov, R., Ruggero, C. J., Watson, D., Zimmerman, M., & Krueger, R. F. (2017). Delineating the joint hierarchical structure of clinical and personality disorders in an outpatient psychiatric sample. *Comprehensive Psychiatry*, *79*, 19–30. https://doi.org/10.1016/j.comppsych.2017.04.006

Forbes, M. K., Neo, B., Nezami, O. M., Fried, E. I., Faure, K., Michelsen, B., Twose, M., & Dras, M. (2024). Elemental psychopathology: Distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychological Medicine*, *54*(5), 886–894. https://doi.org/10.1017/S0033291723002544

Forbush, K. T., & Watson, D. (2013). The structure of common and uncommon mental disorders. *Psychological Medicine*, *43*(1), 97–108. https://doi.org/10.1017/S0033291712001092

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, *105*(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Fox, C. A., Lee, C. T., Hanlon, A. K., Seow, T. X., Lynch, K., Harty, S., Richards, D., Palacios, J., O'Keane, V., Stephan, K. E., & Gillan, C. M. (2023). An observational treatment study of metacognition in anxious-depression. *eLife*, *12*(RP87193).

Fox, C. A., McDonogh, A., Donegan, K. R., Teckentrup, V., Crossen, R., Hanlon, A., Gallagher, E., Rouault, M., & Gillan, C. (2023). Reliable, rapid, and remote measurement of metacognitive bias. *Scientific Reports*, *14*(1), 14941. https://doi.org/10.1038/s41598-024-64900-0

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., Gabbard, G. O., Gau, S. S.-F., Javitt, D. C., & Oquendo, M. A. (2013). The initial field trials of DSM-5: New blooms and old thorns. *American Journal of Psychiatry*, *170*(1), 1–5.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Fried, E. I., Greene, A. L., & Eaton, N. R. (2021). The p factor is the sum of its parts, for now. *World Psychiatry*, *20*(1), 69–70. https://doi.org/10.1002/wps.20814

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

Friedemann, M., Fox, C. A., Hanlon, A., Tighe, D., Yeung, N., & Gillan, C. (2023). Confidence biases in problem gambling. *PsyArXiv.* https://doi.org/10.31234/osf.io/j59ds

Gagne, C., Zika, O., Dayan, P., & Bishop, S. J. (2020). Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife*, *9*, e61387. https://doi.org/10.7554/eLife.61387

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science*, *8*(6), 651–662. https://doi.org/10.1177/1745691613504115

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Garakani, A., Murrough, J. W., Freire, R. C., Thom, R. P., Larkin, K., Buono, F. D., & Iosifescu, D. V. (2020). Pharmacotherapy of anxiety disorders: Current and emerging treatment options. *Frontiers in Psychiatry*, *11*, 595584. https://doi.org/10.3389/fpsyt.2020.595584

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: Psychometric features and clinical correlates. *Psychological Medicine*, *12*(4), 871–878. https://doi.org/10.1017/s0033291700049163

Gawęda, Ł., Li, E., Lavoie, S., Whitford, T. J., Moritz, S., & Nelson, B. (2018). Impaired action self-monitoring and cognitive confidence among ultra-high risk for psychosis and first-episode psychosis patients. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *47*, 67–75. https://doi.org/10.1016/j.eurpsy.2017.09.003

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*, *33*(2), 271–286. https://doi.org/10.1080/13854046.2018.1535662

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gillan, C. M., Fineberg, N. A., & Robbins, T. W. (2017). A trans-diagnostic perspective on obsessive-compulsive disorder. *Psychological Medicine*, *47*(9), 1528–1548. https://doi.org/10.1017/S0033291716002786

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. (2020). Comparison of the association between goal-directed planning and self-

reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry*, *77*(1), 77–85. https://doi.org/10.1001/jamapsychiatry.2019.2998
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. https://doi.org/10.7554/eLife.11305
Gillan, C. M., & Rutledge, R. B. (2021). Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience*, *44*(1), 129–151. https://doi.org/10.1146/annurev-neuro-101220-014053
Gillan, C. M., & Seow, T. X. F. (2020). Carving out new transdiagnostic dimensions for research in mental health. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(10), 932–934. https://doi.org/10.1016/j.bpsc.2020.04.013
Goldin, P. R., Jazaieri, H., Ziv, M., Kraemer, H., Heimberg, R. G., & Gross, J. J. (2013). Changes in positive self-views mediate the effect of cognitive-behavioral therapy for social anxiety disorder. *Clinical Psychological Science*, *1*(3), 301–310. https://doi.org/10.1177/2167702613476867
Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., Ortega, B. N., Zaiko, Y. V., Roach, E. L., Korgaonkar, M. S., Grieve, S. M., Galatzer-Levy, I., Fox, P. T., & Etkin, A. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA Psychiatry*, *72*(4), 305. https://doi.org/10.1001/jamapsychiatry.2014.2206
Gøtzsche, P. C., & Dinnage, O. (2020). What have antidepressants been tested for? A systematic review. *The International Journal of Risk & Safety in Medicine*, *31*(3), 157–163. https://doi.org/10.3233/JRS-200006
Gorsuch, R., & Nelson, J. (1981). CNG screen test: An objective procedure for determining the number of factors. *Presented at the Annual Meeting of the Society for Multivariate Experimental Psychology.*
Goupil, L., & Kouider, S. (2016). Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Current Biology*, *26*(22), 3038–3045. https://doi.org/10.1016/j.cub.2016.09.004
Guggenmos, M. (2021). Measuring metacognitive performance: Type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1), niab040. https://doi.org/10.1093/nc/niab040
Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*, e13388. https://doi.org/10.7554/eLife.13388
Haeffel, G. J., Jeronimus, B. F., Fisher, A. J., Kaiser, B. N., Weaver, L. J., Vargas, I., Goodson, J. T., Soyster, P. D., & Lu, W. (2022). The Hierarchical Taxonomy of Psychopathology (HiTOP) is not an improvement over the DSM. *Clinical Psychological Science*, *10*(2), 285–290. https://doi.org/10.1177/21677026211068873
Haeffel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., Vargas, I., Goodson, J. T., & Lu, W. (2022). Folk classification and factor rotations: whales, sharks, and the problems with the Hierarchical Taxonomy of Psychopathology (HiTOP). *Clinical Psychological Science*, *10*(2), 259–278. https://doi.org/10.1177/21677026211002500
Haller, H., Cramer, H., Lauche, R., Gass, F., & Dobos, G. J. (2014). The prevalence and burden of subthreshold generalized anxiety disorder: A

systematic review. *BMC Psychiatry*, *14*(1), 128. https://doi.org/10.1186/1471-244X-14-128
Hansen, S. J., Christensen, S., Kongerslev, M. T., First, M. B., Widiger, T. A., Simonsen, E., & Bach, B. (2019). Mental health professionals' perceived clinical utility of the *ICD-10* vs. *ICD-11* classification of personality disorders. *Personality and Mental Health*, *13*(2), 84–95. https://doi.org/10.1002/pmh.1442
Haslam, N., Holland, E., & Kuppens, P. (2012). Categories *versus* dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903–920. https://doi.org/10.1017/S0033291711001966
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing ourselves together: The cultural origins of metacognition. *Trends in Cognitive Sciences*, *24*(5), 349–362. https://doi.org/10.1016/j.tics.2020.02.007
Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational psychiatry needs time and context. *Annual Review of Psychology*, *73*(1), 243–270.
Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1
Hoogervorst, K., Banellis, L., Fardo, F., Xue, K., Rahnev, D., & Allen, M. (2024). Gender differences in metacognition: Global and local contrasts in bias and efficiency. *PsyArXiv*. https://doi.org/10.31234/osf.io/2gwky
Hopkins, A., Gillan, C. M., Roiser, J., Wise, T., & Sidarus, N. (2022). Optimising the measurement of anxious-depressive, compulsivity and intrusive thought and social withdrawal transdiagnostic symptom dimensions. *PsyArXiv*. https://doi.org/10.31234/osf.io/q83sh
Hopwood, C. J., Bagby, R. M., Gralnick, T., Ro, E., Ruggero, C., Mullins-Sweatt, S., Kotov, R., Bach, B., Cicero, D. C., Krueger, R. F., Patrick, C. J., Chmielewski, M., DeYoung, C. G., Docherty, A. R., Eaton, N. R., Forbush, K. T., Ivanova, M. Y., Latzman, R. D., Pincus, A. L., … Zimmermann, J. (2020). Integrating psychotherapy with the Hierarchical Taxonomy of Psychopathology (HiTOP). *Journal of Psychotherapy Integration*, *30*(4), 477–497. https://doi.org/10.1037/int0000156
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447
Horvath, A., & Todd, J. (2023). Shared depressive symptomatology across mental disorders: Implications for mental health assessments. *Preprint on OSF*. https://doi.org/doi:10.31234/osf.io/v7zcs
Hoven, M., Denys, D., Rouault, M., Luigjes, J., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nature Mental Health*, 1–9. https://doi.org/10.1038/s44220-023-00062-8
Hoven, M., Lebreton, 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), 268. https://doi.org/10.1038/s41398-019-0602-7

Hoven, M., Rouault, M., Van Holst, R., & Luigjes, J. (2023). Differences in metacognitive functioning between obsessive–compulsive disorder patients and highly compulsive individuals from the general population. *Psychological Medicine*, *53*(16), 7933–7942. https://doi.org/10.1017/S003329172300209X

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(1), 3–19. https://doi.org/10.1038/s41386-020-0746-4

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*(1), 155–179. https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Hyman, S. E. (2021). Psychiatric Disorders: Grounded in human biology but not natural kinds. *Perspectives in Biology and Medicine*, *64*(1), 6–28. https://doi.org/10.1353/pbm.2021.0002

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*(7), 748-751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jelinek, L., Van Quaquebeke, N., & Moritz, S. (2017). Cognitive and metacognitive mechanisms of change in metacognitive training for depression. *Scientific Reports*, *7*(1), Article 1. https://doi.org/10.1038/s41598-017-03626-8

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. https://doi.org/10.1016/j.beth.2015.12.002

Kadlec, J., Walsh, C. R., Sadé, U., Amir, A., Rissman, J., & Ramot, M. (2024). A measure of reliability convergence to select and optimize cognitive tasks for individual differences research. *Communications Psychology*, *2*(1), 64. https://doi.org/10.1038/s44271-024-00114-4

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179. https://doi.org/10.1038/mp.2012.105

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, *148*, 105137. https://doi.org/10.1016/j.neubiorev.2023.105137

Karyotaki, E., Efthimiou, O., Miguel, C., Bermpohl, F. M. G., Furukawa, T. A., Cuijpers, P., Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, Riper, H., Patel, V., Mira, A., Gemmil, A. W., Yeung, A. S., Lange, A., Williams, A. D., Mackinnon, A., Geraedts, A., van

Straten, A., Meyer, B., Björkelund, C., … Forsell, Y. (2021). Internet-based cognitive behavioral therapy for depression: A systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*, *78*(4), 361–371. https://doi.org/10.1001/jamapsychiatry.2020.4364

Katz, J. A., Rufino, K. A., Werner, C., McIngvale, E., & Storch, E. (2020). OCD in ethnic minorities. *Clinical and Experimental Psychology*, *6*(1), 1-6.

Keidel, K., Lu, X., Suzuki, S., Murawski, C., & Ettinger, U. (2024). Association of temporal discounting with transdiagnostic symptom dimensions. *NPJ Mental Health Research*, *3*(1), 13. https://doi.org/10.1038/s44184-024-00060-3

Kendler, K. S. (2008). Explanatory Models for Psychiatric Illness. *American Journal of Psychiatry*, *165*(6), 695–702. https://doi.org/10.1176/appi.ajp.2008.07071061

Kendler, K. S. (2009). An historical framework for psychiatric nosology. *Psychological Medicine*, *39*(12), 1935–1941. https://doi.org/10.1017/S0033291709005753

Kendler, K. S., Aggen, S. H., Knudsen, G. P., Røysamb, E., Neale, M. C., & Reichborn-Kjennerud, T. (2011). The structure of genetic and environmental risk factors for syndromal and subsyndromal common DSM-IV Axis I and all Axis II Disorders. *American Journal of Psychiatry*, *168*(1), 29–39. https://doi.org/10.1176/appi.ajp.2010.10030340

Kendler, K. S., Muñoz, R. A., & Murphy, G. (2010). The development of the Feighner criteria: A historical perspective. *American Journal of Psychiatry*, *167*(2), 134–142. https://doi.org/10.1176/appi.ajp.2009.09081155

Kendler, K. S., & Solomon, M. (2016). Expert consensus vs. evidence-based approaches in the revision of the DSM. *Psychological Medicine*, *46*(11), 2255–2262. https://doi.org/10.1017/S003329171600074X

Kessler, R. C., Avenevoli, S., Costello, E. J., Georgiades, K., Green, J. G., Gruber, M. J., He, J., Koretz, D., McLaughlin, K. A., & Petukhova, M. (2012). Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Archives of General Psychiatry*, *69*(4), 372–380.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 617. https://doi.org/10.1001/archpsyc.62.6.617

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., Feusner, J. D., Garfinkel, S. N., Lane, R. D., Mehling, W. E., Meuret, A. E., Nemeroff, C. B., Oppenheimer, S., Petzschner, F. H., Pollatos, O., Rhudy, J. L., Schramm, L. P., Simmons, W. K., Stein, M. B., … Zucker, N. (2018). Interoception and mental health: A roadmap. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(6), 501–513. https://doi.org/10.1016/j.bpsc.2017.12.004

Koen, J. D., Lewis, L., Rugg, M. D., Clementz, B. A., Keshavan, M. S., Pearlson, G. D., Sweeney, J. A., Tamminga, C. A., & Ivleva, E. I. (2023). Supervised machine learning classification of psychosis biotypes based on brain structure: Findings from the Bipolar-Schizophrenia network for intermediate phenotypes (B-SNIP). *Scientific Reports*, *13*(1), 12980. https://doi.org/10.1038/s41598-023-38101-0

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288

Kornell, N. (2009). Metacognition in humans and animals. *Current Directions in Psychological Science*, *18*(1), 11–15. https://doi.org/10.1111/j.1467-8721.2009.01597.x

Kotov, R., Cicero, D. C., Conway, C. C., DeYoung, C. G., Dombrovski, A., Eaton, N. R., First, M. B., Forbes, M. K., Hyman, S. E., Jonas, K. G., Krueger, R. F., Latzman, R. D., Li, J. J., Nelson, B. D., Regier, D. A., Rodriguez-Seijas, C., Ruggero, C. J., Simms, L. J., Skodol, A. E., … Wright, A. G. C. (2022). The Hierarchical Taxonomy of Psychopathology (HiTOP) in psychiatric practice and research. *Psychological Medicine*, *52*(9), 1666–1678. https://doi.org/10.1017/S0033291722001301

Kotov, R., Krueger, R. F., & Watson, D. (2018). A paradigm shift in psychiatric classification: The Hierarchical Taxonomy of Psychopathology (HiTOP). *World Psychiatry*, *17*(1), 24–25. https://doi.org/10.1002/wps.20478

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., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The Hierarchical Taxonomy of Psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology*, *17*(1), 83–108. https://doi.org/10.1146/annurev-clinpsy-081219-093304

Kotov, R., Ruggero, C. J., Krueger, R. F., Watson, D., Yuan, Q., & Zimmerman, M. (2011). New dimensions in the quantitative classification of mental illness. *Archives of General Psychiatry*, *68*(10), 1003–1011. https://doi.org/10.1001/archgenpsychiatry.2011.107

Kozak, M. J., & Cuthbert, B. N. (2016). The NIMH Research Domain Criteria Initiative: Background, issues, and pragmatics. *Psychophysiology*, *53*(3), 286–297. https://doi.org/10.1111/psyp.12518

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*(10), 921. https://doi.org/10.1001/archpsyc.56.10.921

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., Bach, B., Bagby, R. M., Bornovalova, M. A., Carpenter, W. T., Chmielewski, M., Cicero, D. C., Clark, L. A., Conway, C., DeClercq, B., DeYoung, C. G., … Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282–293. https://doi.org/10.1002/wps.20566

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology*, *2*(1), 111–133. https://doi.org/10.1146/annurev.clinpsy.2.022305.095213

Lafrance, M. N., & McKenzie-Mohr, S. (2013). The DSM and its lure of legitimacy. *Feminism & Psychology*, *23*(1), 119–140. https://doi.org/10.1177/0959353512467974

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, *121*(4), 971–977. https://doi.org/10.1037/a0028355

Lahey, B. B., Rathouz, P. J., Van Hulle, C., Urbano, R. C., Krueger, R. F., Applegate, B., Garriock, H. A., Chapman, D. A., & Waldman, I. D. (2008). Testing structural models of DSM-IV symptoms of common forms of child and adolescent psychopathology. *Journal of Abnormal Child Psychology*, *36*(2), 187–206. https://doi.org/10.1007/s10802-007-9169-5

Lee, C. T., Palacios, J., Richards, D., Hanlon, A. K., Lynch, K., Harty, S., Claus, N., Swords, L., O'Keane, V., Stephan, K. E., & Gillan, C. M. (2023). The Precision in Psychiatry (PIP) study: Testing an internet-based methodology for accelerating research in treatment prediction and personalisation. *BMC Psychiatry*, *23*(1), 25. https://doi.org/10.1186/s12888-022-04462-5

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., & Witte, J. S. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*(9), 984–994.

Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, *22*, 141–173. https://doi.org/10.1159/000414022

Lilienfeld, S. O., & Treadway, M. T. (2016). Clashing diagnostic approaches: DSM-ICD Versus RDoC. *Annual Review of Clinical Psychology*, *12*(1), 435–463. https://doi.org/10.1146/annurev-clinpsy-021815-093122

Liu, Y., Tang, C., Hung, T., Tsai, P., & Lin, M. (2018). The efficacy of metacognitive training for delusions in patients with schizophrenia: a meta-analysis of randomized controlled trials informs evidence-based practice. *Worldviews on Evidence-Based Nursing*, *15*(2), 130–139. https://doi.org/10.1111/wvn.12282

Luciano, J. V., Sanabria-Mazo, J. P., Feliu-Soler, A., & Forero, C. G. (2020). The pros and cons of bifactor models for testing dimensionality and psychopathological models: A commentary on the manuscript "A systematic review and meta-analytic factor analysis of the depression anxiety stress scales". *Clinical Psychology: Science and Practice*, *27*(4). https://doi.org/10.1111/cpsp.12386

Lysaker, P. H., Gagen, E., Moritz, S., & Schweitzer, R. D. (2018). Metacognitive approaches to the treatment of psychosis: A comparison of four approaches. *Psychology Research and Behavior Management*, *11*, 341–351. https://doi.org/10.2147/PRBM.S146446

Maj, M. (2005). 'Psychiatric comorbidity': An artefact of current diagnostic systems? *British Journal of Psychiatry*, *186*(3), 182–184. https://doi.org/10.1192/bjp.186.3.182

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings.

Consciousness and Cognition, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the Apathy Evaluation Scale. *Psychiatry Research*, *38*(2), 143–162. https://doi.org/10.1016/0165-1781(91)90040-v

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856–879. https://doi.org/10.1037/a0023678

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

Mayes, R., & Horwitz, A. V. (2005). DSM-III and the revolution in the classification of mental illness. *Journal of the History of the Behavioral Sciences*, *41*(3), 249–267. https://doi.org/10.1002/jhbs.20103

McGrath, J. J., Lim, C. C. W., Plana-Ripoll, O., Holtz, Y., Agerbo, E., Momen, N. C., Mortensen, P. B., Pedersen, C. B., Abdulmalik, J., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., Bunting, B., De Almeida, J. M. C., De Girolamo, G., De Vries, Y. A., Florescu, S., … De Jonge, P. (2020). Comorbidity within mental disorders: A comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences*, *29*, e153. https://doi.org/10.1017/S2045796020000633

McNally, R. J. (2021). Network analysis of psychopathology: Controversies and challenges. *Annual Review of Clinical Psychology*, *17*(1), 31–53. https://doi.org/10.1146/annurev-clinpsy-081219-092850

McTeague, L. M., Huemer, J., Carreon, D. M., Jiang, Y., Eickhoff, S. B., & Etkin, A. (2017). Identification of common neural circuit disruptions in cognitive control across psychiatric disorders. *American Journal of Psychiatry*, *174*(7), 676–685. https://doi.org/10.1176/appi.ajp.2017.16040400

McWilliams, A., Bibby, H., Steinbeis, N., David, A. S., & Fleming, S. M. (2023). Age-related decreases in global metacognition are independent of local metacognition and task performance. *Cognition*, *235*, 105389. https://doi.org/10.1016/j.cognition.2023.105389

Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D., & Kotov, R. (2021). Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clinical Psychology Review*, *86*, 102025. https://doi.org/10.1016/j.cpr.2021.102025

Moeller, S. J., Fleming, S. M., Gan, G., Zilverstand, A., Malaker, P., d′Oleire Uquillas, F., Schneider, K. E., Preston-Campbell, R. N., Parvaz, M. A., Maloney, T., Alia-Klein, N., & Goldstein, R. Z. (2016). Metacognitive impairment in active cocaine use disorder is associated with individual

differences in brain structure. *European Neuropsychopharmacology*, *26*(4), 653–662. https://doi.org/10.1016/j.euroneuro.2016.02.009
Mogoașe, C., Cobeanu, O., David, O., Giosan, C., & Szentagotai, A. (2017). Internet-based psychotherapy for adult depression: What about the mechanisms of change? *Journal of Clinical Psychology*, *73*(1), 5–64. https://doi.org/10.1002/jclp.22326
Molenaar, P. C. M. (2010). Latent variable models are network models. *The Behavioral and Brain Sciences*, *33*(2–3), 166. https://doi.org/10.1017/S0140525X10000798
Molenaar, P. C. M. (2019). Granger causality testing with intensive longitudinal data. *Prevention Science*, *20*(3), 442–451. https://doi.org/10.1007/s11121-018-0919-0
Morey, L. C., Skodol, A. E., & Oldham, J. M. (2014). Clinician judgments of clinical utility: A comparison of DSM-IV-TR personality disorders and the altemative model for DSM-5 personality disorders. *Journal of Abnormal Psychology*, *123*(2), 398–405.
Moritz, S., Andreou, C., Schneider, B. C., Wittekind, C. E., Menon, M., Balzan, R. P., & Woodward, T. S. (2014). Sowing the seeds of doubt: A narrative review on metacognitive training in schizophrenia. *Clinical Psychology Review*, *34*(4), 358–366. https://doi.org/10.1016/j.cpr.2014.04.004
Moritz, S., Kolbeck, K., & Andreou, C. (2016). Antipsychotics decrease response confidence. *Journal of Psychopharmacology*, *30*(8), 831–833. https://doi.org/10.1177/0269881116650404
Moritz, S., Vitzthum, F., Randjbar, S., Veckenstedt, R., & Woodward, T. S. (2010). Detecting and defusing cognitive traps: Metacognitive intervention in schizophrenia. *Current Opinion in Psychiatry*, *23*(6), 561–569. https://doi.org/10.1097/YCO.0b013e32833d16a8
Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A., & Cuthbert, B. N. (2022). Revisiting the seven pillars of RDoC. *BMC Medicine*, *20*(1), 220. https://doi.org/10.1186/s12916-022-02414-0
Mottron, L., & Bzdok, D. (2020). Autism spectrum heterogeneity: Fact or artifact? *Molecular Psychiatry*, *25*(12), 3178–3185. https://doi.org/10.1038/s41380-020-0748-y
Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. H. (2002). The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *The British Journal of Psychiatry: The Journal of Mental Science*, *180*, 461–464. https://doi.org/10.1192/bjp.180.5.461
Newby, J. M., Williams, A. D., & Andrews, G. (2014). Reductions in negative repetitive thinking and metacognitive beliefs during transdiagnostic internet cognitive behavioural therapy (iCBT) for mixed anxiety and depression. *Behaviour Research and Therapy*, *59*, 52–60. https://doi.org/10.1016/j.brat.2014.05.009
Nicholson, T., Williams, D. M., Lind, S. E., Grainger, C., & Carruthers, P. (2021). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology: General*, *150*(2), 206–220. https://doi.org/10.1037/xge0000878
Noreen, S., & MacLeod, M. D. (2015). What do we really know about cognitive inhibition? Task demands and inhibitory effects across a range of memory and behavioural tasks. *PLOS ONE*, *10*(8), e0134951. https://doi.org/10.1371/journal.pone.0134951

Normann, N., & Morina, N. (2018). The efficacy of metacognitive therapy: A systematic review and meta-analysis. *Frontiers in Psychology*, *9*, 2211. https://doi.org/10.3389/fpsyg.2018.02211

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol*, *73*, 719–748.

Noyes, B. K., Munoz, D. P., Khalid-Khan, S., Brietzke, E., & Booij, L. (2022). Is subthreshold depression in adolescence clinically relevant? *Journal of Affective Disorders*, *309*, 123–130. https://doi.org/10.1016/j.jad.2022.04.067

Ohlsson, H., & Kendler, K. S. (2020). Applying causal inference methods in psychiatric epidemiology: A review. *JAMA Psychiatry*, *77*(6), 637. https://doi.org/10.1001/jamapsychiatry.2019.3758

Okasha, A. (2009). Would the use of dimensions instead of categories remove problems related to subthreshold disorders? *European Archives of Psychiatry and Clinical Neuroscience*, *259*(S2), 129–133. https://doi.org/10.1007/s00406-009-0052-y

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology*, *123*(2), 452.

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. https://doi.org/10.1016/j.concog.2014.06.007

Palminteri, S., & Chevallier, C. (2018). Can we infer inter-individual differences in risk-taking from behavioral tasks? *Frontiers in Psychology*, *9*, 2307. https://doi.org/10.3389/fpsyg.2018.02307

Passell, E., Strong, R. W., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., Grinspoon, L., & Germine, L. (2021). Cognitive test scores vary with choice of personal digital device. *Behavior Research Methods*, *53*(6), 2544–2557. https://doi.org/10.3758/s13428-021-01597-3

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1

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. https://doi.org/10.1016/j.biopsych.2018.06.018

Philipp, R., Kriston, L., Kühne, F., Härter, M., & Meister, R. (2020). Concepts of metacognition in the treatment of patients with mental disorders. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *38*(2), 173–183. https://doi.org/10.1007/s10942-019-00333-3

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Prado, C. E., Watt, S., & Crowe, S. F. (2018). A meta-analysis of the effects of antidepressants on cognitive functioning in depressed and non-depressed

samples. *Neuropsychology Review*, *28*(1), 32–72. https://doi.org/10.1007/s11065-018-9369-5
Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist*, *76*(9), 1445–1453. https://doi.org/10.1037/amp0000937
Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods. *PsyArXiv*. https://doi.org/10.31234/osf.io/waz9h
Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1
Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, *170*(1), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999
Rief, W., Hofmann, S. G., Berg, M., Forbes, M. K., Pizzagalli, D. A., Zimmermann, J., Fried, E., & Reed, G. M. (2023). Do we need a novel framework for classifying psychopathology? A discussion paper. *Clinical Psychology in Europe*, *5*(4), e11699. https://doi.org/10.32872/cpe.11699
Rivers, M. L., Fitzsimmons, C. J., Fisk, S. R., Dunlosky, J., & Thompson, C. A. (2021). Gender differences in confidence during number-line estimation. *Metacognition and Learning*, *16*(1), 157–178. https://doi.org/10.1007/s11409-020-09243-7
Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, *28*(24), 4014-4021.e8. https://doi.org/10.1016/j.cub.2018.10.053
Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003
Rosenblat, J. D., Kakar, R., & McIntyre, R. S. (2015). The cognitive effects of antidepressants in major depressive disorder: A systematic review and meta-analysis of randomized clinical trials. *The International Journal of Neuropsychopharmacology*, *19*(2), pyv082. https://doi.org/10.1093/ijnp/pyv082
Ross, C. A., & Margolis, R. L. (2019). Research Domain Criteria: Strengths, weaknesses, and potential alternatives for future psychiatric research. *Complex Psychiatry*, *5*(4), 218–236. https://doi.org/10.1159/000501797
Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3
Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, e17. https://doi.org/10.1017/pen.2018.16
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. https://doi.org/10.1016/j.biopsych.2017.12.017

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J. L., Chmielewski, M., Conway, C. C., De Clercq, B., Dornbach-Bender, A., … Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069–1084. https://doi.org/10.1037/ccp0000452

Rush, J., & Hofer, S. M. (2014). Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment*, *26*(2), 462–473. https://doi.org/10.1037/a0035666

Rutledge, R. B., Smittenaar, P., Zeidman, P., Brown, H. R., Adams, R. A., Lindenberger, U., Dayan, P., & Dolan, R. J. (2016). Risk taking for potential reward decreases across the lifespan. *Current Biology*, *26*(12), 1634–1639. https://doi.org/10.1016/j.cub.2016.05.017

Salagre, E., Solé, B., Tomioka, Y., Fernandes, B. S., Hidalgo-Mazzei, D., Garriga, M., Jimenez, E., Sanchez-Moreno, J., Vieta, E., & Grande, I. (2017). Treatment of neurocognitive symptoms in unipolar depression: A systematic review and future perspectives. *Journal of Affective Disorders*, *221*, 205–221. https://doi.org/10.1016/j.jad.2017.06.034

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 (Abingdon, England)*, *88*(6), 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

Schaeuffele, C., Schulz, A., Knaevelsrud, C., Renneberg, B., & Boettcher, J. (2021). CBT at the crossroads: The rise of transdiagnostic treatments. *International Journal of Cognitive Therapy*, *14*(1), 86–113. https://doi.org/10.1007/s41811-020-00095-2

Schilling, L., Moritz, S., Kriston, L., Krieger, M., & Nagel, M. (2018). Efficacy of metacognitive training for patients with borderline personality disorder: Preliminary results. *Psychiatry Research*, *262*, 459–464. https://doi.org/10.1016/j.psychres.2017.09.024

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537), eabf4740. https://doi.org/10.1126/science.abf4740

Schurr, R., Reznik, D., Hillman, H., Bhui, R., & Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01814-x

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00111

Seow, T. X. F., Benoit, E., Dempsey, C., Jennings, M., Maxwell, A., O'Connell, R., & Gillan, C. M. (2021). Model-based planning deficits in compulsivity are linked to faulty neural representations of task structure. *The Journal of Neuroscience*, *41*(30), 6539–6550. https://doi.org/10.1523/JNEUROSCI.0031-21.2021

Seow, T. X. F., Fleming, S. M., & Hauser, T. U. (2024). Metacognitive biases in anxious-depression and compulsivity extend across perception and memory. *PsyArXiv*. https://doi.org/10.31234/osf.io/avyph

Seow, T. X. F., & Gillan, C. M. (2020). Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Scientific Reports*, *10*(1), 2883. https://doi.org/10.1038/s41598-020-59646-4

Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological Psychiatry*, *90*(7), 436–446. https://doi.org/10.1016/j.biopsych.2021.05.013

Sha, Z., Wager, T. D., Mechelli, A., & He, Y. (2019). Common dysfunction of large-scale neurocognitive networks across psychiatric disorders. *Biological Psychiatry*, *85*(5), 379–388. https://doi.org/10.1016/j.biopsych.2018.11.011

Sharma, V., Sagar, R., Kaloiya, G., & Mehta, M. (2022). The scope of metacognitive therapy in the treatment of psychiatric disorders. *Cureus*. https://doi.org/10.7759/cureus.23424

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Skapinakis, P., Caldwell, D. M., Hollingworth, W., Bryden, P., Fineberg, N. A., Salkovskis, P., Welton, N. J., Baxter, H., Kessler, D., Churchill, R., & Lewis, G. (2016). Pharmacological and psychotherapeutic interventions for management of obsessive-compulsive disorder in adults: A systematic review and network meta-analysis. *The Lancet. Psychiatry*, *3*(8), 730–739. https://doi.org/10.1016/S2215-0366(16)30069-4

Smoller, J. W., Andreassen, O. A., Edenberg, H. J., Faraone, S. V., Glatt, S. J., & Kendler, K. S. (2019). Psychiatric genetics and the structure of psychopathology. *Molecular Psychiatry*, *24*(3), 409–420. https://doi.org/10.1038/s41380-017-0010-4

Spielberger, C., Gorsuch, R., Lushene, R., Vagg, P., & Jacobs, G. (1983). Manual for the State-Trait Anxiety Inventory (Form Y1 – Y2). In *Palo Alto, CA: Consulting Psychologists Press; Vol. IV*.

Stein, D. J., Palk, A. C., & Kendler, K. S. (2021). What is a mental disorder? An exemplar-focused approach. *Psychological Medicine*, *51*(6), 894–901. https://doi.org/10.1017/S0033291721001185

Stephan, K. E., Bach, D. R., Fletcher, P. C., Flint, J., Frank, M. J., Friston, K. J., Heinz, A., Huys, Q. J. M., Owen, M. J., Binder, E. B., Dayan, P., Johnstone, E. C., Meyer-Lindenberg, A., Montague, P. R., Schnyder, U., Wang, X.-J., & Breakspear, M. (2016). Charting the landscape of priority problems in psychiatry, part 1: Classification and diagnosis. *The Lancet Psychiatry*, *3*(1), 77–83. https://doi.org/10.1016/S2215-0366(15)00361-2

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Strand, M. (2011). Where do classifications come from? The DSM-III, the transformation of American psychiatry, and the problem of origins in the sociology of knowledge. *Theory and Society*, *40*(3), 273–313. https://doi.org/10.1007/s11186-011-9142-8

Strauss, G. P., & Cohen, A. S. (2017). A transdiagnostic review of negative symptom phenomenology and etiology. *Schizophrenia Bulletin*, *43*(4), 712–719. https://doi.org/10.1093/schbul/sbx066

Strawbridge, R., Young, A. H., & Cleare, A. J. (2017). Biomarkers for depression: Recent insights, current challenges and future prospects. *Neuropsychiatric Disease and Treatment*, *Volume 13*, 1245–1262. https://doi.org/10.2147/NDT.S114542

Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology*, *27*(1), 1–18. https://doi.org/10.1037/pha0000235

Sun, X., Zhu, C., & So, S. H. W. (2017). Dysfunctional metacognition across psychopathologies: A meta-analytic review. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *45*, 139–153. https://doi.org/10.1016/j.eurpsy.2017.05.029

Tamminga, C. A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., & Thaker, G. (2014). Bipolar and schizophrenia network for intermediate phenotypes: Outcomes across the psychosis continuum. *Schizophrenia Bulletin*, *40*(Suppl 2), S131–S137. https://doi.org/10.1093/schbul/sbt179

Taylor, D. (2016). Prescribing according to diagnosis: How psychiatry is different. *World Psychiatry*, *15*(3), 224–225. https://doi.org/10.1002/wps.20343

The Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., … Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360*(6395), eaap8757. https://doi.org/10.1126/science.aap8757

Thome, J., Liebke, L., Bungert, M., Schmahl, C., Domes, G., Bohus, M., & Lis, S. (2016). Confidence in facial emotion recognition in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *7*(2), 159–168. https://doi.org/10.1037/per0000142

Tiego, J., Martin, E. A., DeYoung, C. G., Hagan, K., Cooper, S. E., Pasion, R., Satchell, L., Shackman, A. J., Bellgrove, M. A., Fornito, A., the HiTOP Neurobiological Foundations Work Group, Abend, R., Goulter, N., Eaton, N. R., Kaczkurkin, A. N., & Nusslock, R. (2023). Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. *Nature Mental Health*, *1*(5), 304–315. https://doi.org/10.1038/s44220-023-00057-5

Vernon, J. A., Grudnikoff, E., Seidman, A. J., Frazier, T. W., Vemulapalli, M. S., Pareek, P., Goldberg, T. E., Kane, J. M., & Correll, C. U. (2014). Antidepressants for cognitive impairment in schizophrenia—A systematic review and meta-analysis. *Schizophrenia Research*, *159*(2–3), 385–394. https://doi.org/10.1016/j.schres.2014.08.015

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (2021). Citizen science in the natural sciences. In K. Vohland (Eds.), *The Science of citizen science* (pp. 79-96). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4

Von Glischinski, M., Von Brachel, R., Thiele, C., & Hirschfeld, G. (2021). Not sad enough for a depression trial? A systematic review of depression measures and cut points in clinical trial registrations. *Journal of Affective Disorders*, *292*, 36–44. https://doi.org/10.1016/j.jad.2021.05.041

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-based control in dimensional psychiatry. *Biological Psychiatry*, *82*(6), 391–400. https://doi.org/10.1016/j.biopsych.2017.04.006

Vrizzi, S., Najar, A., Lemogne, C., Palminteri, S., & Lebreton, M. (2023). Comparing the test-retest reliability of behavioral, computational and self-reported individual measures of reward and punishment sensitivity in relation to mental health symptoms. *PsyArXiv*. https://doi.org/10.31234/osf.io/3u4gp

Waszczuk, M. A., Eaton, N. R., Krueger, R. F., Shackman, A. J., Waldman, I. D., Zald, D. H., Lahey, B. B., Patrick, C. J., Conway, C. C., Ormel, J., Hyman, S. E., Fried, E. I., Forbes, M. K., Docherty, A. R., Althoff, R. R., Bach, B., Chmielewski, M., DeYoung, C. G., Forbush, K. T., … Kotov, R. (2020). Redefining phenotypes to advance psychiatric genetics: Implications from Hierarchical Taxonomy of Psychopathology. *Journal of Abnormal Psychology*, *129*(2), 143–161. https://doi.org/10.1037/abn0000486

Waszczuk, M. A., Zimmerman, M., Ruggero, C., Li, K., MacNamara, A., Weinberg, A., Hajcak, G., Watson, D., & Kotov, R. (2017). What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Comprehensive Psychiatry*, *79*, 80–88.

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Watts, A. L., Green, A. L., Bonifay, W., & Fried, E. I. (2024). A critical evaluation of the p-factor literature. *Nat Rev Psychol*, *3*(108–122). https://doi.org/10.1038/s44159-023-00260-2

Webb, C. A., Rosso, I. M., & Rauch, S. L. (2017). Internet-based cognitive-behavioral therapy for depression: Current progress and future directions. *Harvard Review of Psychiatry*, *25*(3), 114–122. https://doi.org/10.1097/HRP.0000000000000139

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271. https://doi.org/10.1016/j.concog.2013.01.004

Wells, A., & Cartwright-Hatton, S. (2004). A short form of the metacognitions questionnaire: Properties of the MCQ-30. *Behaviour Research and Therapy*, *42*(4), 385–396. https://doi.org/10.1016/S0005-7967(03)00147-5

Wells, A., Fisher, P., Myers, S., Wheatley, J., Patel, T., & Brewin, C. R. (2012). Metacognitive therapy in treatment-resistant depression: A platform trial.

Behaviour Research and Therapy, *50*(6), 367–373. https://doi.org/10.1016/j.brat.2012.02.004

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*, 1–13. https://doi.org/10.33151/ajp.8.3.93

Winter, N. R., Leenings, R., Ernsting, J., Sarink, K., Fisch, L., Emden, D., Blanke, J., Goltermann, J., Opel, N., Barkhau, C., Meinert, S., Dohm, K., Repple, J., Mauritz, M., Gruber, M., Leehr, E. J., Grotegerd, D., Redlich, R., Jansen, A., … Hahn, T. (2022). Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities. *JAMA Psychiatry*, *79*(9), 879. https://doi.org/10.1001/jamapsychiatry.2022.1780

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, *11*(1), 4179. https://doi.org/10.1038/s41467-020-17977-w

Wise, T., Robinson, O., & Gillan, C. (2023). Identifying transdiagnostic mechanisms in mental health using computational factor modeling. *Biological Psychiatry*, *0*(0). https://doi.org/10.1016/j.biopsych.2022.09.034

Wise, T., Sookud, S., Michelini, G., & Mobbs, D. (2024). Transdiagnostic mental health symptom dimensions predict use of flexible model-based inference in complex environments. *PsyArXiv*. https://doi.org/10.31234/osf.io/f49gr

Wittchen, H., & Beesdo-Baum, K. (2018). “Throwing out the baby with the bathwater”? Conceptual and methodological limitations of the HiTOP approach. *World Psychiatry*, *17*(3), 298–299. https://doi.org/10.1002/wps.20561

Wittchen, H., Beesdo-Baum, K., Gloster, A. T., Höfler, M., Klotsche, J., Lieb, R., Beauducel, A., Bühner, M., & Kessler, R. C. (2009). The structure of mental disorders re-examined: Is it developmentally stable and robust against additions? *International Journal of Methods in Psychiatric Research*, *18*(4), 189–203. https://doi.org/10.1002/mpr.298

World Health Organization. (2019). *International statistical classification of diseases and related health problems (11th ed.). https://icd.who.int/*.

Wright, A. G. C., Ringwald, W. R., & Zimmermann, J. (2023). Measuring Multidimensional hierarchies of psychopathology in daily life. *PsyArXiv*. https://doi.org/10.31234/osf.io/h8npk

Wright, L., Lipszyc, J., Dupuis, A., Thayapararajah, S. W., & Schachar, R. (2014). Response inhibition and psychopathology: A meta-analysis of go/no-go task performance. *Journal of Abnormal Psychology*, *123*(2), 429–439. https://doi.org/10.1037/a0036295

Wu, K. D., & Wyman, S. V. (2016). Examination of racial differences in assessment of OCD symptoms and obsessive beliefs. *Journal of Obsessive-Compulsive and Related Disorders*, *10*, 10–18. https://doi.org/10.1016/j.jocrd.2016.04.008

Xia, C. H., Ma, Z., Ciric, R., Gu, S., Betzel, R. F., Kaczkurkin, A. N., Calkins, M. E., Cook, P. A., García De La Garza, A., Vandekar, S. N., Cui, Z., Moore, T. M., Roalf, D. R., Ruparel, K., Wolf, D. H., Davatzikos, C., Gur, R. C., Gur, R. E., Shinohara, R. T., … Satterthwaite, T. D. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, *9*(1), 3003. https://doi.org/10.1038/s41467-018-05317-y

Xu, I., Passell, E., Strong, R. W., Grinspoon, L., Wu, C. Y., Jung, L., Wilmer, J. B., & Germine, L. (2022). No evidence for consistent reliability across 36 variations of the emotional dot probe task in 9,000 participants. *PsyArXiv*. https://doi.org/10.31234/osf.io/58z4n

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, 103196. https://doi.org/10.1016/j.concog.2021.103196

Xue, K., Zheng, Y., Papalexandrou, C., & Rahnev, D. (2023). No gender difference in confidence or metacognitive ability in perceptual decision making. *PsyArXiv*. https://doi.org/10.31234/osf.io/drvk2

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Ye, Q., Zou, F., Lau, H., Hu, Y., & Kwok, S. C. (2018). Causal evidence for mnemonic metacognition in human precuneus. *The Journal of Neuroscience*, *38*(28), 6379–6387. https://doi.org/10.1523/JNEUROSCI.0660-18.2018

Yip, S. W., Barch, D. M., Chase, H. W., Flagel, S., Huys, Q. J. M., Konova, A. B., Montague, R., & Paulus, M. (2023). From computation to clinic. *Biological Psychiatry Global Open Science*, *3*(3), 319–328. https://doi.org/10.1016/j.bpsgos.2022.03.011

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

Young, G., Lareau, C., & Pierre, B. (2014). One quintillion ways to have PTSD comorbidity: Recommendations for the disordered DSM-5. *Psychological Injury and Law*, *7*(1), 61–74. https://doi.org/10.1007/s12207-014-9186-y

Zachar, P., & Kendler, K. S. (2017). The philosophy of nosology. *Annual Review of Clinical Psychology*, *13*(1), 49–71. https://doi.org/10.1146/annurev-clinpsy-032816-045020

Zbozinek, T. D., Rose, R. D., Wolitzky-Taylor, K. B., Sherbourne, C., Sullivan, G., Stein, M. B., Roy-Byrne, P. P., & Craske, M. G. (2012). Diagnostic overlap of generalized anxiety disorder and major depressive disorder in a primary care sample. *Depression and Anxiety*, *29*(12), 1065–1071.

Zhang, R., Peng, X., Song, X., Long, J., Wang, C., Zhang, C., Huang, R., & Lee, T. M. C. (2023). The prevalence and risk of developing major depression among individuals with subthreshold depression in the general population. *Psychological Medicine*, *53*(8), 3611–3620. https://doi.org/10.1017/S0033291722000241

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., & Dalrymple, K. (2015). How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry*, *56*, 29–34. https://doi.org/10.1016/j.comppsych.2014.09.007

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: a narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004

Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, *7*(10), 1667–1681. https://doi.org/10.1038/s41562-023-01640-7

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry*, *12*, 63–70. https://doi.org/10.1001/archpsyc.1965.01720310065008

7 – Supplementary Material

7.1 Supplementary material for Chapter 2

7.1.1 Supplementary methods for Chapter 2

Self-reported psychiatric questionnaires. Participants completed nine standard self-report questionnaires at baseline and four-week follow-up that assess a variety of psychiatric symptoms, including depression (Zung Self-Rating Depression Scale) (Zung, 1965), trait anxiety (State Trait Anxiety Inventory) (Spielberger et al., 1983), schizotypy (Short Scales for Measuring Schizotypy) (Mason et al., 2005), impulsivity (Barratt Impulsiveness Scale 11) (Patton et al., 1995), obsessive-compulsive disorder (OCD) (Obsessive- Compulsive Inventory-Revised) (Foa et al., 2002), social anxiety (Liebowitz Social Anxiety Scale) (Liebowitz, 1987), eating disorders (Eating Attitudes Test) (Garner et al., 1982), apathy (Apathy Evaluation Scale) (Marin et al., 1991), and alcohol misuse (Alcohol Use Disorders Identification Test) (Saunders et al., 1993). The 'Anxious-depression', 'Compulsivity and Intrusive Thought', and 'Social Withdrawal' dimensions were derived as psychiatric symptom dimensions by applying 209 item weights that were previously conducted from factor analysis on these scales (Gillan et al., 2016). Additionally, participants completed the Working and Social Adjustment Scale (WSAS) (Mundt et al., 2002) at baseline and follow-up. Individuals in the iCBT and antidepressant arms received the original WSAS scale, which asks them to rate their level of functional impairment from mental health problems. Individuals in the control group received an adapted version of the WSAS, which asks them to rate their level of impairment from 'their problems' more generally. Each WSAS item was scored from 0 'not at all' to 8 'very severely', with overall scores ranging from 0 to 40. Higher WSAS scores indicated higher levels of functional impairment (Mundt et al., 2002).

Metacognitive task design. On each trial, participants were shown a fixation cross for 1000 milliseconds (ms), followed by two sunflowers with a varying number of seeds for 300 ms. Participants then had unlimited time to make a judgement on which of two sunflower stimuli contained more seeds and then rated their

confidence in each judgement, on a scale from '1=Guessing' to '6=Certain'. There was a total of 210 trials, divided equally into five blocks. While participants were given feedback on which sunflower they had chosen for 500 ms, there was no trial-level feedback provided on their performance. Split-half reliability was high for baseline mean confidence, as indicated by the correlation between odd and even trials across each participant in the iCBT arm ($r(647)=0.98$, $p<0.001$).

Mean accuracy was tightly controlled in this task using a 'two-down one-up' staircase procedure, in which equally sized changes in dot difference occurred after each incorrect response and after two consecutive correct responses. This maintained objective performance across all participants at a desired level of 70% correct, crucial for estimating metacognition without any confound of real performance differences. Changes in dot difference were calculated using log-space, with a start log difference of 4.2 (+70 dots). Differences in step size changed by ± 0.4 for the first five trials, ± 0.2 for the next five trials and ± 0.1 for the remainder of the task. Dot difference on each trial could range from six dots (1.79 in log-space, the most difficult to discriminate) to 81 dots (4.39 in log-space, the easiest to discriminate).

For analyses, mean reaction time to stimuli choice across trials was measured in seconds and task accuracy was calculated as the mean proportion of correct responses across trials. Task difficulty was operationalised as the mean dot difference between the sunflower stimuli across trials, with a greater dot difference reflecting lower task difficulty.

Metacognitive efficiency (M-Ratio). Metacognitive sensitivity (meta-d') is the extent to which confidence ratings discriminate between correct and incorrect trials, while d' is indicated as objective performance. Metacognitive efficiency (M-Ratio) is calculated as the ratio of metacognitive sensitivity to objective performance (i.e., meta-d'/d'). M-Ratio was calculated in a hierarchical Bayesian framework using a freely available toolbox (HMeta toolbox) (Fleming, 2017), http://github.com/smfleming/HMM, accessed June 2022. An M-Ratio value of 1 indicates that confidence was fully informed by accessing the total perceptual information available. In the iCBT arm, we found low split-half reliability for

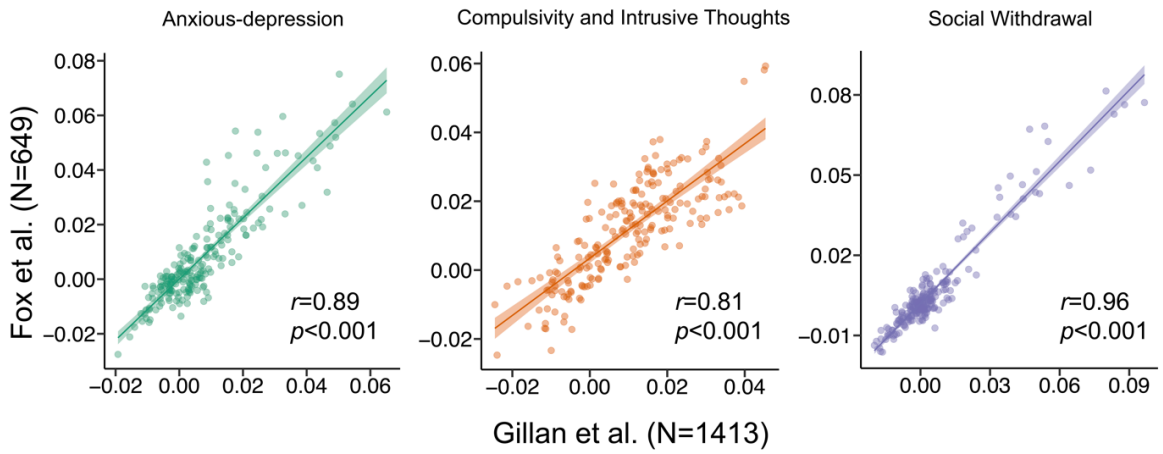
baseline M-Ratio, when correlating odd and even task trials ($r(647)=0.30$, $p<0.001$). The split-half reliability of M-Ratio may have been impacted by the staircase procedure, as there was a significant negatively association between mean accuracy on odd and even trials ($r(647)=-0.25$, $p<0.001$). Additionally, M-Ratio values were confounded by task accuracy ($\beta=-3.19$, $SE=0.62$, $p<0.001$), which violated the assumption that M-Ratio is independent of accuracy. Given that our task has 210 trials, these findings are consistent with the recent report that existing metacognitive efficiency measures are neither independent of type 1 performance nor reliable when tasks have less than 400 trials (Guggenmos, 2021). Thus, all analyses pertaining to M-Ratio should be considered in the context of these limitations.

Treatment data. Every week, participants self-reported if they had started any new concurrent medications and/or psychological treatments for mental health. Participants also reported if they had adhered to iCBT or their antidepressant medications each week. Starting new medication and/or psychological treatments and treatment nonadherence during the study were not grounds for exclusion in any group.

In relation to iCBT programs undertaken by participants in the iCBT arm, ‘Space from Depression’ was the most common iCBT program (n=158, 24.7%), followed by ‘Space from Generalised Anxiety Disorder’ (n=96, 15.0%), ‘Life Skills Online’ (n=89, 13.9%), ‘Space from Anxiety’ (n=84, 13.1%), and ‘Space from Depression and Anxiety’ (n=77, 12.0%). Given the overlapping content of ‘Space from Generalised Anxiety Disorder’ and ‘Space from Anxiety’, these programs were merged into one category ‘Space from Anxiety’ (n=180, 28.1%) for analyses. The ‘Other’ category (n=136, 21.3%) comprised of the following iCBT programs: ‘Space from Stress’ (n=33, 5.2%), ‘Space from Social Anxiety’ (n=20, 3.1%), ‘Space for Resilience’ (n=20, 3.1%), ‘Space for Perinatal Wellbeing’ (n=18, 2.8%), ‘Space in Chronic Pain from Depression and Anxiety’ (n=12, 1.9%), ‘Space from Health Anxiety’ (n=9, 1.4%), ‘Space from OCD’ (n=7, 1.1%), ‘Space from Panic’ (n=6, 0.9%), ‘Space from Phobia’ (n=5, 0.8%), ‘Space for Sleep’ (n=4, 0.6%), ‘Space in Lung Conditions from Depression and Anxiety’ (n=1,

0.2%) and ‘Space from Money Worries’ (n=1, 0.2%), which were merged together into one category for analysis due to low N in each program.

Factor analysis: Transdiagnostic symptom dimensions. Using baseline data from the iCBT arm, we aimed to replicate the three-factor structure for transdiagnostic dimensions of items from nine psychiatric questionnaires reported previously (Gillan et al., 2016). A heterogenous correlation matrix of the 209 items from the nine questionnaires was generated using the hector function in the polycor package in R. We conducted maximum likelihood factor analysis using the fa function from the psych package in R. We specified that three factors would be extracted using regression with an oblique rotation. We applied the same labels previously used for the three factors extracted from our data: 1) Anxious-depression; 2) Compulsivity and Intrusive Thought, and 3) Social Withdrawal (Gillan et al., 2016). Items from scales measuring depression, apathy, and anxiety loaded most strongly onto the Anxious-depression factor. Items relating to OCD, impulsivity, and schizotypy symptoms had the highest loading for Compulsivity and Intrusive Thought. Items that loaded on to Social Withdrawal came largely from the social anxiety scale. The item weights from our data were highly correlated with item weights reported by Gillan and colleagues (Gillan et al., 2016) for Anxious-depression ($r(207)=0.89$, $p<0.001$), Compulsivity and Intrusive Thought ($r(207)=0.81$, $p<0.001$), and Social Withdrawal ($r(207)=0.96$, $p<0.001$), which supported replication of the dimension factor structure among clinical patients in the present study (Supplementary Figure S7.1.1).



Supplementary Figure S7.1.1. Correlation between items weights from factor analysis of transdiagnostic dimensions. Item weights generated from factor analysis in the Gillan et al. (2016) sample were strongly positively correlated with item weights generated from factor analysis in this study, suggesting that the factor structure of psychopathology generated in a general population sample is evident to a clinical, treatment-seeking sample.

Data quality checks: Task and clinical questionnaires. We employed a number of exclusion criteria to ensure high data quality from the metacognitive task. Firstly, we planned to exclude participants if they selected the right or left sunflower on greater than 95% of trials, but none met this criterion. Second, we excluded participants with mean accuracy less than 0.60 and greater than 0.85, indicating the staircase procedure did not converge within acceptable bounds (n=38 at baseline and n=34 at follow-up excluded in the iCBT arm; n=8 at baseline and n=7 at follow-up excluded in the antidepressant arm; n=6 at baseline and n=5 at follow-up excluded in the control group).

To determine the proportion of careless/inattentive responders on the self-report clinical questionnaire, we included 'catch' questions that were embedded in the OCI-R (*If you are paying attention to these questions, please select "A little")* and the WSAS (*If you are paying attention to these questions, please select "Not at all")* at baseline and in the WSAS at all subsequent assessments. In the iCBT arm, 54 (8.3%) participants failed at least one of the checks, with just nine (1.4%) failing more than one. Additionally, eight (9.8%) individuals in the antidepressant arm and seven (8.0%) in the control group failed at least one of the catch items. Given the small number of participants that failed more than one attention checks and that all those participants passed the task exclusion criteria, these individuals

were retained for subsequent analyses. Additionally, excluding those that failed more than one catch item in the iCBT arm did not affect the significance of results, including the change in confidence (β=0.16, SE=0.02, p<0.001), change in Anxious-depression (β=-0.32, SE=0.03, p<0.001), and the association between change in confidence and change in Anxious-depression (r(638)=-0.10, p=0.011).

7.1.2 Supplementary results for Chapter 2

Cross-sectional baseline findings: iCBT. At baseline, age was not significantly associated with mean confidence (β=0.003, SE=0.003, p=0.401). Males had significantly higher confidence (M=3.91, SD=0.81), than females (M=3.75, SD=0.86) (β=0.08, SE=0.04, p=0.044). Mean confidence was significantly lower among those with educational attainment above undergraduate level (M=3.73, SD=0.89), when compared to participants with attainment below undergraduate (M=3.94, SD=0.87) (β=0.21, SE=0.10, p=0.036). Mean confidence did not significantly differ when comparing those who had educational attainment above undergraduate (M=3.73, SD=0.89) to those who had some or had completed an undergraduate degree (M=3.74, SD=0.83) (β=0.01, SE=0.08, p=0.921). When comparing task measures, mean confidence was not significantly associated with mean accuracy (β=-1.29, SE=1.42, p=0.366), mean dot difference (task difficulty) (β=0.002, SE=0.002, p=0.406), mean response time (β=6.47, SE=15.13, p=0.669), or M-Ratio (β=0.11, SE=0.09, p=0.218).

Controlling for age, gender and education in the following baseline analyses, none of the transdiagnostic psychiatric dimensions were associated with overall task accuracy (Anxious-depression, β=0.001, SE=0.001, p=0.260; Compulsivity and Intrusive Thought, β<-0.001, SE=0.001, p = 0.948; Social Withdrawal, β<-0.001, SE=0.001, p=0.656), consistent with the functioning of the staircase. While mean dot difference (task difficulty) was not associated with Anxious-depression (β=-0.58, SE=0.47, p = 0.218) or Social Withdrawal (β=0.28, SE=0.48, p = 0.560), task difficulty was lower among those with higher levels of Compulsivity and Intrusive Thought (β=1.02, SE=0.47, p = 0.032). Mean response time was not associated with Anxious-depression (β=-0.04, SE=0.09, p=0.673),

Compulsivity and Intrusive Thought (β=-0.01, SE=0.09, p=0.882) or Social Withdrawal (β=-0.11, SE=0.09, p=0.214). M-Ratio was also not associated with any of the transdiagnostic psychiatric dimensions (Anxious-depression, β=0.01, SE=0.02, p=0.743; Compulsivity and Intrusive Thought, β=-0.004, SE=0.02, p = 0.783; Social Withdrawal, β=-0.03, SE=0.02, p=0.106).

Treatment findings: iCBT. All scale scores and dimensions significantly improved from baseline to four-week follow-up, except for impulsivity (Supplementary Table S7.1.1, as referenced in the main results). Additionally, functional impairment reduced, as indicated by a decrease in WSAS score from baseline (M=18.78, SD=6.35) to follow-up (M=17.06, SD=8.41) (β=-1.71, SE=0.30, p<0.001). Mean accuracy did not change from baseline (M=0.71, SD=0.02) to follow-up (M=0.71, SD=0.02) (β=0.001, SE=0.001, p=0.299). Mean response time for stimuli choice reduced from baseline (M=0.93, SD=2.22) to follow-up (M=0.69, SD=0.91) (β=-0.24, SE=0.09, p=0.010). Metacognitive efficiency increased from baseline (M=0.81, SD=0.38) to follow-up (M=0.86, SD=0.51) (β=0.05, SE=0.02, p=0.007).

Change in confidence scaled with changes in Anxious-depression (β=-0.07, SE=0.03, p=0.005), trait anxiety (β=-0.08, SE=0.02, p=0.002), depression (β=-0.06, SE=0.02, p=0.011) and alcohol misuse (β=-0.05, SE=0.02, p=0.037), but not significantly with any other psychiatric dimensions or scales (Supplementary Table S7.1.2, as referenced in the main results). Confidence changes did not differ across the different types of iCBT programs, when the Space from Depression and Anxiety program was compared pairwise to Space from Depression (β=-0.09, SE=0.17, p=0.594), Space from Anxiety (β=-0.15, SE=0.17, p=0.376), Life Skills (β=-0.15, SE=0.19, p=0.435) or the ‘other’ category comprising miscellaneous iCBT programs (β=-0.03, SE=0.18, p=0.881).

Supplementary Table S7.1.1 Changes in Psychiatric Dimensions and Scale Scores from Baseline to Follow-up in the iCBT Arm.

Outcome	β	SE	t	p	effect size (r^2)
Dimensions					
AD	-0.41	0.03	-12.80	<0.001	0.04
CIT	-0.28	0.02	-11.85	<0.001	0.02
SW	-0.21	0.03	-7.84	<0.001	0.01
Scale scores					
Trait Anxiety	-0.53	0.04	-14.53	<0.001	0.07
Depression	-0.47	0.03	-13.94	<0.001	0.05
OCD	-0.41	0.03	-14.80	<0.001	0.04
Apathy	-0.25	0.03	-8.34	<0.001	0.02
Social Anxiety	-0.19	0.03	-6.32	<0.001	0.009
Schizotypy	-0.14	0.03	-5.65	<0.001	0.005
Alcohol Misuse	-0.08	0.02	-3.97	<0.001	0.002
Eating Disorders	-0.05	0.02	-2.14	0.033	<0.001
Impulsivity	-0.04	0.02	-1.58	0.114	<0.001

β = standardised beta coefficient, SE = standardised error, t = t-value, p = p-value (unadjusted), AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, OCD = Obsessive compulsive disorder.

Supplementary Table S7.1.2. The Interaction Effect of Time and Psychiatric Dimension/Scale Change on Mean Confidence in the iCBT Arm.

Outcome	β	SE	t	p
Dimensions				
AD	-0.07	0.03	-2.83	0.005
CIT	-0.01	0.03	-0.36	0.723
SW	-0.004	0.03	-0.13	0.894
Scale scores				
Trait Anxiety	-0.08	0.02	-3.18	0.002
Depression	-0.06	0.02	-2.54	0.011
Alcohol Misuse	-0.05	0.02	-2.09	0.037
Apathy	-0.04	0.02	-1.74	0.082
Schizotypy	-0.03	0.02	-1.34	0.180
OCD	-0.02	0.02	-0.99	0.320
Social Anxiety	-0.02	0.02	-0.78	0.434
Eating Disorders	-0.001	0.02	-0.05	0.959
Impulsivity	-0.001	0.02	0.02	0.981

β = standardised beta coefficient, SE = standardised error, t = t-value, p = p-value (unadjusted), AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, OCD = Obsessive compulsive disorder.

Treatment adherence and concurrent treatment. Treatment adherence was high by week 3 in both clinical arms, with over ≥95% of the iCBT arm still undergoing treatment (i.e., 98% at weekly check-in (WCI) 1, 97% at WCI 2 and 95% at WCI 3) and ≥99% of the antidepressant group reported still taking antidepressant medication (i.e., 100% at WCI 1 and WCI 2 and 99% at WCI 3). In the antidepressant arm, four participants altered the dosage of their medication during the study participation (n=2 took less than prescribed and n=2 took more than prescribed). In terms of concurrent treatment, 175 (27.0%) in the iCBT arm were receiving another treatment during the study, of which 48 (7.4%) were taking concurrent medication for a mental health problem and 145 (22.3%) were

receiving a concurrent form of psychotherapy. For the antidepressant group, 33 (40.2%) were receiving another treatment during the study, with 6 (7.3%) taking at least one other medication for a mental health problem and 2 (39.0%) were receiving some form of psychotherapy. Thus, there were partial overlaps in the treatments received across the two clinical arms. In the antidepressant arm, the effect of time on confidence was not dependent on receiving concurrent treatment (β=0.13, SE=0.16, p=0.403). Within the control group, two participants (2.3%) started taking medication during the study and none initiated psychotherapy.

Antidepressant and control arms: Clinical change. All scale scores and dimensions significantly improved from baseline to follow-up in the antidepressant arm, except impulsivity (Supplementary Table S7.1.3, as referenced in the main results). Additionally, WSAS scores significantly reduced among those in the antidepressant arm (β=-0.43, SE=0.12, p<0.001). Among controls, OCD symptoms and alcohol misuse reduced at follow-up (Supplementary Table S7.1.3, as referenced in the main results). WSAS scores, which indicated impairments due to general life among controls, significantly increased from baseline to follow-up (β=0.55, SE=0.11, p<0.001).

Supplementary Table S7.1.3. Changes in Psychiatric Dimensions and Scale Scores from Baseline to Follow-up in Antidepressant and Control Arms.

	Antidepressant Arm				Controls			
Outcome	β	SE	t	p	β	SE	t	p
Dimensions								
AD	-0.70	0.11	-6.65	<0.001	0.13	0.08	1.68	0.096
CIT	-0.30	0.06	-5.22	<0.001	-0.12	0.07	-1.65	0.103
SW	-0.35	0.07	-5.27	<0.001	-0.16	0.09	-1.83	0.071
Scale scores								
Trait Anxiety	-0.82	0.12	-6.83	<0.001	0.11	0.08	1.32	0.189
Depression	-0.87	0.10	-8.68	<0.001	0.06	0.09	0.66	0.511
OCD	-0.43	0.07	-5.73	<0.001	-0.24	0.09	-2.58	0.012
Apathy	-0.36	0.11	-3.40	0.001	0.14	0.09	1.50	0.137
Social Anxiety	-0.32	0.07	-4.39	<0.001	-0.17	0.09	-1.90	0.060
Schizotypy	-0.13	0.06	-2.00	0.049	-0.10	0.07	-1.44	0.152
Alcohol Misuse	-0.13	0.05	-2.67	0.009	-0.12	0.05	-2.37	0.021
Eating Disorders	-0.12	0.06	-2.04	0.045	-0.05	0.08	-0.64	0.522
Impulsivity	-0.08	0.07	-1.03	0.306	0.06	0.06	1.01	0.314

β = standardised beta coefficient, SE = standardised error, t = t-value, p = p-value (unadjusted), AD = Anxious-depression, CIT = Compulsivity and Intrusive Thought, SW = Social Withdrawal, OCD = Obsessive Compulsive Disorder.

7.2 Supplementary material for Chapter 3

7.2.1 Supplementary methods for Chapter 3

Self-report psychiatric questionnaires items. Participants completed the following 49 items taken from six self-report questionnaires that assess a variety of psychiatric symptoms (Wise & Dolan, 2020), including depression (Zung Self-Rating Depression Scale) (Zung, 1965), trait anxiety (State Trait Anxiety Inventory) (Spielberger et al., 1983), impulsivity (Barratt Impulsiveness Scale 11) (Patton et al., 1995), obsessive-compulsive disorder (Obsessive-Compulsive Inventory-Revised) (Foa et al., 2002), eating disorders (Eating Attitudes Test) (Garner et al., 1982), apathy (Apathy Evaluation Scale) (Marin et al., 1991):

Apathy Evaluation Scale (AES)

2. I get things done during the day. Reverse
7. I approach life with intensity. Reverse
17. I have initiative. Reverse
18. I have motivation. Reverse

AES response scores:
Not at all characteristic (1) gets a zero
Slightly characteristic (2)
Somewhat characteristic (3)
Very characteristic (4)

Barratt Impulsiveness Scale (BIS)

1. I plan tasks carefully. Reverse
6. I have "racing" thoughts.
9. I concentrate easily. Reverse
13. I plan for job security. Reverse
14. I say things without thinking.
15. I like to think about complex problems. Reverse
17. I act "on impulse".
20. I am a steady thinker. Reverse
22. I buy things on impulse.
25. I spend or charge more than I earn.
26. I often have extraneous thoughts when thinking.

BIS response scores:
Rarely/never (1)
Occasionally (2)
Often (3)
Almost always/Always (4)

Eating Attitudes Test (EAT)

1. I am terrified about being overweight.
11. I am preoccupied with a desire to be thinner.
12. I think about burning up calories when I exercise.
14. I am preoccupied with the thought of having fat on my body.

EAT response scores:
Always (3)
Usually (2)
Often (1)
Sometimes (0)
Rarely (0)
Never (0)

Obsessive Compulsive Inventory-Revised (OCI)

1. I have saved up so many things that they get in the way.
2. I check things more often than necessary.
4. I feel compelled to count while I am doing things.
6. I find it difficult to control my own thoughts.
7. I collect things I don't need.
9. I get upset if others change the way I have arranged things.
11. I sometimes have to wash or clean myself simply because I feel contaminated.
12. I am upset by unpleasant thoughts that come into my mind against my will.
13. I avoid throwing things away because I am afraid I might need them later.
16. I feel that there are good and bad numbers.
18. I frequently get nasty thoughts and have difficulty in getting rid of them.

OCI response scores:
Not at all (0)
A little (1)
Moderately (2)
A lot (3)
Extremely (4)

Zung Self-Rating Depression Scale (SDS)

11. My mind is as clear as it used to be. Reverse
12. I find it easy to do the things I used to. Reverse
13. I am restless and can't keep still.
14. I feel hopeful about the future. Reverse
16. I find it easy to make decisions. Reverse
17. I feel that I am useful and needed. Reverse
18. My life is pretty full. Reverse
20. I still enjoy the things I used to do. Reverse

SDS response score:

A little of the time (1)
Some of the time (2)
Good part of the time (3)
Most of the time (4)

State Trait Anxiety Inventory, Trait Subscale (STAI)

1. I feel pleasant. Reverse
3. I feel satisfied with myself. Reverse
5. I feel like a failure.
8. I feel that difficulties are piling up so that I cannot overcome them.
9. I worry too much over something that really doesn't matter.
10. I am happy. Reverse
12. I lack self-confidence.
13. I feel secure. Reverse
16. I am content. Reverse
19. I am a steady person. Reverse
20. I get in a state of tension or turmoil as I think over my recent concerns and interests.

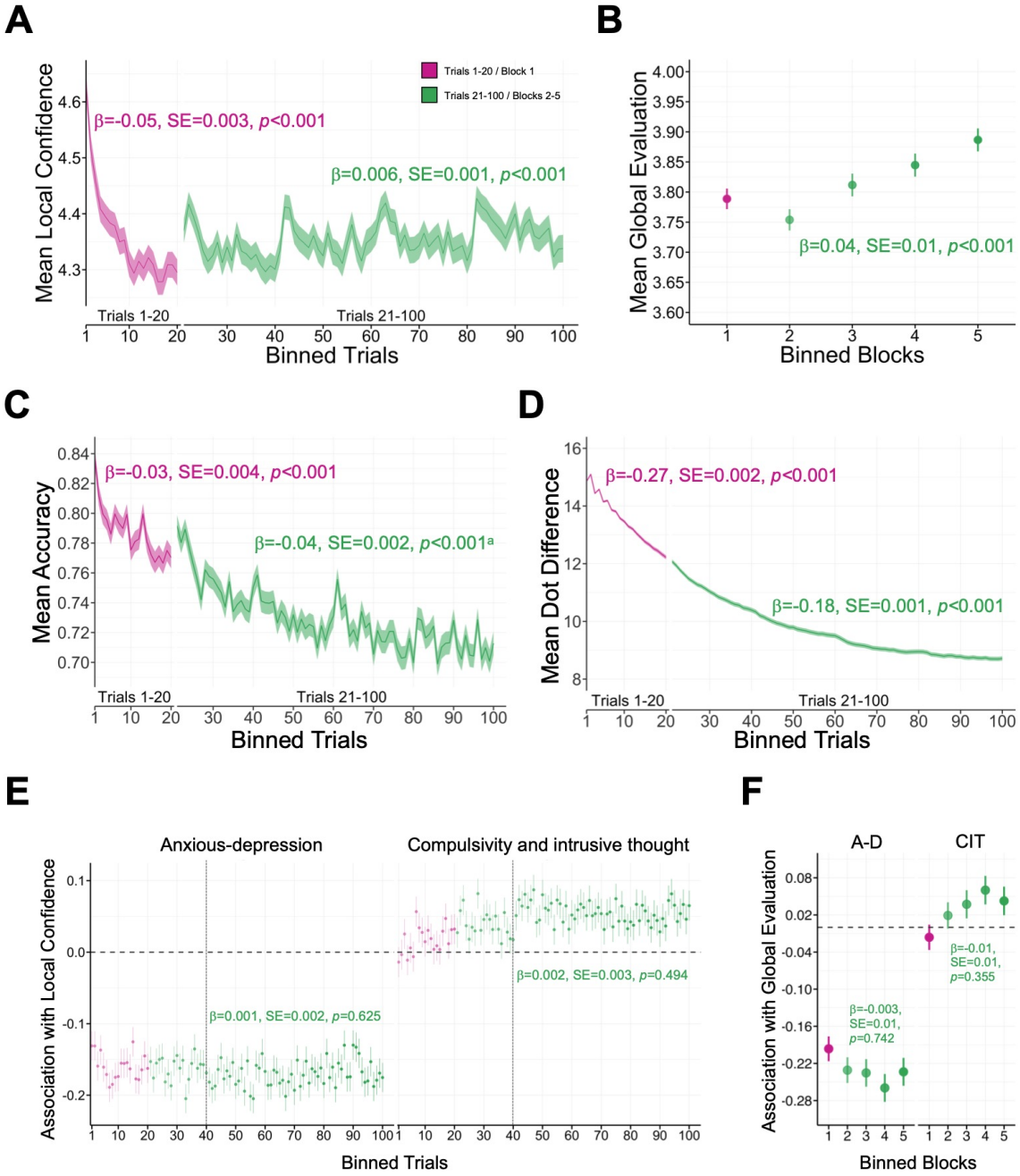
STAI response items:
Almost never (1)
Sometimes (2)
Often (3)
Almost always (4)

7.2.2 Supplementary results for Chapter 3: The impact of increasing trial number of measurements of metacognition with binned trials.

When examining the first 20 burn-in trials, there was a significant decrease in mean confidence for binned trials (β=-0.05, SE=0.003, t=-17.53, p<0.001) (Supplementary Figure S7.2.1A, trials 1-20). For the subsequent 80 main game trials, mean confidence increased for binned trials (β=0.006, SE=0.001, t=3.93, p<0.001) (Supplementary Figure S7.2.1A, trials 21-100). Excluding the burn-in block rating, binned mean global evaluations increased from blocks 2 to 5 (β=0.04, SE=0.01, t=8.48, p<0.001) (Supplementary Figure S7.2.1B). In contrast, accuracy significantly declined during for the binned burn-in trials (β=-0.03, SE=0.004, t=-8.94, p<0.001) and seemed to continue to decline for binned trials in the main game (β=-0.04, SE=0.002, t=-19.77, p<0.001) (Supplementary Figure S7.2.1C), although the model failed to converge due to low variance in main game accuracy values. Mean dot difference (task difficulty level) declined during

the burn-in trials (β=-0.27, SE=0.002, t=-118.9, *p*<0.001) and main game trials (β=-0.18, SE=0.001, t=-173.9, *p*<0.001) (Supplementary Figure S7.2.1D).

Supplementary Figure S7.2.1E shows the association between mean confidence with anxious-depression and compulsivity and intrusive thought across trial numbers, controlling for age, gender, levels of education in the model. The association between local confidence and anxious-depression remained stable from the 40th to the 100th trial, with no significant interaction effect of anxious-depression and trial number on mean confidence (β=0.001, SE=0.002, t=0.49, *p*=0.625) (Supplementary Figure S7.2.1E). Similarly, the positive association between local confidence and compulsivity and intrusive remained stable from the 40th to 100th binned trial, with no significant interaction effect of trial number and compulsivity and intrusive thought on mean confidence (β=-0.002, SE=0.003, t=-0.68, *p*=0.494) (Supplementary Figure S7.2.1E). Therefore, clinical correlates with local confidence were stable across the binned main game play trials. Excluding the burn-in block, there was no significant interaction effect of binned block number with anxious-depression (β=-0.003, SE=0.01, t=-0.33, *p*=0.742) or compulsivity and intrusive thought (β=0.01, SE=0.01, t=0.93, *p*=0.355) on mean global evaluations, indicating clinical correlates with global evaluations were stable across blocks (Supplementary Figure S7.2.1F).

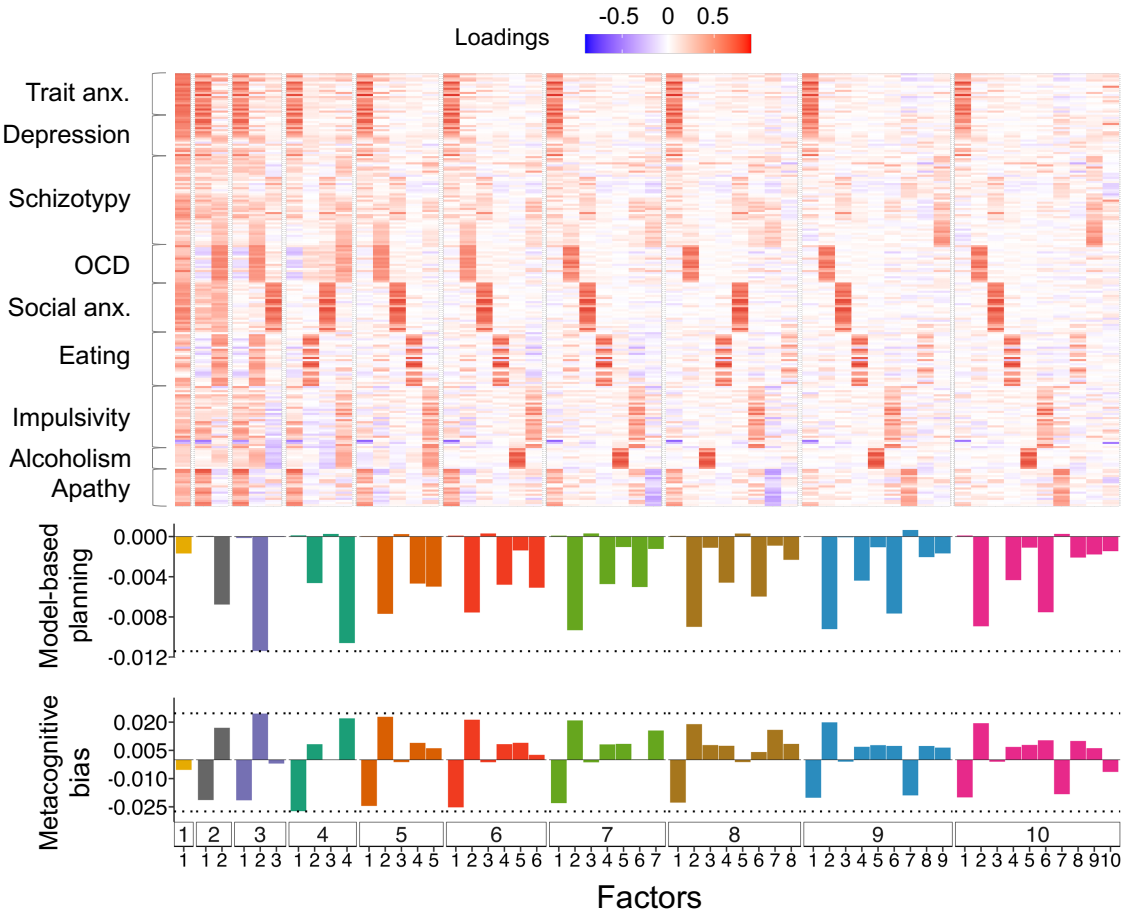


Supplementary Figure S7.2.1. The impact of increasing number of trials with binned trial/block numbers (N=3388). β=standardised beta coefficient, *p*=p-value, [a]=mixed model is singular (did not converge), A-D = Anxious-depression, CIT = Compulsivity and Intrusive Thought. The error bars represent the standard error around the standardised beta coefficient. **(A)** While mean confidence significantly reduced during the burn-in period for the staircase (pink line, trials 1-20), confidence estimates then increased across the remaining 80 main game trials (green line, trials 21-100). **(B)** Following the burn-in block (pink bar, block 1), mean global evaluations increased across blocks (green error bars, blocks 2-5). **(C)** Mean accuracy and **(D)** dot difference (task difficulty) declined for binned burn-in and then continued to decline for main game trials. **(E)** The significant associations between mean local confidence with anxious-depression and compulsivity and intrusive remained stable from the 40th (vertical dotted grey lines) to the 100th trial. **(F)** Following the burn-in block (pink error bar, block 1), there was no significant interaction effects of block number and transdiagnostic dimensions on mean global evaluation, indicating that associations with global evaluations remained stable across binned round-level ratings (green error bars, blocks 2-5).

7.3 Supplementary material for Chapter 4

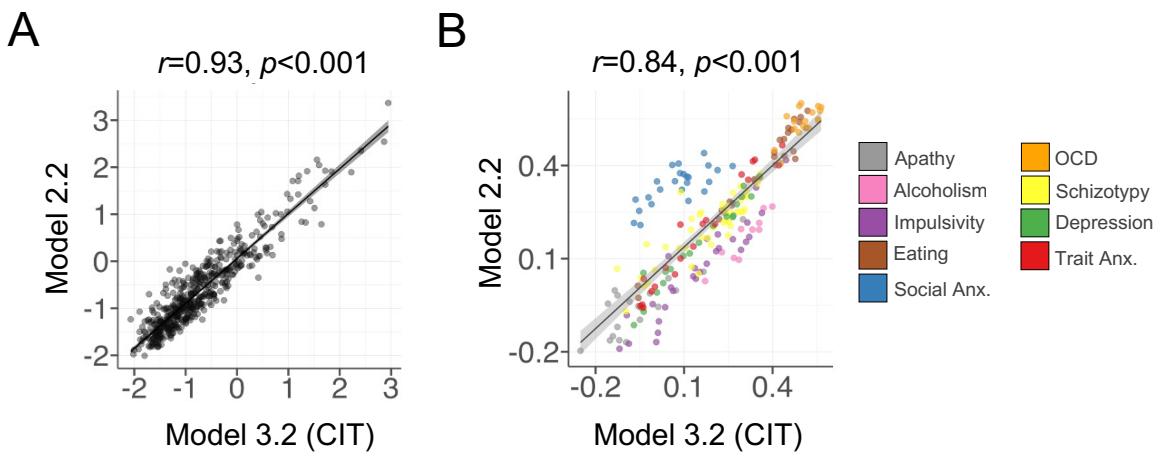
7.3.1 Variation in the number of factors retained

To compare the 55 factors generated from a single to 10-factor solution (Figure 4.3A), we reran regression models per task, including each factor within independent models, along with age, gender and education/IQ. Replicating the findings when factor scores from each solution were included in the same model (Figure 4.3A), no factor outperformed the benchmark of Compulsivity and Intrusive Thought for the association with model-based planning (Supplementary Figure S7.3.1). For associations with metacognitive bias, we found that no factor outperformed the benchmark of Compulsivity and Intrusive Thought in terms of a positive association with metacognitive bias (i.e., higher confidence) (Supplementary Figure S7.3.1), again replicating the top performance of Compulsivity and Intrusive Thought when factors from the same solution were included in the same model (Figure 4.3A). The first factor from a 4-factor solution had a stronger negative association with metacognitive bias than Anxious-Depression (Supplementary Figure S7.3.1), which contrasted the top effect of the first factor in a 2-factor model when all solution factors were included in the regression model (Figure 4.3A). Notably, including the factors from each solution in separate regression models produces top effects that are relatively smaller in magnitude than the effects from models that include all factors within a solution, especially for metacognitive bias (magnitude of effects in Figure 4.3A vs Supplementary Figure S7.3.1). For example, the magnitude of the effect of Compulsivity and Intrusive Thought on confidence bias is nominally larger when factors from the same solution are included in the model (Cohen's f^2= 0.037, Figure 4.3A) compared to when the other factors are not controlled for (Cohen's f^2=0.025, Supplementary Figure S7.3.1). This is also the case for Anxious-depression: Cohen's f^2=-0.030 when other 2 factors from the solution are included in the model (Figure 4.3A), Cohen's f^2=-0.022 when factors are in separate models (Supplementary Figure S7.3.1). This demonstrates the importance of including factors from the same solution in a single regression model, to control for factors with opposing effects on metacognitive bias (e.g., Anxious-Depression vs. Compulsivity and Intrusive Thought).



Supplementary Figure S7.3.1. Re-analyses of variation in the number of factors retained, without controlling for other factors from that model. Fifty-five factors were generated from solutions retaining 1-factor, 2-factors, …, 10-factors. Heatmap indicates the loading of individual items onto each resulting factor (top panel). Weighted effect sizes from 10 linear regression analysis predicting model-based planning, averaged across 5 datasets (N=4990) (middle panel) and metacognition in 4 datasets (N=2575) (bottom panel), when factors are included in separate regression models.

Comparing the 2- and 3-factor solutions, scores on the 2nd factor from this same 2-factor solution were highly correlated with Compulsivity and Intrusive Thought, $r(207)=.93$, $p<0.001$ (Supplementary Figure S7.3.2A), and loadings were correlated at $r(207)=0.84$, $p<0.001$, with the main difference in loadings also being attributed to social anxiety items (Supplementary Figure S7.3.2B), which in a 3-factor solution forms its own factor termed ‘Social Withdrawal’ in prior work (Gillan et al., 2016).



Supplementary Figure S7.3.2. Re-analyses of variation in the number of factors retained, without controlling for other factors from that model.
CIT=Compulsivity and Intrusive Thought. (A) Correlation between scores on the 2nd factor in a 3-factor solution (model 3.2 i.e, Compulsivity and Intrusive Thought) and the 2nd factor from two factor solution (model 2.2) in the discovery dataset. (B) Correlation between factor loadings of Compulsivity and Intrusive Thought and model 2.2 in the discovery dataset.

7.3.2 Variation in the selection of clinical symptoms

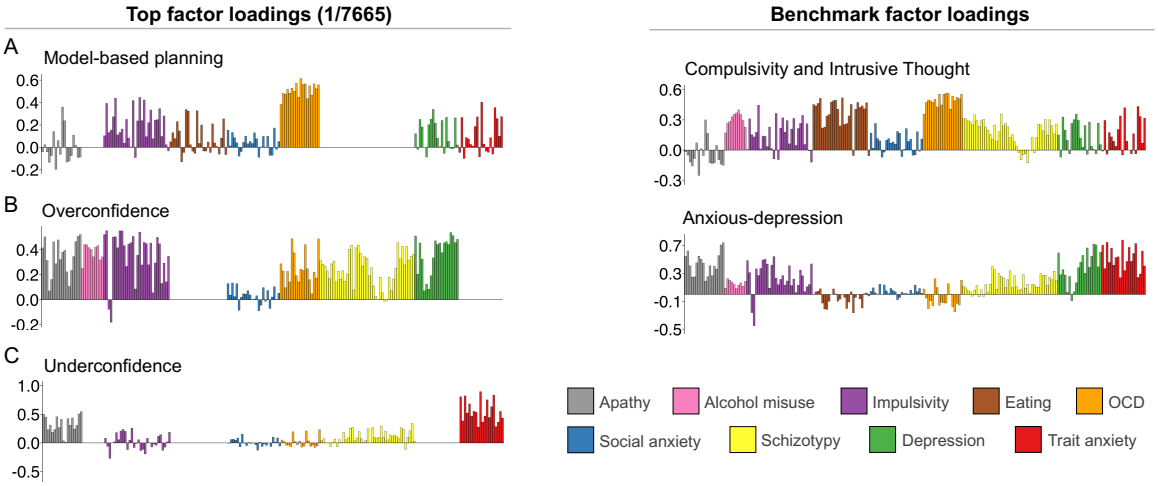
For the association with deficits in model-based planning, the top performing factor and Compulsivity and Intrusive Thought both had highest loadings for OCD items and the lowest loadings for apathy items, with the top factor being generated from a set of questionnaires that omitted the schizotypy and alcohol addiction scales (Supplementary Figure S7.3.3A). Examining loadings for the top performing factor that predicted overconfidence omitted the trait anxiety and eating disorder questionnaires were omitted (Supplementary Figure S7.3.3B). While the top performing factor for associations with underconfidence also omitted the eating disorder questionnaire, this factor had the highest loadings for the trait anxiety items (Supplementary Figure S7.3.3C).

Visualising the item-level loadings across the top 1500 dimensions associated with deficits in model-based planning, items related to obsessive-compulsive disorder appears to have relatively stronger positive loadings across dimensions (Supplementary Figure S7.3.4A). This was confirmed when averaging item-level loadings across the top 100 dimensions (Supplementary Figure S7.3.5A). Items related to obsessive compulsive disorder (M=0.51, SD=0.01), followed by eating disorders (M=0.20, SD=0.07), had the highest average questionnaire-level loadings, while apathy items had the lowest average loading (M=-0.01, SD=0.03)

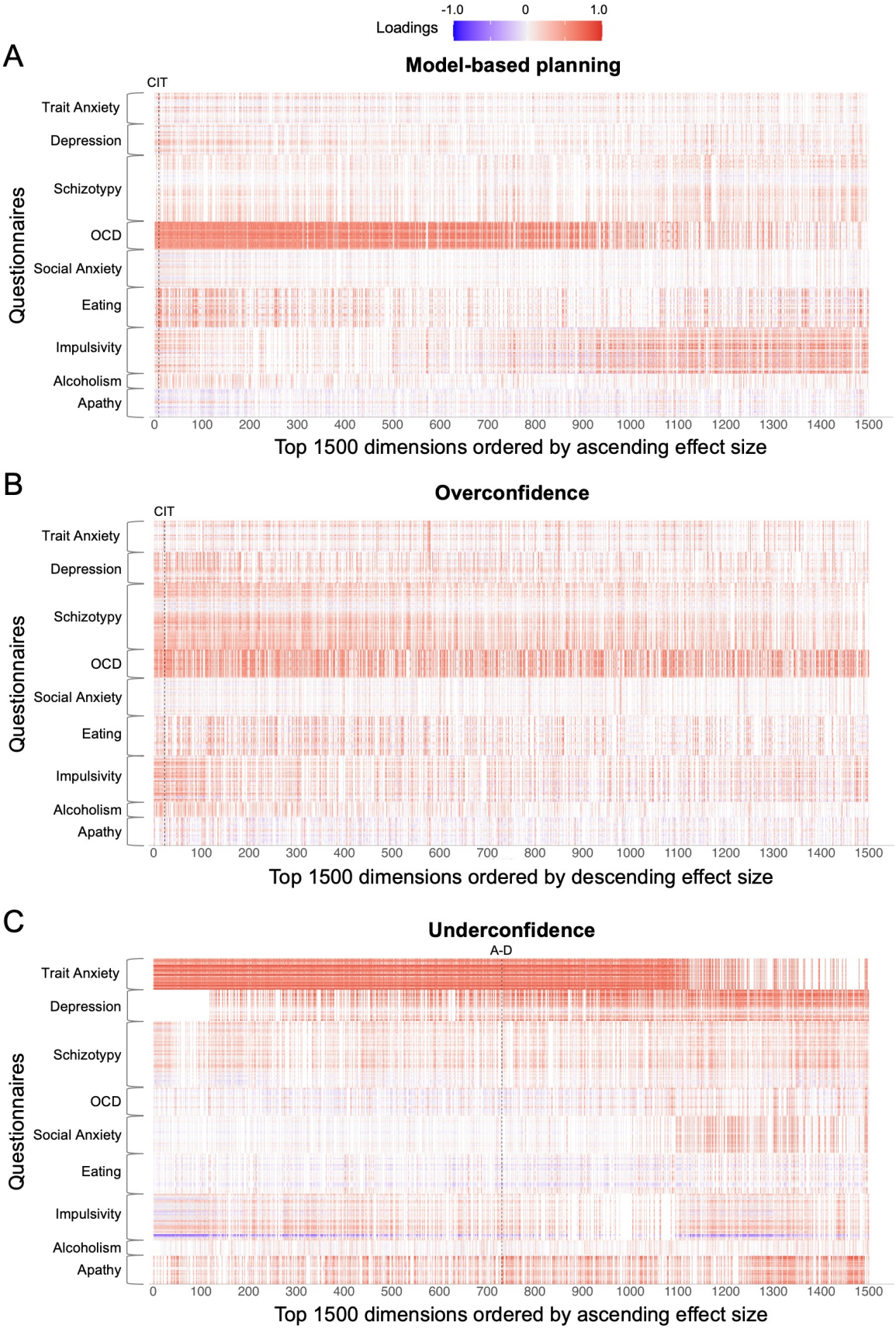
across the top 100 dimensions associated with deficits in model-based planning (Supplementary Figure S7.3.5D). For associations with overconfidence, a heatmap of the top 1500 dimensions illustrated the relative importance of items related to obsessive-compulsive disorder, impulsivity and schizotypy (Supplementary Figure S7.3.4B). Similar to the top dimensions associated with deficits in model-based planning, items from the obsessive compulsive disorder questionnaire contributed most predominantly to the top 100 dimensions (Supplementary Figure S7.3.5B) and had the highest questionnaire-level average loadings (M=0.36, SD=0.02) (Supplementary Figure S7.3.5D). Additionally, items related to impulsivity (M=0.22, SD=0.04) and schizotypy (M=0.20, SD = 0.03) had relatively higher average loadings on the top 100 dimensions that explained individual differences in overconfidence (Supplementary Figure S7.3.5B & S7.3.5D).

When examining negative associations with confidence bias, trait anxiety items had strong positive loadings across the top 1500 performing dimensions (Supplementary Figure S7.3.4C). The relative importance of trait anxiety items was confirmed by averaging loadings across the top 100 dimensions, at the item-level (Supplementary Figure S7.3.5C) and questionnaire-level ((M=0.59, SD=0.02) (Supplementary Figure S7.3.5D). In contrast to trait anxiety, the depression questionnaire was completely omitted from the top 100 dimensions (Supplementary Figures S7.3.4C & S7.3.5C), indicating that removal of the depression questionnaire generates dimensions sensitive to individual differences in underconfidence. The relative importance of trait anxiety was specific to underconfidence, as trait anxiety had lower loadings across the top 100 dimensions predicting model-based planning (M=0.06, SD=0.05) and overconfidence (M=0.06, SD=0.05). While obsessive compulsive disorder items had the highest loadings across the top 100 dimensions for overconfidence (M=0.36, SD=0.02) (Supplementary Figures S7.3.5B and S7.3.5D), obsessive compulsive disorder items did not load relatively highly across dimensions for underconfidence (M=0.02, SD=0.03) (Supplementary Figures S7.3.5C-D). This suggested that over- and under-confidence have differential associations with psychopathology (i.e., overconfidence with OCD vs. underconfidence with trait

anxiety), rather than merely representing effects with opposing directionality on the same clinical phenomena.



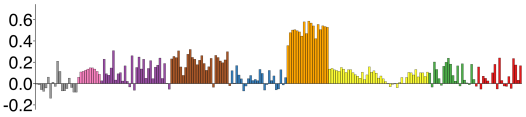
Supplementary Figure S7.3.3. Top (1/7665) performing dimensions from varying the selection of clinical symptoms. OCD = obsessive compulsive disorder. The left side of the panel includes loadings for the 209 questionnaire items across the top performing factors with largest effect sizes for the associations with (A) model-based planning, (B) overconfidence, and (C) underconfidence (left-side of panel). On the right side are loadings for the 209 items on the benchmark dimensions (Compulsivity and Intrusive Thought and Anxious-depression).



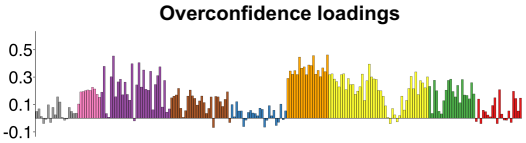
Supplementary Figure S7.3.4. Heatmaps of top 1500/7665 performing dimensions from varying the selection of clinical symptoms. CIT=Compulsivity and Intrusive Thought. A-D=Anxious-depression, OCD= Obsessive compulsive disorder.

Heatmaps for top 1500 dimensions with largest weighted average effect sizes for associations with (A) model-based planning, (B) overconfidence (positive mean confidence), and (C) underconfidence (negative mean confidence).

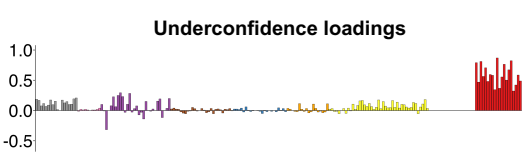
A



B



C



D

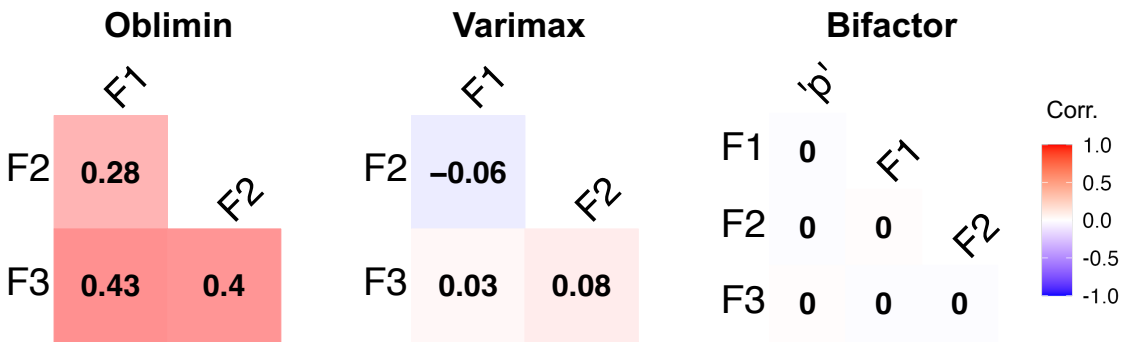
Top 100 dimensions: Questionnaire-level average loadings

Questionnaire	Model-based planning	Over-confidence	Under-confidence
Apathy	-0.02 (0.03)	0.04 (0.03)	0.12 (0.07)
Alcoholism	0.12 (0.03)	0.19 (0.03)	0.01 (0.01)
Impulsivity	0.12 (0.05)	0.22 (0.04)	0.04 (0.02)
Eating	0.20 (0.07)	0.12 (0.05)	0.004 (0.02)
Social Anxiety	0.04 (0.02)	0.03 (0.02)	0.004 (0.02)
OCD	**0.51 (0.01)**	**0.36 (0.02)**	0.02 (0.03)
Schizotypy	0.08 (0.05)	0.20 (0.03)	0.06 (0.04)
Depression	0.10 (0.04)	0.18 (0.06)	
Trait Anxiety	0.06 (0.05)	0.06 (0.06)	**0.59 (0.02)**

Supplementary Figure S7.3.5. Average across top 100/7665 performing dimensions from varying the selection of clinical symptoms. OCD = Obsessive compulsive disorder. Average loadings for the 209 questionnaire items across the top 100 dimensions with largest effect sizes for the associations with (A) model-based planning, (B) overconfidence, and (C) underconfidence. (D) The average questionnaire-level loadings across the top 100 dimensions associated with cognitive outcomes. Questionnaire colours in (D) correspond to item colours in (A-C) (e.g., grey = apathy item-/questionnaire-level loadings).

7.3.3 Higher- versus first-order factor rotation

The within-solution correlations between factor scores verified the oblique and orthogonal nature across rotation types. The oblique rotation 'oblimin' produces correlations between factors 1,2 and 3 ranging from r=0.28 to 0.43, compared to the orthogonal rotation 'varimax', which ranged from r=-.06 to .08 (Supplementary Figure S7.3.6). Applying a bifactor rotation produced factor structures that were internally orthogonal, with correlations between factors 1,2, 3 and the hierarchical general 'G' factor all at r=0 (Supplementary Figure S7.3.6).

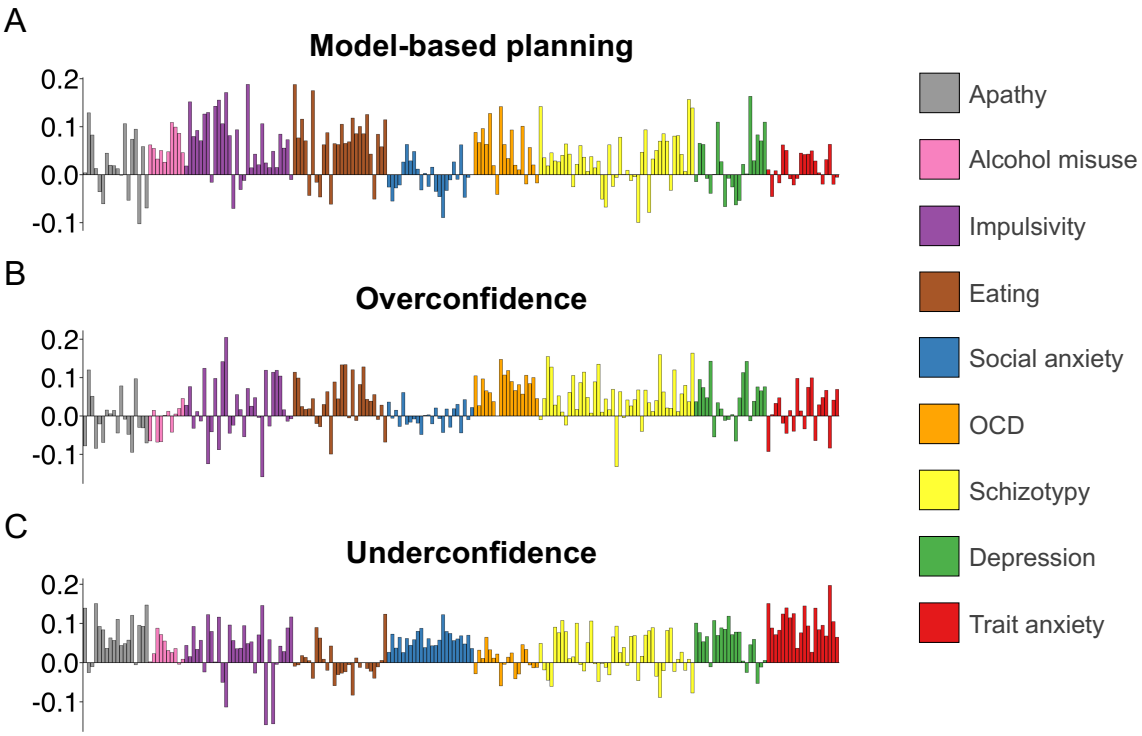


Supplementary Figure S7.3.6. Higher- versus first-order factor rotation: The within-solution correlations between factor scores. F1= factor 1, F2 = factor 2, F3 = factor 3, 'p' = general factor. The within-solution correlations indicated strong positive correlations within the oblimin (oblique) solution, and very weak/no correlation between factors within the orthogonal solutions 'varimax' and the bifactor model.

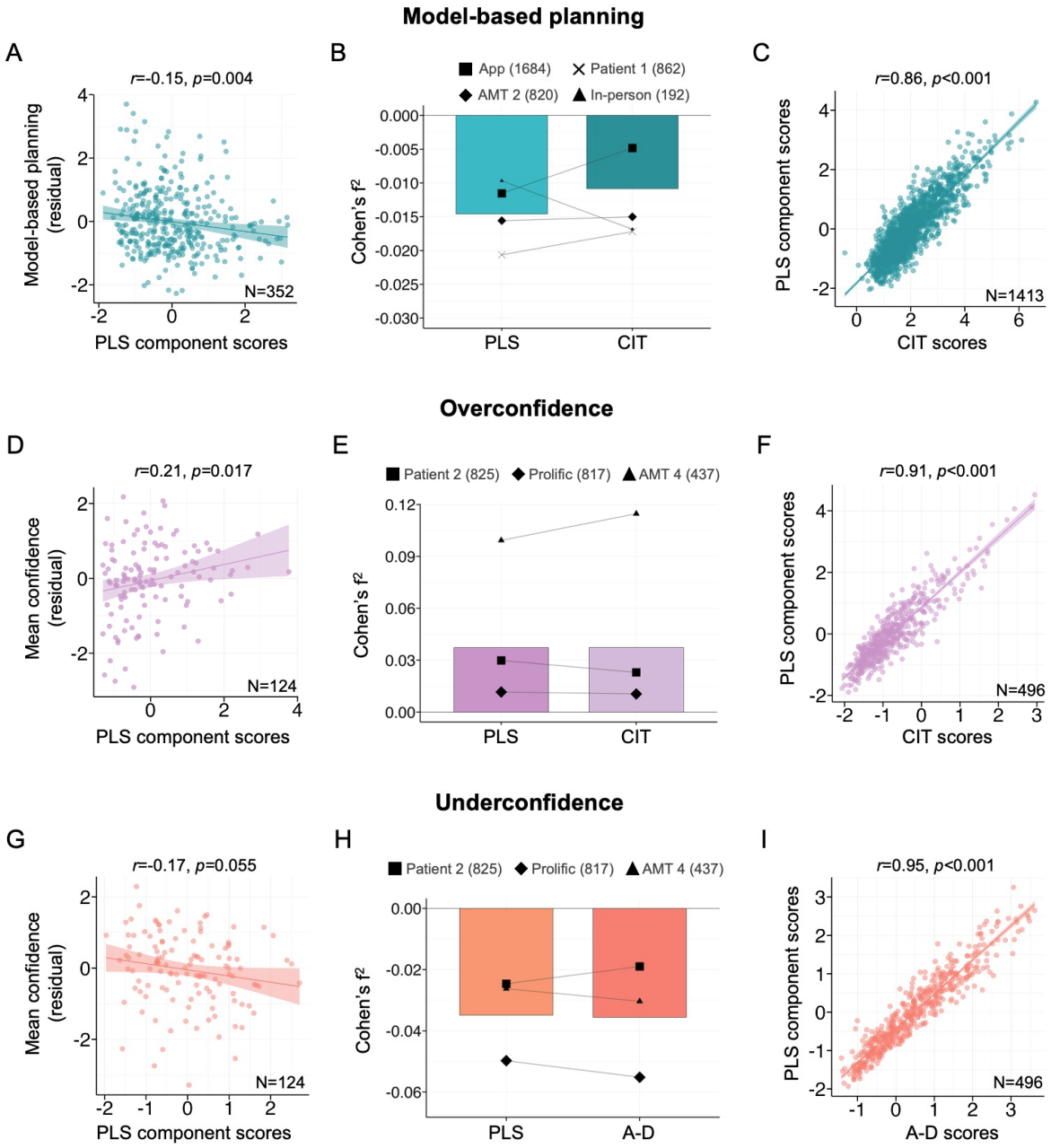
7.3.4 Partial Least Squares Regression

Examining the transdiagnostic components generated from partial least squares regression analyses, item loadings varied across the components used to predict individual differences in model-based planning (Supplementary Figure S7.3.7A), overconfidence (Supplementary Figure S7.3.7B) and underconfidence (Supplementary Figure S7.3.7C).

As detailed in the main text, the results from the partial least squares regression analyses, and comparisons with factor analysis solutions, are illustrated in Supplementary Figure S7.3.8.



Supplementary Figure S7.3.7. Loadings of items across components generated with Partial Least Squares Regression. OCD = Obsessive compulsive disorder. The 209 questionnaire item loadings for the components generated with partial least squares regression to predict individual differences in (A) model-based planning, (B) overconfidence, and (C) underconfidence.



Supplementary Figure S7.3.8. Partial Least Squares Regression. PLS = Partial least squares regression, CIT = Compulsivity and Intrusive Thought, AD = Anxious-depression. Model-based planning. (A) The correlation between PLS component scores and the residual of model-based planning (controlling for age, gender and IQ) in the held out 25% of the data (N=352). (B) Weighted average effect sizes across out-of-sample datasets (bars) and individual dataset effect sizes (dots) for the PLS component and CIT associations with model-based planning. The overall gain in the weighted average effect size for the PLS component was negligible and inconsistent across datasets. (C) Correlation between PLS component scores and CIT scores in the full discovery dataset (N=1413). Overconfidence. (D) The correlation between PLS component scores and the residual of confidence (controlling for age, gender, IQ and A-D) in the held out 25% of the data (N=124). (E) Weighted average effect sizes across out-of-sample datasets (bars) and individual dataset effect sizes (dots) for the PLS component and CIT associations with confidence. The weighted average effect size of the PLS component was not nominally greater than the CIT effect, and the PLS component inconsistently outperformed CIT across datasets. (F) Correlation between PLS component scores and CIT scores in the full discovery dataset (N=496). Underconfidence. (G) The correlation between PLS component scores and the residual of confidence (controlling for age, gender, IQ and CIT) in the held out 25% of the data (N=124). (H) Weighted average effect sizes across out-of-sample datasets (bars)

and individual dataset effect sizes (dots) for the PLS component and AD associations with confidence. The weighted average effect size was higher for AD compared to the PLS component. (I) Correlation between PLS component scores and AD scores in the full discovery dataset (N=496).