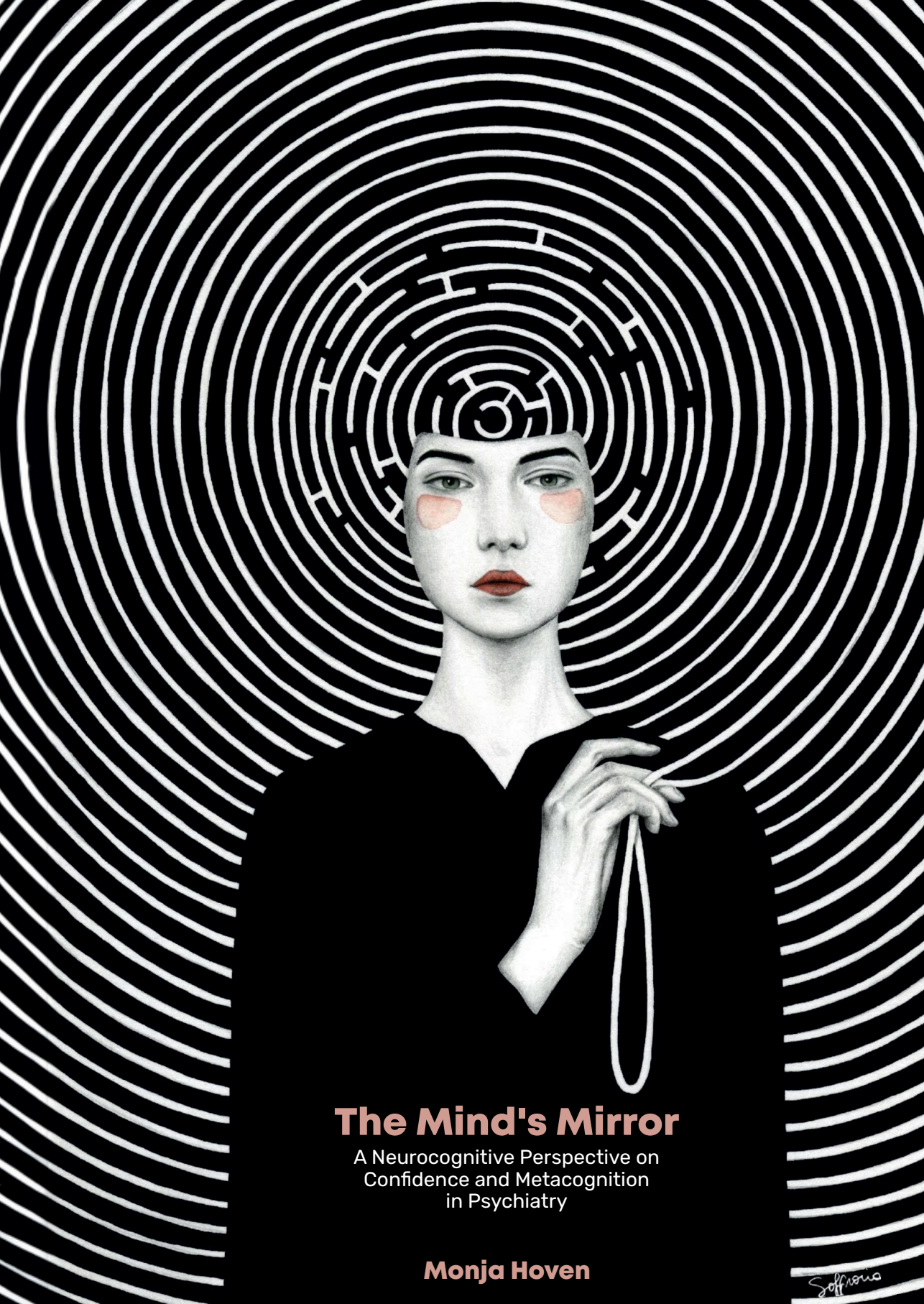
The mind's mirror

A neurocognitive perspective on confidence and metacognition in psychiatry

Hoven, M.

Publication date
2024

Document Version
Final published version



The Mind's Mirror

A Neurocognitive Perspective on Confidence and Metacognition in Psychiatry

Monja Hoven

THE MIND'S MIRROR

A NEUROCOGNITIVE PERSPECTIVE ON CONFIDENCE AND METACOGNITION IN PSYCHIATRY

MONJA HOVEN

Colophon

The research in this thesis was financially supported by grants from the Dutch Research Council (NWO Veni Fellowship (916-18-119)), from Amsterdam Brain and Cognition (ABC, Personal Grants), and an NWO Aspasia Grant (2019/SGW/00764779).

Author Monja Hoven
Cover art Eudoxia by Sofia Bonati © www.sofiabonati.com
Printed by ProefschriftMaken.nl

THE MIND'S MIRROR
A NEUROCOGNITIVE PERSPECTIVE ON CONFIDENCE AND METACOGNITION IN PSYCHIATRY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 8 februari 2024, te 10.00 uur

door Monja Hoven
geboren te Deventer

Promotiecommissie

Promotores

Prof. dr. D.A.J.P. Denys	AMC-UvA
Dr. R.J. van Holst	AMC-UvA

Co-promotor

Dr. J. Luigjes	AMC-UvA

Overige leden

Dr. M. Rouault	CNRS
Prof. dr. A.E. Goudriaan	AMC-UvA
Dr. S de Wit	Universiteit van Amsterdam
Dr. H. Visser	GGz Centraal
Dr. S. van Gaal	Universiteit van Amsterdam
Prof. dr. K.R. Ridderinkhof	Universiteit van Amsterdam
Prof. dr. E.R.A. de Bruijn	Universiteit Leiden

Faculteit der Geneeskunde

Paranimfen

Nora Runia
Laurens van de Mortel

Table of Contents

1

General introduction

Imagine you find yourself on the side of a dark and foggy road. You need to cross to the other side, but visibility is poor. You have to make a decision: is a car approaching in the distance, or not? You look towards the lights, trying to discern whether they are coming from bright headlamps of a car or simply from street lights. In deciding whether to cross the road, you will ultimately rely on your feeling of *confidence* about your decision that the light originates from street lights rather than from a car, making it safe to cross. We cannot help having feelings of confidence about all sorts of decisions we make, and it is important that the confidence you feel about your decision aligns with the actual true state of the world. If you are too overconfident and step into the road when a car is actually approaching, the consequences could be fatal. However, if you keep hesitating when in fact the coast is clear, you will never get to the other side. This process of evaluating and reflecting on one's decisions, judgments, ideas or other mental operations is known at *metacognition*, a remarkable and essential ability in our unpredictable and ever-changing world. In this thesis, I will explore the metacognitive construct of *confidence judgments*, its' neurobiological basis, the biases that it is subject to, and how it is affected by psychiatric symptoms and disorders.

Metacognition and Confidence

Metacognition is defined as *'thoughts about one's own thinking'* (Flavell, 1979), and is a phenomenon that has been written about and studied since times of Ancient Greece (Spearman, 1923). We constantly monitor our own thinking; it is a process that unfolds whatever we do, whether consciously or unconsciously. Metacognition is an umbrella term for many processes that have been widely studied within various research fields. A seminal framework has described two separate but cooperative functions of metacognition: metacognitive monitoring and metacognitive control (Nelson, 1990). The former involves making judgments about your performance (e.g., not being confident that you will remember to water your plants), while the latter involves using that information to adjust or regulate one's cognition and behavior (e.g., setting an alarm to remind you to water your plants).

This thesis focuses on the metacognitive phenomenon of *confidence* from a neurocognitive perspective. Confidence can be defined as the subjective feeling about the probability of being correct about a decision, choice or statement (Pouget et al., 2016). Confidence is a broad concept that exists on many levels of abstraction (Rouault et al., 2019; Seow et al., 2021). It is often assessed retrospectively, in the form of an explicit rating on a scale after a choice has been made. There are different measures to obtain from this rating, such as one's average confidence, over- or under-confidence relative to performance, or the metacognitive sensitivity to correct or incorrect choices,

which are explained in more detail in **Chapter 2**. The majority of research has focused on this form of *local confidence*, i.e., confidence judgments given on trial-by-trial decisions (e.g., "*I am confident that this choice was correct"*). *Global confidence* forms over longer periods of time, integrating more information to form a confidence judgment over one's ability to perform a certain task (e.g., "*I am confident in my performance on this task*"). Higher-order *self-beliefs* are metacognitive beliefs about the self, and span many (personal) domains.

Functions of confidence

Why is confidence an important topic of study and what are its functions? Accurate control over one's behavior requires accurate monitoring of that behavior. Given the self-monitoring nature of confidence judgments, they are well suited to perform a myriad of complex cognitive functions such as adapting, planning and learning in complex and volatile environments. Confidence guides information seeking: if my confidence is low, I will be more likely to gather more information before making a decision (Balsdon et al., 2020; Desender et al., 2018; Pescetelli et al., 2021), which is also visible in neural signatures (Desender et al., 2019). Once a decision has been made, confidence judgments help to re-evaluate these decisions, possibly resulting in changes of mind (Folke et al., 2017; Rollwage et al., 2018; Stone et al., 2022). In doing so, it also guides future decision-making and learning (Cortese, 2022). Confidence serves as a reinforcement signal that promotes learning when no feedback is present (Guggenmos et al., 2016; Rouault et al., 2019), and on the other hand, if I am already highly confident, I am more likely to refrain from additional learning (Nassar et al., 2010). During the learning process, confidence tracks changes in the environment, which helps with flexible adaptation of behavior (Heilbron & Meyniel, 2019) by influencing learning rates (Vinckier et al., 2016) and promoting learning from successes (Cortese et al., 2020), specifically when confidence is low (Lak et al., 2020). In terms of flexible strategy adjustment, confidence plays a role in the ratio between exploration and exploitation, as lower confidence leads to a higher tendency towards exploration (Boldt et al., 2019). Essentially, from an evolutionary standpoint, our survival depends on our ability to accurately monitor the confidence in our actions, as it is essential for guiding optimal behavior.

Studying confidence in psychiatry

This thesis is centered on exploring the overarching question of how confidence is affected by psychiatric symptoms and disorders. To address this, this thesis incorporates different sections dedicated to investigating specific aspects of this

question. Prior to introducing these sections, I will underscore the significance and relevance of studying confidence within the field of psychiatry.

Given the omnipresence of confidence in our daily lives, dysfunction in confidence can have great significance. It is crucial for our behavioral control that your feeling of being correct (i.e., confidence) aligns with your actual performance. Inaccurate confidence judgments could contribute to pathological behavior and decision-making observed in psychiatric disorders. For example, being underconfident in locking the door properly could trigger compulsive checking behavior in patients with obsessive-compulsive disorder (OCD), whilst being too overconfident in incorrect beliefs may go hand in hand with, for example, continued gambling in patients suffering from gambling disorder (GD).

Metacognition, or more generally beliefs about one's abilities, have been a promising target for treatment. Several forms of metacognitive interventions have been developed in recent years that have shown to be effective in alleviating symptoms of different psychiatric disorders (Philipp et al., 2019). This thesis aids in getting more insight into deficits of confidence in psychiatry across various symptoms, contexts and cognitive domains, which hopefully will benefit the further development of these therapies. In this sense, quantification of the confidence disturbances has clinical value and is critical for refining the tools that could improve metacognitive ability in patients and at-risk populations.

Part I: Confidence and its Biases in Psychiatry

In the first part of this thesis, I will give an overview of confidence in psychiatric disorders and symptoms in healthy, subclinical and clinical samples, with a special focus on the behavioral and neurobiological mechanisms of confidence biases in two compulsive disorders: obsessive-compulsive disorder (OCD) and gambling disorder (GD).

Compulsivity and confidence in OCD and GD

Compulsive behavior is defined as "*repetitive acts that are characterized by the feeling that one 'has to' perform them while being aware that these acts are not in line with one's overall goal.*" (Luigjes et al., 2019), and is a hallmark of both OCD and addiction (Figee et al., 2016). Patients with OCD suffer from intrusive obsessions that cause distress and anxiety. In an attempt to reduce distress, patients perform compulsions: ritualistic compulsive behaviors that impair functioning in social and work environment (American Psychiatric Association, 2013). The estimated prevalence of OCD is 2 to 3%

(Ruscio et al., 2010; Stein et al., 2019), with a tenfold higher prevalence of sub-clinical symptoms (Fullana et al., 2009), indicating a substantial contribution to global burden of disease (Baxter et al., 2014).

GD is defined as "*persistent and recurrent problematic gambling behavior leading to clinically significant impairment or distress*" (American Psychiatric Association, 2013). Prevalence estimates of problem gambling are up to 5.8% worldwide and 3.4% in Europe, with a three-to-four fold higher prevalence of subclinical gambling problems and related harm (Calado & Griffiths, 2016), which has been shown to be comparable to that of alcohol dependence (Abbott, 2020).

Even though patients suffering from OCD or GD share symptoms of compulsivity, they also often show very distinct behavior, with evidence for risk avoidance and loss sensitivity in OCD (Shephard et al., 2021), but for risk seeking and reward sensitivity in GD (Clark et al., 2019). Dysfunctions in confidence could help explain these distinct profiles of compulsive behaviors. Clinical representations of these disorders suggest that the confidence patients have would lie on opposites sides of the spectrum, with lower confidence in OCD and higher confidence in GD. Lower confidence in one's actions in OCD could go together with compulsive checking behavior, whereas higher confidence in one's actions in GD could be accompanied by increased risk-taking behavior and gambling. In this way, dysfunctions of confidence could lie at the heart of compulsive behavior and makes it especially relevant to collectively study confidence and the biasing effect of incentives on confidence in these two compulsive disorders.

While the study of confidence in psychiatry had been actively pursued for many years when I started my PhD project in 2019, no topical overview on abnormalities in confidence in psychiatry existed. As a starting point, in **Chapter 2** we reviewed studies investigating confidence in psychiatry, both in clinical patient samples and subclinical samples that focused on psychiatric symptoms in either prodromal phases of disorder or the general population. We focused on OCD, schizophrenia, addiction, anxiety disorders and depression. This chapter provided us with a framework for our own future empirical investigations of confidence in psychiatry.

The neurobiology of confidence

Beyond behavioral and psychological assessment, many years of research have been devoted to studying the neurobiological basis of confidence. There is a strong consensus for a key role of the prefrontal cortex (PFC), as activity in both medial and lateral regions of the PFC, such as the dorsolateral PFC, rostrolateral PFC and ventromedial PFC (vmPFC) is modulated by confidence judgments (Fleming & Dolan, 2012). Next to the PFC, confidence judgments elicit activity in a wider spread network,

among which the dorsal anterior cingulate cortex (dACC), ventral striatum (VS), precuneus, insula and parietal cortex (Rouault, McWilliams, et al., 2018; Vaccaro & Fleming, 2018).

The vmPFC and VS are also involved in valuation processes such as reward anticipation and the encoding of desirability and expected values (Bartra et al., 2013; Lebreton et al., 2009). Behaviorally, it had been shown that monetary incentives bias confidence judgments (Lebreton et al., 2018). Building on this work, in **Chapter 3**, we describe a functional magnetic resonance imaging (fMRI) study in a sample of healthy participants investigating the neural basis of this incentive confidence bias, where we aimed investigate whether motivational signals could disrupt metacognitive signals in the brain.

In **Chapter 4** we further aimed to investigate the behavioral and neurobiological underpinnings of the incentive bias on confidence in OCD and GD using fMRI. We hypothesized that GD patients would show exaggerated overconfidence when a gain was at stake, while OCD patients would show exaggerated underconfidence when a loss was at stake.

Transdiagnostic approach

The diagnostic and statistical manual of mental disorders (DSM) is a standard classification that is most often used in clinical practice for diagnosing patients with psychiatric symptoms (American Psychiatric Association, 2013). Historically, most neurocognitive research in psychiatry has been performed using case-control studies, where clinical patient samples are compared to samples of healthy control participants on some neurocognitive phenomenon. There are, however, some concerns using this type of design. It is well-known that there is much heterogeneity of symptoms within a disorder, and much overlap of symptoms between disorders, together with high comorbidity (Insel et al., 2010). Therefore, a call has been made for a transdiagnostic approach to psychiatry research, in which neurocognitive processes are instead related to disorder-transcending continuous psychiatric symptoms in large general population samples rather than to categorical group adherence (patients versus healthy controls) in smaller clinical samples. Using this approach, studies have shown that a transdiagnostic symptom dimension of anxiety and depression related to decreased confidence, while a symptom dimension of compulsivity and intrusive thoughts instead related to increased confidence (Rouault, Seow, et al., 2018).

In **Chapter 5**, we have applied a transdiagnostic approach to study how various levels of confidence (i.e., local confidence, global confidence and self-beliefs) are related, and how they relate to psychiatric symptoms in a large general population sample.

PART II: Confidence in OCD

The second part of the thesis will focus on our investigations of confidence in OCD in various different contexts and in relation to different cognitive processes. First, we aimed to answer the open question whether the lower *local* confidence that is repeatedly found in patients with OCD compared to healthy participants would generalize to decreases in more *global* forms of confidence in **Chapter 6**. Moreover, psychiatry research in large general population samples is getting more popular and has especially gained momentum since the COVID-19 pandemic, using so-called *analog studies*. These studies, in OCD, are based on the assumption that highly compulsive individuals from the general population resemble clinical OCD patients in terms of symptoms and, importantly, in terms of the (meta)cognitive process under investigation (Abramowitz et al., 2014). Yet, hardly any studies test this assumption by directly comparing these groups, while contradicting metacognitive patterns have been found in these samples, indicating decreased confidence in patients with OCD, but increased confidence in highly compulsive individuals from the general population. This raised the important question whether findings regarding dysfunction of confidence can be generalized from general population samples to clinical samples, which we addressed in **Chapter 6**.

In **Chapter 7** we investigated the role of confidence in learning in a volatile environment using a predictive inference confidence task. Usually, when someone is very confident about their choices they are less likely to change their strategy and less open to learn from new evidence. It was, however, not clear if this type of metacognitive control was functioning well in patients with OCD, since previous research (both in case-control and transdiagnostic studies) using this same task had found mixing results (Marzuki et al., 2022; Seow & Gillan, 2020; Vaghi et al., 2017). Again, we compared our sample of OCD patients to both a healthy and highly compulsive sample from the general population.

PART III: Confidence in GD

The final part of this thesis is dedicated to investigating dysfunction of confidence in GD relating to gambling relevant contexts. First, in **Chapter 8**, using the same task as was used in **Chapter 7**, we investigated the role of confidence in learning in a volatile environment in GD.

Next to metacognitive processes, attentional processes also play an active role in decision-making (Orquin & Mueller Loose, 2013), and attentional biases towards

gambling cues have been described in GD (Anselme & Robinson, 2020; Brevers et al., 2011; Ciccarelli et al., 2016; Hønsi et al., 2013). In **Chapter 9** we took a more direct approach to studying the role of attention in risky decision-making and confidence by employing eye-tracking in a sample of GD patients and healthy controls during a mixed-gamble task.

As an extension to this work, in **Chapter 10**, we focused on studying the effects of making risky versus safe decisions on confidence judgements in GD, with the hypothesis that patients with GD would be more confident while making risky versus safe choices, especially when the gains at stake were high, a dysfunction that could lie at the heart of their compulsive gambling behavior.

Outline of this thesis

This thesis investigates the overarching topic of confidence dysfunctions in psychiatry, with a special focus on compulsive disorders OCD and GD, across different contexts using behavioral and neuroimaging methods.

In part I of this thesis, we start with an inquiry into the dysfunctions of confidence that are related to various clinical psychiatric disorders and subclinical psychiatric symptoms in **Chapter 2**, where we review the literature in terms of established confidence measures. In **Chapter 3** we used fMRI to investigate the behavioral and neurobiological biasing effects of incentives on confidence in a sample of healthy participants. We extended this to compulsive disorders in **Chapter 4** to investigate the incentive bias on confidence in patients with OCD and patients with GD. In **Chapter 5** we took a transdiagnostic approach to investigate the interrelations between confidence levels, as well as their relationship to (transdiagnostic) psychiatric symptoms in a large general population sample using a computational approach.

In part II of this thesis we set forth to explore dysfunctions in confidence in various contexts in OCD. In **Chapter 6** we explored confidence at local and global levels and their interplay in patients with OCD compared to both healthy and highly compulsive subjects from the general population. In **Chapter 7** the link between learning, action and confidence was investigated in OCD patients compared to both healthy and highly compulsive subjects, using computational modelling.

In part III of this thesis we shifted our focus towards confidence in GD. In **Chapter 8** we studied the relationships between learning, action and confidence in GD patients compared to a healthy control group. In **Chapter 9** we used an eye-tracking experiment to assess the relationship between attention, risky decision-making and confidence in

GD compared with healthy controls. This was followed up by **Chapter 10** which focused on the influence of making risky choices on the feeling of confidence. Finally, **Chapter 11** provides a summary and general discussion of the findings presented in this thesis.

Part I

Confidence and its Biases in Psychiatry

2

Abnormalities of confidence in psychiatry: an overview and future perspectives

Hoven M

Lebreton M

Engelmann JB

Denys D

Luigjes J*

van Holst RJ*

* shared last authorship

Translational Psychiatry, 2019, 9(1), 1-18

Abstract

Our behavior is constantly accompanied by a sense of confidence and its' precision is critical for adequate adaptation and survival. Importantly, abnormal confidence judgments that do not reflect reality may play a crucial role in pathological decision-making typically seen in psychiatric disorders. In this review, we propose abnormalities of confidence as a new model of interpreting psychiatric symptoms. We hypothesize a dysfunction of confidence at the root of psychiatric symptoms either expressed subclinically in the general population or clinically in the patient population.

Our review reveals a robust association between confidence abnormalities and psychiatric symptomatology. Confidence abnormalities are present in subclinical/prodromal phases of psychiatric disorders, show a positive relationship with symptom severity, and appear to normalize after recovery. In the reviewed literature, the strongest evidence was found for a decline in confidence in (sub)clinical OCD, and for a decrease in confidence discrimination in (sub)clinical schizophrenia. We found suggestive evidence for increased/decreased confidence in addiction and depression/anxiety, respectively.

Confidence abnormalities may help to understand underlying psychopathological substrates across disorders, and should thus be considered transdiagnostically. This review provides clear evidence for confidence abnormalities in different psychiatric disorders, identifies current knowledge gaps and supplies suggestions for future avenues. As such, it may guide future translational research into the underlying processes governing these abnormalities, as well as future interventions to restore them.

Introduction

Metacognition refers to our ability to think about, reflect, and comment upon our own thinking. Confidence judgment is one such metacognitive operation, and is described as the subjective feeling of being correct about a choice, decision or statement (Pouget et al., 2016). Not only is this feeling of confidence critical to re-evaluate previous decisions, it can also guide future decision-making and drive reasoning and social interactions (Fleming, Dolan, et al., 2012). Producing accurate confidence judgments is an individual ability, which seems stable across different sensory modalities (Ais et al., 2016; Faivre et al., 2018; Rahnev et al., 2015; Song et al., 2011), time-points (Fleming et al., 2016), and across cognitive domains (Rouault, McWilliams, et al., 2018) (but see (Kelemen et al., 2000; Morales et al., 2018)).

The hypothesis that inaccurate confidence judgments can lead to detrimental decision-making - bearing extensive negative consequences for society and the individual - is supported by both theoretical and experimental consensus (Berner & Graber, 2008; Broihanne et al., 2014; Croskerry & Norman, 2008). Systematically inaccurate confidence judgments could contribute to persistent pathological decision-making observed in psychiatric disorders. For example, underconfidence in memory may result in compulsory checking behavior as observed in patients suffering from obsessive-compulsive disorder (OCD). On the other hand, overconfidence in erroneous beliefs could underpin delusional thinking as observed in schizophrenia patients. Yet, to date an overview of abnormalities in confidence judgments across psychiatric disorders is lacking.

Here, we review studies of confidence in subclinical and clinical psychiatric populations to apprehend the associations between confidence abnormalities and psychiatric disorders. Our review focuses on OCD, schizophrenia, addiction, anxiety, and depression, and includes studies in both subclinical and clinical populations. This is because psychiatric disorders have been proposed to be characterized by both qualitative and quantitative shifts in behavior (Wright, 2011), which can be represented by the visible part of a continuum of symptom severity, the lower end of which would be subclinical (Hankin et al., 2005; Krueger et al., 2005; Lincoln, 2007; Stip et al., 2009). Finally, we discuss the benefits of transdiagnostic approaches to investigate confidence and psychiatric symptoms in the general population. Insight into confidence abnormalities could reveal new targets for early interventions. Overall, this review provides a comprehensive framework for the investigation of confidence in psychiatry. It also highlights the methodological challenges and limitations present in this line of research, and delineates suggestions for future avenues of research.

Targeting confidence abnormalities in psychiatry could help alleviate symptoms and improve treatment outcomes.

Methods

Two separate systematic literature searches for subclinical and clinical populations were conducted through the electronic database PubMed in October 2018, using the following key terms:

(1) ("confiden*" OR "metacogniti*" OR "meta-cogniti*") AND ("psychiatr*" OR "impulsiv"* OR "complusiv*" OR "transdiagnostic**" OR "trans-diagnostic*" OR "individual differences" OR "symptom*" OR "healthy"). (862 hits)

(2) ("confiden*" OR "metacogniti*" OR "meta-cogniti*") AND ("depressi*" OR "schizophr*" OR "obsessive compulsive*" OR "OCD" OR "obsessive-compulsive" OR "addict*" OR "substance*" OR "psychiatr*" OR "eating" OR "MDD" OR "gambl*" OR "anxiety*"). (811 hits)

The search was not limited regarding year of publication. We chose not to include autism spectrum disorder (ASD) and attention-deficit hyperactivity disorder (ADHD) for reasons of clarity. Exclusion criteria were non-English manuscripts; studies using questionnaires to assess confidence, and clinical trials assessing effectiveness of metacognitive therapy. All duplicates were removed, abstracts were screened and full texts of relevant studies were reviewed. From the reference lists of selected papers, additional studies and relevant reviews or meta-analyses were included.

Results

We identified 83 studies that met inclusion criteria. Table 1 shows an overview of the task domains, the metacognitive measures and the most commonly used paradigms in these studies. Briefly, three types of confidence measures are often evaluated. Retrospective confidence judgements assess the correctness of a choice (Pannu & Kaszniak, 2005). Feeling of Knowing (FOK) and Judgments of Learning (JOL) are prospective confidence judgments about one's ability to later retrieve knowledge about a specific subject (FOK) or about a learned cue or cue association (JOL). However, retrospective- and prospective judgments are considered to be different (Fleming et al., 2016; Siedlecka et al., 2016), since they rely on distinct cognitive resources and are influenced by separate parameters, and should therefore not be used interchangeably.

In the current review we mostly focus on retrospective judgments, but for the sake of completeness we also include studies using prospective judgements. Confidence accuracy measures can be derived from comparing retrospective confidence judgements to objective task performance (Figure 1). Confidence judgments are deemed more accurate when correct choices are held with higher confidence than incorrect choices (discrimination), and when average confidence matches average performance (calibration). Yet, confidence measures can be confounded by changes in first-order performance (Figure 2). Therefore, recently bias free measures of confidence have been developed that rest on the foundations of signal detection theory (i.e. metacognitive sensitivity, or meta-d') (Fleming, 2017; Fleming & Lau, 2014; Maniscalco & Lau, 2012), which measures the ability to discriminate between correct and incorrect choices with confidence judgments while controlling for confounds. Moreover, metacognitive efficiency, or meta-d'/d', measures how efficiently perceptual information is used to form a metacognitive report. For further details on confidence accuracy metrics, see Figure 1.

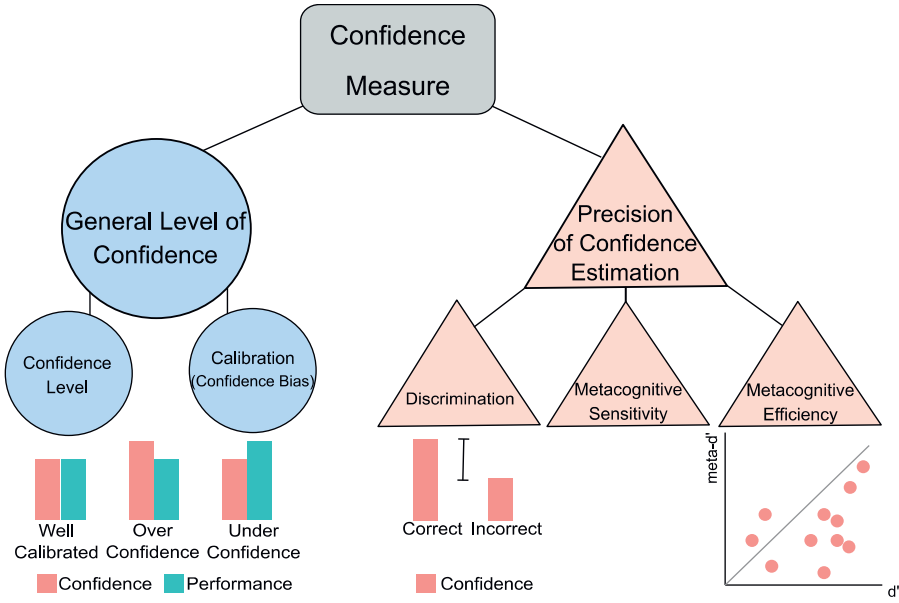


Figure 1: Measures of confidence. Confidence measures can be divided into *general* measures of confidence level and *precision* measures of confidence estimation. To assess someone's general level of confidence, *confidence level* or *calibration* can be analyzed. Calibration (or confidence bias) is usually calculated as the difference between mean task performance and confidence. This results in overconfidence when confidence levels are higher than performance levels, and underconfidence vice versa. To assess someone's precision of confidence estimation, *confidence discrimination*, *metacognitive sensitivity* or *metacognitive efficiency* can be analyzed. Confidence discrimination refers to the difference in confidence levels between correct and incorrect choices. The larger this difference, the higher the discriminatory accuracy of confidence, signaling an increased ability to recognize accurate from inaccurate performance by using one's metacognitive report. Confidence discrimination is sometimes referred to as 'the confidence gap'. Confidence bias and discrimination are two independent aspects of metacognition: an individual might be underconfident, but still be highly sensitive to discriminate between accurate and inaccurate performance with their confidence. Similar to discrimination, metacognitive sensitivity, also referred to as parameter *meta-d'*, aims to measure the ability of a metacognitive observer to discriminate between correct and incorrect trials with their confidence judgments. Yet, it uses a more sophisticated calculation that is bias free, and controls for performance confounds. On the other hand, metacognitive efficiency, referred to as meta-*d'*/*d'*, indicates how well perceptual information (*d'*) is used to form a metacognitive report (*meta-d'*). When meta-*d'*/*d'*, or the *M-ratio*, equals 1 (i.e., indicated by the line in the graph), this signals a metacognitively ideal observer that uses all perceptual information captured in *d'* for the formation of a metacognitive report. When meta-*d'*/*d'* < 1, not all information was used to form a metacognitive report, corresponding to lower metacognitive efficiency. When meta-*d'*/*d'* >1, the observer retrieved additional information to form a metacognitive report, corresponding to higher metacognitive efficiency.

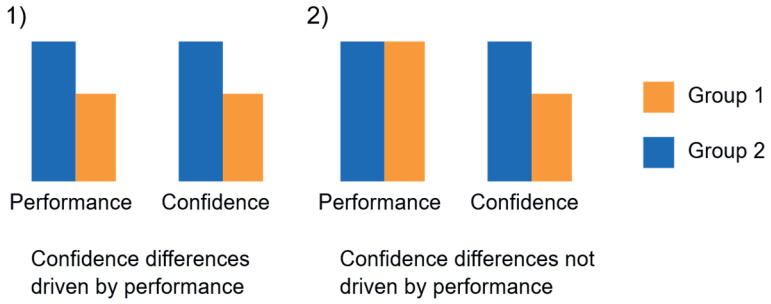


Figure 2: Confidence differences confounded by intergroup differences in first-order performance. 1) The difference in first-order performance between groups might result in untrue differences of confidence between groups. 2) First-order performance is equal between groups and therefore specific effects of group identity on confidence are isolated. This figure illustrates the need for bias free measures such as meta-d' and metacognitive efficiency, which control for performance differences between groups.

Table 1: Most commonly studied cognitive domains, paradigms and measures

Domain	Paradigm	Metacognitive measure	Description of paradigm
Memory	Repeated Checking Task	Confidence level (N-BF)	Participants manipulate different objects (e.g., light switches) and rate their memory confidence. The effects of repeated checking on memory confidence are assessed.
	Repeated Cleaning Task	Confidence level (N-BF)	Participants clean different objects and rate their memory confidence in cleaning those objects. The effects of repeated cleaning on memory confidence are assessed.
	Verbal Memory Task	Confidence level & FOK/JOL measures (N-BF) *	Participants memorize words and after a time interval perform a recall or recognition and rate their memory confidence.
	Visual Memory Task	Confidence level & FOK/JOL measures (N-BF) *	Participants memorize visual stimuli and after a time interval perform a recall or recognition and rate their memory confidence.
	False-Memory Task	Confidence level, confidence in errors & discrimination (N-BF)	Most studies made use of the Deese-Roediger-McDermott (DRM) paradigm. Word lists are presented and after a time interval a recognition test with old and new words (i.e., lure words) is administered and memory confidence is asked.
	Source-Monitoring Task	Confidence level, confidence in errors & discrimination (N-BF)	A wordlist is presented and participants create semantic associations for each word. Afterwards, participants recognize original (old) and self-created (new) words, their source (i.e., experimenter or self) and rate their memory confidence.
Perception	Perceptual Decision Making Task	Confidence level (N-BF), metacognitive sensitivity (i.e., meta-d') and efficiency (i.e., meta-d'/d') (BF)	Participants make a two-alternative decision about perceptual stimuli (i.e., which box contains most dots) and rate their confidence in each decision.

General Knowledge	General Knowledge Task	Confidence level (N-BF)	Participants answer general knowledge questions and rate their level of confidence.
Action	Muscle Tension Task	Confidence level (N-BF)	Participants produce certain levels of muscle tension and rate their confidence about their subjective muscle tension estimates.
Other	Predictive Inference Task	Confidence level (N-BF)	Participants predict the position of a certain particle and state their confidence in their prediction, while the environment is changing over time.
	Wisconsin Card Sorting Task	Confidence level (N-BF)	Participants figure out a sorting rule and rate their confidence in this rule. The sorting rule changes over time and the participants have to relearn the rule.
	Emotion Task	Confidence level (N-BF)	Participants recognize facial emotions and state their confidence.

Most tasks involve retrospective confidence judgements after every decision or action. FOK: feeling of knowing, JOL: judgement of learning, N-BF: non bias free, BF: bias free. *: Task paradigm that uses both prospective and retrospective confidence judgments.

OCD

OCD is a psychiatric condition associated with repetitive and functionally impairing actions (i.e. compulsions, such as checking behaviors), mostly performed to alleviate distress induced by intrusive thoughts (i.e. obsessions) (Figee et al., 2016; Fineberg et al., 2014).

Subclinical: obsessive-compulsive tendencies and compulsivity

Individuals can express compulsivity or obsessive-compulsive tendencies at varying levels of severity without receiving a diagnosis for OCD. Thirteen studies assessing the link between confidence and subclinical OCD symptoms were identified (Table 2A). Two studies found lowered confidence associated with high obsessive-compulsive (OC) tendencies using a false bio-feedback task in which participants evaluated their muscle tension (Lazarov et al., 2012; Zhang et al., 2017). High OC individuals showed more reliance on false feedback and lower confidence in evaluating their muscle tension while the influence of feedback on muscle tension was similar between high and low OC groups. Other studies have not found direct differences in confidence ratings or calibration between individuals with high and low OC tendencies (Ben Shachar et al., 2013; Hauser, Allen, et al., 2017; Rouault, Seow, et al., 2018), but a

subset of these studies has identified other metacognitive effects. Hauser et al., (2017) used a motion detection task and found lower metacognitive efficiency (meta-d'/d') in highly compulsive participants, suggesting that high OC subjects do not utilize all accessible information to form a metacognitive report. Ben Shachar et al., (2013) did not find any differences between high and low OC groups in any confidence measure they used (i.e., confidence level, calibration and discrimination) in a general knowledge task. However, they report that high OC participants were more reluctant to report their answers implicating that they required a higher level of confidence to act on their answer.

Another way of investigating the relationship between confidence and OCD features (such as repetitive checking, cleaning or doubt) is by testing the effect of manipulating confidence on OCD features or vice versa. In particular, this has been done for confidence in memory (i.e. 'metamemory'). Van Den Hout & Kindt (2003) were the first to show that OCD-like checking behavior leads to a decline in memory confidence levels in OCD-relevant scenarios (e.g., involving cleaning or checking), while memory performance was unaffected. Multiple studies have replicated these findings since, both for real life scenarios and mental checks (Ashbaugh & Radomsky, 2007; Coles et al., 2006; Radomsky et al., 2006; Radomsky & Alcolado, 2010; Van Den Hout & Kindt, 2003b). Following the same hypothesis, another study using a repeated cleaning procedure found that memory confidence significantly increases over time for control items, yet remains stable for repeatedly cleaned items, while memory performance was equal for both items (Fowle & Boschen, 2011). Instead of examining the effect of compulsive behavior on memory confidence, Cuttler et al., (2013) studied the effect of manipulating memory confidence on compulsive behavior and found that participants whose memory confidence is diminished, experience a higher level of doubt and more urges to check in a prospective memory task. Moreover, using the same false bio-feedback task as Lazarov et al., (2012), Zhang et al., (2017) found that the group with experimentally undermined confidence was more susceptible to distortions of confidence due to a higher reliance on the false feedback compared with the control group.

In sum, there is substantial evidence that engaging in OC behaviors lowers memory confidence, and that decreasing confidence can increase OC tendencies, supporting the idea of a link between low confidence and subclinical OC tendencies, specifically in OCD-relevant situations (Ashbaugh & Radomsky, 2007; Coles et al., 2006; Cuttler et al., 2013; Fowle & Boschen, 2011; Radomsky et al., 2006; Radomsky & Alcolado, 2010; Van Den Hout & Kindt, 2003a, 2003b). Moreover, there are multiple indications of confidence abnormalities associated with subclinical OC tendencies in the cognitive

domains of interoception and perception (Lazarov et al., 2012; Zhang et al., 2017), such as a decrease in metacognitive efficiency (Hauser, Allen, et al., 2017), although this is not supported by all studies (Ben Shachar et al., 2013; Rouault, Seow, et al., 2018). These contradictory results cannot be further clarified by performance confounds, since all studies showed equal performance levels between groups. Concluding, subclinical OC tendencies are mostly associated with a decrease in confidence or metacognitive efficiency, both in OCD-relevant contexts as well as neutral task environments.

Clinical OCD

Of the 23 studies investigating confidence in OCD patients, most have focused on metamemory tasks (Table 2B). The pioneering study by McNally & Kohlbeck (1993) showed that OCD patients express lower confidence than healthy participants, whereas memory performance was equal between groups. Many studies have since replicated these findings, using both OCD-relevant and neutral tasks or stimuli (Cougle et al., 2007; Foa et al., 1997; Karadag et al., 2005; MacDonald et al., 1997; Moritz & Jaeger, 2018). Two studies reported that the low confidence observed in OCD patients was associated with a decrease in memory performance (Tuna et al., 2005; Zitterl et al., 2001). Although memory performance deficits might have been the driving force behind some reported confidence deficits (Figure 2), many studies still find an impaired confidence in OCD patients in the absence of memory deficits (Foa et al., 1997; Karadag et al., 2005; MacDonald et al., 1997; Moritz & Jaeger, 2018). This association does not consistently replicate, however (Bucarelli & Purdon, 2016; Moritz et al., 2011; Moritz, Jacobsen, et al., 2006; Moritz, Kloss, et al., 2009; Moritz, Ruhe, et al., 2009; Tekcan et al., 2007). To explain these contradictory results, it has been suggested that the metamemory problems in OCD are amplified by contextual factors such as a heightened subjective feeling of responsibility (Boschen & Vuksanovic, 2007; Moritz et al., 2007; Radomsky et al., 2001). Furthermore, declining confidence levels with repetition of checks have been found in clinical OCD populations, also when controlling for anxiety levels, linking reduced memory confidence to typical OCD checking behavior (Boschen & Vuksanovic, 2007; Tolin et al., 2001).

Declines in confidence in OCD patients have also been found in tasks evaluating perception and action (Hermans et al., 2008), general knowledge (Dar, 2004; Dar et al., 2000), and interoception (Lazarov et al., 2014). A recent study found no differences in the dynamic course of confidence between OCD and healthy controls in a volatile reinforcement-learning task, but did show a dissociation between confidence and action in OCD patients (Vaghi et al., 2017). However, the authors did not analyze group differences for confidence precision or confidence calibration.

Overall, most evidence points to a decrease in confidence in OCD patients in multiple cognitive domains (i.e. memory, perception and interoception) (Cougle et al., 2007; Dar, 2004; Dar et al., 2000; Foa et al., 1997; Hermans et al., 2008; Karadag et al., 2005; Lazarov et al., 2014; MacDonald et al., 1997; McNally & Kohlbeck, 1993; Moritz & Jaeger, 2018; Tuna et al., 2005; Zitterl et al., 2001). This has been linked to checking behavior (Boschen & Vuksanovic, 2007; Tolin et al., 2001), where repetitions of actions are associated with a greater distortion of confidence levels. It is, however, not fully established whether decreases in confidence, in addition to OCD-relevant situations, also extend to neutral situations. Conflicting evidence exists, such that some studies did find decreases in confidence in OCD patients using neutral tasks (Cougle et al., 2007; Dar, 2004; Dar et al., 2000; Karadag et al., 2005; Lazarov et al., 2014; MacDonald et al., 1997; Zitterl et al., 2001), whereas others did not (Moritz et al., 2011; Moritz, Jacobsen, et al., 2006; Moritz, Kloss, et al., 2009; Moritz, Ruhe, et al., 2009; Tekcan et al., 2007). None of these studies actively controlled for performance differences between groups, but most studies did nevertheless show equal levels of performance between groups. Importantly, confidence abnormalities are likely dependent on contextual factors, since multiple studies have reported decreases in confidence in OCD patients in OCD-relevant scenarios, or specifically when patients experience heightened responsibility (Boschen & Vuksanovic, 2007; Bucarelli & Purdon, 2016; Hermans et al., 2008; Moritz et al., 2007; Radomsky et al., 2001; Tolin et al., 2001). To our knowledge, no studies have yet investigated abnormalities in metacognitive sensitivity or efficiency in clinical OCD populations. To conclude, decreases in confidence have been found in OCD for various cognitive domains within both neutral and OCD-relevant contexts (Figure 3). However, some studies did not find differences within the OCD population.

Schizophrenia

Schizophrenia is a psychiatric disorder defined by positive symptoms, including hallucinations and delusions, and negative symptoms, comprising flattening of affect, loss of pleasure and social withdrawal (Schultz et al., 2007). Next to these symptoms, schizophrenic patients suffer from cognitive impairment (Bowie & Harvey, 2006).

Subclinical: non-psychotic help-seeking individuals and delusion proneness

Most patients experience a prodromal phase in which symptoms gradually develop into schizophrenia or psychosis (Schultz et al., 2007). One of the predictors of transition into psychosis is cognitive impairment, with high-risk individuals exhibiting moderate to severe deficits in cognitive abilities (Seidman et al., 2010). Next to the cognitive deficits,

metacognition also seems to be impaired in schizophrenia; however, the nature of the impairment is not yet fully understood.

Eight studies investigating the link between confidence and subclinical schizophrenia were identified (Table 2C). Two studies evaluated confidence in verbal memory, executive functioning, and social functioning tasks as possible neuropsychological markers in early pre-psychotic stages of schizophrenia in help-seeking adolescents (Koren et al., 2017; Scheyer et al., 2014). Scheyer et al., (2014) found no differences in either cognitive or metacognitive abilities between individuals with high versus low risk for future psychosis; yet, confidence was a significant predictor for psychosocial functioning above and beyond cognitive abilities alone. Koren et al., (2017) assessed the relationship between confidence and self-disturbance in help-seeking adolescents with or without attenuated psychotic syndrome (APS), which is considered a prodromal phase of schizophrenia. Self-disturbance is a risk factor for developing psychosis, defined as the disruption of the sense of being a self-present subject of experience and action (Raballo et al., 2016). Results showed that confidence monitoring (i.e., the correlation between confidence and actual performance) had a significant positive relationship with self-disturbance, beyond neurocognitive functioning and APS symptoms alone. This indicates that a higher level of self-disturbance was related to increased metacognitive abilities.

Regarding delusion proneness, three studies using false memory and reasoning tasks found that delusion prone subjects are more overconfident (McKay et al., 2006; Warman, 2008), especially in errors (Laws & Bhatt, 2005). Likewise, individuals with a high level of paranoia exhibited lower confidence discrimination in a visual task (Moritz, Göritz, et al., 2014). The authors argue that overconfidence in errors is induced by 'liberal acceptance', when partial information is deemed sufficient for having high confidence in a decision (Moritz & Woodward, 2006b). In turn, this liberal acceptance of false memories or unlikely events may promote delusions and paranoid ideation. Another study, using a general knowledge task, confirmed overconfidence in errors in individuals with high paranoia levels, but also showed that it was dependent on subjective competence and perceived difficulty (Moritz et al., 2015). They found that overconfidence in errors is exaggerated when subjects feel highly competent or deemed the question easy. However, a recent study using a perceptual task did not find any direct relationships between self-reported schizotypy symptoms and confidence level or metacognitive efficiency (Rouault, Seow, et al., 2018).

In sum, prior subclinical studies have produced mixed results. One study reports no differences between high and low risk groups (Scheyer et al., 2014), and one even shows improvement of metacognitive abilities with higher schizotypal symptoms

(Koren et al., 2017). Nevertheless, most of the studies, which were the most extensive in terms of participants, reported that delusion prone or highly paranoid individuals showed an overconfidence effect for errors, resulting in a diminished confidence discrimination within various cognitive domains (i.e. memory, perception and reasoning) (Laws & Bhatt, 2005; McKay et al., 2006; Moritz, Göritz, et al., 2014; Warman, 2008). Of note, a recent study indicates that this effect might also be moderated by subjective level of competence (Moritz et al., 2015). None of the studies actively controlled for performance differences.

Clinical Schizophrenia

Similar to research in OCD, the most considerable evidence for confidence abnormalities in schizophrenia has come from metamemory studies. Most of the 23 identified studies have either performed a source-monitoring or a false memory task (Table 2D). The majority reports that schizophrenia patients exhibit higher confidence for incorrect answers, resulting in a confidence discrimination deficit (Bhatt et al., 2010; Eifler et al., 2015; Gaweda et al., 2012; Kircher et al., 2007; Moritz et al., 2003, 2004, 2005, 2008; Moritz, Woodward, & Rodriguez-Raecke, 2006; Moritz & Woodward, 2002). Schizophrenia, OCD and post-traumatic stress disorder (PTSD) patients all exhibited lower memory performance than healthy controls, but schizophrenia patients showed a specific impairment in discrimination compared with both OCD and PTSD control groups, due to a higher confidence in errors (Moritz & Woodward, 2006a). Moritz, Woodward, & Chen (2006) used the source-monitoring paradigm (Table 1) to study the developmental trajectory of confidence problems in first-episode psychosis patients (FEP). They found a confidence discrimination deficit in the FEP group due to overconfidence in errors. These results were replicated more recently in both FEP patients and high risk groups using a source-monitoring and false memory task (Eisenacher et al., 2015; Gawęda et al., 2018). Together, these findings reinforce the notion that an overconfidence in errors may serve as a risk factor for developing schizophrenia.

The inflated confidence in errors, in the absence of performance differences, was also reported in other cognitive domains, such as emotion perception (Köther et al., 2012; Moritz et al., 2012; Peters et al., 2013). In the perceptual domain, at similar levels of performance, schizophrenia patients showed inflated confidence in errors compared with both a healthy and an OCD control group (Moritz, Ramdani, et al., 2014). Moreover, the amount of high confident errors significantly correlated with self-rated levels of current paranoia. Similarly, Davies et al., (2018) found that FEP patients have a significantly lower metacognitive sensitivity (meta-d') compared with healthy subjects, despite similar performance and confidence levels, suggesting that schizophrenia

patients are impaired in discriminating between correct and incorrect trials with their confidence judgments. However, two studies did not find such a discrimination impairment, although one did report decreased metacognitive performance in schizophrenia patients (Bruno et al., 2012). The other reported higher confidence levels in errors for healthy controls, and more high confident source misattributions in schizophrenia patients (Peters et al., 2007).

Lastly, a study using a FOK task paired with confidence judgments found no differences in confidence level between schizophrenia patients and healthy subjects, while FOK judgments were lower in the patient group (Bacon et al., 2001). This finding was replicated using a memory task (Bacon & Izaute, 2009).

In sum, the most consistent finding in schizophrenia patients is an inflated retrospective confidence in errors resulting in reduced confidence discrimination within multiple cognitive domains (i.e. memory, visual and emotional perception) (Figure 3) (Bhatt et al., 2010; Eifler et al., 2015; Gaweda et al., 2012; Kircher et al., 2007; Köther et al., 2012; Moritz et al., 2008, 2012; Moritz, Woodward, & Rodriguez-Raecke, 2006; Moritz et al., 2003, 2004, 2005; Moritz & Woodward, 2002, 2006b; Peters et al., 2013). This reduced discrimination may be attributed to a deficit in metacognitive sensitivity (Davies et al., 2018). Furthermore, these abnormal confidence levels are already found, albeit less consistently, in early stages of the disorder (i.e. at risk populations and FEP patients) (Eisenacher et al., 2015; Gawęda et al., 2018; Moritz, Woodward, & Chen, 2006). Concluding, schizophrenia patients show abnormal confidence discriminatory abilities induced by overconfidence in errors.

Addiction

Addictions can be roughly divided in two categories: dependency to a substance (i.e., substance-use dependency; SUD) or to an activity (such as gambling disorder; GD). Addictions are characterized by persistent drug use or maladaptive behavior despite negative consequences (Koob & Volkow, 2010). SUDs and behavioral addictions have a common underlying neural mechanism that governs the development and sustenance of these disorders (Limbrick-Oldfield et al., 2013). Next to classic symptoms of habit forming and craving, addicted individuals are also impaired in a broad spectrum of cognitive functions (van Holst et al., 2010).

Subclinical Addiction

Three studies investigating confidence in subclinical addiction were identified (Table 2E). Two studies divided a student population into probable pathological gamblers,

problem gamblers and no-problem gamblers and used a general knowledge task (Goodie, 2005; Lakey et al., 2007). Goodie (2005) found that pathological gamblers have significantly higher confidence, but also lower task performance, compared with the other groups, resulting in higher overconfidence. Similarly, Lakey et al., (2007) showed that non-problem gamblers were less overconfident than the other two groups, with no differences between the pathological and problem gamblers. Both studies also found a significant positive correlation between gambling severity and overconfidence. Considering SUD, a recent study using a perceptual task found no direct relationship between self-reported alcoholism symptoms and either confidence level or metacognitive efficiency in the general population (Rouault, Seow, et al., 2018).

Taken together, these few studies showed some evidence for confidence abnormalities in subclinical GD within the semantic memory domain, pointing to increased overconfidence in a general context (Goodie, 2005; Lakey et al., 2007) (Figure 3). However, task performance was not held equal between groups, rendering it difficult to draw firm conclusions. Furthermore, these findings did not extend to links between alcoholism symptoms and confidence within the perceptual domain (Rouault, Seow, et al., 2018). The link between confidence abnormalities and subclinical symptoms of addiction is therefore not yet apparent.

Clinical Addiction

A total of five studies have investigated confidence in addiction (Table 2F). One study assessed confidence in GD patients and healthy controls using a non-gambling grammar task and reported similar confidence levels in both groups, while GD patients exhibit lower performance (Brevers et al., 2014). However, confidence correlated with performance in healthy controls, but not in GD patients, suggesting an abnormal confidence processing in gamblers. Considering SUD, Le Berre et al., (2010) studied confidence in alcohol-use disorder patients using a memory task with a prospective FOK measure. Results showed that alcohol use disorder patients had a significantly worse memory performance than healthy controls, and were less accurate regarding their FOK judgments as they overestimated their recognition performance. Moreover, a significant positive correlation was found between memory deficits, executive dysfunction and metamemory impairment in alcohol use disorder patients. In another study, using a visuo-perceptual task in which performance was held constant, active cocaine addicted individuals displayed a decreased metacognitive efficiency compared with remitted cocaine users and healthy subjects (Moeller et al., 2016). Interestingly, the remitted group did not differ from the healthy controls. Both cocaine user groups did not differ with regards to peak drug usage, suggesting that the results cannot be attributed to a greater lifetime addiction severity in active users.

To date, two studies have examined confidence in a population of opiate dependent patients receiving methadone maintenance treatment. Mintzer & Stitzer (2002) found that patients reported significantly higher confidence for incorrect choices in a memory task compared with healthy subjects, resulting in worse confidence discrimination. Recently, Sadeghi et al., (2017) found lower metacognitive efficiency for patients using a perceptual task, while no differences in mean confidence levels or performance could be detected. In the memory domain, however, patients exhibited lower performance but similar metacognitive efficiency than controls. These findings suggest that separate metacognitive systems might exist for different cognitive domains.

Summing up, a single study in GD patients showed a disconnection between confidence and accuracy, indicating a deficiency in metacognition (Brevers et al., 2014). Replications using bias free measures of confidence are needed in order to confirm this effect. In SUD patients, multiple studies correcting for performance differences and using bias-free confidence measures reported inflated retrospective confidence for errors and thus decreased confidence discrimination, as well as diminished metacognitive efficiency. This abnormality was found in both memory and perceptual domains (Mintzer & Stitzer, 2002; Sadeghi et al., 2017), and improved in remitted patients (Moeller et al., 2016). Replications and direct comparisons between addiction subtypes are needed to confirm the generalizability of these findings. Concluding, multiple bias-free studies reported a decrease in confidence discrimination and metacognitive efficiency in SUD patients (Figure 3). However, for GD patients, more research is needed.

Anxiety and Depression

Major depressive disorder (MDD) and anxiety disorders are common disorders with a lifetime prevalence of 16.2% and 28.8%, respectively (Kessler et al., 2003, 2005). Since they are both classified as mood disorders and are highly comorbid, they are considered jointly. MDD and anxiety disorders share a negativity bias in information processing, reflecting a greater focus on negative input (J. B. Engelmann et al., 2017; McClintock et al., 2011; McLaughlin & Nolen-Hoeksema, 2011; Williams et al., 2009). While general deficits in cognition are established symptoms in these disorders (Ferreri et al., 2011; Rock et al., 2014), studies investigating confidence disorders are scarce. However, the well-known hallmarks of both disorders: negative self-concepts, rumination and indecisiveness (McClintock et al., 2011), suggest that patients show a negative confidence bias.

Subclinical Anxiety and Depression

Subclinical levels of depression and anxiety are common among the general population (Goldney et al., 2004). Five studies researching subclinical anxiety or depression were identified (Table 2G). Stone et al., (2001) used a general knowledge task in four groups from a general population sample: (1) non-depressed non-anxious, (2) non-depressed anxious, (3) depressed non-anxious, and (4) depressed anxious. They reported lower confidence levels in depressed non-anxious individuals compared with the control group (non-depressed, non-anxious), in the absence of performance differences. Surprisingly, the depressed anxious group did not differ from the control group on any measure, suggesting that the presence of anxiety itself might counterbalance the confidence abnormalities found in depression. Soderstrom et al., (2011) divided a non-clinical sample into non-, mild- and moderate depression groups and used a memory task with a JOL measure (i.e., prospective confidence). While results showed overconfidence in all three groups, mildly depressed subjects exhibited significantly lower overconfidence than the other groups. No differences in calibration were found between the non- and moderately-depressed groups. However, caution must be taken when interpreting these results, as performance levels were significantly different between the groups. The authors of a third study divided a large group of undergraduates into depressed and non-depressed groups and asked participants to predict future events (Dunning & Story, 1991). They reported overconfidence in the depressed group, but this was fully driven by differences in prediction performance: while reporting similar levels of confidence, depressed individuals showed a decreased performance in predicting future events compared with the non-depressed group. Moreover, the lack of confidence differences between groups could be explained by the use of valenced life events rather than a neutral task: since depressed subjects commonly have a negative self-concept and a general focus on negative events (McClintock et al., 2011), they may have a high confidence that negative events could happen.

One study did not detect any association between depression and/or anxiety symptoms and various confidence measures obtained via several cognitive tasks assessing executive functioning, memory and social emotional functioning (Quiles et al., 2015). However, Rouault, Seow, et al., (2018) did find a significant negative relationship between self-reported depression and anxiety symptoms and confidence level in the general population, indicating that individuals with higher depression or anxiety symptom scores report lower levels of confidence.

Together, the research on metacognition in mood disorders remains inconclusive to date due to contradictory results. Two studies reported underconfidence in the

subclinical depressed groups within perceptual and semantic memory domains (Rouault, Seow, et al., 2018; Stone et al., 2001); two studies showed overconfidence due to performance deficits (Dunning & Story, 1991; Soderstrom et al., 2011) using prediction and memory tasks, and one study reported null findings in various cognitive domains (i.e. executive functioning, memory and emotional processing) (Quiles et al., 2015). Moreover, individuals with both depression and anxiety symptoms did not show confidence abnormalities. However, some of these studies were confounded by differences in performance, which could have caused false reports of overconfidence. Regarding only the studies that did correct for performance differences and used retrospective confidence judgments (Rouault, Seow, et al., 2018; E. R. Stone et al., 2001), all reported an effect of underconfidence.

Clinical Anxiety and Depression

In MDD patients, four studies were identified that mostly reported underconfidence compared with healthy controls using different paradigms (Table 2H). One study found decreased confidence discrimination in both current and recovered MDD patients using a general knowledge task (Hancock et al., 1996). This effect significantly correlated with depression severity, such that patients with more severe depression showed lower confidence levels and discrimination. A second study using four different decision tasks (i.e. an episodic memory, general knowledge, perceptual discrimination and a social judgment task) found that MDD patients reported lower confidence levels than the control group, whereas recovered patients did not (Fu et al., 2005). In both studies, performance was equal between the groups. In a third study, MDD patients exhibited lower performance in a memory task than a control and a chronic-fatigue syndrome patient group. This was accompanied by greater underconfidence in the MDD group, both when judgments were made after every single trial and after a block of trials (Szu-Ting Fu et al., 2012). Lastly, a recent study using an emotional perception task found no interaction between group and confidence in a model explaining incorrect responses (Fieker et al., 2016). However, in line with previous findings, the authors did find a significant association between low confidence levels and high depression severity scores.

To our knowledge, there are no studies to date examining confidence focusing solely on anxiety patients versus healthy controls. However, a few studies investigating OCD used anxiety disorder patients as a clinical control group. Two studies found no difference between anxiety or panic disorder patients and healthy controls regarding confidence (Dar et al., 2000; Lazarov et al., 2014), whereas another study showed that anxious controls had lower confidence levels (Tolin et al., 2001). A recent study, which

did not include a healthy control group, found that anxious and OCD patients had similar levels of memory confidence (Bucarelli & Purdon, 2016).

In summary, most studies showed a reduction of confidence levels in MDD in different cognitive domains (i.e. memory, visual and social perception) (Fu et al., 2005; Hancock et al., 1996; Szu-Ting Fu et al., 2012). Furthermore, some studies showed greater levels of underconfidence for current versus recovered MDD patients (Fu et al., 2005), whereas other studies did not report any differences (Fieker et al., 2016). Mixed results were found for anxiety disorders: two studies showed decreased confidence levels similar to OCD when compared to healthy controls within the memory domain (Bucarelli & Purdon, 2016; Tolin et al., 2001), whereas two other studies did not find such differences using general knowledge and interoception paradigms (Dar et al., 2000; Lazarov et al., 2014). Concluding, depression patients mostly showed an effect of underconfidence, whereas this effect was not clear-cut for anxiety patients (Figure 3).

Transdiagnostic Psychiatry

Transdiagnostic psychiatry is an emerging scientific field which attempts to decipher the cognitive, affective and neurobiological processes underlying complex behavior by relating them to symptom dimensions. Since this approach transcends traditional diagnostic categories, it has the potential to refine the current nosology-based clinical classifications beyond the classical Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnostic criteria (Huys et al., 2016; Stephan & Mathys, 2014). The underlying idea of this approach is that cognitive and brain-related functions (e.g., those relating to confidence processing) might map more closely onto symptomatology than DSM diagnoses.

A recent study by Rouault, Seow, et al., (2018) leveraged such a transdiagnostic psychiatry approach to investigate the relationship between confidence and psychiatric symptomatology in the general population. A large sample from the general population performed a perceptual decision-making task and answered self-report questionnaires spanning a range of psychiatric symptoms, including depression, general anxiety, schizotypy, impulsivity, OCD, social anxiety, eating disorders, apathy and alcohol dependency (Experiment 1: n = 498. Experiment 2: n = 497. See table 2A,2C,2E,2G). The relationships between accuracy, decision parameters, confidence and metacognitive efficiency (meta-d'/d') were examined. Results showed that the symptoms were not associated with decision parameters, but that higher levels of depression and anxiety symptoms were significantly associated with decreased

confidence. Furthermore, a factor analysis was carried out to retrieve a parsimonious latent structure that best explained the variance at the item level of all questionnaires, which identified three symptom dimensions: Anxious-Depression (AD), Compulsive Behavior and Intrusive Thought (CIT) and Social Withdrawal (SW). The AD dimension was significantly associated with lower confidence and higher metacognitive efficiency, whereas the CIT cluster was related to higher confidence and a lower metacognitive efficiency. The metacognitive efficiency results did, however, not survive correction for multiple comparisons and must be interpreted with caution. Lastly, none of the three symptom dimensions showed a relationship with decision parameters, indicating that psychiatric symptoms are related to shifts in confidence, but not in performance. Therefore, changes in confidence may represent a specific behavioral correlate of subclinical psychopathology that could be an important component of transdiagnostic psychiatry.

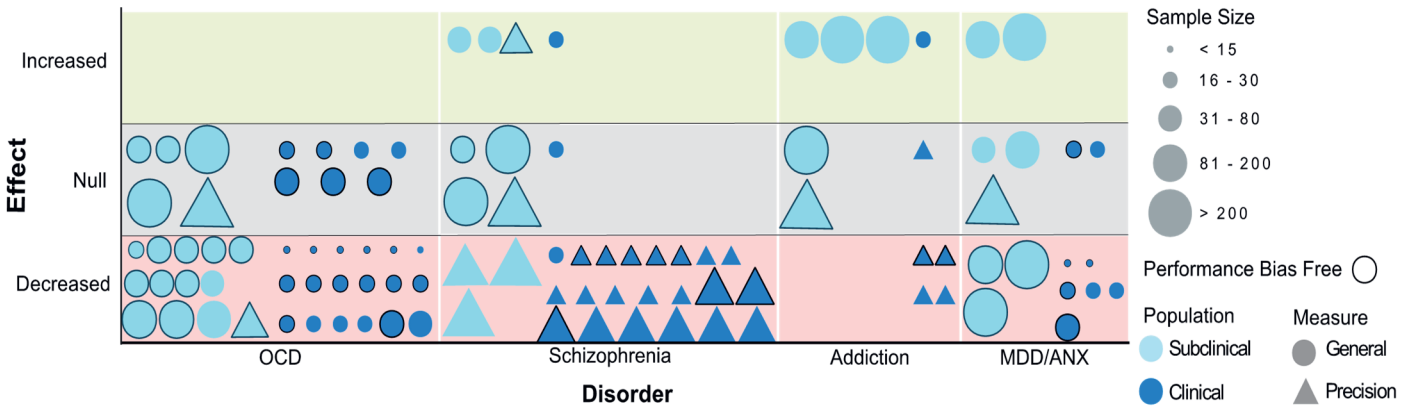


Figure 3: Overall confidence abnormalities in (sub)clinical psychiatry. This figure shows the overall abnormalities in confidence processes in different (sub)clinical psychiatric disorders (versus healthy controls in clinical patient groups). Every study is represented by one data point (circle or triangle). When a study existed of multiple experiments testing different populations, multiple data points were used. For all clinical studies, the sample size of the patient group is displayed. Different colors are used for subclinical (light blue) and clinical (dark blue) populations. Different symbols represent increases (on upper line) no change (middle line) or decreases (lower line) of general confidence level (circles) or precision of confidence estimation (triangles). Studies that controlled for performance biases, be it by using the bias-free meta-d' framework, or by showing (or actively keeping) equal performance levels between groups, are outlined. For studies investigating schizophrenia that found both an increase in confidence for errors as well as a decrease in discrimination, the latter effect is displayed in this figure. The subclinical study by Rouault et al. (2018)[30] is included in all four disorder categories. For explanation of the different confidence measures, see Figure 1: OCD = obsessive-compulsive disorder, MDD/ANX = depression/anxiety disorders.

Discussion

In this review we sought to obtain an answer to the question whether confidence judgments are abnormal across psychiatric disorders. We found evidence for confidence abnormalities across a variety of psychiatric disorders, which take specific directions for the different populations (Figure 3). For (sub)clinical OCD, the most consistent finding is a decrease in confidence level, especially related to typical OCD contexts, such as checking behavior. Regarding (sub)clinical schizophrenia, we primarily found increased confidence in errors resulting in a decrease of discrimination and metacognitive sensitivity. This diminished discriminatory ability between correct (real) and incorrect (imagined) situations fits core schizophrenia symptoms such as delusions and hallucinations, and was recently also found to be dependent on subjective competence. In clinical addiction, an increase in confidence – leading to a decrease in confidence discrimination and metacognitive efficiency – was found, which corresponds to the symptomatic lack of self-insight in this population (Goldstein et al., 2009). Subclinical addiction has not been studied as extensively, but overconfidence was found in subclinical GD. In clinical anxiety and depression, reductions in confidence levels were found, which fit with the negative information processing bias observed in mood disorders (McLaughlin & Nolen-Hoeksema, 2011). However, subclinical studies show mixed results and no studies using anxiety patients as the primary group of interest have been performed to date. Together, these results demonstrate that clinical and subclinical studies generally show similar results.

While these results suggest that there are abnormalities in confidence estimations in psychiatric patients, another important question is how these abnormalities relate to psychiatric disorders. Are these abnormalities closely linked or even underlying psychiatric symptoms? Are they a result of the disorder or perhaps only a byproduct without any significance for symptomatology? The studies discussed in this review indicate that there is a close interplay between psychiatric symptoms and confidence. For instance, several studies found that abnormal levels of confidence are already present in non-clinical populations with psychiatric tendencies or subclinical prodromal populations (Davies et al., 2018; Eisenacher et al., 2015; Gawęda et al., 2018; Hauser, Allen, et al., 2017; Lazarov et al., 2012; Moritz, Woodward, & Chen, 2006; Zhang et al., 2017). Moreover, a normalization of confidence abnormalities was found in three studies after patients recovered (Fu et al., 2005; Hancock et al., 1996; Moeller et al., 2016). Furthermore, four studies found direct correlations between confidence abnormalities and symptom severity (Fieker et al., 2016; Goodie, 2005; Hancock et al., 1996; Lakey et al., 2007). The interaction between psychiatric symptoms and confidence abnormalities was also demonstrated by studies showing that engaging in

compulsive behaviors lowered confidence levels, whereas undermining confidence lead to increases in compulsive tendencies (Ashbaugh & Radomsky, 2007; Coles et al., 2006; Cuttler et al., 2013; Fowle & Boschen, 2011; Radomsky et al., 2006; Radomsky & Alcolado, 2010; Van Den Hout & Kindt, 2003a, 2003b), indicating that confidence and pathological behavior are coupled. While the evidence for the strong relationship between confidence and psychiatric symptoms is convincing, the directionality of the effect is not unequivocal and should therefore be further explored in future studies using causal manipulations of confidence or longitudinal designs.

These findings raise many questions and give way to research advancing our understanding of confidence abnormalities in psychiatry. Confidence is not a unitary construct, since confidence abnormalities are differently expressed in various contexts (Moritz et al., 2007, 2015), and the role of context in confidence abnormalities should be further identified. For example, it is possible that confidence abnormalities aggravate in a symptom-related context. For instance, a gambler might be overconfident in general, but show an even increased overconfidence during gambling. Another interesting future avenue would be to study if normalization of confidence deviations would translate into decreased symptom severity, and vice versa. Interestingly, a recent paper showed that adaptive training can cause a domain-general enhancement of metacognitive abilities in the general population (Carpenter et al., 2019). Up to now, several forms of metacognitive training have been developed as treatment for psychiatric patients. Importantly, recent meta-analyses indicated that they were effective in reducing symptoms within a wide range of psychiatric disorders (Liu et al., 2018; Philipp et al., 2019). Furthermore, metacognitive training, as well as antipsychotic medication, have been shown to attenuate overconfidence in errors in schizophrenia patients (Köther et al., 2017; Moritz et al., 2017). Future work should focus on translating current knowledge about confidence abnormalities in psychiatry to new treatment interventions, tailored to specific confidence abnormalities. Furthermore, it remains uncertain whether confidence abnormalities in psychiatry generalize over different cognitive domains and contexts. Few studies have systematically and directly studied the transfer of confidence abnormalities across different domains within a population and showed mixed results favoring either domain-general (Hermans et al., 2008) or domain-specific (Fu et al., 2005; Sadeghi et al., 2017) views. However, the majority of the discussed studies used metamemory tasks; therefore, more research is needed to establish the generalizability of confidence disruptions to other cognitive domains. More knowledge about the relationship between confidence abnormalities in various domains and psychiatric disorders may eventually allow for personalized therapies focusing on individual deficits.

Next to using the traditional DSM diagnostic categories, it is important to study confidence using a transdiagnostic approach focusing on the level of symptoms. Recently, Rouault, Seow, et al., (2018) used a transdiagnostic approach and found that a symptom cluster of compulsivity and intrusive thoughts is related to heightened confidence, whereas an anxiety and depression cluster is related to lowered confidence in a large sample of the general population. Importantly, their results were less pronounced when symptoms were related to confidence abnormalities in the traditional diagnostic categorical (i.e., disorder-specific) way. This may indicate that confidence abnormalities are better explained by specific symptom clusters than disorder categories that are heterogeneous in their display of symptoms, because they show overlap with other disorders. For example, there might be large individual variety in the role that anxiety (Weinstein et al., 2015) and compulsivity play in psychiatric disorders such as addictions and OCD, resulting in different propensities for under- or overconfidence. Currently, it is not clear if and how these transdiagnostic findings generalize to clinical groups, although our findings seem to suggest that confidence abnormalities are similar between clinical and subclinical populations. An interesting avenue for future work is to apply transdiagnostic approaches to clinical groups and investigate whether symptom-based classification improves correlations with confidence abnormalities compared to classical DSM-based classification. Moreover, in addition to the data-driven transdiagnostic techniques adopted by Rouault, Seow, et al., (2018), other theory-driven techniques fitting the Research Domain Criteria (RDoC) framework should be used to further explore confidence abnormalities in psychiatric populations (Insel et al., 2010). Bearing in mind the advantages of the transdiagnostic approach, new treatment interventions focusing on treatment of confidence abnormalities related to specific symptom clusters instead of DSM classifications could be a promising new avenue. Furthermore, next to confidence being an important transdiagnostic factor associated with psychiatric disorders, many other factors have been shown to be of transdiagnostic value, such as neurocognitive deficits and motivation (Bora et al., 2010; Romanowska et al., 2018; Whitton et al., 2015). These factors may also contribute to confidence deviations within psychiatric populations (Eifler et al., 2015), which makes for an important area of future research.

Confidence can be viewed as a broader concept than the cognitive operationalization reviewed in this paper, relating to themes relevant to psychiatry such as trust and self-confidence (Borkowski et al., 1990; Sowislo & Orth, 2013). In order to gain a wider perspective on the role of confidence in psychiatry it would be interesting to explore how these themes are related and investigate the phenomenology of confidence abnormalities in these disorders.

The reviewed studies also indicate that there are methodological shortcomings in the field. Most of the reported studies suffered from (one of) two limitations. First, they did not account for performance differences between groups of interest. Performing better at a task leads to an increase in confidence (Maniscalco & Lau, 2012), and there is growing evidence that confidence judgments guide future behavior (Fleming & Daw, 2017). It is thus crucial to control for performance differences to isolate effects in confidence. Second, they did not use bias free measures next to the conventional measures of confidence level, such as calibration and discrimination. Bias free measures account for performance differences and response biases and provide more in-depth information about one's metacognitive abilities. Future work would benefit from using tasks that control for potential performance differences and use bias free measures such as meta-d' (although these measures require a considerable amount of trials to obtain sufficient statistical power (Rouault, McWilliams, et al., 2018)). Furthermore, a discrepancy exists in how confidence is assessed inside and outside the clinical fields, with more effort toward a normative definition of confidence1, operationalization using (Bayesian) computational frameworks (Fleming & Daw, 2017; Kepecs & Mainen, 2012) and confidence evaluation, incentivization or assessment (Hollard et al., 2016) outside of clinical fields. Adopting these standards in clinical research could help improving our knowledge about confidence abnormalities in psychiatry. Lastly, there is more and more research into the neurobiological basis of confidence, which shows that brain areas such as the lateral and medial prefrontal cortex and insula are related to confidence encoding (Vaccaro & Fleming, 2018). Interestingly, these brain areas also play a central role in the various psychiatric disorders discussed in this review (Chai et al., 2011; J. B. Engelmann et al., 2017; Goldstein & Volkow, 2011; Namkung et al., 2017; Yücel et al., 2007). Therefore, studying the neural mechanisms responsible for the confidence abnormalities observed in these populations is an important future research endeavor.

Table 2. Overview of reviewed studies

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
A) Overview of subclinical OCD studies					
Ashbaugh & Radomsky	2007	152 HC	Repeated Checking Task **	↓ confidence high-checkers vs low-checkers	-
Ben Shachar et al.	2013	47 HC; high & low OC tendencies	General Knowledge Task	== confidence high vs low OC tendencies	+
Coles, Radomsky & Horng	2006	S1: 51 HC S2: 81 HC	Repeated Checking Task **	S1 & S2: ↓ confidence with repeated checking	+
Cuttler et al.	2013	199 HC	Prospective Memory Task	↓ confidence undermined group	+
Fowle & Boschen	2011	60 HC	Repeated Cleaning Task **	no increase in confidence for repeatedly cleaned items, increase in confidence for non-repeatedly cleaned items	+
Hauser et al.	2017	40 HC; high & low OC tendencies	Global Motion Detection Task	↓ metacognitive efficiency high compulsive group	++
Lazarov et al.	2012	38 HC; high & low OC tendencies	False Feedback Muscle Tension Task	↓ confidence high compulsive group	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Radomsky, Gilchrist & Dussault	2006	55 HC	Repeated Checking Task **	↓ confidence with repeated checking	+
Radomsky & Alcolado	2010	62 HC	Repeated Mental Checking Task **	↓ confidence with repeated checking	-
Rouault, Seow, Gillan & Fleming	2018	S1: 498 HC S2: 497 HC	Perceptual Decision-Making Task	S1: no relationship OCD symptoms and confidence S2: no relationship OCD symptoms and confidence or metacognitive efficiency AD symptom dimension ↓ confidence and ↑ metacognitive efficiency, CIT symptom dimension ↑ confidence and ↓ metacognitive efficiency	++
Van den Hout & Kindt	2003a	S1: 39 HC S2: 40 HC	Repeated Checking Task **	S1 & S2: ↓ confidence with repeated checking	+
Van den Hout & Kindt	2003b	40 HC	Repeated Checking Task **	↓ confidence with repeated checking	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Zhang et al.	2017	S1: 30 HC S2: 32 HC	False Feedback Muscle Tension Task	S1 & S2: ↓ confidence high compulsive group	+
B) Overview of clinical OCD studies					
Boschen & Vuksanovic	2007	15 OCD, 40 HC	Repeated Checking Task **	↓ confidence OCD vs HC ↓ confidence with repeated checking	+
Bucarelli & Purdon	2016	30 OCD, 18 anxious controls	Repeated Checking Task **	== confidence OCD vs anxious controls	-
Cougle, Salkovskis & Wahl	2007	39 OCD checkers, 20 OCD non-checkers, 22 anxious controls, 69 HC	Memory Task	↓ confidence OCD vs HC and anxious controls	-
Dar et al.	2000	20 OCD checkers, 29 PD, 23 HC	General Knowledge Task	↓ confidence OCD vs HC	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Dar	2004	S1: 20 OCD checkers, 20 PD, 20 HC S2: 15 OCD checkers, 15 HC S3: 6 OCD checkers, 6 HC	General Knowledge Task	S1, S2 & S3: ↓ confidence OCD vs both control groups S1, S2 & S3: ↓ confidence with repeated checking	+
Foa et al.	1997	15 OCD, 15 HC	Memory Task	↓ confidence OCD vs HC	+
Hermans et al.	2008	16 OCD, 16 clinical controls, 16 HC	Repeated Actions Task **	↓ confidence OCD vs both control groups	-
Karadag et al.	2005	32 OCD, 31 HC	Memory Task	↓ confidence OCD vs HC	+
Lazarov et al.	2014	20 OCD, 20 anxious controls, 20 HC	False Feedback Muscle Tension Task	↓ confidence OCD vs HC and anxious controls	+
MacDonald et al.	1997	10 OCD checkers, 10 OCD non-checkers, 10 HC	Memory Task	↓ confidence OCD checkers vs non-checkers and HC	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
McNally & Kohlbeck	1993	12 OCD checkers, 12 OCD non-checkers, 12 HC	Reality Monitoring Task	↓ confidence OCD vs HC	+
Moritz, Jacobsen, Willenborg, Jelinek & Fricke	2006	17 OCD checkers, 10 OCD non checkers, 51 HC	Source Memory Task	== confidence OCD vs HC	+
Moritz, Wahl, Zurowski, Jelinek, Hand & Fricke	2007	28 OCD, 28 HC	Memory Task **	↓ confidence OCD vs HC under high responsibility	+
Moritz, Kloss, von Eckstaedt & Jelinek	2009	43 OCD, 46 HC	Memory Task	== confidence OCD vs HC	+
Moritz, Ruhe, Jelinek & Naber	2009	32 OCD, 32 HC	Memory Task	== confidence OCD vs HC	+
Moritz et al.	2011	30 OCD, 20 HC	Memory Task	== confidence OCD vs HC	+
Moritz & Jaeger	2018	26 OCD, 21 HC	Memory Task	↓ confidence OCD vs HC	+
Radomsky, Rachman & Hammond	2001	11 OCD	Repeated Checking Task **	↓ confidence under high responsibility	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Tekcan, Topçuoglu & Kaya	2007	25 OCD checkers, 16 OCD non-checkers, 27 HC	Memory Task	== confidence OCD vs HC	+
Tolin et al.	2001	14 OCD, 14 anxious controls, 14 HC	Repeated Memory Task **	↓ confidence OCD vs both control groups with repetition	+
Tuna, Tekcan & Topçuoglu	2005	17 OCD, 16 subclinical checkers, 15 HC	Memory Task	↓ confidence OCD vs HC	-
Vaghi et al.	2017	24 OCD, 25 HC	Predictive Inference Task	== confidence OCD vs HC	-
Zitterl et al.	2001	27 OCD, 27 HC	Memory Task	↓ confidence OCD vs HC	-
C) Overview of subclinical schizophrenia studies					
Koren et al.	2017	61 help seeking adolescents	Verbal Memory, Executive – and Social Functioning Tasks	Positive relationship self-disturbance and meta-cognitive control	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Laws & Bhatt	2005	105 HC	Memory Task	↑ confidence in errors high delusion-proneness ↓ discrimination high delusion-proneness	-
McKay, Langdon & Coltheart	2006	58 HC	Reasoning Task	↑ confidence high delusion-proneness	-
Moritz, Göritz, van Quaquebeke, Andreaou, Jungclaussen & Peters	2014	2008 HC	Visual Perception Task	↑ confidence in errors high paranoia ↓ discrimination high paranoia	-
Moritz et al.	2015	2321 HC	General Knowledge Task**	↑ confidence in errors high paranoia, exaggerated with high competence or easy questions ↓ discrimination high paranoia, exaggerated with high competence or easy questions	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Rouault, Seow, Gillan & Fleming	2018	S1: 498 HC S2: 497 HC	Perceptual Decision-Making Task	S1: No relationship SCZ symptoms and confidence S2: No relationship SCZ symptoms and confidence or metacognitive efficiency AD symptom dimension ↓ confidence and ↑ metacognitive efficiency, CIT symptom dimension ↑ confidence and ↓ metacognitive efficiency	++
Scheyer et al.	2014	78 help seeking adolescents	Verbal memory, executive functioning & social functioning tasks	== confidence high vs low psychosis-prone groups	+
Warman	2008	70 HC	Decision-making task	↑ confidence high delusion-proneness	-
D) Overview of clinical schizophrenia studies					
Bacon et al. *	2001	19 SCZ, 19 HC	General Knowledge Task	== confidence SCZ vs HC ↓ FOK ratings SCZ vs HC	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Bacon & Izaute*	2009	21 SCZ, 21 HC	Memory Task	↓ FOK ratings SCZ vs HC	-
Bhatt, Laws & McKenna	2010	25 SCZ, 20 HC	False-Memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Bruno et al.	2012	28 SCZ, 14 HC	Emotional and Non-Emotional WCST	== discrimination SCZ vs HC, but ↓ metacognitive performance SCZ vs HC	+
Davies et al.	2018	41 FEP, 21 HC	Perceptual Decision-Making Task	↓ meta-d' FEP vs HC	++
Eifler et al.	2015	32 SCZ, 25 HC	False-memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Eisenacher et al.	2015	34 at risk patients, 21 FEP, 38 HC	Verbal Recognition Task	↑ confidence in errors at risk and FEP vs HC ↓ discrimination at risk and FPE vs HC	+
Gaweda, Moritz & Kokoszka	2012	32 SCZ, 32 HC	Source-Monitoring Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Gaweda et al.	2018	36 at risk patients, 25 FEP, 33 HC	Source-Monitoring Task	↑ confidence in errors UHR and FEP vs HC ↓ discrimination UHR and FEP vs HC	-
Kircher et al.	2007	27 SCZ, 19 HC	False-Memory Task	↑ confidence (more so in errors) SCZ vs HC	+
Köther et al.	2012	76 SCZ, 30 HC	Emotion Recognition Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz & Woodward	2002	23 SCZ, 15 HC	Source-Monitoring Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz, Woodward & Ruff	2003	30 SCZ, 21 HC	Source-Monitoring Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz et al.	2004	20 SCZ, 20 HC	False-Memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz et al.	2005	30 SCZ, 15 HC	Source-Monitoring Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Moritz & Woodward	2006b	31 SCZ, 48 psychiatric controls, 61 HC	Source-Monitoring Task	↑ confidence in errors SCZ vs both control groups ↓ discrimination SCZ vs both control groups	+
Moritz, Woodward & Rodriguez-Raecke	2006	35 SCZ, 34 HC	False-Memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz, Woodward & Chen	2006	30 FEP, 15 HC	Source-Monitoring Task	↑ confidence in errors FEP vs HC ↓ discrimination FEP vs HC	-
Moritz et al.	2008	68 SCZ, 25 HC	False Visual Memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
Moritz et al.	2012	23 SCZ, 29 HC	Emotion Perception Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	+
Moritz, Ramdani, Klass, et al.	2014	55 SCZ, 58 OCD, 45 HC	Perceptual Decision-Making Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	+
Peters et al.	2007	23 SCZ, 20 HC	False-Memory Task	↑ confidence in errors HC vs SCZ ↓ discrimination SCZ vs HC	+

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Peters et al.	2013	27 SCZ, 24 HC	Emotional Memory Task	↑ confidence in errors SCZ vs HC ↓ discrimination SCZ vs HC	-
E) Overview of subclinical addiction studies					
Goodie	2005	S1: 200 HC S2: 384 HC	General Knowledge Task	S1 & S2: ↑ overconfidence problem and possible pathological gamblers	-
Lakey, Goodie & Campbell	2007	221 HC	General Knowledge Task & Iowa Gambling Task	↑ overconfidence problem and possible pathological gamblers	-
Rouault, Seow, Gillan, Fleming	2018	S2: 497 HC	Perceptual Decision-Making Task	S2: No relationship alcoholism symptoms and confidence or metacognitive efficiency AD symptom dimension ↓ confidence and ↑ metacognitive efficiency, CIT symptom dimension ↑ confidence and ↓ metacognitive efficiency	++

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
F) Overview of clinical addiction studies					
Brevers et al.	2014	25 GD, 25 HC	Grammar Task	Disconnection confidence and accuracy GD	-
Le Berre et al.*	2010	28 AUD, 28 HC	Memory Task	↑ FOK judgments AUD vs HC	-
Mintzer & Stitzer	2002	18 MMP, 21 HC	Memory Task	↑ confidence for errors MMP vs HC ↓ discrimination MMP vs HC	-
Moeller et al.	2016	14 remitted CUD, 8 active CUD, 13 HC	Perceptual Decision-Making Task	↓ metacognitive efficiency active CUD vs remitted CUD and HC	++
Sadeghi et al.	2017	23 MMP, 24 HC	Memory & Perceptual Task	↓ metacognitive efficiency MMP vs HC perceptual task, but not memory task	++
G) Overview of subclinical depression/anxiety studies					
Dunning & Story	1991	S1: 164 HC S2: 259 HC	Future Prediction Task	S1 & S2: ↑ confidence depressed vs non-depressed	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Quiles, Prouteau & Verdoux	2015	50 HC	WCST, Digit Span, Memory Task & Emotion Recognition Task	No relationship confidence and depression/anxiety symptoms	-
Rouault, Seow, Gillan, Fleming	2018	S1: 498 HC S2: 497 HC	Perceptual Decision-Making Task	S1: Negative relationship confidence levels and depression/anxiety symptoms S2: Negative relationship confidence levels and anxiety symptoms, no relationship with metacognitive efficiency AD symptom dimension ↓ confidence and ↑ metacognitive efficiency. CIT symptom dimension ↑ confidence and ↓ metacognitive efficiency	++

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Soderstrom, Davalos & Vásquez*	2011	97 HC	Memory Task	↓ calibration based on JOL mildly depressed vs HC == calibration based on JOL moderate depressed vs HC	-
Stone, Dodrill & Johnson	2001	200 HC	General Knowledge Task	↓ confidence depressed group	+
H) Overview of clinical depression/anxiety studies					
Bucarelli & Purdon	2016	30 OCD, 18 ANX	Repeated Checking Task	== confidence ODC vs ANX	-
Dar et al.	2000	20 OCD checkers, 29 PD, 23 HC	General Knowledge Task	== confidence PD vs OCD and HC	+
Fieker et al.	2016	45 MDD, 30 HC	Emotional Perception Task	Negative correlation confidence and depression severity	+
Fu et al.	2005	15 MDD, 15 recovered MDD patients, 22 HC	Memory, General Knowledge, Perceptual & Social Judgment Task	↓ confidence MDD vs HC == confidence recovered MDD vs HC and MDD	-

Authors	Year	Sample size and study populations	Task	Results	Performance bias free
Hancock, Moffoot & O'Carroll	1996	14 MDD, 14 recovered MDD patients, 14 HC	General Knowledge Task	↓ confidence for correct answers in MDD vs HC == confidence recovered MDD vs HC	+
Lazarov et al.	2014	20 OCD, 20 ANX, 20 HC	False Feedback Muscle Tension Task	↓ confidence OCD vs ANX and HC	+
Szu-Ting Fu et al.	2012	23 MDD, 22 dysphoria patients, 32 HC	Memory Task	↓ confidence MDD vs HC and dysphoria	-
Tolin et al.	2001	14 OCD, 14 ANX, 14 HC	Memory Task	↓ confidence ANX vs HC	+

This table shows a summary of all studies assessing confidence in the different psychiatric disorders included in this review. In the various subparts, studies using the following populations are described: A) subclinical OCD, B) clinical OCD, C) subclinical schizophrenia, D) clinical schizophrenia, E) subclinical addiction, F) clinical addiction, G) subclinical depression/anxiety and H) clinical depression/anxiety. The results are schematically represented with ↓ signaling a significant decrease, ↑ significant increase and == no differences. Regarding the performance bias, the signs indicate the following: **++**: Study used bias free measures such as meta-d' and/or actively kept performance equal between groups (e.g., by using a staircase procedure), **+**: The assessed groups had equal levels of performance, -: Study did not use bias free measures and did not control for performance differences between groups, or did not report accuracy measures. For more information about the most frequently used tasks, see Table 1. The asterisks represent the following: *: This study used a prospective confidence measure, **: This study has taken into account moderators (i.e., OCD-relevant contexts, responsibility level or subjective competence). Abbreviations: HC = healthy controls, OC = obsessive-compulsive, OCD = obsessive-compulsive disorder, AD = anxious-depressive, CIT= compulsive behavior and intrusive thought, PD = panic disorder, SCZ = schizophrenia, FEP = first-episode psychosis, FOK = feeling of knowing, GD = gambling disorder, AUD = alcohol use disorder, MMP = methadone maintenance patients, CUD = cocaine use disorder, ANX = anxiety disorder, MDD = major depressive disorder, S1 = study 1, S2 = study 2.

Acknowledgements

M.H. and R.J.v.H. were supported by Amsterdam Brain and Cognition Project Grant (University of Amsterdam). J.L. was supported by an NWO Veni Fellowship (grant 916-18-119). M.L. was supported by an NWO Veni Fellowship (grant 451-15-015) and by a Swiss National Fund Ambizione Grant (PZ00P3_174127). We want to thank Nina de Boer for proofreading the paper.

Disclosure statement

None of the authors have any conflicts of interest to declare

3

Motivational signals disrupt metacognitive signals in the human ventromedial prefrontal cortex

Hoven M

Brunner G

de Boer NS

Goudriaan AE

Denys D

van Holst RJ*

Luigjes J*

Lebreton M*

* shared last authorship

Communications Biology, 2022, 5(1), 1-13

Abstract

A growing body of evidence suggests that, during decision-making, BOLD signal in the ventromedial prefrontal cortex (VMPFC) correlates both with motivational variables – such as incentives and expected values – and metacognitive variables – such as confidence judgments, which reflect the subjective probability of being correct. At the behavioral level, we recently demonstrated that the value of monetary stakes bias confidence judgments, with gain (respectively loss) prospects increasing (respectively decreasing) confidence judgments, even for similar levels of difficulty and performance. If and how this value-confidence interaction is reflected in the VMPFC remains unknown. Here, we used an incentivized perceptual decision-making fMRI task that dissociates key decision-making variables, thereby allowing to test several hypotheses about the role of the VMPFC in the value-confidence interaction. While our initial analyses seemingly indicate that the VMPFC combines incentives and confidence to form an expected value signal, we falsified this conclusion with a meticulous dissection of qualitative activation patterns. Rather, our results show that strong VMPFC confidence signals observed in trials with gain prospects are disrupted in trials with no – or negative (loss) monetary prospects. Deciphering how decision variables are represented and interact at finer scales seems necessary to better understand biased (meta)cognition.

Introduction

Over the past decades, a growing number of neurophysiological studies in human and non-human primates have established that the neural signals recorded during learning and decision-making tasks in the orbito-medial parts of the prefrontal cortex (OMPFC) – the medial orbitofrontal cortex (OFC) and the ventromedial prefrontal cortex (VMPFC) – correlate with key concepts from theories of motivation and decision-making (Kable & Glimcher, 2009; Padoa-Schioppa, 2007; Rangel & Hare, 2010). For instance, in Pavlovian conditioning tasks, activity of neurons in the non-human primate OFC correlate with the anticipatory value of upcoming rewards, with neural activity predicting the monkeys' subjective preferences (Tremblay & Schultz, 1999). In economic decision-making tasks, neuronal activity in the same region of the OFC correlates with the subjective value of available options (Padoa-Schioppa & Assad, 2006). In humans, similar results have been derived from functional neuroimaging studies. Blood oxygen level-dependent (BOLD) signal in the VMPFC scales with the anticipation of upcoming rewards (Kahnt et al., 2011; Knutson et al., 2003); the subjective pleasantness and desirability attributed to different stimuli (Lebreton et al., 2009); the willingness to pay for different types of goods (Chib et al., 2009; Levy & Glimcher, 2011; Plassmann et al., 2007), and the expected value (EV) of prizes, performance incentives and economic bundles such as lotteries (Gläscher et al., 2009; Hare et al., 2008; Knutson et al., 2005; McNamee et al., 2013). Overall, together with the midbrain and the ventral striatum (VS), the VMPFC seems to form a 'brain valuation system' (Bartra et al., 2013; Haber & Behrens, 2014; Pessiglione & Lebreton, 2015), whose activity automatically indexes the value of available options so as to guide value-based decision-making (Lebreton et al., 2009; Levy & Glimcher, 2011) and motivate motor and cognitive performance (Pessiglione & Lebreton, 2015).

Recently, a set of human neurophysiological studies have suggested that activity in the VMPFC is also related to metacognitive processes (Fleming, Huijgen, et al., 2012; Vaccaro & Fleming, 2018). In particular, both single neuron and BOLD activity in the VMPFC correlates with participants' confidence in their own judgments and choices (De Martino et al., 2012; Lebreton et al., 2015; Lopez-Persem et al., 2020; Shapiro & Grafton, 2020). Confidence is a metacognitive variable that can be defined as one's subjective estimate of the probability of a given choice being correct (Fleming & Daw, 2017; Pouget et al., 2016). Just like values, confidence judgments seem to be automatically represented in the VMPFC, for different types of judgments and choices (Abitbol et al., 2015; Lopez-Persem et al., 2020; Morales et al., 2018). Confidence signals could be useful for the flexible adjustment of behavior – such as monitoring and reevaluating previous decisions (Folke et al., 2017), tracking changes in the

environment (Heilbron & Meyniel, 2019; Vinckier et al., 2016), adapting future decisions (Boldt et al., 2019; Folke et al., 2017), or arbitrating between different strategies (Daw et al., 2005; Donoso et al., 2014).

Interestingly, at the behavioral level, values and confidence seem to interact. For instance, a handful of studies in psychology and economics have documented that positive incentive values, operationalized as prospects of monetary bonuses, increase subjective estimates of confidence (Giardini et al., 2008). Similar confidence boosts have been reported with higher state values, operationalized as positive incidental psychological states such as elevated mood (Koellinger & Treffers, 2015), absence of worry (Massoni, 2014) and emotional arousal (Allen et al., 2016; Jönsson et al., 2005; Kuhnen & Knutson, 2011). Recently, we designed an incentivized perceptual decision-making task to demonstrate that monetary incentives bias confidence judgments, with gain (respective loss) prospects increasing (respectively decreasing) confidence judgments, even for similar levels of difficulty and performance (Lebreton et al., 2018). This result was also replicated in a reinforcement-learning context (Lebreton, Bacily, et al., 2019; C. C. Ting et al., 2020). We explicitly hypothesized that this interaction would stem from the concurrent neural representation of – hence putative interaction between – incentive values and confidence in the VMPFC (Lebreton et al., 2018).

Here, we used a functional neuroimaging adaptation of our original perceptual decision-making paradigm that allows for investigation of the overlap in neural correlates between incentive value and confidence (Lebreton et al., 2018). Our first set of analyses did not show the hypothesized overlap of incentive value and confidence signals in the VMPFC at the expected statistical threshold ($p < 0.05$ whole-brain corrected family wise error (FWE) at the cluster level), nor in other regions of interest (ROI) that have been linked with value, motivation and confidence in the past - such as the VS and the anterior cingulate cortex (ACC). Therefore, we formulated an alternative hypothesis, positing that VMPFC integrates confidence and incentive signals into a probabilistic EV signal. We ran several quantitative and qualitative analyses that thoroughly compared the relative merits of these different hypotheses for the neural basis of the value-confidence interaction. Our results ultimately depict a complex picture, suggesting that motivational signals (notably prospects of loss) can disrupt metacognitive signals in the VMPFC.

Results

To investigate the neurobiological basis of the interactions between incentives and confidence, we modified the task used in Lebreton et al. (2018) to make it suitable for functional magnetic resonance imaging (fMRI) (Figure 1A). Basically, this task is a simple perceptual task (contrast discrimination), featuring a 2-alternative forced choice followed by a confidence judgment. Then, we experimentally manipulated the available monetary outcomes, defining several incentive conditions: at each trial, participants could win (gain context) or lose (loss context) points – or not gain or lose anything (neutral context) – depending on the correctness of their choice. Incentives were presented in an interleaved fashion, in order to avoid contextualization of outcomes (rather than in a blocked design, where absence of gain could be reframed as relative loss in a gain block, or vice versa). Importantly, this incentivization was implemented after the moment of choice and before confidence rating. Consequently, by design, there should not be any incentivization effects on either accuracy or reaction times as they develop during the choice. Note that this design corresponds to the simplest implementation of the task – corresponding to Experiment 2 in Lebreton et al. (2018) –, which otherwise conditioned monetary outcomes to confidence rating precision rather than choice accuracy (for details see (Lebreton et al., 2018)). Yet, our previous results suggested that this task still reveals an effect of incentives on confidence, while keeping instructions simpler – a desirable feature especially for clinical and fMRI studies.

Behavioral results

To start, we verified that our task generated the incentive-confidence interaction at the behavioral level. First, using an approach similar to Lebreton et al. (2018), we used linear mixed-effect models to evaluate the effects of our experimental manipulation of incentives (i.e. the incentive condition) on behavioral variables (see Methods). More specifically, we defined and tested the incentives' biasing effects (i.e., the net incentive value, or in other words, the linear effect of incentives coded as -1, 0 and +1) and incentives' motivational effects (i.e., the absolute incentive value, or in other words, the mere presence of incentives, indicating whether something is at stake coded as 0 and +1. Replicating our previous results, we found a significant positive effect of incentive net value on confidence ($\beta = 0.78 \pm 0.32$, $t_{4317} = 2.43$, $p = 0.015$; Figure 1B, 1C) and no effect of incentive absolute value ($\beta = -0.32 \pm 0.55$, $t_{4317} = -0.58$, $p = 0.565$; Figure 1C). This result alone validates the presence of an incentive-confidence interaction at the behavioral level. Importantly, this effect was not driven by any net incentive value

effects on accuracy or reaction time (RT) (accuracy: β = 0.38 ± 0.93, t_{4317} = 0.41, p = 0.685; RT: β = 13.75 ± 19.22, t_{4317} = 0.72, p = 0.474). Moreover, we did not find evidence for an effect of absolute incentive value on both accuracy and RT (accuracy: β = 1.86 ± 1.45, t_{4317} = 1.28, P = 0.199; RT: β = -25.24 ± 29.17, t_{4317} = -0.87, p = 0.387). Next, to confirm the robustness of our main effect of net incentive value on confidence, we ran several full linear mixed-effects models, which included additional control variables that could influence confidence as well (evidence, accuracy, reaction times, et cetera, see Appendix A). Overall, the incentive-confidence interaction remained significant after accounting for those other potential sources of biases and confounds.

At last, we tested for an incentive effect on metacognitive sensitivity – a metric that measures the efficacy with which subjects discriminate between correct and incorrect answers using their confidence ratings (see Methods for details on its' computation). Replicating earlier findings (Lebreton et al., 2018), we found that incentive condition did not have a significant effect on metacognitive sensitivity ($F(2,62) = 0.25$, $p = 0.783$. Loss: 5.5973 ± 1.2106, neutral: 4.8572 ± 1.0515, gain: 5.2797 ± 0.8692).

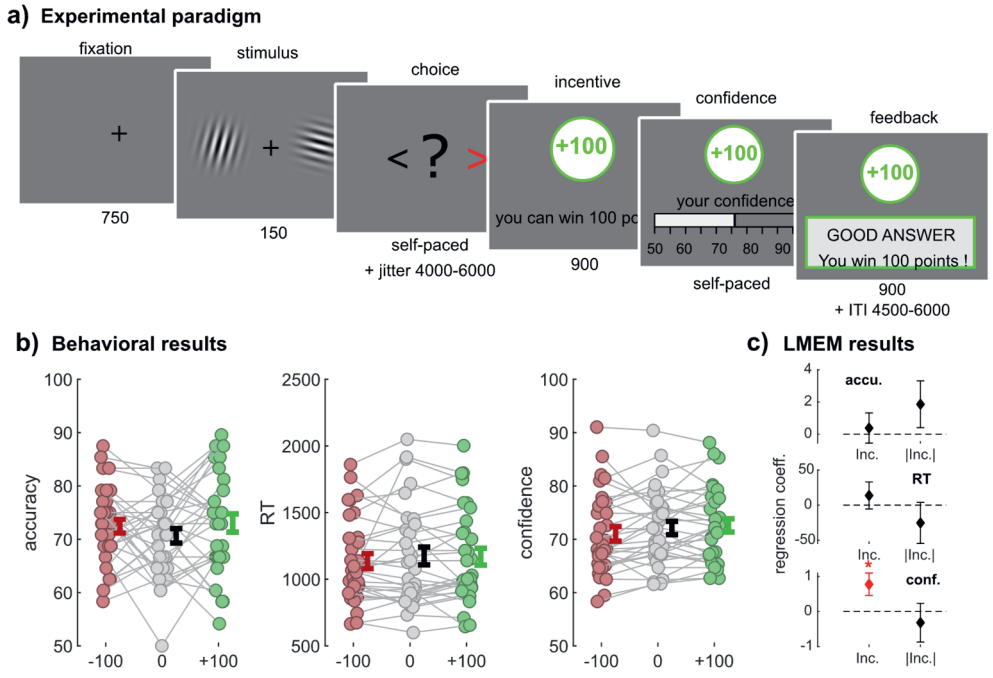


Figure 1: Experimental Design and Behavioral Results. a) Experimental paradigm. Participants viewed two Gabor patches on both sides of the screen (150 ms) and then chose which had the highest contrast (left/right, self-paced). After a jitter of a random interval between 4500 to 6000 ms, the incentive condition was shown (900 ms; green frame for win trials, grey frame for neutral trials, red frame for loss trials). Afterwards, participants were asked to report their confidence in the earlier made choice on a scale ranging from 50% to 100% with steps of 5%. The initial position of the cursor was randomized between 65% and 85%. Finally, subjects received feedback. The inter trial interval (ITI) had a random duration between 4500 and 6000 ms. The calibration session only consisted of Gabor discrimination, without confidence rating, incentives or feedback and was used to adjust difficulty so that every individual reached a performance of 70%. **b)** Behavioral results. Individual-averaged accuracy (left), reaction times (middle) and confidence (right) as a function of incentive condition (-100/red, 0/grey, +100/green). Colored dots represent individuals (N=32), grey lines highlight within subject variation across conditions. Error bars represent sample mean ± standard error of the mean. Note that for confidence and accuracy, we computed the average per incentive level per individual, but that for reaction times, we computed the median for each incentive condition rather than the mean due to their skewed distribution. **c)** Generalized linear mixed-effect regression (GLMER) results. Graph depict fixed-effect regression coefficients (β) for incentive condition (Inc.) and absolute incentive condition (|Inc.|) predicting performance (top), reaction-times (middle) and confidence (bottom). Error bars represent standard errors of fixed effects. * $p < 0.05$

fMRI results

Having established the presence of a robust confidence-incentive interaction at the behavioral level, we next turned to the analysis of the functional neuroimaging data. Critically, our task allowed us to temporally distinguish the moment of stimulus presentation and choice – where the decision value and an implicit estimation of (un)certainty are expected to build up – from the incentive presentation and confidence rating moment – where the explicit, metacognitive confidence signal is expected to interact with the incentive (Figure 2A,B).

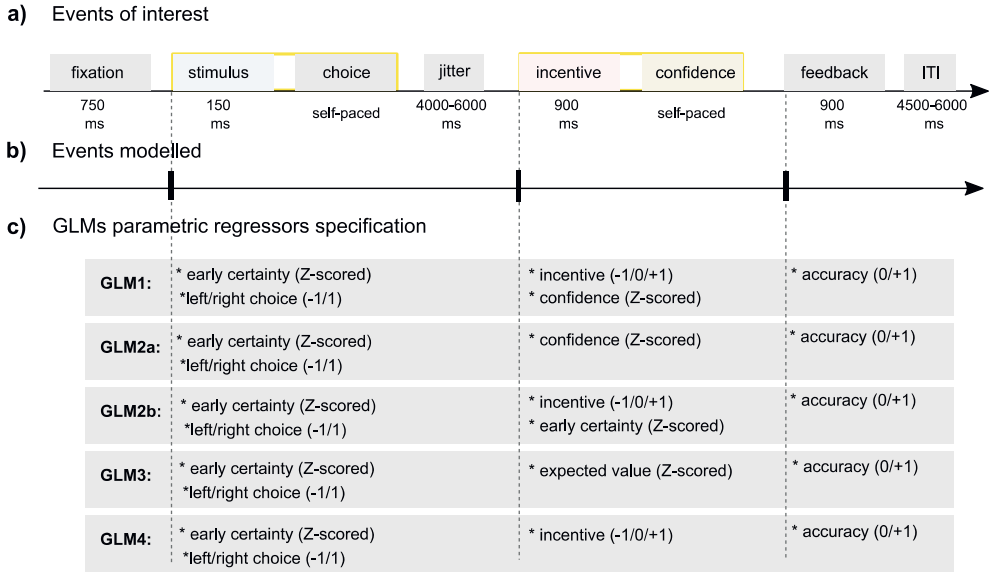


Figure 2: Overview of General Linear Models for fMRI Analyses. a-b) Events of interest. The timeline depicts the succession of events within a trial **a)** Yellow boxes highlight the two events/timing of interest (stimulus/choice and incentive/confidence), that are modelled as stick function for the functional magnetic resonance imaging (fMRI) analysis. We also modelled the feedback event as a stick function. **c)** general linear models (GLMs) parametric regressors specification. The graph displays the different combination of parametric modulators of each event of interest for all GLMs used to analyze the fMRI data.

BOLD signal in the VMPFC correlates significantly with early certainty and incentives but weakly with confidence

Our original hypothesis proposes that incentives bias confidence because those two variables are both correlated to activity in the same brain area – presumably the VMPFC (De Martino et al., 2012; Lebreton et al., 2015). To test this hypothesis, we built a first

fMRI GLM (GLM1) which modeled 1) early certainty during stimulus and choice, and 2) both incentives and confidence ratings during incentive/rating (Figure 2C). Early certainty was defined and computed as the precursor of confidence (i.e., an incentive bias-free signal of confidence), that builds up before the commitment to a choice (see Methods for details). During choice, early certainty positively correlated with activation in the VMPFC and the posterior cingulate cortex (PCC) (Figure 3A). This replicates several studies that have reported an early and automatic (i.e., without explicit instructions) encoding of confidence in the VMPFC (De Martino et al., 2017; Lebreton et al., 2015; Shapiro & Grafton, 2020). Negative correlations of early certainty were observed in a widespread network including the bilateral dorsolateral prefrontal cortex (DLPFC) and rostro-lateral prefrontal cortex (RLPFC), bilateral anterior insula, right putamen, right inferior frontal gyrus, supplementary motor area (SMA), mid- and anterior cingulate cortex and bilateral inferior parietal lobe. This large network has already been implicated in uncertainty and metacognition (Vaccaro & Fleming, 2018).

During the incentive/rating moment, we found positive correlations between incentive value and activity in the VMPFC - extending to clusters in the dorsomedial prefrontal cortex (DMPFC) (Figure 3B). This is in line with our hypothesis and with a large body of neuro-economics literature (Bartra et al., 2013). A small cluster was detected in the occipital lobe, which negatively correlated with incentives.

Finally, regarding subjective confidence, we found significant positive effects in a large, lateralized visuo-motor network including the left primary motor cortex, left putamen and left para-hippocampal gyrus, as well as the right cerebellum and right visual cortex (Figure 3C). All those activations were mirrored in the negative correlation with confidence (although with lower and sometimes subthreshold significance), suggesting these brain regions are part of the visuo-motor network that processes the movement of the cursor on the rating scale (remember that movements of the cursor were operationalized with the left (respective right) index finger to move the cursor toward the left (respective right).

Outside those visuo-motor areas, activity in a large cluster in the dorsal anterior cingulate cortex (dACC) and the mid cingulate cortex (MCC) was found to positively correlate with confidence. Interestingly, an adjacent region of the dACC negatively correlated with early certainty in the choice period (Figure 3A).

To our surprise, and in contradiction with our hypothesis, no whole-brain significant cluster was found in the VMPFC at our a priori defined statistical threshold. There were, however signs of sub-threshold activations (Figure 3C).

As observed with confidence activations, motor-related activity can be an important confound. To ensure that our activity patterns of interest (i.e., early certainty, incentive and confidence) were not related to motor processes, we replicate our analyses using an exclusive motor-related mask, generated from large-scale automated meta-analyses (see Methods for more details). Importantly, those control analyses revealed that most activations – with the exception of the visuo-motor activations identified in the confidence activation maps – remain significantly associated to our variables of interest (for whole-brain activation tables when using this exclusive mask, see Appendix A, Table A13).

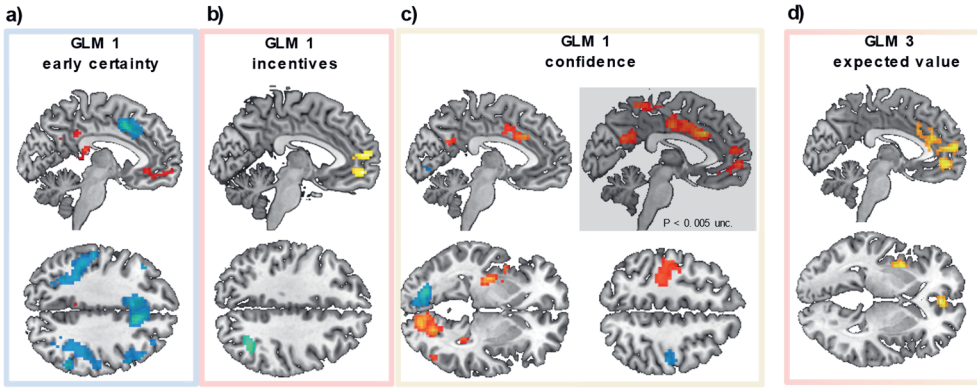


Figure 3: Whole Brain fMRI Results. a-c) Whole brain statistical blood-oxygen level dependent (BOLD) activity correlating with general linear model 1 (GLM1) 'early certainty' (**a**), incentives (**b**) and confidence (**c**). **d)** Whole brain statistical maps of BOLD activity correlating with GLM3 'expected value'. N=30. Unless otherwise specified, all displayed cluster survived $p < 0.05$ family wise error (FWE) cluster correction. Voxel-wise cluster-defining threshold was set at $p < 0.001$, uncorrected. Red/yellow clusters: positive activations. Blue clusters: negative activations. For whole-brain activation tables see Appendix A, Table A12.

Accounting for incentive bias in confidence does not restore VMPFC confidence activations

Next, we attempted to understand the absence of strong correlations with confidence in the VMPFC, despite the same region robustly encoding early certainty and incentives (i.e., precursors of confidence). We reasoned that because confidence is biased by incentive, the shared variance between those two variables could have decreased our chances to reveal clear confidence signals during confidence ratings. We therefore built two control GLMs, which differed in how the incentive/rating period was modelled (Figure 2C): GLM2a only included confidence as a parametric modulator, while GLM2b included incentive and early certainty (i.e., the precursor of confidence devoid of incentive shared variance). We defined an anatomical VMPFC ROI (see Methods and

Figure 4A), and extracted individual standardized regression coefficients (*t*-values) corresponding to the confidence variable in those three GLMs (GLM1, GLM2a, GLM2b) (see Methods). We then tested whether the difference in the GLM specifications had an impact on these activations at the rating period (GLM1 and 2a: confidence; GLM2b: certainty) using repeated measure ANOVAs. Results showed that activations for GLM2a-confidence and GLM2b-early certainty during incentive/rating period were indistinguishable from GLM1-confidence (ANOVA, the main effect of GLMs: $F(2,29) = 0.68$; $p = 0.509$), falsifying the hypothesis that the weak confidence activations in VMPFC observed with GLM1 were due to an ill-specified GLM.

BOLD signal in the VMPFC strongly correlates with expected value

Having established that BOLD activity in the VMPFC only weakly correlates with confidence after the incentive display, we proposed an alternative hypothesis – namely that the VMPFC encodes a signal commensurate to an EV. The rationale of this hypothesis is twofold. First, because confidence represents a subjective probability of being correct, it may be combined with information about the prospective monetary bonus to generate a representation of EV, once this reward information is revealed. Second, activity in the VMPFC has been repeatedly shown to correlate with EV in different contexts (lotteries, et cetera) (Gläscher et al., 2009; Hare et al., 2008; Knutson et al., 2005; McNamee et al., 2013). To test this hypothesis, we built another fMRI GLM similar to the previous ones, but that instead modeled EV at the time of incentive/rating (GLM 3; see Figure 2C).

Whole-brain results showed massive positive correlations between EV and signal in the VMPFC stretching into the anterior medial prefrontal cortex, as well as the ventral and dorsal part of the anterior cingulate cortex and the mid cingulate cortex (Figure 3D, Appendix A, Table A12). There were no activation clusters negatively related to EV.

BOLD signal in the VMPFC correlates better with expected value than with other variables

Although these results seem to validate our second hypothesis, our observation of more activations (wider cluster, lower *p*-values) at the whole-brain level for EV than for confidence does not constitute a formal statistical test that VMPFC signals might rather correlate with EV than with confidence. These results may be owing to incentives and EV being highly correlated – in other words, VMPFC activations to EV could simply be a result of VMPFC activations to incentives. To rule out these hypotheses, we built an additional GLM (GLM4), which only included incentive at the incentive/rating period (Figure 2C). Again, we extracted VMPFC individual standardized regression coefficients (*t*-values) corresponding to the early certainty, incentive and confidence-related

activations in all available GLMs. We tested whether the different specifications had an impact on those activations using repeated measure ANOVAs, and post-hoc *t*-tests (Figure 4, Table 1). Although activations for early certainty during choice moment were similar for all GLMs (ANOVA, main effect of GLM; $F(4,29) = 0.24$, $p = 0.916$; Figure 4B), GLM specification had an impact on both the incentive activations (ANOVA, main effect of GLM; $F(3,29) = 10.67$, $p = 4.837\times10^{-6}$; Figure 4C) and the confidence activations (ANOVA, main effect of GLM; $F(3,29) = 3.22$, $p = 0.027$; Figure 4D) during incentive/rating moment. In both cases, post-hoc *t*-tests showed that *t*-values extracted from the GLM3 that related to the EV regressor were significantly higher than from other GLMs with a different coding of incentives (GLM1 vs GLM3: $t_{29} = 3.90$, $p = 5.306\times10^{-4}$; GLM2b vs GLM3: $t_{29} = 3.38$, $p = 0.002$, GLM4 vs GLM3: $t_{29} = 2.97$, $p = 0.006$), and marginally higher from other GLMs with a different coding of confidence (GLM1 vs. GLM3: $t_{29} = 1.92$, $p = 0.064$; GLM2a vs. GLM3: $t_{29} = 1.72$, $p = 0.096$; GLM2b vs. GLM3: $t_{29} = 2.36$, $p = 0.025$). Overall, these analyses suggest that the VMPFC combines incentive and confidence signals in the form of an EV signal.

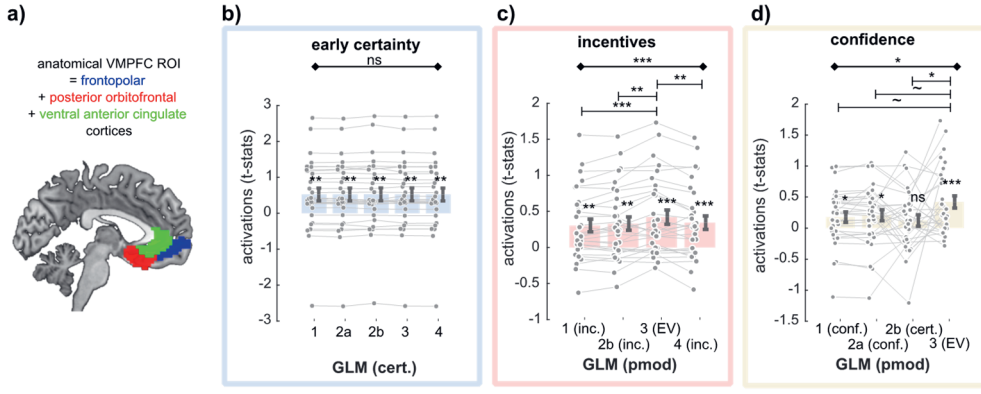


Figure 4: Activation in Ventromedial Prefrontal Cortex Across Models. a) Anatomical ventromedial prefrontal cortex (VMPFC) region of interest (ROI). **b-d** Comparison of VMPFC activations to different specifications of early certainty during choice moment (b), incentives during incentive/rating moment (c) and confidence during incentive/rating moment (d), as implemented in the different GLMs. Dots represent individual activations; bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different specifications. N=30. Cert: early certainty; Inc.: incentives; conf.: confidence; EV: expected value. Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired t-tests. ~ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 1: Comparison of ventromedial prefrontal cortex (VMPFC) parametric activity (*t*-values) as a function of model specification (GLMs)

Early certainty	**GLM1**	**GLM2a**	**GLM2b**	**GLM3**	**GLM4**
	0.52 ± 0.18 t_{29} = 2.92 p = 0.007	0.52 ± 0.18 t_{29} = 2.91 p = 0.007	0.53 ± 0.18 t_{29}= 2.93 p = 0.007	0.53 ± 0.18 t_{29} = 2.93 p = 0.007	0.52 ± 0.18 t_{29} = 2.90 p = 0.007
	RM ANOVA				
	F(4,29) = 0.24 p = 0.916				
Incentive	**GLM1**		**GLM2b**	**GLM3**	**GLM4**
	0.30 ± 0.09 t_{29} = 3.45 p = 0.002		0.33 ± 0.09 t_{29} = 3.60 p = 0.001	0.42 ± 0.10 t_{29} = 4.26 p = 1.981×10^{-4}	0.34 ± 0.09 t_{29} = 3.68 p = 9.433×10^{-4}
	RM ANOVA		*t*-test [3 vs 1]	*t*-test [3 vs 2b]	*t*-test [3 vs 4]
	F(3,29) = 10.67 p = 4.837×10^{-6}		0.12± 0.03 t_{29} = 3.90 p = 5.306×10^{-4}	0.09 ± 0.03 t_{29} = 3.38 p = 0.002	0.08 ± 0.03 t_{29} = 2.97 p = 0.006
Confidence	**GLM1**	**GLM2a**	**GLM2b**	**GLM3**	
	0.18± 0.08 t_{29} = 2.14 p = 0.041	0.21 ± 0.09 t_{29} = 2.30 p = 0.028	0.12 ± 0.09 t_{29} = 1.35 p = 0.187	0.42 ± 0.10 t_{29} = 4.26 p = 1.981×10^{-4}	
	RM ANOVA		*t*-test [3 vs 1]	*t*-test [3 vs 2a]	*t*-test [3 vs 2b]
	F(3,29) = 3.22 p = 0.027		0.24 ± 0.13 t_{29} = 1.92 p = 0.064	0.21 ± 0.12 t_{29} = -1.72 p = 0.096	0.30 ± 0.13 t_{29} = 2.36 p = 0.025

The table reports descriptive and inferential statistics on VMPFC region of interest (ROI) parametric activations with three different variables of interest: early certainty effects at choice moment, incentive effects at rating moment and confidence effects at rating moment (see Figure 4). Per effect of interest, results of one-sample *t*-tests against zero, repeated-measure (RM) ANOVAs on the main effect of GLMS, and post-hoc *t*-test results are shown.

Qualitative falsification of the EV model of VMPFC activity

At last, in order to confirm the conclusions drawn from our quantitative comparison of VMPFC activations, we ran a qualitative falsification exercise (Palminteri et al., 2017). Leveraging the factorial design of our experiment, we could draw qualitative patterns of activations that would be expected under different hypotheses underlying VMPFC activation (Figure 5A).

To this end, we designed a final GLM (GLM5) that divided the task in two time points (stimulus/choice and incentive/rating), and three incentive conditions, and that incorporated a baseline and a regression slope with confidence judgment for all these events. We then extracted the VMPFC activations for all these regressors using our ROI,

and compared them with the theorized qualitative patterns we would expect if the VMPFC encoded one of these variables (Figure 5B,C and Table 2, Table 3). As expected, at the moment of the stimulus/choice, there was no effect of incentive conditions on VMPFC baseline activity, nor on its correlation with confidence – "slope" (ANOVA baseline: $F(2,29) = 0.36$, $p = 0.701$; ANOVA correlation with confidence: $F(2,29) = 0.56$, $p = 0.574$). Basically, the slopes were significantly positive in all three incentive conditions (Loss: $t_{29} = 2.10$, $p = 0.045$; Neutral: $t_{29} = 2.43$, $p = 0.021$; Gain: $t_{29} = 3.04$, $p = 0.005$), confirming that the VMPFC encodes an early certainty signal.

At rating moment, incentive conditions had an effect on both VMPFC baseline activity, and on the correlation of VMPFC activity with confidence (ANOVA baseline: $F(2,29)= 8.56$, $p = 5.543\times10^{-4}$; ANOVA correlation with confidence: $F(2,29)= 5.26$, $p = 0.008$). Post-hoc testing revealed that VMPFC baseline activity was significantly larger in gain versus loss ($t_{29} = 3.47$, $p = 0.002$) and in gain versus neutral conditions ($t_{29} = 3.17$, $p = 0.004$), but not in neutral versus loss condition ($t_{29} = 0.43$, $p = 0.673$) (see Table 3). This constitutes a deviation from a standard linear model of incentives, and suggest that different regions might process incentives in gains and loss contexts (Palminteri & Pessiglione, 2017).

Moreover, we found that the correlation of VMPFC activity with confidence is significantly positive in the gain condition only ($t_{29} = 3.29$, $p = 0.003$), and not in the loss ($t_{29} = -0.75$, $p = 0.457$) nor neutral ($t_{29} = 0.70$, $p = 0.491$) conditions. The correlation with confidence was therefore significantly higher in gain versus loss ($t_{29} = 3.13$, $p = 0.004$) and in gain versus neutral conditions ($t_{29} = 2.02$, $p = 0.053$), but not in neutral versus loss condition ($t_{29} = 1.03$, $p = 0.313$). Although the absence of correlation in the neutral condition would be expected if the VMPFC encodes EV, the lack of correlation in the loss condition was not predicted by any of our models (Figure 5A). Because VMPFC confidence activations were robustly observed in the gain domain, as well as VMPFC early certainty activations in all three conditions, we suggest that the lack of VMPFC confidence activations in the neutral and loss conditions is a feature of the VMPFC signal, rather than a failure of our design to elicit those activations (e.g., due to limited statistical power or excessive statistical noise).

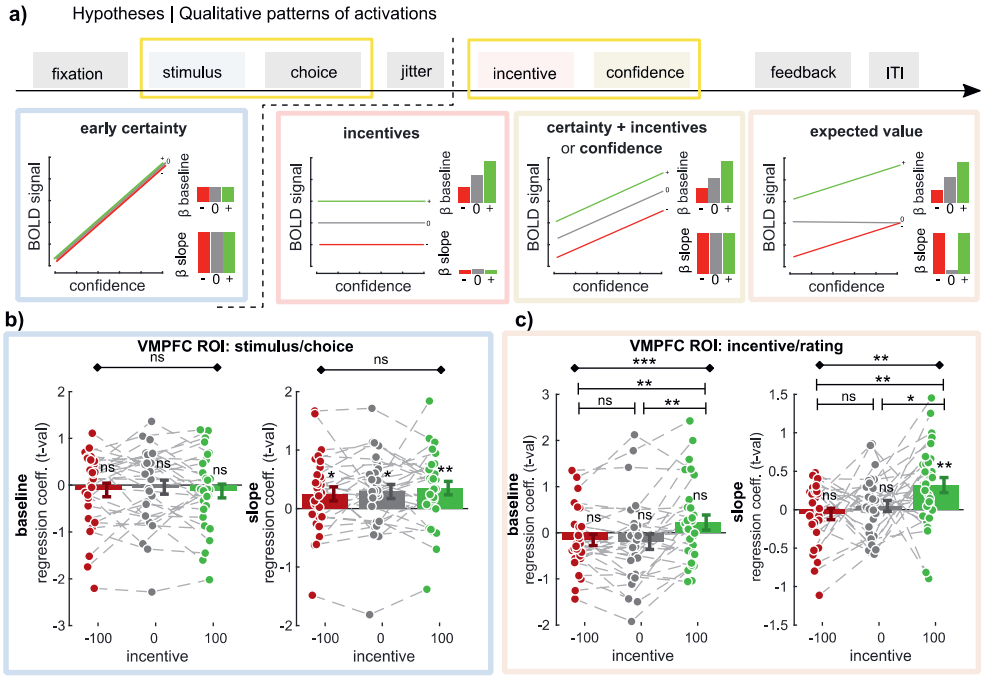


Figure 5: Activation in Ventromedial Prefrontal Cortex across Incentives and Timepoints. a) Qualitative ventromedial prefrontal cortex (VMPFC) activation patterns predicted under different models. The different boxes present how blood-oxygen level dependent (BOLD) signal should vary with increasing confidence in the three incentive conditions (green: +100; grey: 0; red: -100), under different hypotheses (i.e., encoding different variables), at different time points. Bar graphs in insets summarize these relationships as expected intercepts (or baseline – top) and slope (bottom). **b-c)** VMPFC region of interest (ROI) analysis (N=30). *T*-values corresponding to baseline and regression slope were extracted in the three incentive conditions, and at the two time-points of interest (b: stimulus/choice; c: incentive/rating). Dots represent individual activations; bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different incentive conditions. Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired *t*-tests. ns: P>0.05; * P<0.05; ** P<0.01; *** P < 0.001

Table 2: Comparison of ventromedial prefrontal cortex (VMPFC) activity at choice moment (*t*-values), as a function of incentive condition

Choice/Stim	**baseline**	Inc. -100	Inc. 0	Inc. +100	RM ANOVA
		-0.10 ± 0.15 t_{29} = -0.70 p = 0.490	-0.04 ± 0.15 t_{29} = -0.30 p = 0.770	-0.13 ± 0.15 t_{29} = -0.85 p = 0.400	$F(2,29)$ = 0.36 p = 0.701
	slope	Inc. -100	Inc 0	Inc. +100	RM ANOVA
		0.250 ± 0.12 t_{29} = 2.10 p = 0.045	0.29 ± 0.12 t_{29} = 2.43 p = 0.021	0.35 ± 0.12 t_{29} = 3.04 p = 0.005	$F(2,29)$ = 0.56 p = 0.576

The table reports descriptive and inferential statistics on VMPFC region of interest (ROI) parametric activations in our three incentive conditions during choice moment, for both baseline activity as well as the correlation with early certainty (i.e., slope) (see Figure 5B). Results of repeated measures (RM) ANOVAs and one-sample *t*-tests against 0 are shown. Inc. = incentive.

Table 3: Comparison of ventromedial prefrontal cortex (VMPFC) activity at rating moment (*t*-values), as a function of incentive condition

Incentive/rating	**baseline**	Inc. -100	Inc. 0	Inc. +100	RM ANOVA
		-0.16 ± 0.12 t_{29} = -1.31 p = 0.20	-0.20 ± 0.17 t_{29} = -1.19 p = 0.25	0.22 ± 0.16 t_{29} = 1.37 p = 0.18	$F(2,29)$ = 8.56 p = 5.543×10^{-4}
		t-test [-100 vs 0]	*t*-test [0 vs 100]	*t*-test [-100 vs 100]	
		0.04 ± 0.09 t_{29} = 0.43 p = 0.673	-0.42 ± 0.13 t_{29} = 3.17 p = 0.004	-0.38 ± 0.11 t_{29} = 3.47 p = 0.002	
	slope	Inc. -100	Inc. 0	Inc. +100	RM ANOVA
		-0.06 ± 0.07 t_{29} = -0.75 p = 0.457	0.05 ± 0.07 t_{29} = 0.70 p = 0.491	0.32 ± 0.10 t_{29} = 3.29 p = 0.003	I(2,29) = 5.26 p = 0.008
		t-test [-100 vs 0]	*t*-test [0 vs 100]	*t*-test [-100 vs 100]	
		-0.11 ± 0.10 t_{29} = 1.03 p = 0.313	-0.27 ± 0.13 t_{29} = 2.02 p = 0.053	-0.38 ± 0.12 t_{29} = 3.13 p = 0.004	

The table reports descriptive and inferential statistics on VMPFC region of interest (ROI) parametric activations in our three incentive conditions during rating moment, for both baseline activity as well as the correlation with confidence (i.e., slope) (see Figure 5C). Results of one-sample *t*-tests against 0, repeated measures (RM) ANOVAs and post-hoc *t*-tests are shown. Inc. = incentive.

To evaluate whether the lack of robust confidence activation in the neutral and loss condition could be caused by the rough averaging of the VMPFC signal over the anatomical ROI, we also performed a finer-grained analysis. We extracted confidence activations in the three conditions and two time-points at the voxel level in a large anatomical area covering most of the medial prefrontal cortex, averaged those activations over two dimensions (respectively X and Z, and X and Y), and assessed how activations unfold over the last dimension – respectively Y and Z (Figure 6). This last analysis confirmed three main facts: first, the early certainty activations are robustly observed in the same portion of the VMPFC, and – as expected – with similar effect sizes in the three conditions; second, the confidence activations in the gain condition are observed at similar levels as the early certainty activations, confirming that our experimental design elicits robust activations at the incentive/confidence rating time-point; third, no confidence activations can be detected at this finer-grained level in the neutral or loss condition, in the VMPFC. If anything, it seems that the confidence activations in the loss condition trend toward a negative correlation between VMPFC BOLD signal and confidence.

Overall, these results initially explain why EV appears a better model of VMPFC activation than confidence and/or incentive (correct pattern in gains and neutral conditions), but ultimately falsify this account by demonstrating the absence of positive correlation between VMPFC activation and confidence in the loss condition.

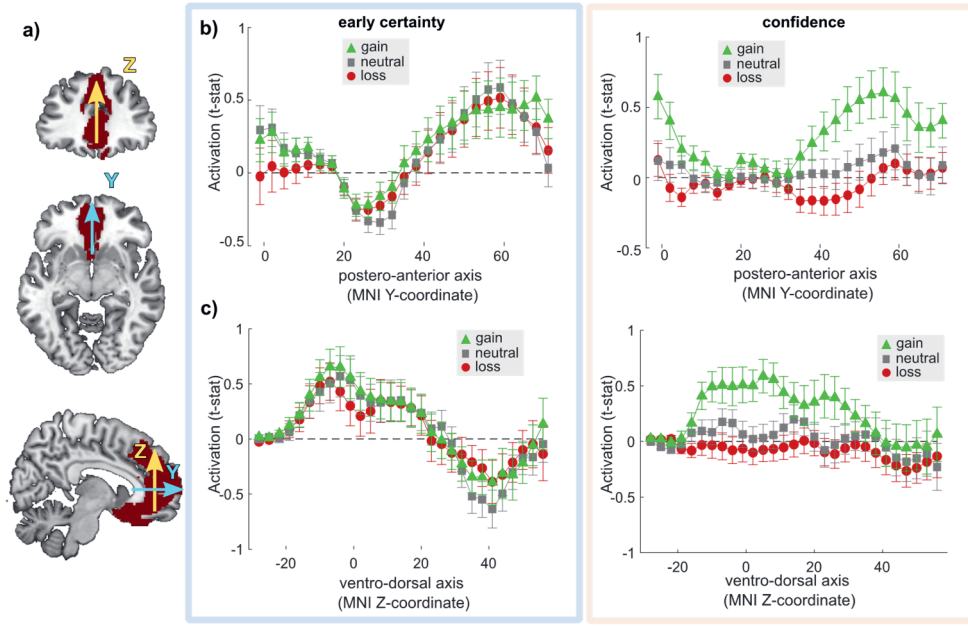


Figure 6: Activation in Ventromedial Prefrontal Cortex across Y and Z dimensions. a) Large anatomical medial prefrontal cortex region of interest (ROI). The Y (blue) and Z (yellow) arrows indicate the dimensions over which the signal is extracted and marginalized – respectively corresponding to the postero-anterior axis and ventro-dorsal axis. **b-c.** MPFC region of interest (ROI) analysis of confidence activations, at the voxel-level, marginalized over the Y (**b**) and Z (**c**) dimensions. Voxel-wise *T*-values corresponding to regression slope were extracted in the three incentive conditions (green: +100; grey: 0; red: -100), and at the two time-points of interest (left: stimulus/choice; right: incentive/rating), averaged over two dimensions and plotted as a function of the third dimension. Dots and error bars indicate sample mean ± standard error of the mean (N=30).

Discussion

In this study, we set out to investigate the neural signature of incentive bias on confidence estimations, using an fMRI-optimized version of an incentivized perceptual decision-making task (Lebreton et al., 2018). First, at the behavioral level, we replicated the biasing effect of incentives on confidence estimation, in the form of higher confidence in gain contexts and lower confidence in loss context – despite equal difficulty and performance. This result is the fourth independent replication of this bias, initially revealed in perceptual decision making and later generalized in a reinforcement-learning task (Lebreton, Bacily, et al., 2019; C. C. Ting et al., 2020). Note, however, that the bias' effect size remains small – a few average confidence percentage points at the population level –, what a priori limits our ability to dissect its precise

neurophysiological basis with current (correlational) functional neuroimaging techniques.

Our initial goal and hypothesis were therefore quite simple and modest. In the literature, it is now well established that the BOLD signal in the VMPFC correlates with confidence and/or values in a variety of tasks (De Martino et al., 2012, 2017; Lebreton et al., 2015; Lopez-Persem et al., 2020; Morales et al., 2018; Shapiro & Grafton, 2020). We reasoned that if we could provide evidence for the presence of both incentive and confidence signals in the VMPFC during our task, this would reinforce the intuition that the VMPFC has a role in the observed behavioral phenomenon, i.e., the incentive bias on confidence. Our neuroimaging predictions were that 1) the VMPFC should correlate with early certainty before and during choice, regardless of the context, and 2) the VMPFC should integrate confidence and incentive after the choice and the revealing of the incentive condition. Our broader, speculative neural hypothesis was that during this last confidence judgment step, a third-party metacognitive region or network would sample signal in the VMPFC (Meyniel, Sigman, et al., 2015; Shekhar & Rahnev, 2018), and incidentally end up with a biased confidence estimate incorporating incentive signal. Our limited sample size combined with some known limits of brain-behavior analyses (Lebreton, Bavard, et al., 2019) restricted a priori any ambition to validate a neurobiological model of the observed confidence bias by running inter-individual correlations between VMPFC activations and the confidence bias estimated at the behavioral level.

Our fMRI investigation of the neural correlates of early certainty confirms our first prediction: BOLD activity in the VMPFC positively correlates with early certainty in all conditions. This result replicates and extends previous studies demonstrating this area to be associated to the initial and automatic processing of confidence during choice (De Martino et al., 2012; Lebreton et al., 2015; Shapiro & Grafton, 2020). In parallel with this positive correlation in the VMPFC, we also observed wide-spread negative correlations in the DLPFC, DMPFC and insula, a network robustly associated with both metacognition and uncertainty (Molenberghs et al., 2016; Morales et al., 2018; Vaccaro & Fleming, 2018). Contrary to our second prediction, we only found weak evidence (i.e., at a lower statistical threshold than the one we defined a priori) for confidence encoding in the VMPFC. Robust activations were nonetheless observed in the dACC, a region known to be recruited in metacognitive judgments (Bang & Fleming, 2018; Fleming, Huijgen, et al., 2012).

Given that the lack of robust confidence signal in the VMPFC is somewhat in contradiction with what we expected from our previous work, as well as numerous other reports in the literature (De Martino et al., 2012, 2017; Lebreton et al., 2015;

Lopez-Persem et al., 2020; Morales et al., 2018; Shapiro & Grafton, 2020), we formulated an alternative hypothesis: we proposed that VMPFC could encode a signal commensurate to an expected reward (or EV), i.e., incorporating the subjective probability of being correct with the potential incentive bonus when revealed. Whole-brain activations and ROI quantitative analyses clearly showed that this second hypothesis seems to give a better account of VMPFC BOLD activations. EV signals are frequently reported in the VMPFC, but mostly in reinforcement-learning contexts, where they are critical to both choices between available options and learning – i.e., value updating, through the computation of prediction errors (Chase et al., 2015). In the present perceptual task, there is no learning, therefore no explicit need to encode EV.

Because quantitative comparisons of hypotheses are notoriously hard to interpret, we decided to leverage the factorial aspect of our design to proceed to a qualitative hypothesis falsification, to validate – or falsify – the EV account of VMPFC activity (Palminteri et al., 2017). In short, different hypotheses about what should be contained in VMPFC signal (EV, confidence and/or incentives) predict different patterns of activations (baseline and correlation with confidence) in our different incentive conditions. From activity extracted from an anatomical VMPFC ROI, it is clear that VMPFC activity correlates with confidence only in the gain context, once the incentive has been revealed. This finding explains why the EV hypothesis obtained stronger quantitative support than the confidence and/or incentives hypotheses (as the VMPFC activity pattern is similar to the EV predictions in the gain and neutral context). However, it also ultimately falsifies this EV hypothesis as well, as VMPFC activity does not seem to correlate with confidence in the loss context. Interestingly, VMPFC does correlate with early certainty – a precursor of confidence – in all conditions before the incentives are revealed. Therefore, it does not seem that the VMPFC fails to activate in the neutral and loss conditions, but rather that the signal is actively suppressed once those contexts are explicit. Moreover, the fact that we do not observe confidence activations in neutral or loss condition is also not due to the fact that participants are less focused on evaluating confidence in those conditions compared to the gain condition, as we showed that the confidence sensitivity is identical in all incentive conditions. In summary, we believe that our results show a complex picture of disruptions of confidence signals within the VMPFC in response to motivational signals.

The absence of VMPFC confidence signal in the neutral condition might seem at odds with other studies that report such signal in non-incentivized tasks such as pleasantness or desirability ratings (Lebreton et al., 2015). One possible explanation is that VMPFC confidence signals, like attentional modulation of evidence integration (Sepulveda et al., 2020), are primarily observed for behavior or conditions that are

relevant to participants' goals: in non-incentivized tasks such as pleasantness or desirability ratings, participants still have a goal, which is to provide ratings that are as accurate as possible. In our task, if the goal of participants is to maximize their score, the neutral condition might not be goal-relevant, which could result in a disrupted VMPFC confidence signal. Note that because our design features interleaved (rather than blocked) conditions, the valence manipulation is somewhat exacerbated, as the succession of the different conditions limit the contextualization of outcomes (whereby the absence of loss could be reframed as a relative gain in a loss-block). Also, because trials featuring gains, losses and neutral incentives follow each-others in a pseudorandomized order, the interleaved design also prevent any systematic bias or confound for the valence effects (at the behavioral or neurobiological levels) that could be due to the processing of the feedbacks (gains, losses or nothing).

The notion that there are different brain networks which execute symmetric computations in gains versus loss contexts is increasingly popular (Palminteri & Pessiglione, 2017; Seymour et al., 2015). Because the positive, gain context network also typically includes the ventral striatum (VS; see e.g. (Bartra et al., 2013; Knutson et al., 2005), we replicated all analyses using an anatomical VS ROI (see Appendix A). These analyses qualitatively rendered very similar results to what we observed in the VMPFC. In the present dataset though, we did not find any region correlating either positively or negatively with confidence in the loss context, even when exploring the whole brain level with very lenient statistical thresholds. The dACC is a promising area, since it has repeatedly been associated with loss anticipation and correlated positively with subjective confidence in our data. However, when we performed a similar falsification exercise within the dACC as we used within the VMPFC (see Appendix A), the results were similar to the VMPFC activation patterns: dACC activity only correlated with confidence within the gain contexts. In summary, it remains an open question what the neurobiological correlates of confidence judgments in loss contexts are.

Our results constitute a stepping stone and have important implications for studying clinical populations where these (meta)cognitive processes go awry. It shows that motivational processes can influence confidence, and when there are discrepancies between one's behavior and confidence in that behavior, this could give rise to pathological decision making. Indeed, several psychiatric disorders such as addiction, obsessive-compulsive disorder and schizophrenia have been associated with disrupted incentive processing (Admon et al., 2012; Choi et al., 2012; Clark et al., 2019; Koob & Volkow, 2016; Strauss et al., 2014) and studies have additionally demonstrated distorted confidence estimations in these groups (Hoven et al., 2019). Our study indicates that the VMPFC is a key region involved in the interaction between motivation

and metacognition, and VMPFC function is also often affected in many psychiatric disorders (Hiser & Koenigs, 2018). The current study provides a means of studying neurobiological explanations for confidence abnormalities and their interaction with incentive motivation in the clinical population which can potentially impact clinical practice, as it could help treat psychopathology (Hiser & Koenigs, 2018). Therefore, the relationship between motivational processes and confidence estimation and their role in psychopathology warrants future investigation.

In conclusion, we show that although the VMPFC seems to encode both value and metacognitive signals, these metacognitive signals are only present during the prospect of gain and are disrupted in a context with loss or no monetary prospects. Studies targeting this problem within a finer spatial (Kepecs & Mainen, 2012; Lopez-Persem et al., 2020; Middlebrooks et al., 2013) and/or temporal scale (Desender et al., 2016) could help with resolving and better comprehending biased confidence judgments and metacognition overall.

Methods

Participants

We included 33 right-handed healthy participants with normal or corrected to normal vision. Exclusion criteria were an IQ below 80, insufficient command of the Dutch language or MRI contraindications. All experimental procedures were approved by the Medical Ethics Committee of the Academic Medical Center, University of Amsterdam (METC 2015_319) and participants gave written informed consent. Participants were compensated with a base amount of €40 and additional gains based on task performance. Session-level behavioral and fMRI data were excluded when task accuracy was below 60% or when subjects did not show sufficient variation in their confidence reports (standard deviation of confidence judgments < 5 confidence points), and session-level fMRI data when participants showed head movements > 3.5 mm. This led to the inclusion of 32 participants (18/14 females/males, 18-58 years old (sd: 9.76)) for the behavioral analyses and 30 for the fMRI analyses, of which four participants contributed only one of two task sessions.

Decision-making and confidence judgment task

We adapted the task from Lebreton et al. (2018) for use in an fMRI environment with fMRI suitable timing intervals. For an overview and details, see Figure 1A. All tasks used in this study were implemented using MATLAB® (MathWorks Inc., Sherborn, MA, USA) and the COGENT toolbox (www.vislab.ucl.ac.uk/cogent.php).

Study procedure

On the day of testing, subjects were first assessed for clinical and demographic data, after which they performed one practice session (10 trials) outside of the scanner and another one inside the scanner to become acquainted with the task. Subjects were instructed that they would only be rewarded based on their performance (i.e., they should be as accurate as possible to maximize their earnings), and that it was important to give accurate confidence judgments. They were notified that 50% confidence would signal that they made a guess, whereas 100% confidence would signal that they were absolutely certain that they made the correct choice. Thus, performance but not confidence was incentivized. According to our previous findings (Lebreton et al., 2018), this design elicits incentive bias on confidence while keeping confidence sensitivity identical across conditions – an important consideration when interpreting differences in confidence activations between those conditions. All subjects initially performed a 144-trial calibration session inside the scanner to tailor the difficulty levels of the task to each individual and to keep performance constant across subjects. This was done using a staircase procedure, which data were used to estimate a full psychometric function, whose parameters were used to generate stimuli for the main task, spanning three difficulty levels (i.e. 65%, 75% and 85% accuracy, on average) (for details see (Lebreton et al., 2018)).

Two sessions of the main task were performed in the fMRI scanner, each consisting of 72 trials with 24 trials per incentive condition, presented in a random order. The practice task, calibration and main sessions were projected onto an Iiyama monitor in the fMRI environment, which subjects could see through a 45-degree angle mirror fixed to the head coil. After completing the fMRI task, six random trials were drawn (i.e., two of each incentive condition) on which the payment was based. If subjects made an accurate choice, they would either gain or avoid losing points, whereas they would miss out on gaining or losing points when making an error. In the neutral trials, nothing was at stake. Finally, the total amount of points was converted to money.

Behavioral measures

We extracted various trial-by-trial experimental factors (evidence, incentive and difficulty level) and behavioral measures (accuracy, subjective confidence ratings, reaction times). Control analyses were performed to confirm the properties of confidence ratings (Appendix A). Three additional variables were computed as combinations of those experimental factors and behavioral measures: early certainty, EV and metacognitive sensitivity.

Early certainty

We built an "early certainty" variable that represents a confidence signal prior to the biasing effects of incentives. We assume that such an early certainty signal should be encoded automatically at the moment of choice, in turn allowing us to investigate confidence signals with and without incentive bias (Lebreton et al., 2015). Importantly, such a signal should be highly correlated with the later, biased confidence judgment obtained from the subjects, while exhibiting no statistically significant relationship with incentives. Therefore, we used a leave-one-trial-out approach to obtain trial-by-trial estimations of early certainty (Bang & Fleming, 2018). We fitted a generalized linear regression model to each subject's subjective confidence ratings using choice and stimulus features as predictors (i.e., log-transformed reaction times, evidence, accuracy and the interaction between accuracy and evidence), using the whole individual dataset but trial *X*. We then applied this model's estimates to generate predictions about the early certainty in trial *X*, using the choice and stimulus features of trial *X*. This process was repeated for every trial, resulting in a trial-by-trial prediction of early certainty based on stimulus features at choice moment. The resulting early certainty signal featured high correlation with confidence, and no statistical relationship with incentives (see Appendix A for more details). Importantly, since the early certainty signal follows the main properties of confidence judgments (Appendix A, Figure A6), but does not show any incentive bias, this critically enables us to differentiate between non-biased confidence signals during decision-making and biased confidence signals after incentivization.

Expected value

We computed a value-based measure of EV. In our task paradigm, EV was computed as an integrative signal of early certainty (i.e., the non-biased probability of being correct) and the incentive value (i.e., the value-context of the current trial). Early certainty ratings represent the subjects' probability of being correct, and thus the probability of gaining (or avoid losing) the incentive at stake. Thus, EV corresponds to 0 in the neutral condition (no value is expected to be gained or lost), is equal to early certainty in the

gain condition (e.g., being 100% certain results in a maximal EV in a positive incentive environment), and is equal to early certainty – 100 (e.g., being 100% certain in a loss trial results in an EV of 0, as you avoid losing).

Metacognitive sensitivity

Metacognitive sensitivity is a metric that indicates how well an observer's confidence judgments discriminate between their correct and incorrect answers and can be represented using several indexes. For example, discrimination is a metric calculated as the difference between the average confidence for correct answers and the average confidence for incorrect answers, whereas meta-d' is a metric based on the Signal Detection Theory framework (Maniscalco & Lau, 2012). Notably, meta-d' computations are known to be imprecise in designs with a low number of trials per condition (Rouault, McWilliams, et al., 2018). This, together with results from our earlier work (Lebreton et al., 2018) showing high correlations between discrimination and meta-d', as well as identical conclusions with respect to the effects of incentives on these measures, lead to us using the discrimination metric as our measure of metacognitive sensitivity.

fMRI acquisition & preprocessing

fMRI data was acquired by using a 3.0 Tesla Intera MRI scanner (Philips Medical Systems, Best, The Netherlands). Following the acquisition of a T1-weighted structural anatomical image, 37 axial T2*-weighted EPI functional slices sensitive to BOLD contrast were acquired. A multi echo (3 echoes) combine interleaved scan sequence was applied, designed to optimize functional sensitivity in all parts of the brain (Poser et al., 2006). The following imaging parameters were used: repetition time (TR), 2.375 seconds; echo times (TEs), 9.0ms, 24.0ms, and 43.8ms, (total echo train length: 75ms); 3 mm (isometric) voxel size; 37 transverse slices; 3 mm slice thickness; 0.3 mm slice-gap. Two experimental sessions were carried out, each consisting of 570 volumes. All further analyses were performed using MATLAB® with SPM12 software (Wellcome Department of Cognitive Neurology, London, UK).

Raw multi-echo functional scans were weighed and combined into 570 volumes per scan session. During the combining process, realignment was performed on the functional data by using linear interpolation to the first volume. The first 30 dummy scans were discarded. The remaining functional images were co-registered with the T1-weighted structural image, segmented for normalization to Montreal Neurological Institute (MNI) space and smoothed using a Gaussian kernel of 6 mm at full-width at half-maximum.

Due to sudden motion, in combination with the interleaved scanning method, a number of subjects showed artifacts in some functional volumes. In order to reduce those artifacts, the Art-Repair toolbox (Mazaika et al., 2007) was used to detect large volume-to-volume movement and repair outlier volumes. The toolbox identifies outliers by using a threshold for the variation of the mean intensity of the BOLD signal and a volume-to-volume motion threshold. A threshold of 1.5% variation from the mean intensity was used to detect and repair volume outliers by interpolating from the adjacent volumes (n=12).

Statistics and reproducibility: behavioral analyses

All behavioral analyses were performed using MATLAB® and the R environment (RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA). For the statistical analyses reported in the main text, we used linear mixed-effects models (estimated with the fitglme function in MATLAB®) to model accuracy, reaction times and confidence. In order to analyze the effect of the incentive condition (i.e., of our experimental manipulation of incentives), for all three trial-by-trial dependent variables we used the absolute incentive value (i.e., the absolute value of the monetary incentive, |V|, coded as 0 and +1) and the net incentive value (i.e., the linear value of the monetary incentive, V, coded as -1, 0 and +1) as predictor variables. All mixed models included random intercepts and random slopes (N=32). Additional control analyses are reported in Appendix A. For the analysis of metacognitive sensitivity, we performed a repeated measures ANOVA, with net incentive value as within-subject factor.

Statistics and reproducibility: fMRI analyses

All fMRI analyses were conducted using SPM12. All general linear models (GLMs) were estimated on subject-level (N=30) with two moments of interest: the moment of choice (i.e., presentation of the Gabor patches) and the moment of incentive presentation/confidence rating (Figure 2). The rating moment follows the presentation of the incentive after 900 ms, hence the decision to analyze them as a single moment of interest. Moreover, the GLMs also included a regressor for the feedback moment, which was not of interest for analysis, but was intended to explain variance in neural responses related to value and accuracy feedback, but unrelated to the decision-making process.

When using parametric modulators in our GLMs, those were not orthogonalized and competed to explain variance. Nuisance regressors consisting of six motion parameters were included in all GLMs. Regressors were modeled separately for each scan session and constants were included to account for between-session differences in mean activation. All events were modeled by convolving a series of delta functions with the canonical hemodynamic response function (HRF) at the onset of each event and were linearly regressed onto the functional BOLD-response signal. Low frequency noise was filtered with a high pass filter with a cut off of 128 seconds. All contrasts were computed at subject level and taken to a group level mixed effect analysis using one-sample *t*-tests.

We controlled for the number of sessions while making the first-level contrasts. We assessed group-level main effects by applying one-sample *t*-tests against 0 to these contrast images. All whole-brain activation maps were thresholded using family wise error correction for multiple correction (FWE) at cluster level ($p_{FWE\_clu} < 0.05$), with a voxel cluster-defining threshold of $p < .001$ uncorrected.

GLM1: neural signatures of certainty, incentive and confidence

GLM1 consisted of three regressors for the three moments of interest: 'choice', 'incentive/rating' and 'feedback', to which one or more parametric modulators (pmod) were added (Figure 2). The regressors were specified as stick function time-locked to the onset of the events. The choice regressor was modulated by two pmods: early certainty (*z*-scored before entering the GLM) and button press (left/ right choice) in order to control for activity related to motor preparation. The incentive/rating regressor was modulated by two pmods: incentive value and subjective confidence level (*z*-scored). Lastly, the feedback regressor was modulated by a pmod of accuracy.

Importantly, to ensure that our brain activations of interest (i.e., related to early certainty, incentive and confidence) were not confounded by motor-related activations, we performed control analyses that implemented an exclusive masking for motor activations. To do so, we generated the exclusive mask from 'Neurosynth' (a platform for large-scale, automated synthesis of fMRI data (Yarkoni et al., 2011)), using the term 'motor' (https://neurosynth.org/analyses/terms/motor/). This mask represents key regions related to motor processes as identified by an automated meta-analysis of 2565 studies.

GLM2a: control for incentive bias 1

GLM2a consisted of the same regressors as GLM1, except that rating moment was only modulated by confidence judgments (i.e., we deleted the incentive modulator).

GLM2b: control for incentive bias 2

GLM2b consisted of the same regressors as GLM1, except that the pmod of confidence judgments at rating moment was replaced by a pmod for early certainty.

GLM3: neural signatures of expected value

GLM3 consisted of the same regressors as GLM1, except that rating moment was modulated by a single pmod of EV.

GLM4: control for incentive

GLM4 consisted of the same regressors as GLM1, except that rating moment was only modulated by incentives (i.e., we deleted the confidence judgment modulator).

GLM5: qualitative patterns of activations

GLM5 included a regressor for all three incentives at two timepoints of interest: choice and rating moment, as well as a regressor at feedback moment. All regressors at choice moment were modulated by a pmod of early certainty and button press (L/R). All regressors at rating moment were modulated by a pmod of confidence judgment. The feedback regressor was modulated by accuracy. This GLM allowed us to investigate activity related to both baseline and the regression slope with early certainty or confidence judgment for these events.

Regions of interest

To avoid circular inference, we took an independent anatomical ROI of the VMPFC from the Brainnetome Atlas (Fan et al., 2016). We included three areas along the ventral medial axis for the VMPFC ROI. Using this ROI, we extracted individual *t*-statistics (i.e., normalized beta estimates (Lebreton, Bavard, et al., 2019)) from contrasts of interest, and statistically compared them using paired *t*-tests or repeated measure ANOVAs.

Moreover, in order to perform a finer-grained analysis into early certainty and confidence activations, we took a larger anatomical ROI, covering most of the medial prefrontal cortex (MPFC) from the Brainnetome Atlas (Fan et al., 2016). With this ROI, we extracted individual *t*-statistics from our contrasts of interest in GLM5 and averaged those activations over two dimensions (respectively X and Z, and X and Y), so that we could assess the spread of activations over the last dimension, respectively Y (anterior-posterior axis) and Z (ventral-dorsal axis).

Data availability statement

All source data needed to evaluate or reproduce the figures and analyses described in the paper and supplementary materials are available online at 'https://doi.org/10.6084/m9.figshare.19228977'.

Second level neuroimaging maps can be found at 'https://neurovault.org/collections/12221/'.

Code availability statement

All code needed to evaluate or reproduce the figures and analyses described in the paper and supplementary materials are available online at 'https://doi.org/10.6084/m9.figshare.19228977'.

Acknowledgements

Data collection for this work was funded by two independent personal Amsterdam Brain and Cognition (ABC) Talent grants to J.L. and R.J.v.H., and an NWO Veni Fellowship (grant 451-15-015) granted to M.L. M.L. is supported by a Swiss National Fund Ambizione Grant (PZ00P3_174127), J.L. is supported by an NWO VENI Fellowship grant (916-18-119).

Disclosure statement

None of the authors have any conflicts of interest to declare.

4

Metacognition and the effect of incentive motivation in two compulsive disorders: gambling disorder and obsessive-compulsive disorder

Hoven M

de Boer NS

Goudriaan AE

Denys D

Lebreton M

van Holst RJ*

Luigjes J*

* shared last authorship

Psychiatry and Clinical Neurosciences, 2022, 76(9), 437-449

Abstract

Aim

Compulsivity is a common phenotype amongst psychiatric disorders, such as obsessive-compulsive disorder (OCD) and gambling disorder (GD). Deficiencies in metacognition, such as the inability to estimate ones' performance via confidence judgments could contribute to pathological decision-making. Earlier research has shown that OCD patients exhibit underconfidence, while GD patients exhibit overconfidence. Moreover, it is known that motivational states (e.g., monetary incentives) influence metacognition, with gain (respectively loss) prospects increasing (respectively decreasing) confidence. Here, we reasoned that OCD and GD symptomatology might correspond to an exacerbation of this interaction between metacognition and motivation.

Methods

We hypothesized GD's overconfidence to be exaggerated during gain prospects, while OCD's underconfidence to be worsened in loss context, which we expected to see represented in ventromedial prefrontal cortex (VMPFC) blood-oxygen-level-dependent (BOLD) activity. We tested those hypotheses in a task-based functional magnetic resonance imaging (fMRI) design (27 GD, 28 OCD, 55 controls). The trial is registered in the Dutch Trial Register (NL6171).

Results

We showed increased confidence for GD versus OCD patients, that could partly be explained by sex and IQ. Although our primary analyses did not support the hypothesized interaction between incentives and groups, exploratory analyses did show increased confidence in GD patients specifically in gain context. fMRI analyses confirmed a central role for VMPFC in the processing of confidence and incentives, but no differences between the groups.

Conclusion

OCD and GD patients reside at opposite ends of the confidence spectrum, while no interaction with incentives was found, nor group differences in neuronal processing of confidence.

Introduction

Compulsive behaviors are defined as "repetitive acts that are characterized by the feeling that one 'has to' perform them while being aware that these acts are not in line with one's overall goal" (Luigjes et al., 2019). Various psychiatric disorders are associated with compulsivity, of which obsessive-compulsive disorder (OCD) is the most typical (Stein, 2002), but it's also seen in addictive disorders such as gambling disorder (GD) (van Timmeren et al., 2018). Both disorders are characterized by a loss of control over their compulsive behaviors, albeit originating from distinct motivations, serving different purposes and relating to distinct symptoms (Chamberlain et al., 2005; Figee et al., 2016). Hence, compulsivity seems to be a common phenotype in otherwise symptomatically different disorders.

Dysfunctions in metacognition could explain distinct features of compulsive behaviors. Metacognition is the ability to monitor, reflect on, and think about our own behavior (Fleming, Dolan, et al., 2012). One metacognitive computation is the judgment of confidence, defined as the subjective estimate of the probability of being correct about a choice (Pouget et al., 2016). Confidence plays a key role in decision-making and learning (Fleming, Dolan, et al., 2012; Meyniel, Sigman, et al., 2015; Pouget et al., 2016), and therefore in steering our future behavior (Folke et al., 2017; Samaha et al., 2019). It is crucial for behavioral control that one's confidence is in line with reality. Nonetheless, discrepancies between actual behavior (e.g. choice accuracy) and confidence in that behavior (subjective estimate of accuracy) have been consistently described, which could contribute to pathological (compulsive) decision-making as seen in various psychiatric disorders (Hoven et al., 2019). Clinical presentations of OCD and GD indeed suggest confidence abnormalities in the opposite direction, underconfidence and overconfidence, respectively, which could both promote detrimental decision-making, such as checking behavior and compulsive gambling (Fortune & Goodie, 2012; Goodie & Fortune, 2013; Nestadt et al., 2016; Samuels et al., 2017). In a recent review we showed that both people with subclinical and clinical OCD consistently showed a decrease in confidence level, which was especially profound in OCD-symptom contexts (Hoven et al., 2019). Oppositely, in pathological gamblers, there was evidence for overconfidence in rewarding gambling contexts. which was also related to symptom severity (Goodie, 2005; Lakey et al., 2007). In sum, patients with GD and those with OCD seem to function at opposite sides of the confidence continuum, respectively over- and underestimating their performance, which could explain how opposite traits may underlie similar pathological behavior (i.e., compulsive behavior).

Reward processes are important for learning and decision-making and interact with cognition (Pessoa & Engelmann, 2010). Many studies have implicated subcortical

regions such as the ventral striatum (VS) and cortical regions such as the ventromedial prefrontal cortex (VMPFC) in reward processing, forming a "brain valuation system" (Bartra et al., 2013; Lopez-Persem et al., 2020; Rangel et al., 2008) whose activity relates to value-based decision-making (Lebreton et al., 2009) and motivates behavior (Pessiglione & Lebreton, 2015). Both patients with OCD and those with GD show deficits in reward processes and accompanying dysregulated neural circuitries. A recent review on neuroimaging of reward mechanisms by Clark et al. (2019) clearly indicated dysregulated reward circuitries, especially focused on the VMPFC and VS in patients with GD, with mixed evidence regarding the direction of these effects. In patients with OCD, a recent review showed that the ventral affective circuit, consisting of medial frontal cortex and VS was consistently shown to be dysregulated, showing decreased activity in response to rewards, which was increased in response to losses (Shephard et al., 2021). This is particularly relevant to the question of how confidence might contribute to those pathologies' symptoms, as an increasing number of studies show that affective and motivational states can influence confidence (Allen et al., 2016; Koellinger & Treffers, 2015; Massoni, 2014). Recently, we demonstrated that monetary incentives bias confidence judgments in healthy individuals, where prospects of gain (respectively loss) increase (respectively decrease) confidence, while performance levels remained unaffected in both perceptual and reinforcement-learning contexts (Hoven, Brunner, et al., 2022; Lebreton et al., 2018; Lebreton, Bacily, et al., 2019; C. C. Ting et al., 2020).

We therefore reasoned that an interaction between incentive and confidence processing could cause or fuel the compulsive behaviors in GD and OCD. On the one hand, prospects of high monetary incentives could exaggerate overconfidence in patients with GD, leading to continuation of compulsive gambling; on the other hand, in OCD this could lead to exaggerated decreased confidence in negative value context as harm avoidance is considered one of the core motivations of compulsive behavior in patients with OCD (Bey et al., 2017, 2020; Summerfeldt et al., 2014).

On the neurobiological side, a growing number of functional magnetic resonance imaging (fMRI) studies have associated metacognitive processes with activity in the frontal-parietal network (Allen et al., 2017; Baird et al., 2013; Fleming et al., 2010; Hilgenstock et al., 2014; Vaccaro & Fleming, 2018), and activity in the dorsomedial prefrontal cortex (dmPFC), insula and dorsal anterior cingulate cortex (dACC) has been negatively associated with confidence, suggesting a role for these areas in representing uncertainty-related variables (Fleming et al., 2018; Molenberghs et al., 2016; Morales et al., 2018; Rouault & Fleming, 2020; Shenhav et al., 2016). Interestingly, recent studies have also found activity in the VS, the VMPFC, and perigenual anterior cingulate

cortex (pgACC) - to be positively associated with confidence (Bang & Fleming, 2018; De Martino et al., 2012; Gherman & Philiastides, 2018; Hebart et al., 2016; Lebreton et al., 2015; Rouault, McWilliams, et al., 2018; Rouault & Fleming, 2020). Importantly, this latter network has been previously positively associated with value-based processes (Bartra et al., 2013; Haber & Behrens, 2014; Haber & Knutson, 2010; Lopez-Persem et al., 2020). Actually, both confidence judgments and value information seem to be automatically integrated into VMPFC's activity (Gherman & Philiastides, 2015; Lebreton et al., 2009, 2015; Lopez-Persem et al., 2020; Shapiro & Grafton, 2020). Yet, little is known about whether and how the behavioral interaction observed between incentives and confidence can be explained by their shared association with the VMPFC. In an attempt to answer this question, we recently reported an important interaction between incentive and metacognitive signals in the VMPFC in healthy individuals: confidence signals in the VMPFC were observed in trials with gain prospects, but disrupted in trials with no – or negative (loss) monetary prospects (Hoven, Brunner, et al., 2022). This suggests that the VMPFC has a key role in mediating the relationship between incentives and metacognition. Given the crucial roles of the VMPFC and VS in reward processes and metacognition, which were found to be dysregulated in GD and OCD, we hypothesized that both regions would show disrupted activation patterns related to incentive processing and metacognition and their interaction in patients compared with healthy controls (HCs).

Overall, in the present study we investigate metacognitive ability and its interaction with incentive motivation in patients with OCD and those with GD, behaviorally and neurobiologically.

Methods

Ethics

Experimental procedures were approved by the Medical Ethics Committee of the Academic Medical Center, University of Amsterdam. All subjects provided written informed consent.

Participants

We recruited a total of 31 patients with GD, 29 patients with OCD and 55 HCs between the ages of 18 and 65 years. Of our HC sample of 55 participants, 25 participants were included in our earlier work (Hoven, Brunner, et al., 2022). HCs were recruited through

online advertisements and from our participant database. Patients with GD were recruited from a local treatment center (Jellinek Addiction Treatment Center Amsterdam) and were recently diagnosed with GD. Patients with OCD were recruited through the department of psychiatry at the Academic Medical Center in Amsterdam and were diagnosed with OCD.

Exclusion Criteria

After applying all exclusion criteria (see Appendix B), we included 27 patients with GD, 28 patients with OCD and 55 HCs for the behavioral analyses, of which four, two, and two participants contributed only one of two task sessions, respectively. For the fMRI analyses we included 24 patients with GD, 27 patients with OCD and 53 HCs, of which seven, three, and two participants contributed only one of two task sessions, respectively.

Experimental Design and Study Procedure

We used a similar experimental design and study procedure as previously described (Hoven, Brunner, et al., 2022). For details on the experimental design and study procedure, see Hoven et al. (2022) and Figure 1. In sum, participants performed a simple perceptual decision-making task, with a two-alternative forced choice of contrast discrimination followed by a confidence judgment. In each trial, participants could either win (gain context) or lose (loss context) points, or not (neutral context), conditional on the accuracy of the choice in that trial. Importantly, this incentivization was administered after the choice moment but before the confidence rating. The task was implemented using MATLAB (The MathWorks, Inc.) and the COGENT toolbox.

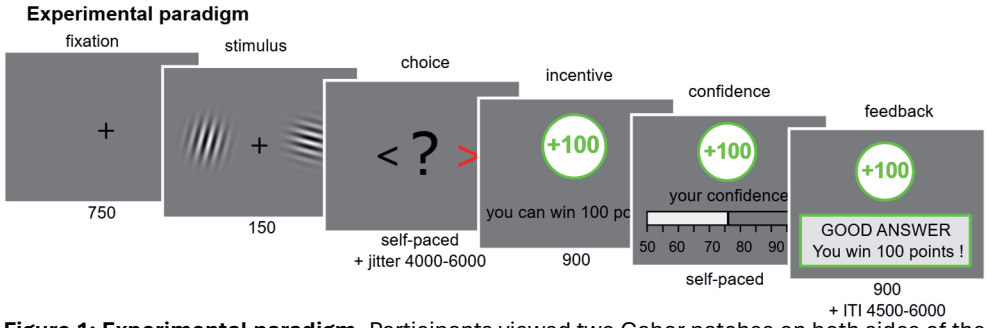


Figure 1: Experimental paradigm. Participants viewed two Gabor patches on both sides of the screen (150 ms) and then chose which had the highest contrast (left/right, self-paced) (for more information, see Hoven et al., 2020). After a jitter of a random interval between 4500 to 6000 ms, the incentive was shown (900 ms; green frame for win trials, grey frame for neutral trials, red frame for loss trials). Afterwards, participants were asked to report their confidence in their choice on a rating scale ranging from 50% to 100% with steps op 5%. The initial position of the cursor was randomized between 65% and 85%. Finally, subjects received feedback. The inter trial interval (ITI) had a random duration between 4500 and 6000 ms. The calibration session only consisted of Gabor discrimination, without confidence rating, incentives or feedback and was used to adjust difficulty so that every individual reached a performance of 70%.

Behavioral Measures

We extracted trial-by-trial experimental factors including incentive condition, evidence, and behavioral measures (accuracy, confidence ratings, reaction times). Evidence was calculated by normalizing the unsigned difference of the two Gabor patches' contrast intensities by their sum to adjust for saturation effects (for more details see (Lebreton et al., 2018)). In addition, we computed an extra *latent* variable: early certainty.

The early certainty variable was computed in order to analyze BOLD activity at choice moment, when the brain encodes a confidence signal that is not yet biased by incentives. This was done by making a trial-by-trial prediction of early certainty based on stimulus features (reaction times, evidence and accuracy) at choice moment. This resulted in an early certainty signal that was highly correlated with confidence, but showed no statistical relationship with incentives (see Appendix B). For more details, see (Hoven, Brunner, et al., 2022).

Next to confidence ratings we also assessed additional metacognitive metrics: (i) confidence calibration - the difference between average confidence and average performance as an indicator of overconfidence or underconfidence, (ii) metacognitive sensitivity - the ability to discriminate between correct answers and errors using confidence judgments (see Appendix B).

Behavioral Analyses

All analyses were performed in the R environment (RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA). We used linear mixed effects models (LMEMs) as implemented in the lmer function from the lme4 and afex packages (Bates et al., 2015; Singmann et al., 2015). To determine p-values for the fixed effects, we performed Type 3 F tests with Satterthwaite approximation for degrees of freedom as implemented in the afex package. When relevant, we used the 'emmeans' package to perform post-hoc tests that were corrected for multiple comparisons using Tukey's method (Lenth et al., 2018).

To answer our main research questions, we built several LMEMs and performed a model selection procedure (Table 1). The final model (Model 1) included fixed effects of incentive, group, accuracy and evidence (z-scored) and interactions between incentive and group, as well as two-way and three-way interactions between evidence, accuracy and group. Moreover, a random subject intercept and a random slope of incentives per subject were included in the final model as well. To confirm that the incentive condition or group did not influence accuracy or reaction time, we modelled additional LMEMs with performance and reaction time as dependent variables (Model 2, Model 3).

Lastly, we added IQ (z-scored) and sex as fixed effects to our original Model 1 (Model 4) to control for differences in the distribution of these demographic variables. Model fit was assessed and compared using Chi-square tests on log-likelihood values. Additional control analyses on the properties of confidence, early certainty, confidence calibration and metacognitive sensitivity are reported in Appendix B.

Due to a technical bug, our design was not fully balanced as the level of perceptual evidence was not equal across the incentive conditions. ANOVA and post-hoc testing indeed showed that evidence was highest in neutral condition, followed by gain and loss. There were no group differences, nor an interaction between group and incentive. These effects cannot account for any group differences we find in our data, since evidence did not differ between groups. Importantly, the evidence differences did not affect performance, since performance is equal across conditions. See Appendix B for more details.

fMRI Analyses

For details on fMRI acquisition and preprocessing see Appendix B and Hoven et al. 2022.

All fMRI analyses were conducted using SPM12. Critically, our design allowed us to distinguish between our two timepoints of interest: 1) the moment of stimulus presentation and choice in which implicit (un)certainty about the choice is formed, and 2) the moment of incentive presentation and confidence rating, in which the value of incentives and the confidence rating are encoded. We built a general linear model (GLM 1) estimated on subject-level with these two moments of interest: the moment of choice (i.e., stimulus presentation) and the moment of incentive presentation/confidence rating. We chose to analyze the incentive presentation and confidence rating as a single timepoint since the rating moment followed the presentation of the incentive after 900 ms, with regressors time-locked to the onset of incentive presentation. We also included a regressor for the moment of feedback to explain variance in neural responses related to feedback on accuracy and value that was not related to the decision-making process, but this regressor was not of interest for the current analyses. All whole-brain activation maps were thresholded using family-wise error correction (FWE) at cluster level ($P_{FWE\_clu} < 0.05$), with a voxel cluster-defining threshold of $p<.001$ uncorrected.

Using GLM 1, with regressors for choice modulated by early certainty, for incentive/rating modulated by incentive and confidence, and for feedback modulated by accuracy we were able to investigate our contrasts of interest: (1) choice moment modulated by early certainty, (2) incentive/rating moment modulated by incentive value and (3) incentive/rating moment modulated by confidence rating. For details see Appendix B.

In order to study the interaction between incentive motivation and metacognitive ability on the neurobiological level we leveraged the factorial design of our task to build GLM 2. We used GLM 2 to explicate the effect of incentive motivation on both the integration of evidence at choice moment, as well as on confidence formation, and compare those between groups. GLM 2 consisted of regressors for each time point (choice and incentive/rating moments) and for each incentive condition, as well as a single regressor at feedback moment, resulting in seven regressors. For all these events we examined both baseline activity and regression slopes relating to their pmod of interest: signed evidence for choice and confidence for incentive/rating. See Appendix B for more details.

Since the results by Hoven et al., 2022 suggested that the VMPFC plays an important role in the interaction between incentive motivation and metacognition, we created a functional region of interest (ROI) that represented the confidence-related activity in the VMPFC cluster from our GLM 1 across groups results (see Figure 3D, Table 5). We then extracted individual *t*-statistics within this ROI (i.e. normalized beta estimates

(Lebreton, Bavard, et al., 2019)) from our contrasts of interest and performed one-sample t-tests against 0 to check for positive or negative activation patterns. Then, we compared them between incentive conditions, groups, and studied their interactions using mixed ANOVAs implemented in the afex package. When appropriate, we performed post-hoc testing using the emmeans package, correcting for multiple comparisons using Tukey's method. Since we also hypothesized that the VS would play a role in the interaction between incentives and metacognition, we performed the same ROI analysis in the VS with a functional ROI that represented the incentive-related activity in the VS cluster from our GLM 1 across group results (see Table 5).

Table 1: Model descriptions and comparison

Model	Model notation	AIC	BIC	Comparison	X^2	P-value	Winning model
A	Confidence ~ Incentive * Group + (Incentive|Subject)	122919	123041				
B	Confidence ~ Incentive * Group + Accuracy + (Incentive|Subject)	122273	122402	A vs. B	648.59	< 2.2e-16	B
C	Confidence ~ Incentive * Group + Accuracy + Evidence + (Incentive|Subject)	122004	122141	B vs. C	271.00	< 2.2e-16	C
D	Confidence ~ Incentive * Group + Accuracy*Evidence + (Incentive|Subject)	121791	121936	C vs. D	214.53	< 2.2e-16	D
E	Confidence ~ Incentive * Group + Accuracy*Evidence* Group + (Incentive|Subject)	121751	121942	D vs. E	52.141	1.747e-09	E
F	Confidence ~ Incentive * Group + Accuracy*Evidence* Group + Sex + IQ + (Incentive|Subject)	121752	121958	E vs. F	2.7018	0.259	E

Shown here are the model notations of all models with their respective Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, as well as model comparison outcomes with corresponding $\chi 2$ and P-values, resulting in the winning model 'E', which is referred to as Model 1 in the manuscript.

Results

Demographics

IQ and sex distributions differed between groups (IQ: $F_{2,107} = 3.222$, p=0.0438; sex: X = 14.483, df = 2, p<.001), with higher IQ scores for HC subjects compared with GD patients (t = 2.53, p=0.014) and with mostly men in the GD group, and relatively more women in the OCD group (Table 2). This corresponds to the natural distribution observed in epidemiological studies for OCD and GD, showing higher prevalence of GD amongst men, and a slightly higher prevalence of OCD in women (Black & Shaw, 2019; Calado & Griffiths, 2016; Mathes et al., 2019; Ruscio et al., 2010). Age did not differ between groups. For post-hoc group differences on questionnaire scores, see Table 2.

Table 2: Demographics

	HC	GD	OCD	Statistics
Age	33.51 ± 12.32	33.22 ± 10.40	31.93 ± 8.21	$F_{2,107} = 0.25$, p = 0.777
IQ*	91.18 ± 10.96	85.22 ± 9.53	89.54 ± 8.32	$F_{2,107} = 3.22$, p = 0.0438 HC vs GD: t(80) = 2.41, p = 0.0181 HC vs OCD: t(81) = 0.70, p = 0.487 GD vs OCD: t(53) = 1.79, p = 0.0791
Y-BOCS***	0.25 ± 1.76	1.19 ± 2.60	20.36 ± 6.15	$F_{2,107} = 322.2$, p<.001 HC vs GD: t(80) = -1.01, p = 0.0592 HC vs OCD: t(81) = -22.64, p<0.001 GD vs OCD: t(53) = 14.97, p<0.001
PGSI***	0.05 ± 0.40	14.85 ± 4.80	0.64 ± 1.91	$F_{2,107} = 380.5$, p<.001 HC vs GD: t(80) = -22.84, p<0.001 HC vs OCD: t(81) = -2.20, p = 0.030 GD vs OCD: t(53) = -14.52, p<0.001
HAMA***	1.09 ± 1.97	3.93 ± 5.88	11.43 ± 6.28	$F_{2,107} = 48.02$, p<.001 HC vs GD: t(80) = -3.24, p = 0.0017 HC vs OCD: t(81) = -11.22, p<0.001 GD vs OCD: t(53) = 4.57, p<0.001
HDRS***	1.31 ± 2.31	5.07 ± 6.24	7.71 ± 4.04	$F_{2,107} = 24.97$, p<.001 HC vs GD: t(80) = -3.97, p<0.001 HC vs OCD: t(81) = -9.19, p<0.001 GD vs OCD: t(53) = 1.87, p = 0.0673
Sex (m/f)***	33 / 22	24 / 3	11 / 17	$X^2(2) = 14.483$, p<.001

Means ± standard deviations of various demographic variables are shown per group, for sex counts are displayed. Statistics for group comparisons are shown, including F and X^2 statistics, degrees of freedom and p-values. IQ= estimated Intelligence Quotient, GD = gambling disorder, HAMA = Hamilton Anxiety Rating Scale, HC = healthy control, HDRS = Hamilton Depression Rating Scale, OCD = obsessive-compulsive disorder PGSI = Problem Gamblers Severity Index, Y-BOCS = Yale-Brown Obsessive Compulsive Scale. *p<.05, ***p<.001

Behavioral Results

To start, we answered our main questions: (1) are there group differences in confidence, and (2) what is the influence of incentive motivation on confidence. Model 1 showed a main effect of group ($F_{2,112}$ = 4.7910, p=.01) and incentive ($F_{2,112}$ = 20.9371, p<.001) on confidence (Figure 2, Appendix B Table B3). We also found a main effect of accuracy ($F_{1,15107}$ = 608.8906, p<0.001), with subjects showing higher confidence for correct answers. Moreover, there was a significant two-way interaction of group and evidence ($F_{2,15099}$ = 3.5094, p=0.02994). As expected, we also found a significant interaction between accuracy and evidence, replicating the 'X-pattern' signature of evidence integration where confidence increases with increasing evidence when correct, and vice versa ($F_{1,15097}$=185.3245, p<0.001)[64]. Interestingly, the evidence integration effect differed per group, as signaled by a significant three-way interaction between accuracy, evidence and group ($F_{2,15094}$ = 3.0533, p=0.04723) (Appendix B Figure B3, Table B3, for post-hoc tests see Appendix B). Lastly, the interaction between incentive and group revealed a trend towards an effect ($F_{4,112}$= 2.2821, p=0.06487).

Post-hoc tests indicated a significantly higher confidence in GD patients versus OCD patients (GD-OCD = 6.38 ± 2.12, Z-ratio = 3.014, p=0.0073), and a trend towards higher confidence in GD compared to HC subjects (GD-HC = 4.30 ± 1.84, Z-ratio = 2.333, p=0.0513), whereas OCD patients did not differ from HC subjects. Moreover, we replicated the parametric effect of incentive value on confidence (loss-neutral =-1.80 ± 0.429, Z-ratio = -4.192, p<0.001; loss-gain =-3.14 ± 0.486, Z-ratio = -6.460, p<0.001; neutral-gain = -1.34 ± 0.363, Z-ratio = -3.683, p<0.001). With regards to the three-way interaction, we found that GD patients' confidence was less influenced by evidence for correct answers compared to both HCs and OCD patients (see Appendix B, Figure B3). Exploratory post-hoc analyses on the group*incentive interaction effect showed that, especially in context of possible gains, GD patients were more confident than OCD patients (GD - OCD = 8.12 ± 2.24, Z-ratio = 3.621, p<0.001) and HC subjects (GD - HC = 5.83 ± 1.95, Z-ratio = 2.989, p=0.0079), with no differences between HC and OCD patients in any incentive condition (Table 3).

As control analyses we estimated Model 2 and 3 with accuracy and reaction time as dependent variables (Table 4). No effect of group, incentive or an interaction effect on accuracy or reaction time were found, as expected from our design (where incentives follow choices), confirming that accuracy and response times cannot confound any effect of incentives that we found on confidence.

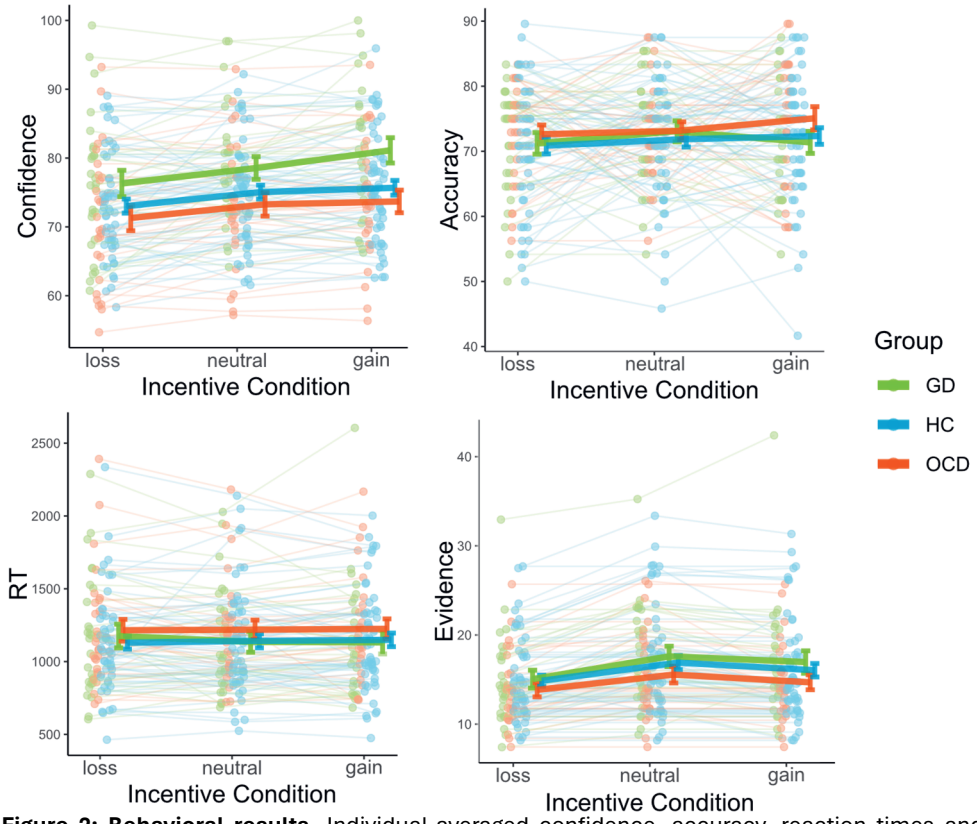


Figure 2: Behavioral results. Individual-averaged confidence, accuracy, reaction times and evidence as a function of incentive condition (loss, neutral and gain) per group. Green dots and lines represent gambling disorder patients, blue dots and lines represent healthy controls and red dots and lines represent obsessive-compulsive disorder patients. Dots represent individuals, and lines highlight within subject variation across conditions. Error bars represent sample mean ± SEM per group. GD = gambling disorder, HC = healthy control, OCD = obsessive-compulsive disorder

Table 3: Results of linear mixed-effects models

Model 1	**Confidence**
Incentive	F(2.00, 112.34) = 20.94, p < .001
Group	F(2.00, 112.51) = 4.79, p = .010
Accuracy	F(1.00, 15107.05) = 608.89, p < .001
Evidence	F(1.00, 15104.05) = 0.04, p = .848
Incentive:Group	F(4.00, 112.10) = 2.28, p = .065
Accuracy:Evidence	F(1.00, 15097.33) = 185.32, p < .001
Group:Accuracy	F(2.00, 15106.28) = 2.27, p = .103
Group:Evidence	F(2.00, 15099.41) = 3.51, p = .030
Group:Accuracy:Evidence.	F(2.00, 15094.35) = 3.05, p = .047
Model 4	**Confidence**
Incentive	F(2.00, 112.34) = 20.93, p < .001
Group	F(2.00, 112.50) = 2.75, p = .068
Sex	F(1.00, 110.26) = 2.88, p = .093
IQ	F(1.00, 109.80) = 0.03, p = .865
Accuracy	F(1.00, 15107.01) = 609.14, p < .001
Evidence	F(1.00, 15104.51) = 0.04, p = .845
Incentive:Group	F(4.00, 112.11) = 2.29, p = .064
Accuracy:Evidence	F(1.00, 15097.16) = 185.42, p < .001
Group:Accuracy	F(2.00, 15106.06) = 2.30, p = .100
Group:Evidence	F(2.00, 15098.91) = 3.45, p = .032
Group:Accuracy:Evidence	F(2.00, 15094.15) = 3.09, p = .046

Shown here are the results of Model 1 (without demographics) and Model 4 (with demographics) acquired using Type 3 F tests with Satterthwaite approximation for degrees of freedom using the afex package. Shown are F values, with corresponding degrees of freedom and P-values.

Table 4: Results of control models

Model 2: Accuracy ~ Incentive*Group + (1+Incentive|Subject)	
Group	$F_{2,109} = 0.5827$, p = 0.5601
Incentive	$F_{2,1591} = 1.0319$, p = 0.3566
Group*Incentive	$F_{4,1586} = 0.8671$, p = 0.4830
Model 3: RT ~ Incentive*Group + (1+Incentive|Subject)	
Group	$F_{2,110} = 0.5207$, p = 0.5956
Incentive	$F_{2,220} = 0.0994$, p = 0.9054
Group*Incentive	$F_{4,219} = 0.4269$, p = 0.7891

Shown here are the results of Model 2 and Model 3 linear mixed-effects models, acquired using Type 3 F tests with Satterthwaite approximation for degrees of freedom using the afex package. Shown are F values, with corresponding degrees of freedom and P-values.

Since sex and IQ were significantly different between the groups, we aimed to control for these variables by adding them as fixed effects, resulting in Model 4. The main effect of group did not remain significant, but showed a trend towards an effect (F2,112 = 2.7465, p=0.06846), while the main effect of incentive did remain significant (F2,112 = 20.9326, p< 0.001). We found no evidence for a significant effect of sex (F1,110 = 2.8776, p=0.09264), or IQ (F1,109 = 0.0291, p=0.86489). The interaction effect between group and incentive remained non-significant at trend-level (F4,112 = 2.2898, p=0.06412). The significant three-way interaction between accuracy, evidence and group persisted (F2,15094 = 3.0871, p=0.04566). Importantly, when performing a Chi-square test on the log-likelihood values of the models excluding and including the demographic variables to compare model fit, the model without demographics showed a better model fit (X2 = 2.7018, df=2, p=0.259), thereby favoring this simpler model. Additionally, to investigate how confidence was differently affected by sex in our healthy controls, we performed a two-sample t-test which showed that males were generally more confident than females (males: 76.51 ± 1.04; females: 71.70 ± 0.77) ($t_{52} = 2.6518$, p-value=0.01057). However, both sex and IQ did not show a significant influence on confidence level in Model 4.

Next to confidence, we also examined calibration and metacognitive sensitivity (see Appendix B). In short, we showed that GD patients were more overconfident than OCD patients, without an effect of incentive condition. No differences in metacognitive sensitivity were found between groups or incentive conditions.

fMRI results GLM 1

We analyzed functional neuroimaging data to test for differences in brain activity between groups for our contrasts of interest: (1) choice moment modulated by early certainty, (2) rating/incentive moment modulated by incentive value, and (3) rating/incentive moment modulated by confidence. The results from the fMRI group analysis revealed no significant differences between the groups for any of our contrasts.

Next, we grouped all subjects together and performed one-sample t-tests on our contrasts of interest to examine the results across groups (cluster-generating voxel threshold $p<.001$ uncorr.; clusterwise correction for multiple comparisons $p_{FWE}<0.05$). During choice, early certainty positively correlated with activation in the precuneus, VMPFC, bilateral VS and putamen, and bilateral visual areas (Figure 3A). The dorsal anterior cingulate cortex, bilateral dorsomedial- and dorsolateral prefrontal cortex, bilateral insula, thalamus, middle frontal gyrus, bilateral sensorimotor cortex, superior and inferior parietal lobe related negatively to early certainty (Figure 3A).

At the moment of incentive presentation, the incentive value correlated positively with activation in the VS and VMPFC stretching into more dorsal areas, as well as the superior temporal gyrus (Figure 3B). Incentive value was negatively related to activity in the right (pre)motor cortex and dorsolateral PFC, as well as the left middle and superior temporal gyrus, left occipitotemporal gyrus, and left middle and inferior frontal gyrus. Moreover, activity in right lateral occipitotemporal gyrus and middle temporal gyrus were negatively related to incentive value (Figure 3B).

During rating moment, confidence was positively related to activity in the VMPFC, left motor cortex and putamen and bilateral visual areas (Figure 3C). The following areas were negatively related to confidence: the left superior and inferior parietal lobes, right dorsolateral PFC, right supramarginal gyrus and thalamus, right motor cortex stretching into the dorsolateral PFC, left visual cortex and cerebellum (Figure 3C). See Table 5 for details of across group fMRI results.

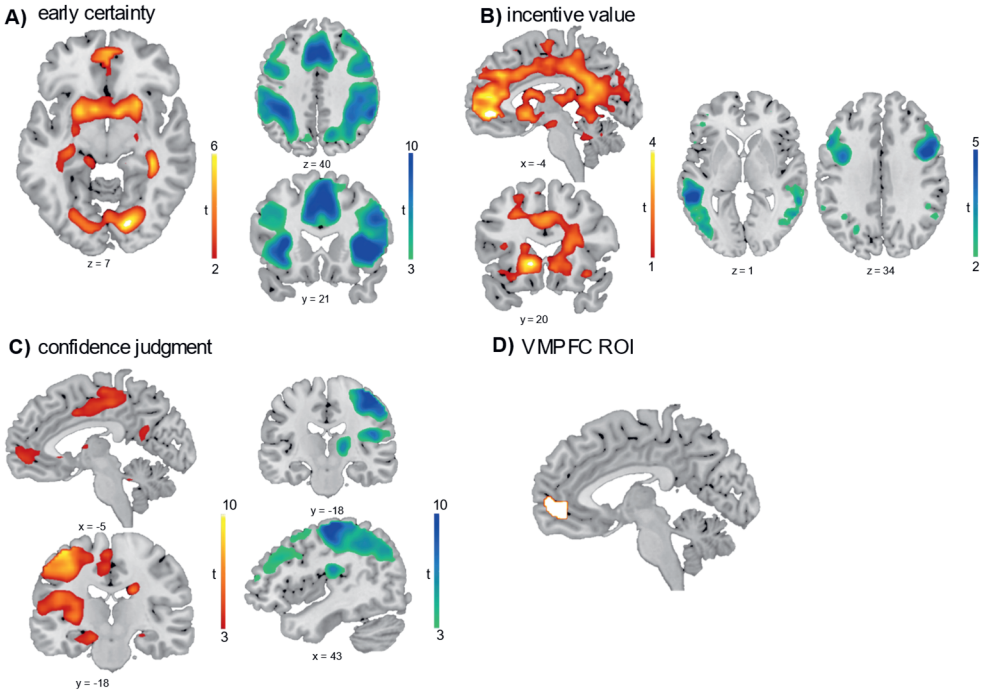


Figure 3: Whole brain statistical bold-oxygen-level-dependent (BOLD) activity across groups. Red/yellow areas represent areas with a positive relationship, while green/blue areas represent areas that have a negative relationship. (A) Areas correlating significantly with early certainty at choice moment. Shown are positive activations in ventromedial prefrontal cortex, ventral striatum and visual cortices. Negative activations in dorsal anterior cingulate cortex, dorsolateral prefrontal cortices, insula, parietal cortices. (B) Areas correlating significantly with incentive value at incentive/rating moment. Shown are positive activations in ventromedial prefrontal cortex, anterior cingulate cortex, ventral striatum. Negative activations in dorsolateral prefrontal cortices and temporal gyri (C) Areas correlating significantly with confidence judgments at incentive/rating moment. Positive actions are shown in ventromedial prefrontal cortex, motor cortex and putamen. Negative clusters in motor cortex and dorsolateral prefrontal cortex. All clusters survived P<0.05 FWE cluster correction. Voxel-wise cluster-defining threshold was set at P<.001, uncorrected. For whole brain activation see Table 5. (D) Region of interest (ROI) of the VMPFC used for GLM2 analyses.

Table 5: Whole brain activation

Effect	Brain Region	k	Peak z-score	P (FWE cluster corrected)	Peak MNI x	y	z	Hemi-sphere
Early Certainty +	Precuneus Ventromedial PFC Ventral Striatum Putamen	2180	6.66	<.001	-6	-34	11	LR
	Lingual gyrus (visual cortex)	154	6.39	<.001	18	-81	-4	R
	Lingual gyrus (visual cortex)	54	4.49	0.045	-21	-79	-4	L
Early Certainty -	Dorsal Anterior Cingulate Dorsomedial PFC Dorsolateral PFC Insula Thalamus Middle Frontal Gyrus Precentral Gyrus Postcentral Gyrus Supramarginal Gyrus Superior Parietal Lobe Inferior Parietal Lobe Calcarine gyrus (visual cortex)	13299	Inf (>8)	<.001	45	14	2	LR
	Middle Occipital Lobe Middle Temporal Gyrus Lateral Occipito-temporal Gyrus	451	7.06	<.001	-30 -48 -45	-91 -67 -61	-4 -1 -10	L
	Right Cerebellum	144	6.64	<.001	33	-55	-31	R
Incentive Value +	Ventral Striatum	74	4.75	.004	-12	11	-4	L
	Ventromedial PFC Dorsomedial PFC	212	4.53	<.001	-3 -9 0	44 50 35	-4 -4 14	LR
	Superior Temporal Gyrus	48	4.25	.026	-45 -39	-16 -22	-1 5	L
Incentive Value -	Precentral gyrus stretching into premotor cortex and dorsolateral PFC	283	5.81	<.001	39 45 48	11 5 14	26 32 29	R
	Middle temporal gyrus	277	5.26	<.001	-54 -51	-43 -52	2 11	L

	Superior temporal gyrus				-48	-25	-7	
	Lateral occipitotemporal gyrus Medial occipitotemporal gyrus	183	5.06	<.001	-45 -24 -24	-61 -73 -82	-13 -7 -10	L
	Middle frontal gyrus Inferior frontal gyrus	299	4.93	<.001	-45 -39 -54	2 17 17	53 23 14	L
	Lateral occipitotemporal gyrus	116	4.90	<.001	42 45	-58 -49	-13 -13	R
	Middle temporal gyrus	47	3.74	.029	57 60 57	-46 -46 -61	11 2 2	R
Confidence +	Middle occipitotemporal gyrus Lateral occipitotemporal gyrus Cerebellum	1947	Inf (>8)	<.001	12 21 15	-73 -70 -52	-10 -7 -16	R
	Motor cortex (precentral gyrus)	993	Inf (>8)	<.001	-36 -36 -54	-25 -19 -16	65 47 47	L
	Putamen Rolandic operculum	968	5.91	<.001	-30 -45 -30	-19 -16 -22	2 20 14	L
	Occipital lobe	65	4.58	.011	42	-67	5	R
	Ventromedial PFC	92	4.39	.002	-3 -12 -19	56 47 41	-4 8 -1	LR
Confidence -	Lingual gyrus (visual cortex) Cerebellum	1144	Inf (>8)	<.001	-9 -15 -24	-79 -52 -67	-7 -22 -28	L
	Motor cortex (precentral gyrus) Stretching into dorsolateral PFC	2421	Inf (>8)	<.001	45 42 39	-16 -37 -52	59 62 41	R
	Supramarginal gyrus Thalamus	262	6.92	<.001	45 15	-19 -22	20 2	R
	Superior parietal lobe Inferior parietal lobe	168	5.09	<.001	-33 -39 -39	-58 -52 -43	41 47 41	L

	Middle frontal gyrus (Dorsolateral PFC)	71	4.49	.007	-45 -45	32 23	32 35	R

Brain activations (whole brain analyses) showing activity related to early certainty at choice moment, as well as activity related to incentive and confidence at incentive/rating moment. All whole-brain activation maps were thresholded using family-wise error correction for multiple correction (FWE) at cluster level (P FWE_clu < 0.05), with a voxel cluster-defining threshold of P<0.001 uncorrected. Activity that positively correlates to given variable is denoted by '+', whereas negative correlations are denoted by '-'. PFC = prefrontal cortex.

Interaction between metacognition and incentives in VMPFC (GLM 2)

Our recent study suggested an important role of the VMPFC in the interaction between incentive-processing and metacognitive signals (Hoven, Brunner, et al., 2022). To investigate how this interaction takes effect in and differs between our clinical groups, we performed an ROI analysis by leveraging our factorial design. We extracted VMPFC activations for both time points (choice and rating), all incentives (loss, neutral and gain), and all groups (HC, OCD and GD), for both baseline activity and a regression slope with (1) signed evidence and (2) confidence judgments (see Figure 3D for the ROI).

First, one-sample t-tests showed that, overall, VMPFC baseline activations were negative at choice and rating moment (choice: t_{100} = -3.611, p<0.001; baseline: t_{100} = -4.9287, p<0.001). The correlations between VMPFC activity and both signed evidence at choice moment and confidence at rating moment, however, were significantly positive (choice: t_{100} = 3.057, p=0.003; baseline: t_{100} = 3.7399, p<0.001) (Figure 4). This implies that the VMPFC represents both confidence judgments and signed evidence (i.e., interaction between accuracy and evidence: increased VMPFC activity with increased evidence when correct and vice versa).

Then, we investigated whether there were effects of incentive condition and group around this general signal. As expected, at choice moment there were no effects of incentive condition on VMPFC baseline activity, nor on its correlation with the signed evidence signal (i.e., slope) (Figure 4, Table 6). Despite the behavioral group effect on evidence integration, we did not find a group nor interaction effect on both baseline VMPFC activity and the correlation with signed evidence. At rating moment, however, incentive condition had a significant effect on both the baseline VMPFC activity, as well as its correlation with confidence. Post-hoc testing showed that the baseline VMPFC activity was higher during gain versus loss (t_{196}: -3.874, p<0.001), and during gain versus neutral (t_{196} = -3.228, p<0.001), but no differences between neutral and loss conditions were found (t_{196} = -0.646, p=0.7948). The correlation of VMPFC activity with confidence

was significantly higher (i.e., increased slope) in gain versus neutral (t_{196} = -3.053, p=0.0072), while no differences between gain and loss, or between neutral and loss were found. Moreover, there was a significant group effect on VMPFC baseline activity during rating moment. The post-hoc tests revealed that OCD subjects had significantly decreased activity compared with HCs, averaged over incentive conditions (t_{98} = -2.515, p=0.0358). No interaction effects between group and incentive were found on baseline activity or its correlation with confidence at rating moment.

Similar analyses using a ROI of the VS were performed (see Appendix B), with similar results: VS activity correlated with signed evidence, but no incentive, group or interaction effects were found at choice moment. Similarly, the correlation of VS activity with confidence was significantly higher in gain versus neutral, with no group difference at rating moment.

Table 6: Results of VMPFC ROI analysis

	Incentive	**Group**	**Incentive x Group**
Choice Baseline	F(1.99, 195.28) = 0.37 p = 0.687	F(2, 98) = 0.54 p = 0.582	F(3.99, 195.28) = 0.41 p = 0.803
Choice Slope 'Signed Evidence'	F(1.99, 195) = 1.15 p = 0.320	F(2, 98) = 0.20 p=0.819	F(3.98, 195) = 0.31 p = 0.869
Rating Baseline	F(1.91, 186.81) = 8.61 p<**0.001**	F(2, 98) = 3.24 p = **0.044**	F(3.81, 186.81) = 0.44 p = 0.771
Rating Slope 'Confidence Judgment'	F(1.92, 187.68) = 4.67 p = **0.012**	F(2, 98) = 0.99 p = 0.375	F(3.83, 187.68) = 1.29 p = 0.277

Shown here are the results of the mixed ANOVAs of t-statistics in the ventromedial prefrontal cortex (VMPFC) region of interest (ROI) using the afex package. Shown are the main effects of incentive condition, group and their interaction effect on the choice and rating time points, focusing on both the baseline activity as well as the slope of signed evidence and confidence judgments, respectively. F-values, with corresponding degrees of freedom and p-values are reported.

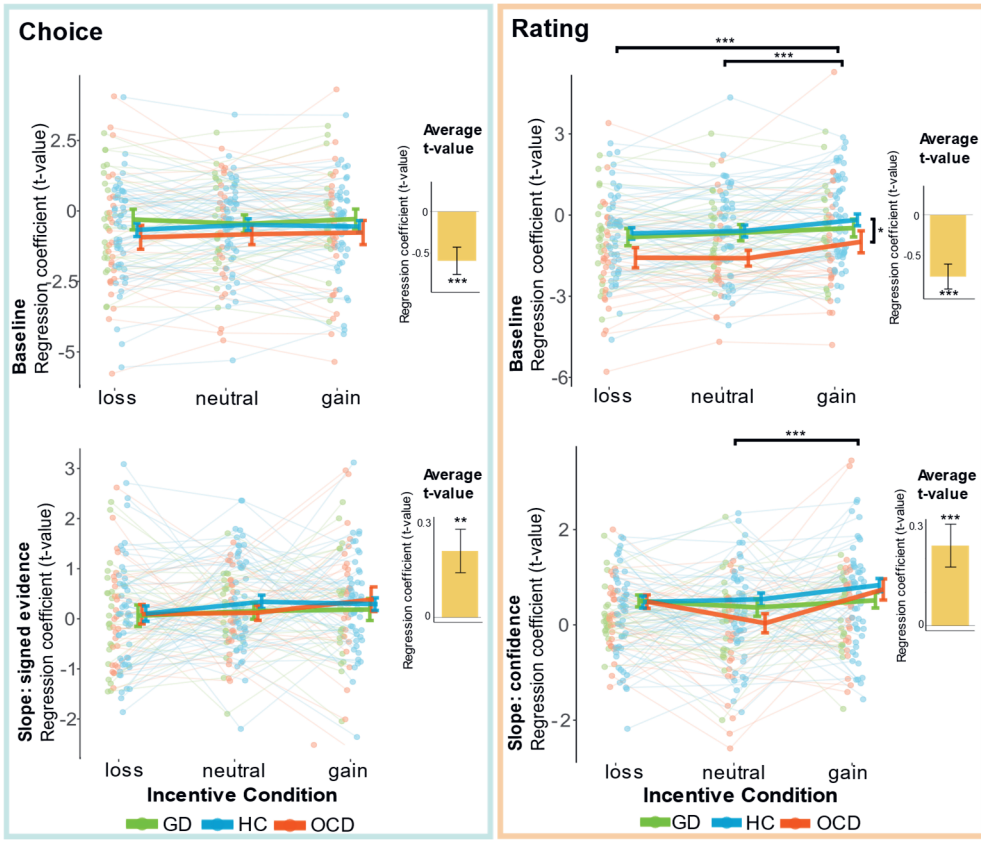


Figure 4: Activation in ventromedial prefrontal cortex across incentives and groups. Ventromedial prefrontal cortex region of interest (ROI) analysis. T-values corresponding to baseline and regression slopes were extracted for all three groups and three incentive conditions, at two time points of interest: choice and incentive/rating moment. Green dots and lines represent gambling disorder patients, blue dots and lines represent healthy controls and red dots and lines represent obsessive-compulsive disorder patients. Dots represent individual t-statistics, and error bars represent sample mean ± SEM per group. Black bars represent significant post-hoc tests. Yellow bars represent average t-values, with corresponding significance level of one-sample t-tests against 0. (* p<0.05, ** p<0.01, *** p<0.001). GD = gambling disorder, HC = healthy control, OCD = obsessive-compulsive disorder.

Discussion

In this study we investigated the (neural signatures of) metacognitive ability and its interaction with incentive motivation in two compulsive disorders: OCD and GD. First, we replicated the biasing effect of incentives on confidence estimation in all groups, showing that confidence was higher in the gain context and lower in the loss context. This is a robust effect, that has now been independently replicated multiple times (Hoven, Brunner, et al., 2022; Lebreton et al., 2018; Lebreton, Bacily, et al., 2019; C. C.

Ting et al., 2020). We initially found evidence for a significantly higher confidence in GD patients versus OCD patients, although this effect diminished after controlling for sex and IQ differences between groups. Hence, we only found moderate evidence for our hypothesis of group differences in confidence, as well as for our hypothesis that incentive motivation would affect confidence judgments differently in the groups. Future research should address the role of the demographic confounding factors more specifically.

When looking into the computational signatures of confidence formation in more detail, GD patients interestingly showed less integration of evidence into their confidence judgments for correct choices compared to both HCs and OCD patients. This suggests that GD patients were less able to use evidence they received to form confidence judgments. This decreased sensitivity to objective evidence could fit GD's symptomatology of cognitive inflexibility (Perandrés-Gómez et al., 2021; van Timmeren et al., 2018), and cognitive distortions (Ledgerwood et al., 2020; Mallorquí-Bagué et al., 2019). Illusion of control leads pathological gamblers to believe they can predict outcomes, rendering them less influenced by objective evidence, which may promote continuation of (overconfident) gambling behavior (Cowley et al., 2015; Goodie & Fortune, 2013).

Notably, our patient groups seemed to be situated on opposite sides of the confidence spectrum, with GD patients being more confident than OCD patients. However, this effect was partly driven by sex and IQ differences between groups. The GD group consisted mostly of males, whereas the OCD group had a more mixed composition, mirroring the prevalence distribution of these disorders (Howe et al., 2019; Subramaniam et al., 2015; Swedo et al., 1989; Welte et al., 2017). Consistent with our findings of increased confidence in HC male subjects, recent studies have shown that males are more confident than females, despite equal performance (Ariel et al., 2018; Rivers et al., 2021). Therefore, the effect of sex might have explained some variance in our data, but does not fully explain the group differences, since we do find a trend toward a group effect. The importance of taking into account sex and gender as factors in both neuroscience and psychiatry research is increasingly recognized and acted upon (Cahill, 2006), since sex differences play a role in the incidence, treatment and manifestation of psychopathology (Cosgrove et al., 2007; Gobinath et al., 2017). The precise role of sex and gender in metacognition deserves more attention and should be characterized further in future research.

Our data shows no convincing evidence for an exaggerated decrease/increase in confidence during loss/gain anticipation in OCD/GD, respectively. However, the group*incentive interaction approached significance, with increased confidence in GD

patients compared to both OCD patients and HCs, specifically in the gain condition. This finding agrees with literature demonstrating increased reward sensitivity in GD (Navas et al., 2017; Van Holst, Veltman, Bchel, et al., 2012). Confidence in OCD patients has been mostly studied using metamemory paradigms, and abnormalities were most profound in OCD-relevant contexts (Boschen & Vuksanovic, 2007; Bucarelli & Purdon, 2016; Hermans et al., 2008; Moritz et al., 2007; Radomsky et al., 2001; Tolin et al., 2001). Earlier studies probing confidence in GD are sparse, and whilst they all did show an effect of overconfidence in (sub)clinical problem gamblers, none of the studies actively controlled for performance differences, making it difficult to draw strong conclusions about confidence biases (Brevers et al., 2014; Goodie, 2005; Lakey et al., 2007).

Since confidence in GD and OCD did not differ from the healthy population we cannot technically speak of confidence 'abnormalities' in GD and OCD. Future work is necessary to study the link between compulsivity and confidence more directly. One interesting method is transdiagnostic research to study metacognition in psychiatry. Transdiagnostic research methods are useful, since (meta)cognition might relate more closely to symptoms than diagnoses, due to high levels of comorbidity and heterogeneity of symptoms within disorders. Indeed, a transdiagnostic factor of 'anxious-depression' was negatively related to confidence, whereas 'compulsive behavior and intrusive thoughts' were positively related to confidence and showed decoupling of confidence and behavior by diminished utilizing of perceptual evidence for confidence judgments (Seow & Gillan, 2020). This latter result is in line with our findings of diminished evidence integration into confidence judgments in GD patients.

The brain areas we found to be related to confidence and incentive processing converge with earlier work. Confidence was found to be positively related to the VMPFC via automatic processing at the choice moment (De Martino et al., 2012; Lebreton et al., 2015; Lopez-Persem et al., 2020; Shapiro & Grafton, 2020). Early certainty processing was also positively related to activity in the VS and precuneus (Hebart et al., 2016; Rouault, McWilliams, et al., 2018; Vaccaro & Fleming, 2018). We also observed a widespread network of areas negatively related to early certainty, containing the dACC, dorsolateral PFC, insula, inferior parietal lobe and midfrontal gyrus, a network repeatedly associated with uncertainty and metacognitive processes (Hebart et al., 2016; Molenberghs et al., 2016; Morales et al., 2018; Vaccaro & Fleming, 2018). Also, well-known relationships between reward processing and activity in both VS and VMPFC were replicated (Bartra et al., 2013; Lebreton et al., 2009). Moreover, we found negative relationships between incentive value and BOLD activity in the central executive network (i.e. lateral PFC and middle frontal gyrus), as well as superior

temporal gyrus (Liu et al., 2011; Wilson et al., 2018). Confidence was found to be related to VMPFC activity, not only at choice moment, but also during rating (De Martino et al., 2012; Lebreton et al., 2015; Lopez-Persem et al., 2020). Overall, our fMRI findings closely resemble activation patterns previously shown in healthy populations.

We also replicated the effect of incentive condition on VMPFC baseline activity and on the correlation of VMPFC activity with confidence, which was highest in gain conditions, which we also found in the VS (Hoven, Brunner, et al., 2022). While we found aberrant evidence integration in GD patients on a behavioral level, we did not find any group differences in evidence processing on neurobiological level. Interestingly, OCD patients showed a decreased baseline VMPFC activity during incentive/rating moment, which fits with earlier work showing neurobiological deficits in a 'ventral motivational circuit' including the VMPFC (Stein et al., 2019; Thorsen et al., 2018). However, we did not find any interactions with incentive condition in the VMPFC activity related to either signed evidence or confidence.

In sum, contrary to our hypotheses, we did not find neurobiological deficits directly related to confidence or to the effects of incentive on confidence in our clinical samples. This might not be surprising, given that the behavioral group effects were small (and disappeared when controlling for demographics), which limited our ability a priori to find impairments in neural circuits mediating confidence processes. Because, to our knowledge, the present study represents the first attempt in investigating the joint neural basis of metacognitive and reward processes in both GD and OCD, further study - e.g. looking into transdiagnostic variations of symptoms - might be more powerful in detecting clinically useful neurocognitive signatures of those processes than the present clinical case-control comparisons (Parkes et al., 2019).

Acknowledgments

Data collection for this work was funded by two independent personal Amsterdam Brain and Cognition (ABC) Talent grants to JL and RJvH, and a NWO Veni Fellowship grant (451-15-015) to ML. ML is supported by a Swiss National Fund Ambizione Grant (PZ00P3_174127) and an ERC Starting Grant (ERC-StG-948671), and JL is supported by a NWO Veni Fellowship grant (916-18-119).

5

How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach

Hoven M

Luigjes J

Denys D

Rouault M

van Holst RJ

Nature Mental Health, 2023, 1, 337-345

Abstract

Confidence is suggested to be a key component in psychiatry and manifests at various hierarchical levels, from confidence in a decision (local confidence), to confidence about performance (global confidence) to higher-order traits such as self-beliefs. Most research focused on local confidence, but global levels may relate more closely to symptoms. Using a transdiagnostic framework, we tested the relationships between self-reported psychopathology, local and global confidence and higher-order self-beliefs in a general population sample(N=489). Here we show contrasting relationships between confidence and psychopathology dimensions. An anxious-depression (AD) dimension related to local and global underconfidence. Contrarily, a compulsive-intrusive-thoughts (CIT) dimension related to increased overconfidence at both levels, and showed a decoupling between (i) higher-order self-beliefs and (ii) local and global task confidence. The strongest predictor of mental health was a self-beliefs dimension. This study examines higher-order confidence in relation to psychiatric symptoms fluctuating in the general population. Critically psychopathological symptoms show distinct associations with confidence.

Introduction

Using confidence judgments, we are able to think about, reflect upon and evaluate our own thoughts, decisions and actions, which is central to our behavior and crucial for adequate adaptation (Boldt et al., 2019; Heilbron & Meyniel, 2019; Vinckier et al., 2016), behavioral control (D. G. Lee & Daunizeau, 2021) and learning (Fleming, Dolan, et al., 2012; Folke et al., 2017; Pouget et al., 2016). A plethora of studies have shown impairments in confidence in various psychiatric populations (Hoven et al., 2019), ranging from over- to underconfidence. The majority of these earlier studies have focused on 'local' confidence, i.e., the level of confidence reported on individual trial-by-trial decisions. Local confidence is usually quantified using distinctive, independent metrics: confidence bias and metacognitive sensitivity. Confidence bias (or calibration) reflects one's overall confidence level relative to one's actual performance, whereas metacognitive sensitivity reflects how well one's confidence judgments distinguish between correct and incorrect choices. However, these metrics are limited regarding their time scale and the scope of behaviors they control.

A new theoretical framework has emerged which posits that confidence manifests at various hierarchical levels of abstraction, from 'local' confidence judgments on a specific trial to more 'global' metacognitive constructs (Seow et al., 2021). In this framework, local constructs are theorized to pertain to isolated choices. In contrast, global constructs are proposed to reflect beliefs formed over extended periods of time and integrate larger amounts of information. For instance, one can develop global confidence about one's ability to perform a certain task. These local and global constructs are likely related to even higher-order feelings of confidence about the self: self-beliefs, such as self-esteem, self-efficacy, mastery and autonomy, which are relatively more stable over time for a given individual. All these constructs focus on the self from distinct angles. Self-esteem relates to one's self-worth (Quiles et al., 2015), self-efficacy to beliefs in one's ability to influence their lives (Bandura, 1977), autonomy to one's ability to live a meaningful life (Bergamin et al., 2022), and mastery to one's beliefs to control their lives (Eklund et al., 2003; O'Kearney et al., 2020).

How these various levels of confidence are mutually related, and how they relate to psychopathology are important open questions. One could argue that individuals suffering from depressive symptoms develop negative self-beliefs, which could give rise to lower confidence levels. These relationships between confidence levels may differ for individuals suffering from other symptomatology, whose symptoms might instead be related to increased confidence. Thus, alterations in local confidence found across psychopathology could go hand-in-hand with alterations in global confidence, and eventually with higher-order levels of confidence, such as self-belief constructs in

a specific manner, dependent on the psychiatric symptoms at hand. In particular, relative to local confidence, higher-order forms of confidence may be more directly related to clinically relevant subjective experiences and psychopathological behavior in daily life. These higher-order constructs could influence behavior over a wide range of contexts and over longer time scales, where local confidence instead would only influence small-scale decisions. For example, patients suffering from major depression could in general have lower self-esteem and lower global confidence about their abilities, resulting in avoidance behavior and negative schemas (Beck, 2003; Korn et al., 2014; Moroz & Dunkley, 2015). Gaining a deeper understanding of the mechanisms of higher-order confidence is hypothesized to provide more insight into clinically relevant psychopathological behavior (Seow et al., 2021). By identifying which specific processes are most impacted, therapies can be tailored and refined to be more effective in treating symptoms.

A recent study operated the formation of global confidence in a behavioral task (Rouault et al., 2019). Participants performed two mini-blocks of perceptual games; after each block, they selected the game in which they thought they performed best, and gave global confidence ratings about their perceived ability at the games; both proxies for global confidence. To decompose the relationships between these various levels of confidence and mental health symptoms, a recent case has been made to employ a transdiagnostic approach (Seow et al., 2021). The central idea is that cognitive processes might relate more closely to transdiagnostic symptomatology than to DSM diagnoses, due to high levels of comorbidity and heterogeneity in symptoms within a disorder (Insel et al., 2010). For example, many OCD patients suffer to a greater or lesser extent from compulsivity, as well as anxiety, which can disguise important findings when only considering classic DSM disorder boundaries. Here, we leveraged a previously validated transdiagnostic approach that allowed to establish three latent transdiagnostic symptom dimensions termed “Anxious-Depression” (AD), “Compulsive Behavior and Intrusive Thought” (CIT) and “Social Withdrawal” (SW) (Gillan et al., 2016). With this approach, previous work has shown that higher scores on the AD dimension related to lower confidence, while higher scores on the CIT dimension related to higher confidence levels (Rouault, Seow, et al., 2018; Seow & Gillan, 2020). Some of these findings were only visible within the transdiagnostic framework. Likewise, it has been argued that opposing effects of certain symptoms on local confidence could be clarified by considering multiple hierarchical levels of confidence at the same time, each of which may map differently onto mental health symptoms (Seow et al., 2021). Local confidence ratings could, for example, be more strongly driven by a prior of low self-beliefs in anxious-depression, whereas in

compulsivity these priors may have less impact. However, studies testing these hypotheses are currently lacking.

Here, we tested this proposition by systematically investigating the relations between and within the various levels of the confidence hierarchy and psychiatric symptoms. As preregistered (https://osf.io/6vbpr), we acquired questionnaire and task data from a large general population sample online, using a previously validated "local and global confidence task" (Rouault et al., 2019) with a transdiagnostic approach. We hypothesized that local confidence informs global confidence, and negative (respectively positive) relationships between the AD (respectively CIT) symptom dimension and both local and global confidence, corresponding to a general under- and overconfidence, respectively. Moreover, based on the proposed hierarchical structure of confidence, we expected positive relationships between local confidence, global confidence and self-belief constructs. Finally, we tested whether there would be dissociations in the relationships between the hierarchical levels of confidence depending on symptom dimensions, and identified which of the hierarchical levels of confidence best explained symptom dimensions.

Results

Experimental Design

In this cross-sectional study, conducted online, human subjects performed a local and global confidence task (Rouault et al., 2019). During this perceptual decision-making task, participants performed short blocks with two randomly interleaved 'games' indicated by two arbitrary color cues (Figure 1). In each trial of a game subjects had to indicate which of two black boxed contained a higher number of white dots. The games involved different conditions: they could either be easy of difficult (i.e., difficulty feature), and deliver either veridical feedback or no feedback (i.e., feedback feature). These features resulted in six possible pairings of conditions within each block, which were repeated twice in randomized order. On all trials without feedback participants had to indicate their *local confidence* about their probability of being correct on that specific trial from 50% - 100%. After each block participants had to choose on which game they believed they performed best (global task choice) for which they were incentivized, and they had to rate their confidence in their overall performance on each of the two games (*global confidence*) on a scale from 50% to 100%. See Methods for more details on the task.

After finishing the task, participants filled in self-report questionnaires assessing psychiatric symptoms and self-belief constructs (see Methods for details). To study the interplay between the various hierarchical levels of confidence (local, global and self-beliefs) and psychopathology, we assessed psychopathology across a large range of different symptoms, as well as various higher order self-belief constructs.

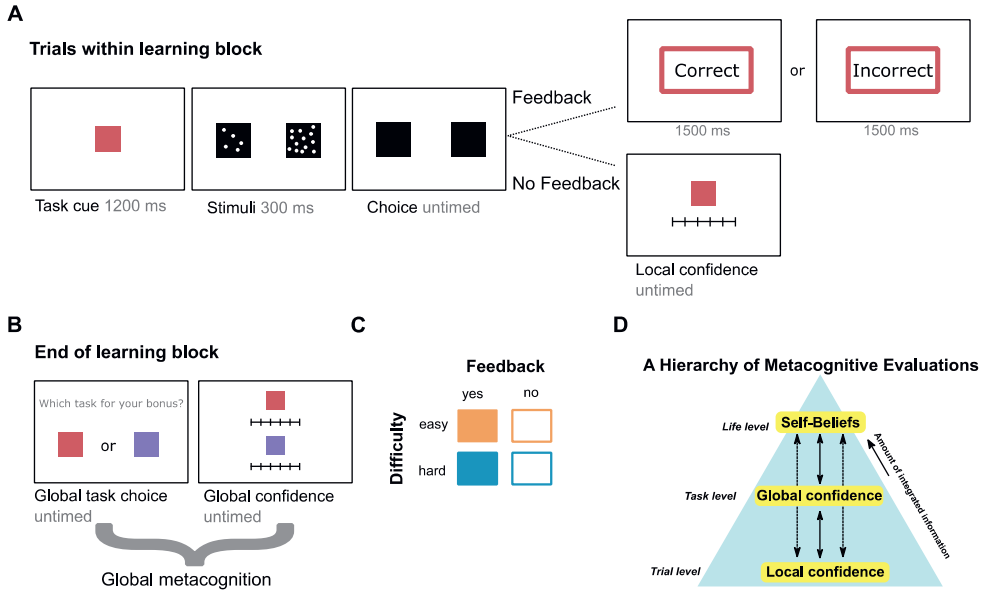


Figure 1: Experimental task design. A Participants performed learning blocks with two randomly alternating trials from two tasks, indicated by a task cue. Each task was either easy or difficult and provided feedback or no feedback (**C**), resulting in six different possible task pairings. Each trial started with the presentation of a color cue, indicating which of the two tasks was presented, after which subjects had to choose which of two boxes contained a higher number of dots. Each judgment was either easy or difficult, dependent on the dot difference between the boxes. After their choice, subjects either received feedback (correct or incorrect) about their choice, or did not receive feedback and instead were asked to provide a *local confidence* rating about the probability of their perceptual judgment being correct. **B** At the end of each learning block participants were asked to choose which task should be used to calculate a bonus (based on their performance; *global task choice*. They also rated their overall ability; *global confidence*. Together, they are measures of global metacognition. **D** The hierarchical levels of metacognitive evaluations, which consist of local confidence at the trial level in isolated decisions, global confidence at the task level and Self-Beliefs, which operate at a level more directly relevant to daily life. All levels include reciprocal interactions with the other levels, in a way that local confidence contributes to global confidence, and eventually to Self-Beliefs, and in turn these levels can impose their influence on local confidence. From trial level information at local confidence levels, to the aggregation of multiple trials at global confidence levels, up to the integration of multiple sources of information encountered in real life situations in one's Self-Beliefs.

Behavioral Variables

In order to relate our confidence measures to our psychopathology measures, various behavioral variables were calculated. First, we calculated *global calibration* (task level), which is the difference between average global confidence and performance on all trials, reflecting how well subjects' overall task confidence matches their overall actual performance. For *local calibration* (trial level), we calculated the difference between average local confidence and performance on no feedback games only. Moreover, the correlation between local and global confidence levels was calculated for the non-feedback condition only (see Methods for more details). Finally, metacognitive efficiency (i.e. meta-d'/d', or M-Ratio) was calculated per subject, which is a confidence precision measure, indexing how well subjects can discriminate between correct and incorrect choices using their local confidence ratings, independently of performance or confidence biases (Fleming, 2017; Maniscalco & Lau, 2012). After applying rigorous exclusion criteria (see Methods), data of 489 participants were used for the analyses.

Transdiagnostic Dimensions

Within a specific psychiatric questionnaire, different items may map onto different latent psychiatric factors, which would be obscured in traditional analyses. In order to obtain a more parsimonious latent transdiagnostic structure that would explain this item-level variation in scores on all psychiatry questionnaires, we performed a factor analysis and tested the relationship between our task variables and these latent transdiagnostic psychiatry dimensions (see Appendix C, Figure C1 and Figure C9 for more details on the factor analysis). Our factor analysis indicated that a three-factor structure provided the best and most parsimonious fit of the variance across the questionnaire items, showing very strong correlations with the item loadings from previous studies (Rouault, Seow, et al., 2018) (Appendix C: Figure C8), indicating that the factor analysis robustly identifies the same latent constructs. For consistency, these three factors were given the same labels as earlier studies, according to the items that loaded the most strongly on each factor, even though these labels are to some extent arbitrary (Appendix C: Figure C1B) (Gillan et al., 2016; Rouault, Seow, et al., 2018): 'Anxious-Depression (AD)', 'Compulsive Behavior and Intrusive Thought (CIT)', and 'Social Withdrawal (SW)'.

Relating Task Variables to Transdiagnostic Dimensions

First, our aim was to test the specific relationships between each transdiagnostic symptom dimension (AD, CIT and SW) and task variables (performance, local confidence, global confidence, local calibration, global calibration, metacognitive efficiency and the correlation between local and global confidence), while controlling for the other transdiagnostic dimensions and demographics. In this way, we tested whether we could replicate previously found relationships between local confidence and transdiagnostic dimensions, and whether these relationships were also found for higher-order levels of confidence.

Results of our multiple regression analyses showed that the AD dimension was significantly positively related with performance, while a significant negative relationship existed for the CIT dimension (Figure 2A, Table C1). Mean local and global confidence were significantly negatively related with scores on the AD dimension. For both local and global calibration, we found a significant negative (i.e., underconfidence) associations with the AD dimension, but positive (i.e., overconfidence) relationships with the CIT dimension.

Because there were effects of symptom dimensions on performance, we performed a sensitivity analysis to examine whether our results were maintained when adding participant's average performance level as a predictor to our regression models when predicting local and global confidence (i.e., Local/Global Confidence ~ AD + CIT + SW + Mean Performance + Age + IQ + Gender). We again found a negative relationships between AD and both local and global confidence (local: $\beta = -0.247 \pm 0.045$, $t = -5.526$, $p_{cor} < 0.001$; global: $\beta = -0.219 \pm 0.040$, $t = -5.473$, $p_{cor} < 0.001$), and also again showed significant positive relationships between CIT and both local and global confidence (local: $\beta = 0.160 \pm 0.046$, $t = 3.508$, $p_{cor} < 0.01$; global: $\beta = 0.125 \pm 0.041$, $t = 3.049$, $p_{cor} < 0.05$), with a type I error rate of .05/7. For additional exploratory analyses on reaction times and inattentiveness see Appendix C.

Moreover, CIT was associated with a lower correlation between local and global confidence, indicating that high CIT scores related to a more distorted coupling between local and global confidence. No associations with metacognitive efficiency were found, and the SW dimension did not relate significantly to any of our task variables (Appendix C: Table C1).

We further asked whether the uncovered relationships between the confidence variables and transdiagnostic dimensions were significantly different from each other. Since the SW dimension did not significantly relate to any of our variables, we did not include it in post-hoc tests. For performance, the regression coefficient of the CIT

dimension was significantly lower than that of the AD dimension (Appendix C:Table C1). For local confidence, local calibration and global calibration, the regression coefficients of CIT were significantly higher than those of AD, indicating that individuals scoring high on CIT show a higher (over)confidence than those scoring high on AD. Moreover, the correlation of local and global confidence was significantly higher for AD than CIT, suggesting that local and global confidence show a better coupling in individuals scoring high on AD versus those scoring high on CIT.

Note that AD and CIT dimensions show a significant positive correlation ($r = 0.24$, $p<.001$, Figure 2D). Therefore, our finding of opposing relationships cannot be attributed to a negative correlation between the factors.

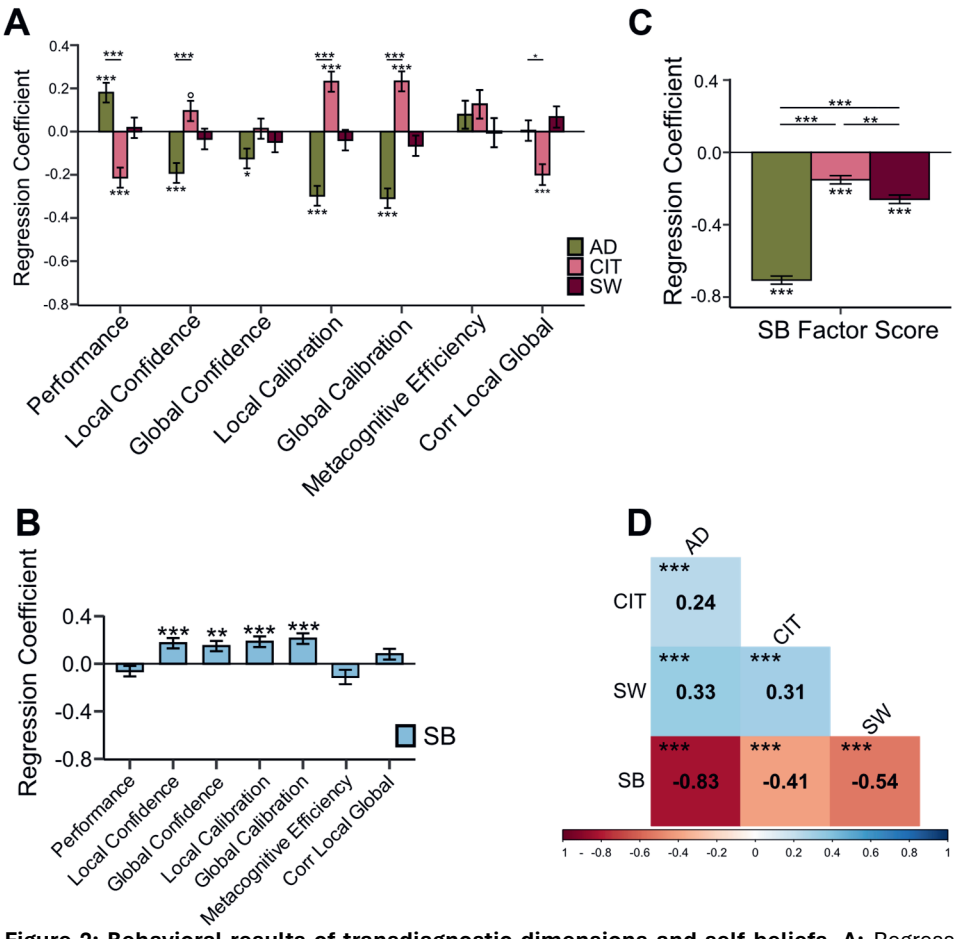


Figure 2: Behavioral results of transdiagnostic dimensions and self-beliefs. A: Regression coefficients of the two-sided multiple regression models of the three latent factors comprising 'Anxious-Depression' (AD), 'Compulsive behavior and Intrusive Thought' (CIT) and 'Social Withdrawal' (SW) and various dependent variables measuring performance and metacognition. All factor scores were entered as predictors for each regression model, together with age, gender and IQ. N = 489 independent subjects. Results are corrected for multiple testing. **B:** Regression coefficients of the two-sided multiple regression models of the latent 'Self-Beliefs' (SB) factor. The regression model containing SB factor scores, together with age, gender and IQ revealed positive associations between SB scores and both local and global confidence and calibration. N = 489 independent subjects. Results are corrected for multiple testing. **C:** Relating the three psychiatry factors to the SB factor using a two-sided regression analysis. All three psychiatric factors are strongly negatively related to SB factor scores, with strongest effects for the AD scores. N = 489 independent subjects. **D:** Pearson correlation matrix of the psychiatric and SB factor scores across subjects. Correlation tests were two-sided and Bonferroni corrected for multiple testing. Error bars represent SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons over the number of dependent variables tested, ° $p < 0.05$ uncorrected. Exact p-values are described in Tables C1 and C2. AD = 'Anxious-Depression', CIT = 'Compulsive behavior and Intrusive Thoughts', SW = 'Social Withdrawal', SB = 'Self-Beliefs', Corr Local Global = correlation between local confidence and global confidence.

Relating Task Variables to Self-Beliefs

Since all self-belief construct questionnaire scores correlated highly (Appendix C: Figure C1C, Figure C5), we aimed to acquire an all-encompassing single latent factor underlying these constructs. To do so, we performed a second factor analysis on the self-belief construct questionnaires using the same methods. All questionnaires showed an average loading of > 0.5 (autonomy M = 0.50 ± 0.16, mastery M = 0.55 ± 0.1, self-esteem (Rosenberg's) M = 0.70 ± 0.08, self-efficacy M = 0.56 ± 0.09, self-esteem (Short Form) M = 0.63 ± 0.09). We labeled this factor the 'Self-Beliefs' (SB) factor (Appendix C: Figure C1D), an even higher hierarchical level of confidence. We then tested the association between this factor (i.e., highest-order level of confidence), performance, local confidence and global confidence using multiple linear regressions (see Methods for details) (Figure 2B). There was a significant positive relationship between SB scores and both local and global confidence, as well as local and global calibration, indicating that subjects with higher scores on the SB factor were more overconfident (Figure 2B, Table C2).

Relating Transdiagnostic Dimensions to Self-Beliefs

After investigating the associations between the transdiagnostic dimensions and both local and global confidence, we examined the associations between the transdiagnostic dimensions and the highest-order level of Self-Beliefs using a linear regression. We found that all three transdiagnostic dimensions related negatively to SB scores, while correcting for demographics (AD: $\beta = -0.706 \pm 0.022$, $p<.001$; CIT: $\beta = -0.151 \pm 0.023$, $p<.001$; SW: $\beta = -0.259 \pm 0.023$, $p<.001$) (Figure 2C). Post-hoc comparisons showed that the regression coefficient of the AD factor was significantly more negative than the SW ($t = 12.257$, $p<.001$) and CIT coefficient ($t = 16.134$, $p<.001$), whereas the coefficient of SW was more negative compared to CIT ($t = 2.949$, $p = 0.003$).

Relationships Between the Various Levels of Confidence

Since we argued that local confidence, global confidence and SB scores each represent various levels of confidence, it is expected that they show positive associations. Indeed, we found that local and global confidence were significantly positively correlated ($r=0.82$, $p < 0.001$). Moreover, local confidence and SB scores, as well as global confidence and SB scores were also significantly positively related ($r=0.16$, $p<0.001$; $r=0.13$, $p = 0.004$, respectively). Although, as expected, the task variables of local and global confidence correlate more strongly together than they both

do with the questionnaire-based SB scores, critically all three confidence levels studied here correlate positively.

Predicting Psychopathology with Levels of Confidence

Even though the various levels of confidence were positively correlated, their influence on symptoms could be divergent. Thus, to assess the influence and relative importance of the hierarchical levels in predicting psychiatric symptoms, we entered all three confidence levels as predictors of psychiatric dimension scores in three separate regressions (one for each dimension, see Methods for details) (Figure 3, Table C3).

SB scores strongly negatively predicted AD and SW scores, whereas neither local confidence nor global confidence was a significant predictor of AD or SW. For CIT, both SB scores and global confidence significantly negatively explained CIT scores, but contrarily, local confidence was a significant positive predictor of CIT scores. For all three psychiatric symptom dimensions, post-hoc tests showed stronger effects of the SB factor compared to both local confidence and global confidence measures (Appendix C: Table C4). Also, a comparison of the standardized absolute regression coefficients of the SB and local confidence predicting CIT indicated that SB was the most important predictor. Since local confidence and global confidence are highly correlated, we also examined (i) regression models where we used the average of local and global confidence as a predictor alongside self-beliefs and (ii) regression models where we only used either local or global confidence as a predictor alongside self-beliefs. Both approaches resulted in similar results as described above, see Appendix C for more detail.

Taken together, these results show that Self-Beliefs is the strongest predictor for all psychiatry dimensions. For CIT, a dissociation was found in the predictions of local confidence on the one hand, and global confidence and Self-Beliefs on the other hand. This further supports the finding of a decoupling between local and global confidence in high CIT individuals.

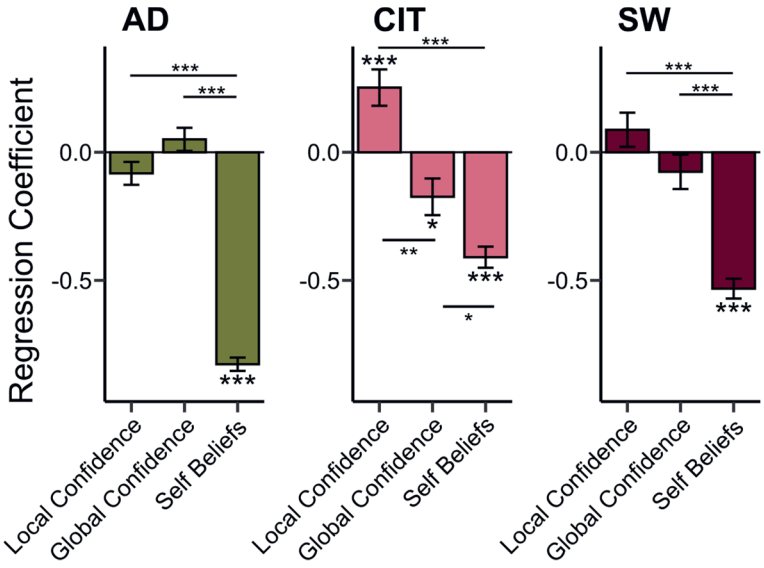


Figure 3: Relationships between hierarchical levels of metacognition and transdiagnostic symptom dimensions. Regression coefficients of the two-sided regression models predicting AD scores, CIT scores or SW scores with the three levels of metacognitive hierarchy: local confidence, global confidence and 'Self-Beliefs' (SB) score. All three levels of the hierarchy were entered as predictors for each model, together with age, gender and IQ. Error bars represent SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Post-hoc test results comparing the strength of the regression coefficients were corrected for multiple comparisons. Exact p-values are described in Tables C3 and C4. AD = 'Anxious-Depression', CIT = 'Compulsive behavior and Intrusive Thoughts', SW = 'Social Withdrawal'.

Discussion

Measures of local confidence, global confidence, and self-beliefs, all concern judgments of the self and one's personal competence, albeit at various levels of abstraction formed over different timescales. Together, they are proposed to form a hierarchical structure of confidence, of which all levels impact our behavior (Rouault et al., 2019; Seow et al., 2021). Even though higher-order levels of confidence are an evident determinant of mental health (Quiles et al., 2015; Silverstone & Salsali, 2003), mostly local confidence has been empirically studied using behavioral tasks. Here, we assessed three key hierarchical levels of confidence across a large spectrum of psychopathology. We found opposite relationships between subclinical psychiatric symptoms and the hierarchical levels of confidence. Interestingly, the levels of confidence differently explained symptom dimensions, depending on the type of psychopathology.

First, our transdiagnostic approach revealed that the "Anxious-Depression" (AD) symptom dimension was associated with local and global underconfidence. Contrarily, the "Compulsive Behavior and Intrusive Thoughts" (CIT) symptom dimension was related to local and global overconfidence, presumably driven by lower performance levels. The current study thus replicates and extends earlier results on local confidence (Rouault, Seow, et al., 2018; Seow & Gillan, 2020), by showing that alterations in local confidence generalize to global confidence. Interestingly, the "Social Withdrawal" (SW) symptom dimension did not significantly relate to either local or global confidence, but had a strong negative relationship with self-beliefs. This suggests that individuals with high SW symptom scores suffer from low self-beliefs, despite being able to properly evaluate their global and local confidence.

Positive associations between local confidence, global confidence, and higher-order self-beliefs were found. This is in line with evidence for associations between confidence and metacognitive beliefs in educational settings (Kleitman & Gibson, 2011; Kleitman & Stankov, 2007), positive associations between local confidence and self-esteem (Moses-Payne et al., 2019), and decreased global confidence in individuals with low versus high self-esteem (Rouault et al., 2022). While previous reports had suggested that self-beliefs may function as a prior informing lower levels of confidence, the current study, for the first time empirically tested their relations. By doing so our study critically extends the literature by showing that higher-order self-beliefs are strongly positively related to confidence and overconfidence at both global and local levels in the same participants, while not being related to performance.

Self-beliefs were a general predictor of mental health, which manifested as strong negative relationships between all three symptom dimensions and SB scores, indicating that self-beliefs are affected across the entire spectrum of psychopathology. Self-beliefs were most strongly affected in the AD dimension, and least affected in the CIT dimension. This difference is reflected in the DSM, in which a feeling of worthlessness is a diagnostic criterion for depression (American Psychiatric Association, 2013). Indeed, self-beliefs such as self-esteem, autonomy and self-efficacy are usually strongly diminished in depression and anxiety disorders (Bachrach et al., 2013; Beck, 2003; Bekker & Belt, 2006; Bekker & Croon, 2010; Muris, 2002; Sowislo & Orth, 2013). In contrast, diminished self-beliefs are not an explicit diagnostic criterion of OCD and schizophrenia, disorders that contain symptoms relating to CIT. Instead, such symptoms might be more related to a lack of self-control and/or a loss of sense of reality, which could secondarily affect self-beliefs (Henriksen & Parnas, 2014; Strayhorn, 2002). Our findings are in line with these observations and suggest that, to the extent that self-beliefs may act as priors, their influence was strongest in AD,

resulting in decreased lower-level confidence. In contrast, for CIT and SW the prior was less strong, and possibly, therefore, did not lead to negative changes in local confidence. Despite these differences between dimensions, self-beliefs are deemed an important factor across many different disorders and symptoms (Silverstone & Salsali, 2003).

Intriguingly, while for AD and SW symptoms there was no additional predictive influence of local and global confidence beyond self-beliefs, local and global confidence did significantly predict CIT symptoms, with a dissociation in the directionality of these effects. Higher CIT symptom scores were negatively related to self-beliefs and global confidence, but positively related to local confidence. It thus seems that higher local confidence in CIT did not generalize to increases in global confidence or self-beliefs, and vice versa that the influence of a negative self-beliefs prior did not lead to decreased local confidence levels. These findings of a disconnect between confidence levels are reinforced by our finding of a decreased correlation between local and global confidence with higher CIT symptoms.

The formation of local confidence is a coalescence of multiple influences: from higher-order confidence priors to local assessments of decision evidence. In CIT, it could be that higher-order confidence priors are not optimally used or simply underweighted to inform local confidence judgments. Alternatively, the assessment of local decision evidence might be erratic in CIT, and therefore less incorporated into higher-order beliefs about self-confidence. Indeed, it has been previously found that CIT is associated with lower sensitivity to environmental evidence while informing confidence, leading to a decoupling between action and confidence (Seow & Gillan, 2020). This sub-optimal transfer of information between different levels of confidence could lead to maladaptive generalization across tasks and domains, resulting in rigid self-beliefs that are not updated appropriately or not at all following decision performance. This suggestion fits with recent theoretical models on the formation of confidence (Rouault, McWilliams, et al., 2018) and self-esteem (N. M. P. De Ruiter et al., 2017) that suggest reciprocal relationships between various levels of self-beliefs, where higher levels can impact or constrain lower level components (i.e. local confidence) and simultaneously these lower levels can exert bottom-up influence on higher-order levels.

We hypothesized that confidence levels higher up the hierarchy would relate more strongly to symptomatology. Indeed, self-beliefs significantly explained the transdiagnostic symptom dimensions more strongly than local or global confidence, implying that they would be the most important determinant of mental health. However, stronger relationships between self-beliefs and symptoms could also arise

since they both rely on questionnaires and focused on phenomenological experiences, whereas local and global confidence are task-based behavioral outcomes. However, the fact that we do find relationships between self-beliefs and local and global confidence encourages the notion that behavioral confidence measures can help characterize the cognitive processes that contribute to important phenomenological experiences, such as self-beliefs and mental health.

Unlike previous findings showing a subthreshold decrease in metacognitive efficiency with CIT (Rouault, Seow, et al., 2018), we found no evidence for such a deviation other than a subthreshold increase of metacognitive efficiency in CIT. This also is in contrast to our review evidencing a worsened metacognitive efficiency in schizophrenia and addiction (Hoven et al., 2019), disorders with symptoms that are thought to overlap with our CIT factor. However, a recent meta-analysis revealed that alterations in metacognitive efficiency in schizophrenia likely have been overestimated, which would better correspond to our present findings (Rouy et al., 2021). Although metacognitive efficiency takes into account task performance, here, by design, performance was not constant for all subjects. This may have impacted the relationship between metacognitive efficiency and symptoms. Moreover, it is worth noting that, due to the relatively low number of trials in the current task, estimates of metacognitive efficiency may not be reliable enough for robustly assessing these associations (Guggenmos, 2021). In the current subclinical sample, however neither symptoms nor symptom dimensions significantly related to metacognitive efficiency, but instead, point to specific disturbances of confidence level and bias.

Our study has limitations. By design, our task did not maintain a constant level of performance across subjects, so we could not strictly isolate confidence changes from performance changes. Using calibration measures, however, we could contrast confidence with objective performance. In addition, controlling for performance differences between subjects did not eradicate the positive relationship between CIT and both local and global confidence. Since trials with excessive reaction times were excluded, inattentive trials could not have influenced our findings of lower performance on the task. Moreover, a positive relationship between CIT scores and reaction times did not survive correction for multiple testing (see Appendix C). Overall, the relatively lower performance in individuals with high CIT scores is likely not due to inattentiveness, and has also been reported recently in a reinforcement learning paradigm (J. K. Lee et al., 2023). This study is cross-sectional and only allowed us to test for associations. Future longitudinal studies need to clarify how confidence develops within each level over time, and how the levels reciprocally influence each other. Longitudinal designs would also help to disentangle the directionality of the influence

between shifts in metacognition and shifts in mental health symptoms, at least in establishing temporal precedence of one onto the other (e.g. (Orth et al., 2008; Rieger et al., 2016)). Nevertheless, our findings advance our understanding of the computational processes underlying the hierarchical framework, which could in turn help to precisely target treatments to the cognitive steps that are distorted and to a patient's specific symptom profile. Even though we are careful not to extrapolate our findings pertaining to fluctuations in (sub)clinical symptoms in the general population to formally diagnosed patients, previous studies have shown the same behavioral effects in diagnosed clinical samples and general population samples using a transdiagnostic approach (Gillan et al., 2011, 2016; Seow & Gillan, 2020; Snorrason et al., 2016; Vaghi et al., 2017), but see (Hoven, Rouault, et al., 2023). Using a general population sample allowed us to study a broad set of symptoms and aspects of mental health rather than focusing on one specific disorder or symptom set. Applying a transdiagnostic approach to large clinical samples would be an important next step for this area of research. In addition, it remains to be tested how our findings generalize to other cognitive domains (Benwell et al., 2022), and to daily decisions. Also, the influence of higher-order cognitive states, such as emotional distress or motivational state on confidence should be taken into consideration, since they could play a possible role as additional underlying factors for changes in metacognitive abilities. Here, we have relied on metacognition as typically defined in the field of cognitive neuroscience, quantifying the correspondence between subjective confidence judgments and objective task performance (Katyal & Fleming, 2023). However, in other fields such as cognitive psychology models, the definition of metacognitions encompasses other processes, including intersubjective processes. A goal for future research is to combine insights from cognitive psychology models, specifically the metacognitive model by Wells (Wells, 2019), which centers around beliefs about thinking (i.e. metacognitions), with findings from cognitive neuroscience studies that examine feelings of confidence as a type of metacognition (Katyal & Fleming, 2023).

The current study provides empirical evidence that aberrancies in local confidence extend to higher-order levels of confidence, that self-beliefs are related to local and global confidence, that the associations between these confidence levels and psychiatric symptom dimensions differ depending on the type of psychopathology and that confidence measures higher up the hierarchy related more strongly to symptoms across psychiatry.

Methods

Ethics Approval and Consent

All participants gave informed consent in accordance with procedures approved by the Ethics Committee of the University of Amsterdam (2020-CP-12416).

Participants

Data of 625 participants were collected online through the Prolific Academic platform (prolific.co). Sample size was based on earlier similar transdiagnostic studies (Rouault, Seow, et al., 2018). Subjects were not screened for psychiatric diagnosis with official clinical interviews, but sampled from the general population focusing on continuous variation across psychiatric symptoms that naturally fluctuate in the general population. All participants were paid a base amount of €7.5 per hour, plus a €1.5 bonus when passing check questions. The whole experiment lasted 1.5 hours on average and was conducted in English. Using Prolific's available demographic information, we only invited participants who reported being fluent in English, regardless of their geographical location. We did not collect data on ethnic or cultural background.

Study Design

The current study is a cross-sectional study. All data were collected between 13 October and 2 November 2020, and each participant was tested once.

Local and Global Confidence Task

The perceptual-decision making task was adapted from *Experiment 3* from Rouault et al. (2019) (Rouault et al., 2019) and was coded in JavaScript, HTML and CSS using jsPsych version 4.3 and hosted on Gorilla (Anwyl-Irvine et al., 2020). The code ensured that browsers were in full screen throughout the experiment.

During this experiment, participants performed short blocks with two randomly interleaved 'games' (6 trials each, pseudo-randomized) indicated by two arbitrary color cues (Figure 1). All games involved perceptual discrimination judgments. In each trial of a game subjects indicated which of two black boxes contained a higher number of white dots. The games involved different conditions: easy or difficult (i.e., difficulty feature), and deliver either veridical feedback or no feedback (i.e., feedback feature),

resulting in six possible pairings of conditions within each block, which were repeated twice in randomized order. On all trials without feedback participants had to indicate their *local confidence* about their probability of being correct on that specific trial on a scale from '50% correct (chance level)' to '100% correct (perfect)'. After each block participants had to choose between the two games, on which one they believed they performed best (*global task choice*), for which they were incentivized. Moreover, subjects were asked to rate their confidence in their overall performance on each of the two games (*global confidence*) on a scale from 50% to 100%. Subjects were not explicitly informed about the conditions of the games. For more information on the task structure, see Appendix C.

Self-report Psychiatric and Self-Belief Questionnaires

The symptoms assessed included alcoholism, apathy, impulsivity, eating disorders, social anxiety, obsessive-compulsive disorder, schizotypy, depression and generalized anxiety. The self-belief constructs assessed included autonomy, self-efficacy, mastery and self-esteem. Moreover, subjects were assessed with a rapid IQ evaluation (Condon & Revelle, 2014). The order of questionnaires was fully randomized. See Appendix C for detailed information on the specific questionnaires used and for distributions of total scores (Appendix C: Figure C6). Both the task and questionnaires were validated in previous research.

Exclusion Criteria

To ensure good data quality, we employed a state-of-the-art approach (Zorowitz et al., 2021), using a combination of multiple task-based and questionnaire-based checks, as preregistered here: https://osf.io/6vbpr. As described in Appendix C, we made sure that subjects who did not perform adequately on the task, failed comprehension tests or attention checks were excluded from analyses. After implementing our criteria, our final sample consisted of 489 subjects that were on average 27.2 years old (± 8.5 years), of which 318 were male. For more details on exclusion criteria, see Appendix C. An a posteriori power analysis using the G*Power toolbox (Faul et al., 2007) was performed using the smallest effect size from previous transdiagnostic research (f^2=0.107)[21], a power of 80% and an alpha levelof .05, which confirmed that a sample of 438 subjects would be well-powered. We took into account an expected exclusion rate of 25% (common in online studies), and our final sample consisted of 489 subjects.

Behavioral Variables

Analyses were performed using R-Studio version 4.0.3. We calculated *global calibration* and *local calibration*. In addition, the relationship between local and global confidence in the non-feedback condition was calculated, resulting in 12 games per subject. We extracted the average local confidence per subject per game, and the corresponding global confidence. A Spearman's correlation was performed between the local and global confidence values per subject, which indicated the degree of correlation between the local and global confidence levels.

Moreover, we computed *metacognitive efficiency* (meta-d'/d', or M-Ratio) per subject. Since this computation relies on a signal detection theory framework and assumes that all trials have a constant signal strength, we calculated metacognitive efficiency separately for easy and hard trials in no-feedback trials (36 trials per difficulty level) and averaged the two values per subject. We used the HMeta-d toolbox to perform individual participant fits to compute metacognitive efficiency (using the fit_meta_d_mcmc function from the toolbox). Due to a small number of trials per subject, some subjects' metacognitive efficiency values were noisy, resulting in values that were negative or >3 standard deviations from the mean. Those subjects were excluded (N=40) only for the regressions with metacognitive efficiency, such that in those regressions we used data of 449 subjects. Finally, we performed several control analyses to corroborate the effects of the task features (feedback and difficulty) on performance, local- and global confidence (Rouault et al., 2019), described in Appendix C (Figure C2, Figure C3).

Relating Task Variables to Questionnaires

As a first step, we studied the relationship between our task variables and scores on the (1) individual psychiatric symptom questionnaires and (2) individual self-belief construct questionnaires, using regressions, as is further detailed in Appendix C (Figure C4, Figure C5).

Relating Task Variables to Transdiagnostic Dimensions

For more details on the factor analysis, see Appendix C, Figure C1 and Figure C9. First, we tested the specific relationships between each psychiatric symptom factor and task variables, while controlling for the other factors and demographics. To do this, we used linear regressions, and controlled for demographics, as such:

Behavioral variable ~ AD + CIT + SW + Age + IQ + Gender

All regressors were z-scored before entering the models, such that we obtained standardized (i.e., comparable) regression coefficients. To correct for multiple testing a Bonferroni correction was applied that took the number of dependent variables into account, following Rouault, Seow, et al. (2018), which resulted in a type I error rate of .05/7. Post-hoc tests on regression coefficients were performed using the *esticon()* function from the *doBy* package in order to test whether the strength of the associations between the three factors and our various dependent variables differed. Post-hoc tests were also corrected using a Bonferroni correction with a type I error rate of .05/7. The distribution of data was assumed to be normal but this was not formally tested (Appendix C: Figure C7).

Relating Task Variables to Self-Beliefs

We tested the association between the SB factor (i.e., highest-order level of confidence) and the task variables using linear regressions:

Behavioral variable ~ SB + Age + IQ + Gender

Again, all regressors were z-scored and Bonferroni correction was applied (type I error rate of .05/7).

Relating Transdiagnostic Dimensions to Self-Beliefs

To directly relate our three psychiatry dimensions to the Self-Beliefs factor, a single linear regression was performed, as such:

SB ~ AD + CIT + SW + Age + IQ + Gender

Regression coefficients were compared post-hoc, and were Bonferroni corrected for the number of tests (type I error rate of .05/3).

Predicting Psychopathology with Levels of Confidence

Finally, we aimed to investigate how the three hierarchical levels of confidence studied (i.e., local confidence, global confidence and Self-Beliefs) may contribute to fluctuations in each psychiatric factor. Importantly, after establishing the relationships between the task variables and each symptom dimension, we aimed to investigate

which hierarchical level would best predict each of our psychiatric symptom dimensions. To achieve this, we set each psychiatric factor as the dependent variable in three separate regressions. The hierarchical levels of confidence were set to compete for variance that allowed us to compare their predictive power in predicting symptom severity, and to identify possible dissociations between the three confidence levels. To do so, we conducted Pearson's correlation tests, and three linear regressions, where we separately regressed the three psychopathology factors (i.e., AD, CIT and SW, z-scored) on local confidence, global confidence, and Self-Beliefs scores, as such: Psychiatry dimension (AD / CIT / SW) ~ local confidence + global confidence + SB + Age + IQ + Gender.

Regression coefficients were compared post-hoc, and were Bonferroni corrected for the number of tests (type I error rate of .05/3).

Data Availability

Fully anonymized task and questionnaire data are available on https://osf.io/ncg4s/.

Code Availability

The R analysis script is available on https://osf.io/ncg4s/.

Acknowledgements

MR is the beneficiary of a postdoctoral fellowship from the AXA Research Fund, and is also supported by the Fondation des Treilles, and by a department-wide grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017, EUR FrontCog) and receives support under the program «Investissements d'Avenir» launched by the French Government and implemented by ANR (ANR-10-IDEX-0001-02 PSL). JL and this work are supported by an NWO VENI Fellowship grant (916-18-119). RjvH is supported by an NWO Aspasia Grant (2019/SGW/00764779).

Disclosure statement

None of the authors have any conflicts of interest to declare.

Part II

Confidence in OCD

6

Differences in metacognitive functioning between obsessive-compulsive disorder patients and highly compulsive individuals from the general population

Hoven M

Rouault M

van Holst RJ*

Luigjes J*

* shared last authorship

Psychological Medicine, 2023, 1-10

Abstract

Background

Our confidence, a form of metacognition, guides our behavior. Confidence abnormalities have been found in obsessive-compulsive disorder (OCD). A first notion based on clinical case-control studies suggests lower confidence in OCD patients compared to healthy controls. Contrarily, studies in highly compulsive individuals from general population samples showed that obsessive-compulsive symptoms related positively or not at all to confidence. A second notion suggests that an impairment in confidence estimation and usage is related to compulsive behavior, which is more often supported by studies in general population samples. These opposite findings call into question whether findings from highly compulsive individuals from the general population are generalizable to OCD patient populations.

Methods

To test this, we investigated confidence at three hierarchical levels: local confidence in single decisions, global confidence in task performance and higher-order self-beliefs in 40 OCD patients (medication-free, no comorbid diagnoses), 40 controls, and 40 matched high-compulsive individuals from the general population (HComp).

Results

In line with the first notion we found that OCD patients exhibited relative underconfidence at all three hierarchical levels. In contrast, HComp individuals showed local and global *over*confidence and worsened metacognitive sensitivity compared with OCD patients, in line with the second notion.

Conclusions

Metacognitive functioning observed in a general highly-compulsive population, often used as analogue for OCD, is distinct from that in a clinical OCD population, suggesting that OC symptoms in these two groups relate differently to (meta)cognitive processes. These findings call for caution in generalizing (meta)cognitive findings from general population to clinical samples.

Introduction

Humans have the ability to monitor and introspect on their own thoughts and cognitive processes, a process referred to as metacognition (Fleming, Dolan, et al., 2012). In our uncertain world, our metacognition, and in particular our sense of confidence, guides our behavior. The feeling of confidence helps us seek information (Balsdon et al., 2020; Desender et al., 2019; Pescetelli et al., 2021; Rollwage et al., 2020), guides our learning (Cortese, 2022; Guggenmos et al., 2016) and changes our mind (Stone et al., 2022), especially when external feedback is lacking (Rouault et al., 2019). There is great variability in how well humans are able to judge their own performance. Given the fundamental function of metacognition in guiding behavior, distortions in metacognitive ability have been associated with pathological behavior (Hoven et al., 2019), such as excessive checking behavior when having low confidence (Baptista et al., 2021).

Traditionally, theories have placed dysfunctions of metacognition at the center of obsessive-compulsive disorder (OCD) aetiology (Purdon & Clark, 1999; Wells & Papageorgiou, 1998). Varying notions about the nature of these dysfunctions have been proposed. A first notion suggests that OCD patients suffer from a negative bias in confidence, resulting in underconfidence relative to healthy control subjects. This underconfidence may not necessarily be a defect in judging one's performance, since it could be an appropriate correction of the usual overconfidence seen in healthy individuals (Johnson & Fowler, 2011). Nevertheless, it could lead to excessive doubts, low self-beliefs and obsessive thoughts which could in turn promote compulsive behaviors, while checking behavior itself can also provoke feelings of low confidence (Jaeger et al., 2021; Radomsky et al., 2006). Indeed, a recent meta-analysis of 19 studies covering a variety of cognitive tasks indicated that patients with OCD showed general underconfidence, in both cognitive domains of memory and perception (i.e., less confident than they should be considering their performance) (Dar et al., 2022). These studies focused mostly on local confidence judgments while doing specific tasks (i.e. trial by trial estimates on the correctness of a decision (Pouget et al., 2016) with the underlying assumption that underconfidence on a local level is related to clinically relevant subjective experiences of doubts such as decreased self-beliefs (i.e. higher order metacognition), but this has not yet been investigated. Recent studies suggest that local confidence and self-beliefs may be linked by more global estimates of confidence (e.g. confidence about performance on multiple decisions or a task) and that investigating the interplay between these hierarchical levels of confidence may bridge this gap (Seow et al., 2021).

A second notion suggests that perhaps not underconfidence, but an impairment in estimating or properly utilizing confidence judgments lies at the heart of OCD symptoms, particularly for compulsive behavior. This might manifest as a decreased sensitivity to identify correct from incorrect decisions using confidence judgments (i.e., decreased metacognitive sensitivity) (Hauser, Allen, et al., 2017; Rouault, Seow, et al., 2018) or a decoupling between levels of metacognition (Hoven, Luigjes, et al., 2023). As a result, patients might be less capable to self-correct and inform their future decisions using their confidence, and thus revert to compulsive behavior.

We will test these two notions using a behavioral protocol probing three hierarchical levels of confidence. The hypothesis put forward by the first notion is that relative underconfidence will be found in OCD patients at all three levels. The expectation that follows from the second notion is an impairment in using confidence judgements to separate correct from incorrect choices (i.e., metacognitive sensitivity). Note that these two notions are not mutually exclusive, and could simultaneously exist. However, following the second notion, a decoupling between different levels of metacognition could be expected which opposes the first notion of underconfidence across the three levels.

The relationship between obsessive-compulsive (OC) symptoms and metacognition has also been studied using general population samples, with the advantage of probing large samples with less time and costs investments, while also sampling larger symptom variability. Three such studies did not find evidence for a direct relationship between local confidence and OC symptoms (Benwell et al., 2022; Hoven, Luigjes, et al., 2023; Rouault, Seow, et al., 2018), while another study did find a positive relationship, indicating that increased OC symptoms related to higher confidence (Seow & Gillan, 2020). Moreover, high OC symptoms in the general population have been related to decreases in metacognitive sensitivity, also without a difference in local confidence (Hauser, Allen, et al., 2017). Overall, there is no evidence for decreased confidence, but some indication of reduced metacognitive sensitivity in these samples. The assumption of these types of studies is that there is a spectrum of OCD symptomatology where highly compulsive individuals resemble (albeit to a lesser extent) OCD patients in terms of possibly disturbed (meta)cognitive processes. However, the comparability of OCD patients and highly compulsive individuals have not been directly tested using carefully matched groups. Since clinical studies and general population studies have reported mixed findings regarding the relationship between OC symptoms and metacognition, these populations might be inherently different. In terms of metacognitive functioning, highly compulsive individuals from the general population could (1) resemble OCD patients (to a lesser extent) regarding both

decreased confidence levels and metacognitive sensitivity, (2) only resemble OCD patients regarding decreased usage of confidence (i.e., decreased sensitivity, decreased coupling between metacognitive levels), or (3) be inherently different from OCD patients.

To test this, here we compared OCD patients not only to healthy subjects, but also to a group of matched highly compulsive individuals, on a wide range of metacognitive functions and their relationship with compulsive symptoms. We investigate both local confidence, global confidence, and higher-order self-beliefs to obtain an inclusive picture of metacognitive abilities in people suffering from OC symptoms. We expect (as preregistered: *https://osf.io/3knjc*) decreased local and global confidence in OCD patients compared to HCs, as well as decreased self-beliefs (i.e., self-esteem, autonomy). Moreover, since OCD patients were found to be more reliant on external feedback when assessing their confidence (Lazarov et al., 2014), we expected that underconfidence in OCD patients would be more pronounced in trials without feedback and with increased symptom severity. Also, we expect lower metacognitive sensitivity in OCD patients, resulting in a decreased ability to use local confidence to differentiate between correct and incorrect answers (i.e., discrimination), and we expect a distorted relationship between local and global confidence in OCD as well. Finally, we test whether abnormalities in metacognition found in OCD resemble those of matched highly compulsive individuals.

Methods

Ethics

All experimental procedures were approved by the Medical Ethics Committee of the Amsterdam University Medical Centre. All participants provided written informed consent before the start of any experimental procedure and were reimbursed for their time.

Participants

In this study we collected data from three groups: HCs, OCD patients and high-compulsive non-clinical subjects. We did not perform an a-priori power analysis for the sample sizes of these three groups. Instead, we based our sample size on similar studies assessing clinical populations (e.g. (Marton et al., 2019; Radomsky et al., 2014; Vaghi et al., 2017)).

OCD patients

45 patients with OCD, aged between 18 and 65 years old were included. They were recruited via various local treatment centers and patient associations across the Netherlands, and previously and/or currently underwent psychotherapy. The average duration of symptoms in the patient group was 19.3 years with an average time since diagnosis of 9.2 years. Severity as measured by the Y-BOCS (mean: 21.88 ± 5.84) indicated to be in the upper range of moderate and lower range of severe symptom strength. Exclusion criteria included diagnoses of any comorbid psychiatric disorders, and the use of any medication for the treatment of psychiatric symptoms, including, but not limited to, selective serotonin reuptake inhibitors, tricyclic antidepressants, or antipsychotics. After applying task-based exclusion criteria of lower than chance level performance or too little variation in confidence judgements (for more extensive description see (Hoven, Luigjes, et al., 2023), our final sample consisted of 40 OCD patients.

Healthy controls

45 HCs were included in this study, between 18 and 65 years old. They were recruited through online advertisements and from our participant database across the Netherlands. HCs were matched to OCD patients on age, sex and education levels. Exclusion criteria included diagnoses of any psychiatric disorder or the use of any psychotropic medication. After applying task-based exclusion criteria (Hoven, Luigjes, et al., 2023), our final sample consisted of 40 HCs.

High-compulsive subjects

As part of a larger previous study, 625 English speaking world-wide participants were collected online via the Prolific Academic platform (www.prolific.co) (see (Hoven, Luigjes, et al., 2023)for more details). Subjects were not screened for psychiatric diagnosis, since our aim was to collect data based on continuous variation in psychiatric symptoms within the general population. We excluded subjects who failed attention and comprehension checks, and used the same task-based exclusion criteria as in the clinical sample, and the final sample consisted of 489 subjects. Then we performed propensity score matching in order to select subjects from our large general population sample (N=489) to match our patient sample in terms of obsessive-compulsive symptoms. Using the MatchIt package in R (Ho et al., 2007) we performed nearest neighbor matching. We matched our OCD patient sample to an equal number of high-compulsive subjects from the general population sample based on Obsessive-Compulsive Inventory Revised (OCI-R) score, age, sex and education level (Foa et al., 2002). Our final sample thus consisted of three sets of 40 subjects: 40 OCD patients,

40 HCs and 40 high-compulsive subjects (HComp) from the general population study. Demographics were compared between groups using two-sample t-tests for continuous measures or Chi-square tests for categorical measures.

Questionnaires

All HCs and OCD patients were subjected to the MINI structured psychiatric interview (Sheehan et al., 1998) to screen for any (comorbid) psychiatric disorders. OCD symptom severity was measured using the Obsessive-Compulsive Inventory – Revised (OCI-R) (Foa et al., 2002). All our 120 subjects were assessed with questionnaires on autonomy (Autonomy Scale Amsterdam: ASA) (Bergamin et al., 2023) and self-esteem (Rosenberg Self-Esteem Scale: rSES) (Rosenberg, 1965) as measures of higher-order self-beliefs. Moreover, anxiety and depression symptoms were assessed using the Depression Anxiety and Stress Scale (DASS) (Parkitny & McAuley, 2010) in the clinical sample (OCD and HC) and using the Generalized Anxiety Disorder-7 questionnaire (Williams, 2014), and Zung's depression scale (Zung, 1965), respectively, in the general population (HComp) sample. Metacognitive beliefs were measured in the clinical sample using the Metacognitions Questionnaire-30 (MCQ-30) (Wells & Cartwright-Hatton, 2004).

Local and Global Confidence Task

The perceptual-decision making task was adapted from Experiment 3 in Rouault et al. (2019) and was coded in JavaScript, HTML and CSS using jsPsych version 4.3 and hosted on Gorilla (gorilla.sc) (Anwyl-Irvine et al., 2020). All subjects performed the task online using their personal computer.

All participants performed blocks with two randomly interleaved perceptual tasks (with 6 pseudo-randomized trials each) indicated by two color cues (Figure 1). Participants had to indicate which of two black boxes contained a higher number of white dots. Two experimental features were implemented: a task could be easy or difficult (i.e., difficulty feature), and could deliver veridical feedback or no feedback (i.e., feedback feature), resulting in six possible pairings of tasks within each block. All six possible pairings occurred twice in randomized order, resulting in 144 trials per participant. On each trial without feedback (72 trials per participant) participants indicated their *local confidence* about their probability of being correct on that specific trial on a scale from '50% correct (chance level)' to '100% correct (perfect)'. At the end of each block participants had to indicate the task in which they believed they performed best.

Moreover, participants rated their confidence in their overall performance on each of the two tasks (*global confidence*) on a scale from 50% to 100%. For more detailed information on the task specifics, see (Hoven, Luigjes, et al., 2023).

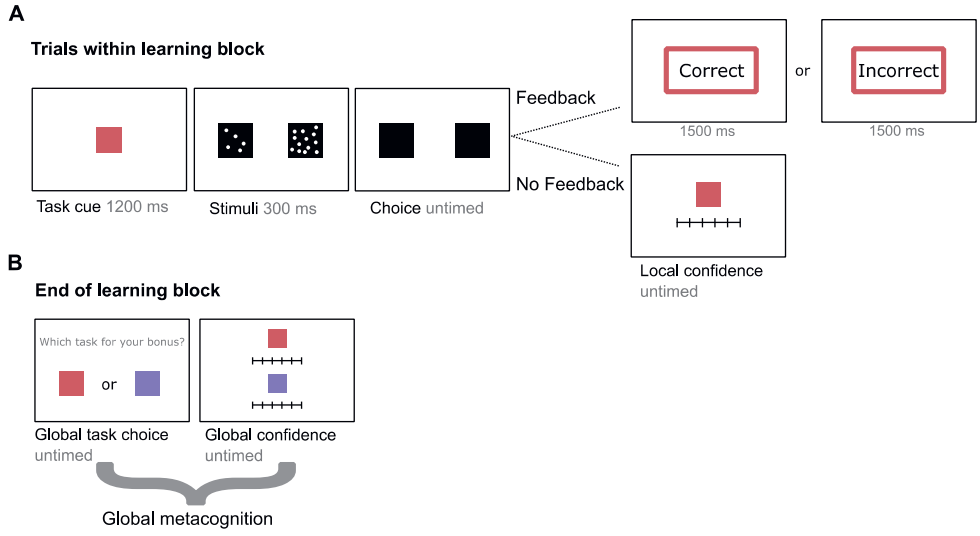


Figure 1: Experimental design. A) Participants performed learning blocks with two randomly alternating trials from two tasks, indicated by a task cue. Each task was either easy or difficult and provided feedback or no feedback (2x2 design), resulting in six different possible task pairings. Each trial started with the presentation of a color cue, indicating which of the two tasks was presented, after which subjects had to choose which of two boxes contained a higher number of dots. Each judgment was either easy or difficult, dependent on the dot difference between the boxes. After their choice, subjects either received feedback (correct or incorrect) about their choice, or did not receive feedback and instead were asked to provide a local confidence rating about the probability of their perceptual judgment being correct. **B)** At the end of each learning block participants were asked to choose which task should be used to calculate a bonus based on their performance; global task choice. They also rated their overall ability; global confidence. Both are measures of global metacognition.

Task-based measures of metacognition

Using local and global confidence, we calculated *local calibration* (decision level), which is the difference between average local confidence and performance on no feedback tasks only. *Global calibration* (task level) was calculated as the difference between average global confidence and performance on all trials. These measures reflect how well one's confidence matches one's actual performance and can be interpreted as overconfidence or underconfidence. We also calculated the direct correlation between average local and global confidence per subject on no feedback tasks only. Note that for one OCD patient this correlation could not be determined due

to a lack of variance in their global confidence. Moreover, we computed *discrimination*, which is a metric of metacognitive sensitivity that indicates how well one's confidence judgments discriminate between their own correct and incorrect choices. It is calculated as the difference between the average confidence for correct and the average confidence for incorrect trials. Another metric to assess metacognitive sensitivity is meta-d' (Fleming, 2017), whose computations are known to be imprecise in designs with a low number of trials per subject per condition (Rouault, McWilliams, et al., 2018) (in our case, 36 trials). Moreover, since results from earlier work (Lebreton et al., 2018) showed high correlations between discrimination and meta-d', we used the discrimination metric as our measure of metacognitive sensitivity in the current study.

Analyses

All analyses were performed using RStudio (version 2022.07.2). Mixed ANOVAs (afex package in R (Singmann et al., 2015)) were used to investigate the effects of group, difficulty and feedback on: accuracy, reaction times, global task choice and global confidence, and to investigate the effects of group and difficulty on local confidence. Using this approach, we investigated whether OCD patients showed metacognitive deviations compared to HCs, and importantly, whether and how metacognitive findings from a general population sample of HComp individuals are comparable to a clinical sample of OCD patients.

Two-sample t-tests were used to compare local calibration, global calibration, discrimination, the correlation of local and global confidence, autonomy and self-esteem between (1) OCD and HC, and (2) OCD and HComp subjects. One sample t-tests against 0 were performed to formally assess the existence over- or underconfidence for both local and global calibration in each of the three groups. Additionally, regression analyses were performed to explore differences between groups in how internal fluctuations in local confidence would predict global confidence, over and above fluctuations in accuracy or reaction times. For these regressions, only blocks without feedback were used (since only these blocks contained local confidence judgments). All predictors were standardized (z-scored). In this analysis we aimed to predict differences in global confidence between tasks using main effects and the interactions between group and the difference in accuracy, RT and local confidence between those tasks, as follows:

Δ global confidence ~ Δ accuracy*group + Δ RT*group + Δ local confidence*group

For all analyses where the measure of local confidence was used (i.e., local calibration, discrimination, correlation of local and global confidence), only the 72 trials from the no-feedback condition were used, since participants only rated their local confidence in those trials. In order to assess if there were differences in the relationship between obsessive-compulsive symptom strength (OCI-R score) and metacognitive abilities between OCD patients and HComp subjects, we performed linear regressions on our metacognition variables with OCI-R score, group and their interaction as predictors.

All analyses codes and anonymized data that will reproduce the figures can be found at https://osf.io/ksfp6/.

Results

Demographics

Demographic and clinical characteristics are given in Table 1. The groups did not differ in terms of age, sex distribution or years of education. OCD patients have significantly higher OCI-R scores than HCs, while OCI-R scores were similar between OCD patients and HComp subjects (Figure 2A). Together, this confirms successful matching of the groups. For details on all descriptive statistics and statistical outcomes, see Table 1. For correlations between questionnaires, see Appendix D Table D1.

Replication Analyses on Task Structure

Using mixed ANOVAs in our clinical sample, we replicated earlier findings investigating the effects of feedback and difficulty on performance and metacognition (Rouault et al., 2019). For performance, reaction times and global confidence we assessed the effects of feedback, difficulty and group, whereas for local confidence we assessed the effects of difficulty, accuracy and group. For none of the analyses interactions between task features and group were found.

In line with previous findings, performance was better for easy versus hard tasks ($F(1,78) = 501.93$, $p<.001$), but did not differ between feedback or no feedback conditions ($F(1,78) = 0.14$, $p = 0.705$). Reaction times were faster for easy versus hard tasks ($F(1,78) = 42.01$, $p<.001$) and tasks that provided feedback versus no feedback ($F(1,78) = 28.45$, $p<.001$).

Global confidence was higher for easy versus hard tasks ($F(1,78) = 87.58$, $p<.001$), and for tasks providing feedback versus no feedback ($F(1,78) = 101.92$, $p<.001$), even

though performance was equal between presence and absence of feedback. The difference in global confidence between feedback and no-feedback tasks was bigger when the tasks were easy ($F(1,78) = 5.10$, $p = 0.0267$). As expected, local confidence was higher for easy versus hard tasks ($F(1,78) = 114.99$, $p<.001$), and for correct versus incorrect trials ($F(1,78) = 217.01$, $p<.001$). Together, these results largely confirm previous observations with this protocol (Hoven, Luigjes, et al., 2023; Rouault et al., 2019).

Comparing OCD patients to healthy controls

In line with our expectations, OCD patients showed significantly lower local calibration compared with HCs, and a trend level of lower global calibration, indicating underconfidence relative to HCs (Table 1, Figure 3A,B). These results were due to significantly decreased local and global confidence levels in OCD compared with HCs, without any performance or reaction time differences (Figure 3C,D,F). One sample t-tests against zero indicated that HCs showed significant local ($t_{39} = 3.42$, $p = .001$) and global overconfidence ($t_{39} = 3.81$, $p<.001$), while local and global calibration did not differ from zero in the OCD group, indicating that the OCD group was well calibrated (local: $t_{39} = -0.09$, $p = 0.928$, global: $t_{39} = 0.86$, $p = 0.397$). Moreover, autonomy (as measured by the ASA), self-esteem (as measured by the rSES) were found to be significantly lower in patients with OCD compared with HCs (Figure 2B,C), while metacognitive beliefs (as measured by the MCQ-30) were significantly more distorted (Table 1).

No significant interactions between task parameters (feedback or difficulty) and group were found, refuting our hypothesis that OCD patients would especially show lower global confidence when feedback was unavailable. Also, no group differences in discrimination or the correlation between local and global confidence were found (Table 1, Figure 3E).

Table 1: Demographics, clinical data and task performance per group and differences between groups

	OCD (N = 40)	HCs (N = 40)	HComp (N=40)	OCD vs. HCs	OCD vs. HComp
Demographics					
Age in years	38.18 (11.22)	38.58 (11.11)	36.53 (12.73)	T = 0.16 P = 0.87	T = 0.61 P = 0.54
Females (%)	26 (65%)	27 (67.5%)	28 (70%)	X2 = 0.81 P = 0.81	X2 = 0.23 P = 0.63
Years of education	10.11 (3.21)	10.20 (3.13)	10.35 (2.64)	T = 0.12 P = 0.90	T = -0.36 P = 0.72
Questionnaire Scores					
OCI-R	23.23 (9.43)	2.90 (2.48)	23.35 (13.18)	T = -13.19 P < .001	T = -0.05 P = 0.96
ASA	133.33 (21.70)	168.13 (19.18)	160.35 (33.99)	T = -7.60 P < .001	T = 4.24 P < .001
rSES	16.95 (4.89)	23.48 (3.94)	18.53 (7.56)	T = -6.57 P < .001	T = 1.11 P = 0.273
DASS	34.2 (17.31)	4.98 (4.23)		T = -10.37 P < .001	
DASS anx	9.05 (6.59)	0.68 (0.92)		T = -7.96 P < .001	
DASS dep	10.38 (8.28)	1.28 (1.47)		T = -6.84 P < .001	
MCQ-30	66.10 (15.45)	42.88 (8.79)		T = -8.26 P < .001	
GAD-7			9.08 (6.20)		
ZungDEP			44.78 (11.11)		
Metacognition					
Accuracy (percent correct)	75.04 (7.00)	76.49 (7.76)	69.90 (8.64)	F = 0.81 P = 0.372 η^2_G = 0.006	F = 8.59 P = 0.004 η^2_G = 0.061
Local Confidence (on 50-100 scale)	74.74 (8.11)	81.14 (8.04)	76.82 (9.58)	F = 12.59 P < .001 η^2_G = 0.129	F = 1.11 P = 0.296 η^2_G = 0.013
Global Confidence	76.24 (7.27)	80.69 (7.34)	76.21 (8.83)	F = 7.42 P = 0.008 η^2_G = 0.069	F = .0002 P = 0.989 η^2_G = $2.1 \cdot 10^{-6}$
Local Calibration	-0.17 (11.83)	4.82 (8.92)	6.63 (11.22)	T = 2.13 P = 0.036 d = 0.48	T = 2.64 P = 0.010 d = 0.59
Global Calibration	1.20 (8.84)	4.20 (6.98)	6.31 (9.62)	T = 1.68 P =0.096 d = 0.38	T = 2.48 P = 0.015 d = 0.55

Correlation Local & Global Confidence	0.51	0.56	0.52	T = -0.80 P = 0.429 d = 0.18	T = 0.18 P = 0.862 d = 0.04
Discrimination	9.40 (5.94)	8.34 (4.77)	6.73 (4.66)	T = 0.88 P = 0.383 d = 0.20	T = -2.24 P = 0.028 d = 0.50

Abbreviations: OCD = Obsessive-Compulsive Disorder, HCs = Healthy Controls, HComp = High-Compulsive subjects, OCI-R: Obsessive-Compulsive Inventory-Revised, ASA: Autonomy Scale Amsterdam, rSES: Rosenberg Self-Esteem Scale, DASS: Depression Anxiety and Stress Scale, DASS anx: Depression Anxiety and Stress – subscale Anxiety, DASS dep: Depression Anxiety and Stress – subscale Depression, GAD-7: Generalized Anxiety Disorder-7 Questionnaire, ZungDEP: Zung's Depression scale, T = T-value from two-sample t-test, F = F-value from ANOVA, P = P-value, η^2_G = Generalized Eta-squared, d = Cohen's d. Data are reported as mean (standard deviation).

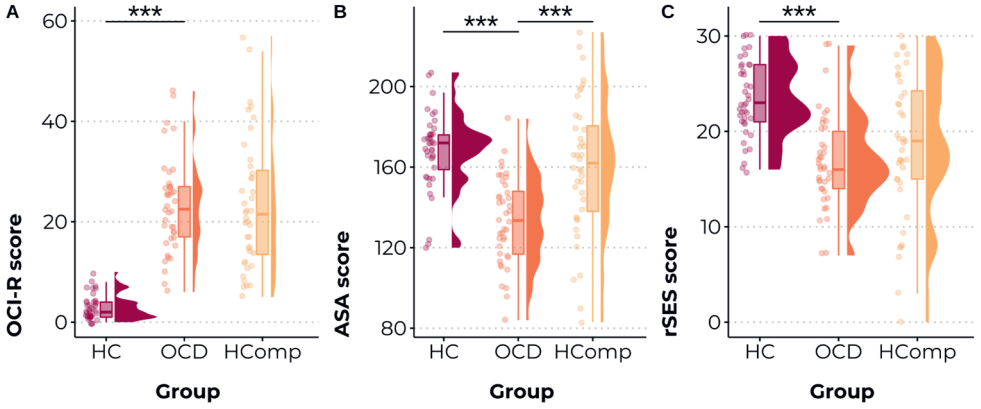


Figure 2: Clinical scores across groups. Scores on the **(A)** OCI-R score, **(B)** ASA score reflecting autonomy and **(C)** rSES score reflecting self-esteem per group. Dots show data from individual participants, boxplots show median and upper/lower quartile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. *p <.05, **p < .01, ***p < .001. HC = Healthy Control subjects, OCD= Obsessive-Compulsive Disorder patients, HComp = High-Compulsive subjects from general population sample, OCI-R = Obsessive-Compulsive Inventory-Revised, ASA = Autonomy Scale Amsterdam, rSES = Rosenberg's Self Esteem Scale.

It has been argued that the findings of decreased confidence in OCD in case-control studies could be driven by comorbid depressive and anxiety symptoms, while compulsivity would contrarily lead to increased (over)confidence (Rouault, Seow, et al., 2018; Seow & Gillan, 2020). We performed regression analyses investigating the effect of group (OCD versus HC) on local and global confidence and calibration, while controlling for anxiety and depression symptoms (DASS scores). The effect of group on all four metacognitive outcome measures remained significant (local confidence: β = -8.508 ± 2.785, p = 0.003; global confidence: β = -6.027 ± 2.526, p = 0.0195; local calibration: β = -11.091 ± 3.521, p = 0.002; global calibration: β = -7.234 ± 2.691, p = 0.009; see Appendix D Table D3 for full regression results). This suggests that in this clinical case-control sample decreases in confidence in OCD compared to HCs were not explained away by comorbid anxiety and depression symptoms.

Comparing OCD patients to high compulsive subjects

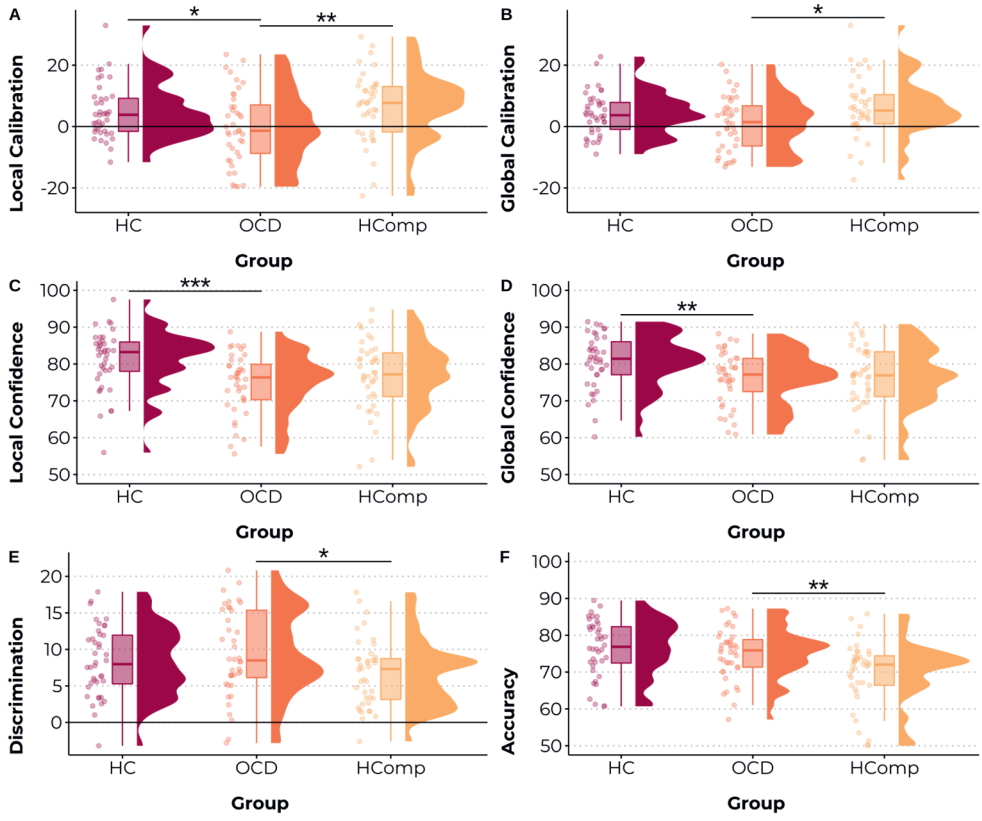
HComp subjects had significantly higher calibration (i.e., more overconfidence) at both local and global levels compared to OCD patients (Table 1, Figure 3A,B). One sample t-tests against zero confirmed that the HComp group showed significant local (t_{39} = 3.73, p<.001) and global overconfidence (t_{39} = 4.15, p<.001). This was due to a significantly worse performance of HComp subjects compared with OCD patients, while local and global confidence levels (and reaction times) did not differ between groups (Figure 3C,D,F). In other words, HComp subjects were just as confident in their decisions as OCD patients, while performing significantly worse, leading to overconfidence. Moreover, autonomy was significantly lower in patients with OCD compared with HComp subjects, but there were no group differences in self-esteem scores (Figure 2B,C).

HComp subjects showed decreased discrimination compared with OCD patients, indicating that the difference in confidence between correct and incorrect choices was smaller in this group, reflecting worse metacognitive sensitivity (Figure 2E). However, no group differences were found in the correlation between local and global confidence. Again, we did not find any significant interaction effects between task parameters (feedback or difficulty) and group.

To deepen our understanding of the relationships between obsessive compulsive symptoms and metacognition beyond group differences, we investigated if OCD patients and HComp subjects showed a different relationship between obsessive compulsive symptom strength and metacognitive ability. Using regression analyses, a trend level interaction effect of OCI-R score and group on local confidence was found

(β = 4.03 ± 2.09, p = 0.057, see Appendix D Table D4 for full regression results). This interaction effect hints at a negative relationship in the OCD patients (i.e., more symptoms reflect lower local confidence), and a positive relationship in the HComp group (i.e., more symptoms reflect higher local confidence), however, post-hoc correlational tests did not show significance for the groups separately (OCD: r = -0.26, t_{38}= -1.63, p = 0.11; HComp: r = 0.18, t_{38}= 1.16, p = 0.25) (Figure 4).

Figure 3: Metacognition and performance across groups.



Local calibration **(A)**, global calibration **(B)**, local confidence **(C)**, global confidence **(D)**, discrimination **(E)**, and accuracy **(F)** data, all in percentages. Dots show data from individual participants, boxplots show median and upper/lower quartile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. For plots A, B and E significance stars represent two-sample t-tests, for plots C, D and F significance stars represent the main effect of group in mixed ANOVAs (see Table 1). *p <.05, **p < .01, ***p < .001. HC = Healthy Control subjects, OCD= Obsessive-Compulsive Disorder patients, HComp = High-Compulsive subjects from the general population sample.

Comparing healthy controls to highly compulsive subjects

For completeness, we performed exploratory analyses to compare the HC and HComp groups using the same methods as were used to compare the other groups. For results, see Appendix D (Table D2).

Interplay between hierarchical levels of metacognition

Using regression analyses we replicated in our clinical sample that differences in local confidence between two tasks significantly inform global confidence differences between those tasks ($\beta = 6.57 \pm 1.21$, $p<.001$), over and above differences in objective accuracy ($\beta = -0.32 \pm 1.05$, $p = 0.761$) or reaction times ($\beta = 0.22 \pm 1.05$, $p = 0.831$). No interaction effects with group were found, suggesting that the relationship between local and global confidence did not differ between OCD patients and HCs, or between OCD patients and HComp subjects. This is in line with non-significant group differences between the correlation coefficients of local and global confidence (Table 1).

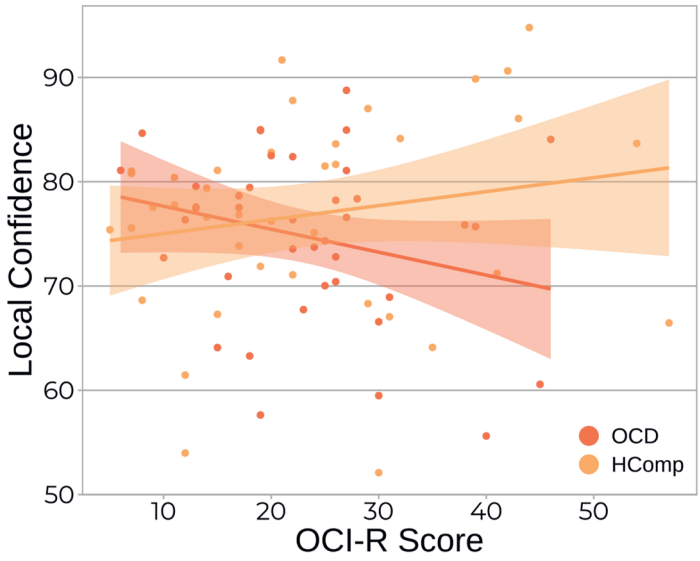


Figure 4: The relationship between local confidence and OCI-R scores in OCD patients and highly compulsive non-clinical subjects. Individual data points showing the relationship between OCI-R score and local confidence, which is negative in the OCD group, and positive in the HComp group. OCD = Obsessive-Compulsive Disorder patients, HComp = High-Compulsive subjects from the general population sample.

Discussion

Human research in psychiatry has historically been carried out by examining either clinical patient samples or psychiatric symptoms at subclinical or clinical levels in samples from the general population. It is assumed, but hardly ever formally tested, that psychological or cognitive processes that play a role in the symptoms in question are comparable between clinical patient samples and general population samples (Abramowitz et al., 2014). The current study tested this assumption by directly comparing carefully matched clinical and analogue groups on cognitive processes central to the development and maintenance of OCD.

In line with our hypotheses and the notion of a negative confidence bias (Dar et al., 2022), the current study shows decreased local confidence in patients with OCD compared to HCs, with no performance differences, where HC are overconfident and OCD patients are relatively more underconfident. Interestingly, this negative bias extended to higher order levels of metacognition, both task-based and questionnaire-based. Patients with OCD compared to HCs had decreases in global confidence, metacognitive beliefs, self-esteem and autonomy. However, critically, OCD patients showed no impairments in confidence estimation or usage: they were just as good in discriminating between correct and incorrect choices using their confidence judgments (i.e., measured using discrimination), did not show specifically decreased confidence in trials without feedback, and showed no distortion of the relationship between local and global confidence. Overall, this supports the notion of a general negative bias across hierarchical metacognitive levels, reflecting the wide-spread nature of these deficits in OCD, with no evidence for disturbances in the estimation and usage of confidence. It remains possible, however, that deficits in metacognitive sensitivity and coupling of metacognitive levels would be more pronounced in clinically relevant contexts than in the current neutral perceptual task (Hoven et al., 2019).

Interestingly, the metacognitive pattern of the high-compulsive general population sample was different from the OCD sample, challenging the assumption that these two sample types are directly comparable. Contrary to the notion of a negative confidence bias in OCD samples, HComp subjects were significantly more overconfident – both at the local (decision) and global (task) levels – than patients with OCD, which was driven by decreased performance with equal confidence. Importantly, the metacognitive aberrancies of HComp did not resemble those of OCD patients. Instead, they were in the opposite direction: HComp individuals had relatively higher overconfidence (albeit not significant, see Appendix D) than HCs. Moreover, directly going against the assumption of similar associations between symptoms and cognitive processes for clinical and general population samples, there were tentative opposite associations

between OC symptoms and local confidence in patients with OCD (negative relationship) and HComp subjects (positive relationship). In line with previous findings of a decreased metacognitive sensitivity (Hauser, Allen, et al., 2017), HComp subjects were worse in discriminating errors from correct answers using their confidence judgments compared with OCD patients.

Unlike most prior case-control studies in OCD, here we controlled for the influence of comorbid symptomatology (e.g., anxiety and depression) on confidence in patients with OCD. Since depression is associated with decreases in confidence (Hoven et al., 2019), it could partly explain lower confidence in OCD. We found, however, that decreases in local and global confidence and calibration levels in OCD compared to HCs remained when controlling for anxiety and depression symptoms. Additionally, anxiety and depression scores in OCD and HComp groups (using the DASS in OCD, and GAD-7 and Zung Depression Scale in HComp) both indicated mild severity. It is thus unlikely that the opposite metacognitive patterns we found are due to strong differences in comorbid symptoms between these samples. In the same line, a possible explanation is that decreased calibration (i.e., relative underconfidence) as found in our OCD sample relates more strongly to (anxiety driven) obsessive symptoms, whereas overconfidence or defects in metacognitive sensitivity would relate more strongly to compulsive symptoms. Yet, obsessive and compulsive symptoms, as measured by the Y-BOCS in the patients, were on average equally severe, going against the idea that more severe obsessions versus compulsions would drive underconfidence.

To account for comorbidities and heterogeneity within OCD and other disorders, a case has been made for transdiagnostic, dimensional approaches (Insel et al., 2010). Studies with large general population samples, found that a symptom cluster of 'Compulsive Behavior and Intrusive Thoughts' (CIT), mostly including symptoms of OCD, schizotypy, eating disorders, alcoholism and impulsivity, was related to increases in local confidence, whereas a symptom cluster of 'Anxious Depression' (AD) was related to decreases in local confidence, while disorder-specific symptoms did not show these associations (Benwell et al., 2022; Rouault, Seow, et al., 2018; Seow & Gillan, 2020). In recent work, we extended these findings showing that CIT symptoms related to local and global overconfidence, while AD symptoms related to local and global underconfidence (Hoven, Luigjes, et al., 2023). In light of previous findings that AD symptoms lead to lower confidence, while CIT symptoms lead to higher confidence, it could be that our current general population sample has higher CIT symptom dimension scores than the OCD sample which may additionally include non-OCD symptoms. Moreover, in the OCD sample we found lower confidence even when

corrected for anxiety and depression symptoms. This questions the idea that the symptom dimensions and their relation with confidence biases may directly translate to a clinical population, at least in the case of OCD and compulsive symptoms. Although caution is warranted in generalizing transdiagnostic findings to clinical populations, transdiagnostic research is valuable in itself (McGorry et al., 2018; Vanes & Dolan, 2021). An impactful step forward would be to apply transdiagnostic research within clinical samples. Recently, within a large patient sample of generalized anxiety disorder and OCD patients, it was found that deficits in goal-directed behavior were more strongly associated with a dimension of compulsivity symptoms than OCD diagnosis status itself (Gillan et al., 2020), supporting the importance of studying both transdiagnostic symptoms and diagnostic criteria in concert in clinical samples.

The current study has to be interpreted in light of its limitations. Because of the difficulty manipulation in the experimental design, we did not use a staircase procedure, and used calibration measures to analyze the strength of correspondence between confidence and performance. Differences in performance between the OCD and HComp group were found, with a negative relationship between OCI-R score and performance in the large general population sample (Hoven, Luigjes, et al., 2023). Including subjects' mean performance in the propensity score matching strongly worsened the matching on our primary variable of interest, the OCI-R score, which is why we did not pursue matching on performance. In next studies it would be useful to keep performance equal between participants to more clearly isolate changes in confidence. Our clinical sample consisted of Dutch OCD patients that were help-seeking, did not use psychotropic medication at time of testing and did not suffer from co-morbid diagnoses. This allowed us to isolate associations with metacognition without these confounds, but could limit the generalizability of our findings to the general OCD patient population, because co-morbidities and medication use are common in OCD (Grabe et al., 2000; Ruscio et al., 2010). Moreover, all subjects were tested online (and originated from a variety of countries), allowing for less control over the environment in which the task was performed. Nevertheless, online testing has many advantages, including lower costs and access to larger and more representative samples. Future studies could investigate metacognition in a more clinically relevant setting, by – for example – studying the effects of symptom provocation on metacognitive abilities, and could study the specific role of obsessions versus compulsions in metacognition. Moreover, metacognition does not only serve monitoring purposes, but also has a controlling function, which should be investigated further in OCD (Vaghi et al., 2017).

Together, these findings argue for being cautious in generalizing metacognitive findings from highly compulsive samples from the general population to clinical samples. In our current samples, with equal OC symptom severity, distinct neurocognitive processes might be at play, relating to OC symptoms in different ways. This caution might not apply similarly to all psychiatric disorders, since for example, both clinical and general population studies have consistently shown decreases in confidence in depression (Hoven et al., 2019; Rouault, Seow, et al., 2018). Overall, the current study showed evidence for decreased local and global confidence, as well as decreased higher order metacognition in OCD patients compared with HCs. Meanwhile, a general population sample with similar OC symptoms showed local and global overconfidence and diminished metacognitive sensitivity compared with OCD patients. The patterns observed in a non-clinical population, used as an analogue for OCD, may thus not necessarily generalize to clinical samples.

Acknowledgments

We would like to thank Katja Cornelissen, Fabiënne Meijboom and Tosca Mulder for their help with data collection. This work was supported by a VENI grant (JL; grant number 916-18-119). JL was supported by a VENI grant (916-18-119), MR is the beneficiary of a postdoctoral fellowship from the AXA Research Fund. MR's work was also supported by the Fondation des Treilles.

Disclosure statement

None of the authors have any conflicts of interest to declare.

7

OCD patients show lower confidence and higher error sensitivity while learning under volatility compared to healthy and highly compulsive samples from the general population

Hoven M

Mulder T

Denys D

van Holst RJ

Luigjes J

Submitted

Abstract

Background

Our sense of confidence guides our actions. A decoupling between confidence and action could relate to compulsive behavior as seen in obsessive-compulsive disorder (OCD). The link between confidence and action in OCD has been investigated in clinical case-control studies and in the general population with discrepant findings. The generalizability of findings from highly-compulsive samples from the general population to clinical OCD samples has been questioned. Here, we address the discrepancies by investigating the relationship between action and confidence in OCD patients compared to healthy controls (HC) and a population sample of matched highly-compulsive subjects (HComp).

Methods

38 medication and comorbid diagnosis free OCD patients, 37 HC and 76 matched HComp participants performed a predictive inference task to investigate action and confidence while learning under volatility. Action-updating, confidence and their coupling in the OCD group were compared to HC and HComp groups. Moreover, computational modeling was performed to compare groups on error sensitivity, and parameters reflecting learning and environmental changes.

Results

OCD patients showed lower confidence and higher learning rates in reaction to (particularly small) prediction errors than both HC and HComp groups, signaling hyperactive error signaling and a negative confidence bias. No evidence was found for differences in action-confidence coupling between groups.

Conclusions

Different behavioral profiles are related to obsessive-compulsive symptoms in different samples, with lower confidence and higher error sensitivity in clinical OCD samples. Overall, the underlying mechanisms of obsessive-compulsive behavior might differ between clinical and highly-compulsive general population samples, resulting in different (meta)cognitive profiles.

Introduction

Throughout daily life we perform actions based on our beliefs and our sense of confidence in those beliefs. For example, if I am confident that I have brought my passport to the airport, I am less likely to engage in actions to double-check it. Following the Bayesian framework of belief updating (Knill & Pouget, 2004; Meyniel & Dehaene, 2017; Parr & Friston, 2017), in addition to new information, confidence in prior beliefs also plays a role in determining the extent to which a belief is updated. The more confident one is in their belief, the less impactful new information is, leading to minimal updates of their existing beliefs. When confidence is low, however, it can motivate the gathering of additional evidence to increase confidence in those beliefs (Boldt et al., 2019; Desender et al., 2018, 2019). In this way, the confidence with which a belief is held shapes our future behavior, particularly in volatile and uncertain environments.

This process of utilizing new information and confidence to guide decision-making can go awry in various disorders psychiatric disorders, among which obsessive-compulsive disorder (OCD). OCD is a psychiatric disorder that is typically characterized by intrusive obsessions and compulsions (American Psychiatric Association, 2013). Moreover, OCD has been associated with abnormalities in confidence (i.e., the subjective feeling of being correct in a choice, belief or decision), with most studies showing lower confidence in patients compared with healthy controls (Dar et al., 2022). A bias toward lower confidence, indicating a disruption in metacognitive monitoring, could lead to doubts and uncertainty that drive compulsive behavior in OCD. An alternative theory on confidence in OCD has suggested that the repetitive nature of compulsions is driven not by underconfidence but by a dissociation between confidence and actions, indicating a disruption in metacognitive control. In this framework, OCD behavior is viewed as a disruption of goal-directed action (Gillan et al., 2011, 2020; Gillan & Robbins, 2014) where information from confidence judgments is not used accurately to inform behavior. For example, patients with OCD have been found to show a weakened association between confidence and updating of actions, while confidence in those actions were not affected (Vaghi et al., 2017). This dissociation of the interaction between action and confidence resembles the clinical presentation of OCD where patients often continue performing actions that they know are disproportionate, and which are by definition ego-dystonic (i.e., not consistent with the persons' beliefs) (Kashyap et al., 2014).

Previous research has examined the relation between confidence and action in individuals with OCD during a volatile learning process (Marzuki et al., 2022; Vaghi et al., 2017). These studies employed a predictive inference task, wherein participants

had to catch particles and assess their confidence in successfully catching the particle. Vaghi and colleagues found that, compared to healthy controls, OCD patients exhibited a dissociation between action and confidence, which was primarily driven by excessive updating of actions in response to small changes in the environment (Vaghi et al., 2017). Specifically, when OCD patients had to guess the location of an upcoming particle using a virtual bucket, they excessively moved their bucket in response to small prediction errors, while their confidence in successfully catching the particle did not differ from healthy controls. In these patients, the usual negative coupling of confidence to the updating of the bucket location was weakened. Using the same paradigm, Marzuki and colleagues also found excessive action updating for small prediction errors in adolescent OCD patients, but did not observe a dissociation between action and confidence (Marzuki et al., 2022).

In addition to clinical samples, OCD is often studied using analogue samples from the general population to assess relationships between (meta)cognitive phenomena and obsessive-compulsive (OC) symptoms (Abramowitz et al., 2014). Seow & Gillan (2020) conducted a study with the same predictive inference task as Vaghi et al. (2017), using a large general population sample and found that OC symptom severity was positively related to a dissociation between action and confidence. However, unlike the results of the study by Vaghi et al. (2017), here the authors reported that subjects with high OC symptom severity showed increased confidence rather than increased action updating. The dissociation between action and confidence was not specific to OC symptoms, however, and was also found for other psychiatric symptoms (e.g., depression, anxiety, psychosis). When using a transdiagnostic approach to consider co-morbid symptoms across psychiatric disorders, it was discovered that a symptom dimension of compulsivity specifically contributed to the action-confidence dissociation through inflated confidence. A recent replication study, however, failed to replicate the associations between OC symptoms and both confidence or the coupling between action and confidence (Loosen et al., 2023).

In general population studies it is often assumed that the (meta)cognitive abilities of highly compulsive individuals resemble those of patients with OCD, albeit to a lesser degree. However, in a recent study we challenged this assumption and showed distinct metacognitive patterns in highly compulsive individuals and patients with OCD (with similar OC symptom severity), suggesting that these groups are inherently different (Hoven, Rouault, et al., 2023). This finding may explain why previous research has shown conflicting results on the cause of the dissociation between action and confidence, with one study finding it was due to increased action updating (Vaghi et al.,

2017), while the other study found that it was due to increased confidence (Seow & Gillan, 2020).

Here we aimed to address the discrepancies between previous studies by investigating action, confidence, and their coupling in a group of OCD patients who were not taking medication and did not have any co-morbid diagnoses. We compared this group to two other samples: (1) a group of healthy controls and (2) a group of highly compulsive individuals matched to the patient group. Moreover, we will compare our subjects' behavior to a reduced Bayesian model used in previous studies, allowing us to compare how patients, healthy controls and highly compulsive subjects from the general population respond differently to various sources of environmental information (i.e., recent outcomes, surprising outcomes, uncertainty, and feedback) in updating their actions and confidence. We found evidence for lower confidence and higher error sensitivity in OCD patients compared to both healthy controls and highly compulsive subjects, but not for group differences in the coupling between action and confidence. Overall, this indicates that caution is warranted in generalizing findings from high compulsive samples from the general population to clinical OCD samples.

Methods

Participants

This study included three groups of participants: patients with OCD, healthy controls (HCs) and highly compulsive individuals from the general population (HComp). The study was approved by the Medical Ethics Committee of the Amsterdam University Medical Centre, and all subjects provided written informed consent before participating and were reimbursed for their time.

Patients with OCD

Recruitment through the psychiatry department of the Amsterdam University Medical Centre and OCD community websites resulted in the inclusion of 43 OCD patients between 18 and 65 years in the study whose diagnosis of OCD was confirmed using structured psychiatric interviews. Exclusion criteria included a current diagnosis of major depressive disorder, (hypo)mania, anxiety disorders, substance use disorders or psychotic disorders, and use of medication for the treatment of psychiatric symptoms during the time of inclusion.

Healthy controls

45 HCs were included and recruited through online advertisements and matched to OCD patients in terms of age, gender, and education.

Highly compulsive individuals from the general population

Data from HComp individuals from the general population was obtained from a previous study by Seow & Gillan (2020). This study collected data from a large general population sample (N=589) through Amazon's Mechanical Turk, with 427 participants remaining after applying exclusion criteria, using similar task-based exclusion criteria as the current study (see '*Subject task-based exclusions'*). These HComp participants completed the exact same online task as the OCD and HC groups, but with 150 trials per participant rather than 300.

We used propensity score matching to select participants from the HComp sample to match the patient sample in terms of OC symptom severity. We used the MatchIt package in R (Ho et al., 2007) to perform optimal pair matching to minimize the sum of the absolute pairwise distances in the matched sample. To balance the number of trials completed, we used a 1:2 ratio of OCD participants to HComp participants in the matching process, resulting in a similar number of data points for both groups. Matching was performed based on the Obsessive-Compulsive Inventory-Revised score (OCI-R (Foa et al., 2002)), age, and sex. Demographics were compared between groups using two-sample t-tests for continuous measures and Chi-square tests for categorical measures. Our final sample, after task-based exclusions (see '*Subject task-based exclusions*' for more information) consisted of 38 patients with OCD, 37 HCs and 76 HComp participants.

Questionnaires

HC and OCD participants were assessed using the MINI structured psychiatric interview to screen for additional psychiatric disorders (Sheehan et al., 1998). OC-symptoms were measured using the Obsessive-Compulsive Inventory - Revised (OCI-R) in all participants (Foa et al., 2002), and symptom severity was additionally assessed in patients with OCD using the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS). Patients with a YBOCS score < 12 were not included in the study. Anxiety and depression symptoms were assessed using the Depression Anxiety and Stress Scale (DASS) (Parkitny & McAuley, 2010) in OCD and HC and the STAI (Marteau & Bekker, 1992) and Zung's self-rating depression scale (Zung, 1965) in HComp participants.

Predictive inference task

We used the web-based version of the predictive inference task, originally described by Nassar et al. (2010), and modified by Seow & Gillan, (2020). The task involved error-driven learning, in which participants had to infer the landing location of a particle based on its previous landing locations. To do this, participants were shown a circle with a center dot and asked to place a "bucket" (represented by a curved rectangle) at the predicted landing location of the particle, which they could update after each trial. After confirming their bucket placement, participants rated their confidence that the particle would land in the bucket on a scale of 1 (not at all confident) to 100 (extremely confident). The confidence scale was randomly initiated at a rating of either 25 or 75, to stimulate participant action (Figure 1).

After confirming the confidence rating, a particle was fired from the center dot. The landing location of the particle was sampled from a Gaussian distribution with a fixed standard deviation (SD) of 12. At certain trials, known as change-points (CP), a new mean was drawn from a uniform distribution over the full range of the circle U(1,360), with a probability of 0.125 (hazard rate, H). Optimal task performance thus required participants to distinguish between signal (change-point) and noise (SD of the generative distribution). Participants were rewarded for correctly catching the particle in their bucket and penalized for missing it via point summations and subtractions, respectively.

The task consisted of 4 blocks of 75 trials, with a practice round that was not included in the final score and not analyzed. Participants were given a quiz after practicing that they had to answer correctly before the task started to ensure they understood the task instructions. Participants were instructed to earn as many points as possible, which would be converted to monetary rewards and could be up to €5. Confidence ratings were not directly incentivized, but participants were instructed to rate their confidence as accurately as possible. If participants had left their confidence rating as the default for more than 70% of the trials at the 20th and 50th trial mark, participants were reverted back to the instructions.

All participants performed the same task, which was coded in JavaScript and hosted on Gorilla for the OCD and HC group (Anwyl-Irvine et al., 2020), and on Amazon's Mechanical Turk for the HComp group. The HComp group completed 150 trials per participant, while the OCD and HC groups completed 300 trials per participant.

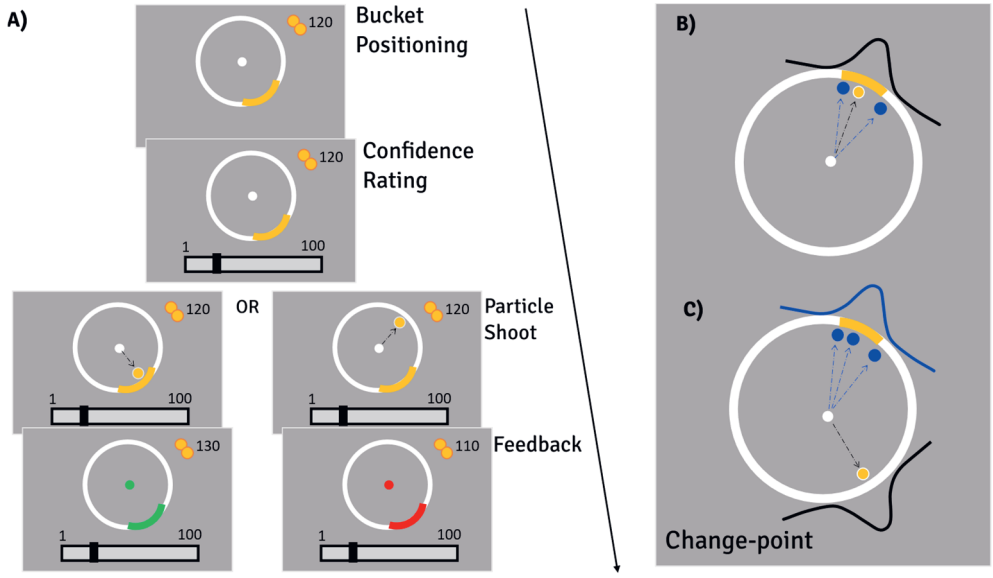


Figure 1: Predictive Inference Task. (A) Trial sequence of the task. Participants had to position their bucket (i.e., yellow bar on the edge of the circle) to catch a flying particle that was released from the center dot to the edge of the circle. After positioning their bucket, participants indicated their confidence in catching the particle. The particle was either caught (bar turned green) or missed (bar turned red), which resulted in gaining or losing points, respectively. **(B)** In every trial, the particle landing position was sampled from a random Gaussian distribution. The trajectories were always straight lines from the center to the landing position. Particles thus landed close together with a small amount of noise. Current trial particle trajectory is marked in black, previous trial particle trajectories are marked in blue. Over time participants thus learn about the Gaussian distribution from which the particle trajectories are drawn. **(C)** During a change-point the mean of the Gaussian distribution abruptly changes to another point in the circle. The particle landing locations are then sampled in a similar manner, until a new change-point occurs. Figure was adapted with permission from Seow et al., (2020).

Task-based exclusions

We preregistered task-based exclusions (https://osf.io/zury3), based on criteria set by Seow & Gillan, (2020). Specifically, participants were excluded if they left their confidence rating at the default score on >60% of trials (n = 3; 1 OCD), if their mean confidence after hits was lower than their mean confidence after misses (n=7, 3 OCD), or if the correlation between confidence rating and the default confidence was >0.5 (n=7, 2 OCD) to ensure subjects sufficiently used the confidence scale. After applying these criteria, the final dataset included data from 75 participants (38 OCD, 37 HC; 26 females in each group). The same exclusion criteria were applied to the dataset from Seow & Gillan (2020). In addition to subject-based exclusions, we also performed trial-based exclusions (see section '*Computational Model*').

Analyses

All data preparation and analyses were conducted using MATLAB (version 2018b) and R (version 4.2.1). We compared the OCD and HC groups, as well as the OCD and HComp groups separately. Our analysis plan for the clinical case-control sample was pre-registered (https://osf.io/zury3), and the same analyses were applied to the comparison between OCD and HComp groups. Additional control analyses can be found in the Appendix E.

Action and confidence

Our first aim was to compare action updating and confidence between groups. We used linear-mixed effects models using the package lme4 (Bates et al., 2015), with either action update (absolute difference in bucket position from trial (t) to trial (t+1)) or confidence as dependent variable and group as predictor, together with random intercepts.

Action-confidence coupling

Our second aim was to assess differences in the strength of action-confidence coupling between groups. We constructed a linear mixed-effects model with trial-by-trial action update as the dependent variable and trial-by-trial confidence (z-scored), group and their interaction as predictors. Random intercepts and slopes of the effect of confidence were added.

In addition, we conducted a Pearson's correlation to examine the relationship between the strength of action-confidence coupling and OCI-R scores in the OCD group, using subject-level β coefficients of the action-confidence coupling.

Computational model

A computational modeling approach was used to examine whether and how the relationship between behavior on the task (i.e., action or confidence) and various environmental parameters differed between groups. In a volatile environment, participants must adjust their learning rate based on recent evidence in order to update their beliefs about the generative distribution. Large prediction errors signal a radical change in the environment, requiring strong belief updating with higher learning rates, while small prediction errors likely are noise and do not require belief updating, in which case learning rates are low.

The human prediction error $\hat{\delta}_t$ (PE) for each trial was calculated as the difference between the current bucket position b_t and the particle landing location X_t.

$$\hat{\delta}_t = X_t - b_t$$

Human learning rate $\hat{\alpha}_t$ (LR) was then calculated as the fraction of PE used for the subsequent action update, which was calculated as the absolute difference in bucket position from trial (t) to trial (t+1):

$$\hat{\alpha}_t = \frac{|b_{t+1} - b_t|}{\hat{\delta}_t}$$

Trials were excluded from all analyses (both model-free and model-based) if the LR exceeded the 95th percentile (4.9% of OCD trials, 5% of HC trials, 5% of HComp trials), which was calculated separately for each group (Marzuki et al., 2022). Due to motor noise in bucket movement when using the keyboard, trials with very small prediction errors (<5) are sensitive to measurement error (McGuire et al., 2014). In this way, trials with small PEs often resulted in very large LRs, even if action update was minimal. Several trials in our sample revealed to have extremely high learning rates due to motor noise, we applied a more stringent exclusion threshold than we reported in our pre-registration, similar to the one used in the paper by Marzuki et al. (2022). In addition, trials with PE = 0 were excluded, since these trials do not drive error-driven learning (1.96% of OCD trials, 2.06% of HC trials, 1.99% of HComp trials). Additionally, the first and last trials within each block were excluded from analyses; in the first trials, there is no error-driven learning yet, and for the last trials no learning rate could be calculated. Finally, due to technical server failure, some trials were not properly recorded and therefore not analyzed (53 OCD trials, 2 HC trials), along with the trials following those corrupted trials, since action updates could not be calculated in these cases. In total, 8.98% of OCD trials, 9.12% of HC trials and 9.09% of HComp were excluded from all analyses.

Error sensitivity

To assess group differences in error sensitivity for learning, linear mixed models were constructed with human LR as the dependent variable and human PE, group and their interaction as predictors. For visualization of the relationship, values of PE were binned into 20 quantiles that each contained an equal fraction of trials. For each quantile, the average LR was computed per subject.

Computational analyses

Additionally, task behavior of participants was analyzed using a quasi-optimal Bayesian observer model that approximates optimal task behavior (Nassar et al., 2010), that was also used in previous research (Marzuki et al., 2022; Seow & Gillan, 2020; Vaghi et al., 2017). The publicly shared data by Seow & Gillan, (2020) included the model parameters derived from the quasi-optimal Bayesian observer model for each HComp subject. Using the same model (publicly available from (Vaghi et al., 2017)) we fitted the particle landing locations of individual subjects to obtain model parameters for the OCD and HC groups.

The model parameters represent statistical characteristics of the environment experienced by participants during the task. In short, these statistical features included the prediction error δ (PE, the absolute difference between model belief and location of the particle), the probability that a change-point occurred (CPP, the likelihood that the sampling distribution of the particle's location has changed, thus that a change-point has occurred), and relative uncertainty (RU, the fraction of uncertainty about the mean that is not due to noise). RU was expressed as its inverse, named model confidence (MC, related to the precision of the model's own beliefs about the mean), to allow for a comparison with confidence reported in the task (Vaghi et al., 2017). For more information on the model see Appendix E.

We assessed how these different Bayesian parameters related to participant behavior, and whether this differed between the groups. Following previous studies, participant behavior (either action or confidence) was regressed against three latent variables computed by the Bayesian model: absolute PE, CPP and (1-CPP)(1-MC), and the categorical variable hits, indicating whether the particle was caught. While PE represents uncertainty regarding the most recent observation, CPP and (1-CPP)(1-MC) represent the model's estimation that a change-point did or did not occur, given the sequence of past observations, respectively. The dependent variable action was calculated as: LR * PE, indicating the bucket update. The predictors in the action model were also interacted with PE for the regression on action (McGuire et al., 2014; Nassar et al., 2019; Seow & Gillan, 2020; Vaghi et al., 2017). Mixed models were constructed with either action or confidence as the dependent variable, and the three model parameters and Hit as fixed-effect predictors (all z-scored), which were all interacted with group. Random intercepts and slopes of all predictors were also included in the model.

In addition, we also performed sensitivity analyses where we calculated the best fitting hazard rate parameter for each subject based on the fit of the model on the participant's behavior (see Appendix E).

Results

Demographics

There were no differences in age or gender distribution between HC and OCD groups, or between HComp and OCD groups. OCD patients had significantly higher OCI-R scores than HCs, while OCI-R scores were similar for OCD and HComp groups, which confirms successful matching of the groups. For details on demographics and clinical data, see Table 1.

Table 1: Demographics, clinical and task-based variables

	OCD	HC	HComp	OCD vs. HC	OCD vs HComp
Age	36.3 (10.9)	38.9 (10.9)	36.8 (11.1)	$t_{73} = 1.02$ $p = 0.31$	$t_{112} = 0.22$ $p = 0.83$
Females (%)	26 (68.4%)	26 (70.3%)	49 (64.5%)	$X^2 = 0.03$ $p = 0.86$	$X^2 = 0.18$ $p = 0.68$
Years of education	3.9 (0.9)	3.8 (0.8)		$t_{73} = -0.72$ $p = 0.47$	
OCI-R	24.0 (10.6)	2.6 (2.2)	24.8 (14.7)	$t_{40.3} = -12.19$ $p < .001$	$t_{112} = 0.31$ $p = 0.75$

Abbreviations: OCD = Obsessive-Compulsive Disorder, HC = Healthy Controls, HComp = High-Compulsive subjects, OCI-R: Obsessive-Compulsive Inventory-Revised. Data are reported as mean (standard deviation). Welch's t-tests were used to compare OCI-R scores between OCD and HC groups, since variances were not equal.

Comparing OCD patients to healthy control subjects

Model-free results

Lower confidence in OCD but no differences in action updating

To investigate whether our groups differed in terms of task behavior, we conducted a mixed-model analysis comparing action and confidence between the OCD and HC groups. Patients with OCD had significantly lower confidence than HCs ($\beta = -18.9$ (4.9), $t = -3.83$, $p<.001$), but there were no differences in the amount of action updating between groups ($\beta = 0.8$ (1.5), $t = 0.56$, $p=0.579$) (Figure 2).

No differences in action-confidence coupling

Next, we evaluated whether the coupling between action update and confidence differed between the groups. As expected, a significant negative relationship between confidence and action update existed across groups, such that higher confidence was related to less action updating (i.e., action-confidence coupling) (β = -9.06 (0.85), t = -10.66, p<.001). However, there was no evidence for a distortion of this action-confidence coupling in OCD, as no interaction between group and confidence on action updating was found (β = 1.06 (1.19), t = 0.89, p=0.379) (Figure 3A). The same results were found when using confidence update from trial t-1 to t as a predictor.

Symptom severity and task behavior

To rule out the possibility that the decrease in confidence in OCD was due to comorbid anxiety and depression symptoms, we conducted a similar mixed-model analysis while controlling for DASS scores. The effect of group on confidence remained significant (β = -27.1 (7.8), t=-3.48, p<.001), while there was no effect of DASS score (β = 0.30 (0.22) t = 1.36, p = 0.179). This suggests that the lower confidence in OCD compared to HCs was not explained by comorbid anxiety and depression symptoms.

Higher learning rates for small prediction errors in OCD

We also assessed differences in error sensitivity between the OCD and HC groups. Across both groups, learning rates increased as a function of prediction error magnitude (β = 0.004 (0.0002), t = 24.00, p<.001), and thus learning rates were highest after large errors. This effect was less pronounced in OCD (significant PE x group interaction effect: β = -0.001 (0.0002), t = -4.89, p<.001). To unpack this effect, a post-hoc mixed-model analysis binning the prediction error in 3 quantiles (i.e., low, medium and high error magnitude), showed that OCD patients specifically had increased learning rates when error magnitude was small (HC-OCD estimate = -0.16 (0.06), Z-ratio: -2.57, p=0.01). Learning rates were not higher for OCD in general (β = 0.113 (0.06), t = 1.78, p = 0.080. This indicates that only when errors were small, the influence of the most recent outcome on subsequent action (i.e., PE) was higher in the OCD compared to the HC group (Figure 4). Moreover, the learning rate at small error magnitude was significantly positively related to OCI-R score in OCD patients (r = 0.34, p = 0.039).

Model-based results

No differences in the effect of Bayesian parameters on behavior

Finally, we assessed whether behavior (action and confidence) was differently predicted by the Bayesian parameters, which represent different forms of uncertainty

and feedback. As expected, action was significantly predicted by all model-derived parameters and hit, such that increases in PE, CPP and (1-CPP)*(1-MC) predicted an increase in action, while a successful catch of the particle predicted a decrease in action. We did not find any evidence for group differences in the strength of these effects (Figure 5). Confidence was, as expected, negatively predicted by both PE, CPP and (1-CPP)*(1-MC), and increased with a successful catch of the particle. Again, we did not find any evidence for group differences in the strength of these effects (Figure 5).

No differences in perceived hazard rates

Sensitivity analyses in which we performed the same analyses as described above, including the subject-specific perceived hazard rate as a covariate (for calculation see Appendix E) indicated that the perceived hazard rate did not differ between the OCD and HC groups. Moreover, none of the significant group differences found between OCD patients and HCs were influenced by differences in perceived hazard rate between groups. For more details, see Appendix E.

Comparing OCD patients to highly compulsive subjects from the general population

Model-free results

Lower confidence and higher action updating in OCD

Patients with OCD showed significantly lower confidence than HComp participants (β = -14.43 (4.50), t = -3.20, p=.0018), and in addition, patients had significantly higher action update than the HComp group (β = 4.81 (1.09), t = 4.41, p<.001) (Figure 2).

No differences in action-confidence coupling

Next, we assessed group differences in the action-confidence coupling. Again, as expected we found a negative relationship between confidence and action update across groups (β = -8.65 (0.75), t = -11.60, p<.001). We did not find an interaction effect between group and confidence (β = 0.65 (1.25), t = 0.52, p=0.60), suggesting similar coupling of action and confidence in OCD patients and HComp subjects (Figure 3A). The same results were found when using confidence update (difference in confidence from trial t-1 to t as used in Vaghi et al. (2017)) instead of confidence as predictor.

Symptom severity and task behavior

To get more insight into the relationship between task behavior and symptom severity, we performed a linear regression to investigate whether the relationship between confidence and OCI-R score differed between the OCD and HComp groups. Indeed, a significant interaction effect was found (β = -0.73 (0.36), t = -2.03, p = 0.04), indicating that there was a significantly positive relationship between OCI-R and confidence for the HComp group (β = 0.824 (0.16), t = 5.12, p < 0.001), while no relationship existed for the OCD group (β = 0.095 (0.32), t = 0.30, p = 0.768). This indicates that HComp subjects with higher symptom severity were *more* confident compared to those with low symptom severity, whereas this is not the case in the patient group. We did not find evidence for a group difference in the relationship between OCI-R score and action updating.

We further aimed to investigate whether the relationship between action-confidence coupling and OCI-R score differed between groups. Using a regression analysis, a significant interaction between OCI-R score and group was found (β = -0.18 (0.08), t = -2.12, p = 0.036), indicating that in HComp there was a positive relationship between OCI-R score and action-confidence coupling (β = 0.188 (0.04), t = 4.98, p<.001), while no relationship existed for the OCD group (β = 0.010 (0.08), t = 0.13, p = 0.896) (Figure 3B). This suggests that subjects from the HComp group with more severe obsessive-compulsive symptoms had a weaker coupling between action and confidence.

Higher learning rates in OCD, regardless of error magnitude

In terms of error sensitivity, we again showed that learning rates increased as a function of prediction error magnitude across groups (β = 0.005 (0.0001), t = 35.66, p<.001). This effect was less pronounced in OCD than HComp (β = -0.002 (0.0002), t = -10.88, p<.001). Moreover, OCD patients had higher learning rates than HComp subjects over the whole range of error magnitudes (main effect of group: β = 0.342 (0.045), t = 7.53, p<.001) (Figure 4).

Model based results

More sensitive to prediction error, but less sensitive to surprising outcomes and relative uncertainty in adjusting actions in OCD

Lastly, we compared the effects of the Bayesian parameters on behavior in the OCD and HComp groups. We found the same main effects of PE, CPP, (1-CPP)*(1-MC) and hit on action and confidence. However, now, the strength of the effects on action differed significantly between the groups. We found a stronger effect of PE on action in OCD compared to HComp, indicating that OCD patients were more sensitive to the

most recent error magnitude in adjusting their action. Contrarily, we found weaker effects of CPP and (1-CPP)*(1-MC) on action in OCD, suggesting that OCD patients were less sensitive to surprising outcomes and relative uncertainty over the course of past observations in adjusting their subsequent actions compared to HComp subjects (Figure 5). We did not find any evidence for group differences in the strength of the effects of the model parameters on confidence (Figure 5).

Higher perceived hazard rates in OCD

Again, we performed sensitivity analyses to account for differences in subject-specific perceived hazard rates. These analyses showed that the group difference in action update between OCD and HComp groups was predominantly driven by a group difference in perceived hazard rates (which was higher in OCD than in HComp), while all other significant effects were unaffected by including hazard rate as a covariate. This indicates that the difference in perceived hazard rate between OCD patients and HComp participants explained their difference in action updating, but not their differences in confidence and learning rate. For more details, see Appendix E.

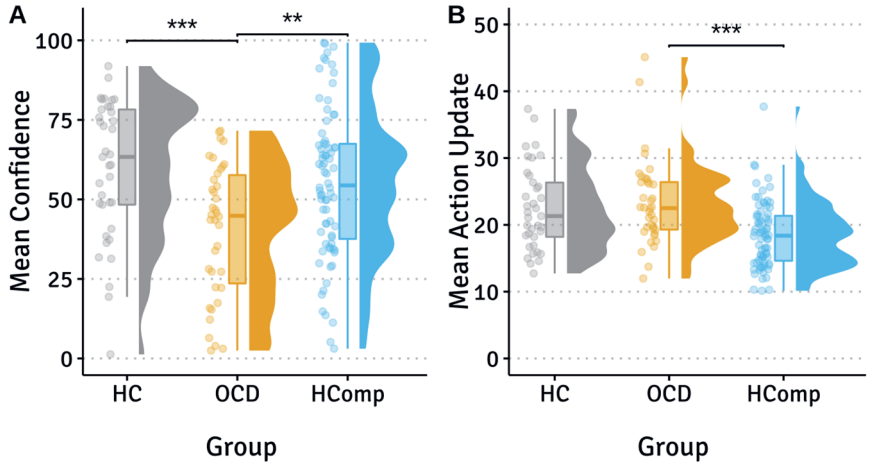


Figure 2: Task behavior across groups. Mean confidence **(A)** and action update **(B)** per group. Dots show data from individual participants, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. Significance starts represent the main effects of group in the respective mixed-effects models. **p<.01, ***p<.001. HC = healthy control subjects, OCD = obsessive-compulsive disorder patients, HComp= highly compulsive subjects from the general population.

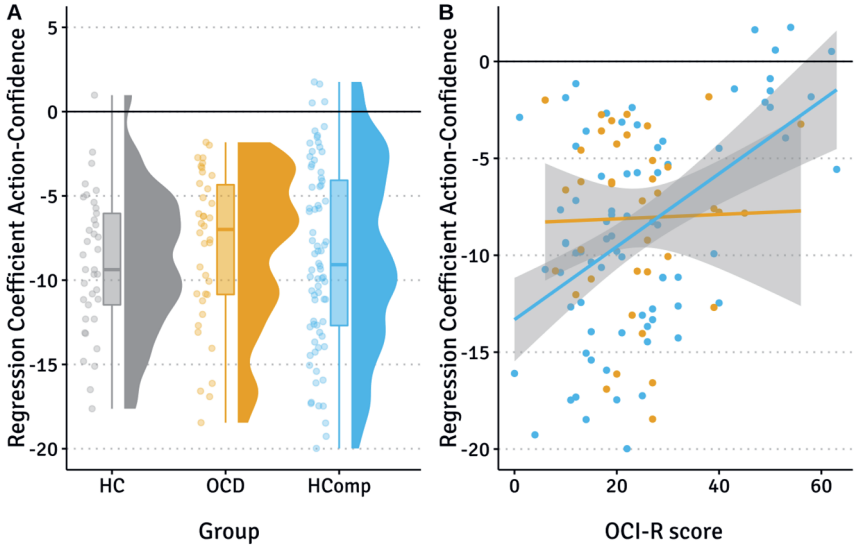


Figure 3: Action-confidence coupling across groups. (A) Results from a regression model where action update was predicted by confidence. As expected, across all groups, regression coefficients were negative indicating that higher confidence was associated with smaller action updates of the bucket location. Dots represent regression coefficients of individual subjects, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. **(B)** The

relationship between symptom severity as measured by the OCI-R and action-confidence coupling in OCD and HComp groups. The shaded areas represent the 95% confidence intervals. A significant interaction effect indicated that a positive relationship exists for the HComp group, but not for the OCD group, indicating that HComp subjects with higher symptom severity had more distorted action-confidence coupling. HC = healthy control subjects, OCD = obsessive-compulsive disorder patients, HComp= highly compulsive subjects from the general population, OCI-R = obsessive-compulsive inventory revised.

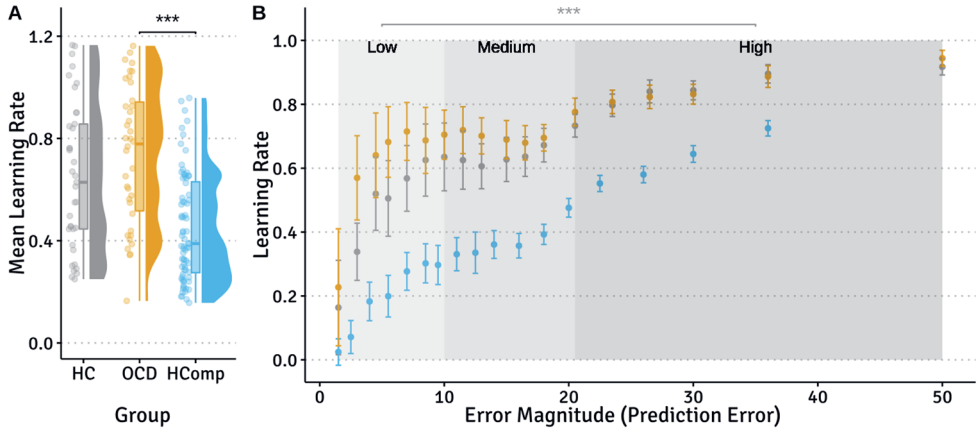


Figure 4: Learning rates and error sensitivity. (A) Mean learning rates per group ($\hat{\alpha}_t$). Patients had significantly increased learning rates compared to the HComp group, but not to the HC group. Dots represent regression coefficients of individual subjects, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. **(B)** The relationship between prediction error magnitude ($\hat{\delta}_t$) and learning rate for each group. Prediction errors were divided in 20 quantiles, of which 18 quantiles are shown here for visualization purposes. Dots represent mean learning rates per group, error bars represent the SEM. All groups' learning rates were higher when prediction errors were larger. Learning rates were higher in the OCD group compared to the HC group at low error magnitudes, and compared to the HComp group across the whole range of error magnitudes. *** p<0.001, represents the main effects of prediction error on learning rate in both mixed-effects models.

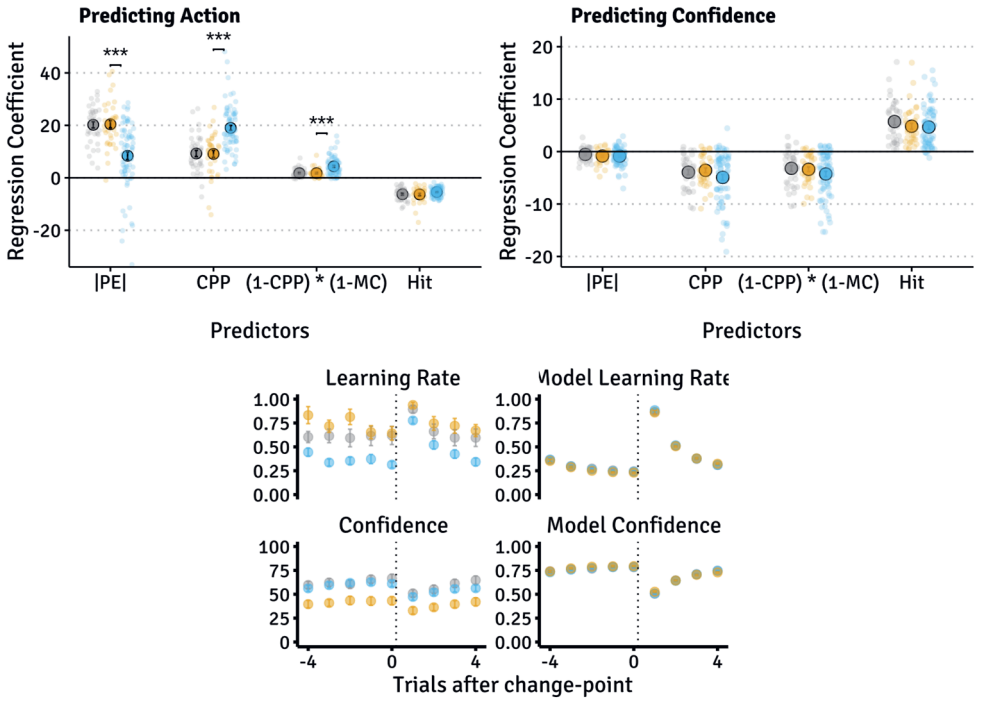


Figure 5: Model-based results on action and confidence. Regression coefficients of the regressions assessing the relationship between the parameters from the computational model and **(A)** human action (i.e., learning rate * absolute prediction error), or **(B)** human confidence. Small dots represent individual regression coefficients, big dots represent mean regression coefficients per group, error bars denote SEM per group. Predictors included absolute prediction error (PE), change-point probability (CPP), model confidence (MC) and a categorical variable representing hits/misses. **(C)** Human learning rate, model learning rate, human confidence and model confidence aligned to change-points (vertical line). Learning rates increased and confidence decreased after a change-point. Confidence was decreased in the OCD group across the entire range of trials, whereas learning rate was decreased in the HComp group across the entire range of trials.

Discussion

In this study we sought to extend our understanding of the relationship between action and confidence in a volatile learning environment, comparing individuals with OCD to both healthy and highly compulsive people from the general population. The current paradigm has been previously used to investigate OCD adult patients (Vaghi et al., 2017), OCD adolescent patients (Marzuki et al., 2022) and non-clinical samples with varying obsessive-compulsive symptoms (Seow & Gillan, 2020). Here, we specifically included medication-free OCD patients without comorbid diagnoses to obtain a more

specific profile of disturbances in action, confidence and their coupling, which we further aimed to compare to a highly compulsive non-clinical sample.

Findings are inconsistent and discrepant between studies examining clinical OCD samples. We did not find evidence that the coupling between action and confidence was disturbed in OCD, in concordance with Marzuki et al. (2022), but in contrast to Vaghi et al. (2017). In addition, model-based results are inconsistent; while earlier studies showed that OCD patients were more (Vaghi et al., 2017) or less (Marzuki et al., 2022) influenced by prediction errors in adapting their action or confidence, respectively, than healthy controls, we did not find any differences in the effects of the model-based parameters on action or confidence between OCD patients and the control group.

A consistent finding between the various studies is an increased learning rate in OCD patients compared to controls (Vaghi et al., 2017), specifically when prediction error is low (Marzuki et al., 2022). The finding that OCD patients are especially prone to excessive action updating in response to non-relevant small errors, without a beneficial effect on their performance, could be indicative of hyperactive error signaling (Norman et al., 2019; Stern et al., 2011). Increased error sensitivity is a well-known endophenotype of OCD (Riesel, 2019), which has been related to an increased risk for developing OCD (Riesel et al., 2011). Overly precise action updating also resembles excessive checking and information gathering behavior typical for OCD, which has been found especially when actions come with no external cost (Banca et al., 2015; Hauser, Allen, et al., 2017; Toffolo et al., 2016).

OCD patients were less confident than controls in their actions, while accuracy and action updating were equal, corroborating previous work (Dar et al., 2022; Hoven et al., 2019). Even when controlling for anxiety and depression symptoms, OCD patients still showed lower confidence than controls, refuting the idea that decreases in confidence in OCD might be driven by comorbid anxiety and depression symptoms (Seow & Gillan, 2020).

After comparing our comorbid and medication-free OCD sample to healthy controls, we aimed to compare the OCD sample to a non-clinical sample with equal obsessive-compulsive symptom strength. Many studies have used analogue samples from the general population to study the relationship between (meta)cognitive phenomena and psychiatric symptoms, with the assumption that these relationships resemble those found in clinical patient samples. However, we recently showed divergent relationships between metacognition and obsessive-compulsive symptoms in patients with OCD and highly compulsive individuals from the general population with similar OCD

symptom severity (Hoven, Rouault, et al., 2023). Here, we corroborate our previous findings, showing that OCD patients had decreased confidence compared to the HComp group, which was specific to the OCD group as exploratory analyses showed no differences between healthy controls and the HComp group. Furthermore, OCD patients had increased learning rates compared to the HComp group and their actions were more strongly predicted by prediction error and less by the change point probability or uncertainty about the landing position. These results, however, were not specific to OCD patients and may relate to a relatively higher perceived volatility of the task as indicated by higher hazard rates in both the OCD and HC groups compared to the HComp group (Figure S2). Interestingly, different relationships were found between OC symptoms and task parameters for the OCD and HComp group: a positive relationship between OC symptom severity and confidence was found in the HComp group, which was not present in OCD patients. Finally, although there was no difference in the coupling of action and confidence between the groups, HComp subjects with more severe OC symptoms had significantly weaker action-confidence coupling (Seow & Gillan, 2020), while this relationship did not exist in OCD patients.

These findings point to the idea that there are different behavioral and (meta)cognitive profiles that go together with obsessive-compulsive symptoms, which might be contingent on the clinical or sub-clinical nature of the sample in question. In line, it is plausible that there are different mechanisms that could relate to similar obsessive-compulsive symptom severity, but to different behavioral manifestations. Our findings support a recent model of OCD proposed by Fradkin et al. (2020) which suggests that OCD patients experience difficulty in using past experiences to inform future actions, resulting in excessive uncertainty about their own actions (e.g. low confidence) and the state of the world. This low confidence may lead to increased reliance on immediate sensory feedback at the expense of prior beliefs. This profile of behavior seems consistent with our profile of lower confidence and higher learning rates during non-relevant small errors in OCD patients. The positive correlation between learning rates during small errors and OCD symptoms further supports this interpretation. On the other hand, Fradkin's model suggests that compulsive behavior can also result from overreliance on prior beliefs at the expense of new evidence, leading to habitual behavior taking precedence. This interpretation is more in line with the profile of the HComp group, which showed a more restricted range of action (lower learning rate), higher confidence and a positive relationship between compulsive symptoms and the decoupling between action and confidence. It suggests that while OCD patients and HComp individuals may present with similar symptoms, the underlying mechanisms of their behavior may differ substantially. However, we acknowledge that alternative explanations for the differences exist, and more research is needed to establish

whether indeed compulsive behavior of these two groups can have different origins. Furthermore, it is worth noting that even though the symptom severity as measured with the OCI-R is similar between groups, the extent to which their symptoms impact daily life may be different. Since the OCI-R measures distress induced by specific and select types of obsession and compulsions, it can confound severity with the type and range of symptoms (Abramovitch et al., 2020). It is plausible that individuals suffering from OCD exhibit a greater frequency of compulsive behaviors on a daily basis, leading to a more pronounced impact on their work, social interactions, and family life compared to those in the HComp group, even if their OCI-R scores are similar. The comparability of the burden of the compulsive symptoms in clinical and analogue samples is a topic worth exploring further using more comprehensive assessment of OCD symptoms.

Compulsivity is a broad concept that is defined as "*repetitive acts that are characterized by the feeling that one 'has to' perform them while one is aware that these acts are not in line with one's overall goal*" (Luigjes et al., 2019). Compulsive behavior is observed in other disorders than OCD, such as (gambling) addiction (Figee et al., 2016), where it instead goes hand-in-hand with increased confidence (Hoven et al., 2019). Moreover, multiple previous studies have shown that a transdiagnostic factor incorporating compulsivity and intrusive thoughts related to increased confidence in sub-clinical samples (Benwell et al., 2022; Hoven, Luigjes, et al., 2023; Rouault, Seow, et al., 2018; Seow & Gillan, 2020). In future studies it would be of interest to compare relationships between transdiagnostic symptom scores and (meta)cognition between non-clinical and clinical samples.

This study has to be seen in light of its limitations. All groups were tested online, but nevertheless received extensive instructions. The OCD and HC groups were not recruited via specialized online research platforms, whereas the HComp group was. It is likely that the HComp group consisted of subjects with more experience in participating in online research, which could relate to the finding that the HComp group was better able to estimate the volatility of the task. For reasons of consistency with previous studies we included the model-based analyses. However, while a recent study indicated that the main measures of confidence and learning rates yield good internal consistency and test-retest reliability, the Bayesian model parameters had poorer psychometric quality (Loosen et al., 2023). Therefore, the model-based measures should be used and interpreted with caution, especially for studying between-subject differences. Improving the ecological validity of the paradigm, using a task where excessive action updating is penalized, or where the context is more symptom-specific, could provide insight into behavior of OCD patients when excessive precision is costly.

Together, we showed that OCD patients have lower confidence and increased error sensitivity than healthy and highly compulsive subjects from the general population, without a dissociation between action and confidence. This points to disturbances in metacognitive monitoring (where confidence is negatively impacted), without disturbances in metacognitive control (i.e., utilizing confidence to inform behavior) in OCD. Importantly, clinical OCD patients have lower confidence than highly compulsive subjects from the general population. It is likely that the underlying mechanisms of compulsive behavior differ substantially between these groups, resulting in contrasting (meta)cognitive behavioral manifestations despite equal OC symptom severity.

Acknowledgements

We would like to thank Katja Cornelissen and Fabiënne Meijboom for help with data collection. We also would like to thank Nathan Evans and Eric-Jan Wagenmakers for their helpful comments and suggestions. This study was funded by NWO VENI fellowship granted to JL (number 916-18-119).

Disclosure statement

None of the authors have any conflicts of interest to declare.

Part III

Confidence in GD

8

Learning and metacognition under volatility in gambling disorder: lower learning rates and distorted coupling between action and confidence

Hoven M

Luigjes J

van Holst RJ

Submitted

Abstract

Decisions and learning processes are under metacognitive control, where confidence in one's actions guides future behaviour. Indeed, studies have shown that being more confident results in less action updating and learning, and vice versa. This coupling between action and confidence can be disrupted, as has been found in individuals with high compulsivity symptoms. Patients with Gambling Disorder (GD) have been shown to exhibit both higher confidence and deficits in learning. In this study, we tested the hypotheses that patients with GD display increased confidence, reduced action updating and lower learning rates. Additionally, we investigated whether the action-confidence coupling was distorted in patients with GD. To address this, 27 patients with GD and 30 healthy controls performed a predictive inference task designed to assess action and confidence dynamics during learning under volatility. Action-updating, confidence and their coupling were assessed and computational modeling estimated parameters for learning rates, error sensitivity, and sensitivity to environmental changes. Contrary to our expectations, results revealed no significant group differences in action updating or confidence levels. Nevertheless, GD patients exhibited a weakened coupling between confidence and action, as well as lower learning rates. This suggests that patients with GD may underutilize confidence when steering future behavioral choices. Ultimately, these findings point to a disruption of metacognitive control in GD, without a general overconfidence bias in neutral, non-incentivized volatile learning contexts.

Introduction

Gambling Disorder (GD) is a recognized psychiatric disorder characterized by a loss of control and an inability to stop gambling despite known adverse consequences. This irrational or compulsive behavior has spurred numerous studies to investigate the decision-making processes underlying this behavior, including reinforcement learning.

Learning in GD has frequently been investigated by feedback-based learning tasks, such as reinforcement learning, reversal learning and model-based learning, revealing various impairments. Using reinforcement learning tasks, patients with GD have shown to have less strategic exploration of choice options, lower non-decision time, more decision noise, and lower learning rates for losses, but higher learning rates for rewards (for a review, see (Hales et al., 2023)). There is also evidence of impairments in probabilistic reversal learning (Boog et al., 2014; M. B. de Ruiter et al., 2009; Perandrés-Gómez et al., 2021; van Timmeren et al., 2018). Studies focusing on model-based learning have also suggested that patients with GD rely more on model-free than model-based learning than healthy controls (Bruder et al., 2021; Wyckmans et al., 2019), however not all studies showed this (Van Timmeren et al., 2023; Wagner et al., 2022). In all, there is evidence that GD is associated with deficits in (reinforcement) learning and decision-making.

Decision-making and learning processes are guided by metacognitive control, a process rooted in metacognition – our capacity to monitor and reflect upon our thoughts and actions. This capacity can be assessed by prompting individuals to evaluate their level of confidence in the accuracy of their choices. Indeed, research has demonstrated that confidence has a guiding role in information seeking, impacting decision-making, reassessment of choices, and learning (Balsdon et al., 2020; Desender et al., 2018; Meyniel, Schlunegger, et al., 2015). Moreover, confidence contributes to the adaptable adjustment of behavior, influencing the balance between exploration and exploitation (Boldt et al., 2019; Heilbron & Meyniel, 2019). Thus, a sense of confidence about one's choices has been demonstrated to be indispensable for optimal decision-making.

An influential Bayesian framework of learning shows that confidence in actions influences behavior (Knill & Pouget, 2004; Meyniel, Sigman, et al., 2015; Parr & Friston, 2017). Crucially, this framework predicts that the impact of new information on subsequent actions depends on the epistemic confidence of the decision-maker. When one is more confident, new information has less impact, resulting in less action updating and less learning. Conversely, lower confidence motivates gathering additional evidence to increase confidence in possible actions and also facilitates

learning. Thus, in healthy populations, there is a strong link between confidence and subsequent action and learning. However, in many psychiatric disorders, confidence judgments are distorted, showing underconfidence or overconfidence relative to performance (Hoven et al., 2019). Specifically, patients with GD have exhibited overconfidence, particularly in contexts involving monetary gains (Goodie, 2005; Hoven, de Boer, et al., 2022). Studies investigating the coupling between confidence and action, and their relationship with psychiatric symptoms have shown that individuals with high compulsive (but not gambling) symptoms have a weakened confidence-action coupling (Seow & Gillan, 2020). This suggests that highly compulsive individuals tend to consider their confidence to a lesser extent when informing their future actions. However, it is currently unknown whether the relationship between confidence and action, and subsequent learning, is affected in GD.

Based on earlier findings, we hypothesized that patients with GD, relative to controls, show higher confidence, less action-updating and lower learning rates. With regard to the coupling of confidence and subsequent actions, we posited two hypotheses. First, patients with GD could have an intact coupling between confidence and action, in line with the Bayesian framework. The alternative hypothesis posited that GD patients (similar to findings of individuals with highly compulsive symptoms) have a weakened confidence-action coupling.

To test these hypotheses, we investigated confidence, action, their coupling and learning by using a predictive inference task originally described by (Nassar et al., 2010), and used in many studies since (Hoven, Mulder, et al., 2023; Seow & Gillan, 2020; Vaghi et al., 2017) in patients with GD and matched healthy controls. Our results revealed that patients with GD have a weaker action-confidence coupling but exhibit similar confidence levels and action updating compared to controls. Moreover, patients demonstrated lower learning rates than controls.

Methods

Participants

27 patients with GD and 30 healthy controls (HCs) were included in this study, matched on age, sex and education. The study was approved by the Ethics Board of the Behavioral Science Laboratory at the University of Amsterdam (2018-DP-9420). All subjects provided written informed consent and were reimbursed for their time. Patients with GD were recruited through patient clinics in the Netherlands and HCs via an online participation pool. All patients with GD had been in treatment for their

gambling problems at least once and had gambled regularly within the past 12 months prior to participating. The HCs did not currently or in the 6 months prior to participation suffer from any psychiatric disorders and did not use any psychotropic medication.

Experimental procedure

Predictive inference task

All participants performed a predictive inference task, similar to the one reported in (Vaghi et al., 2017), implemented using Psychtoolbox in MATLAB.

This task allows for the investigation of the relationship between error-driven learning and confidence, by letting participants infer the landing location of a particle based on its previous landing locations. A circle with a dot in the center was shown to participants, after which they had to place a "bucket" (represented by a curved rectangle) at the location at which they predicted a particle (i.e. a 'coin') would land. The position of the bucket could be updated every trial in response to new information. After confirming the location of the bucket, participants were asked to rate their confidence that they would catch the particle in the bucket on a scale of 1 (not at all confident) to 100 (extremely confident) (Figure 1).

After the confidence rating was confirmed, the particle would fly from the center dot to the edge of the circle. The landing location of the particle was sampled from a Gaussian distribution with a fixed standard deviation (SD) of 12. At certain 'change-point' (CP) trials a new mean for the particle landing location was drawn from a uniform distribution over the full range of the circle U(1,360), with a fixed probability of 0.125 (hazard rate, H). Performing accurately on this task thus required participants to distinguish between actual signals of change (i.e. CP trials) and noise (SD of the generative Gaussian distribution). When the particle landed in the bucket, participants received points, and they were penalized for missing the particle.

The task consisted of 4 blocks of 75 trials, with a practice round that was not included in the analyses. Participants were instructed to earn as many points as possible, which would be converted to a bonus up to €5. Confidence ratings were not directly incentivized, but participants were instructed to rate their confidence as accurately as possible.

Moreover, a subset of the sample (24 GD, 15 HC) additionally performed the predictive inference task at a higher hazard rate of 0.20, corresponding to higher task volatility. As

the main focus of this paper is on the results of the original task, analyses pertaining to the higher volatility task can be found in Appendix F.

Task-based exclusions

Based on exclusion criteria set by (Seow & Gillan, 2020) and our previous study using this task (Hoven, Mulder, et al., 2023), we excluded participants when their mean confidence after hits was lower than their mean confidence after misses (n=6, of which 2 GD). Since this current version of the task (lab-based instead of online (Seow & Gillan, 2020)) did not randomly initialize the confidence rating every trial, we cannot use previously used exclusion criteria pertaining to the deviation of subjects' confidence ratings compared to the initialized confidence rating. After applying the subject-based exclusion criteria, the final dataset included data from 51 participants (25 GD (4 females), 26 HC (6 females)). For one GD subject, data for one out of four blocks was corrupted and thus this subject has data for 225 instead of 300 trials. Since previous studies did not use any exclusion criteria based on accuracy on the task, here we also did not apply accuracy-based exclusion criteria. However, when inspecting the data, one GD participant showed an average accuracy of around 18%, and analyses excluding this subject are detailed in Appendix F. In addition to subject-based exclusions, we also performed trial-based exclusions (see section '*Computational Model'*).

Analyses

All data analyses were conducted using MATLAB (version 2018b) and R (version 4.2.1) using packages lme4, lmerTest, nlme and emmeans (Bates et al., 2015; Kuznetsova et al., 2017; Lenth et al., 2018; Pinheiro et al., 2022), and were similar to our previous case-control work in OCD for consistency (Hoven, Mulder, et al., 2023).

Action and confidence

First, to compare action updating and confidence between groups, separate linear-mixed effects models were fitted with either action update (absolute difference in bucket position from trial (t) to trial (t+1)) or confidence as dependent variable and a fixed effect of group, together with random intercepts per subject.

Action-confidence coupling

Second, differences in the strength of action-confidence coupling between groups were assessed using a mixed-effects model with action update as the dependent variable and confidence (z-scored), group and their interaction as fixed effects together with random intercepts and random slopes of confidence per subject.

In addition, we conducted two Pearson's correlation tests to examine the relationship between the strength of action-confidence coupling (using subject-level β coefficients of the action-confidence coupling model) and PGSI and GBQ scores in the GD group.

Computational model

Third, a computational approach was employed, similar to earlier work (Hoven, Mulder, et al., 2023; Marzuki et al., 2022; Seow & Gillan, 2020; Vaghi et al., 2017), in order to examine whether and how the relationship between behavior on the task (i.e. action or confidence) and various parameters describing the volatile environment differed between groups. In a volatile setting, where the environment is subject to frequent changes, participants need to adjust their learning rate based on recent information to update their beliefs about the generative distribution. When significant discrepancies between predicted and observed outcomes occur (i.e., large prediction errors), indicating a substantial shift in the environment, belief updates need to be strong and learning rates should be higher. Conversely, when prediction errors are small and likely due to random fluctuations, belief updates are less necessary, resulting in lower learning rates.

For each trial, the human prediction error $\hat{\delta}_t$ (PE) was calculated as the difference between the current bucket position b_t and the particle landing location X_t.

$$\hat{\delta}_t = X_t - b_t$$

Subsequently, the human learning rate $\hat{\alpha}_t$ (LR) was calculated as the proportion of PE used for the subsequent action update, which was calculated as the absolute difference in bucket position from trial (t) to trial (t+1):

$$\hat{\alpha}_t = \frac{|b_{t+1} - b_t|}{\hat{\delta}_t}$$

Following earlier studies, trials were excluded from all analyses if the LR exceeded the 99th percentile which was calculated separately for each group (Seow & Gillan, 2020; Vaghi et al., 2017). In addition, trials where PE = 0 were excluded, since these trials do not drive error-driven learning (1.95% of GD trials, 1.97% of HC trials). Additionally, the

first and last trials within each block were excluded from analyses; in the first trials, there is no error-driven learning yet, and for the last trials no learning rate could be calculated. In total, 5.49% of GD trials and 5.52% of HC trials were excluded from analyses.

Error Sensitivity

To assess group differences in error sensitivity in terms of learning, a linear mixed model with human LR as the dependent variable and human PE, group and their interaction as predictors was run. For visualization purposes, PE was binned into 20 quantiles with each an equal fraction of trials, for which the average LR was computed per subject.

Bayesian Observer Model Analyses

Following previous research (Marzuki et al., 2022; Seow & Gillan, 2020; Vaghi et al., 2017), behavior of participants was analyzed using a quasi-optimal Bayesian observer model that approximates optimal task behavior (Nassar et al., 2010). Using the model code that is publicly available (Vaghi et al., 2017), we fitted the particle landing locations of all subjects to obtain individual-level model parameters. These parameters represent various statistical characteristics of the environment experienced by participants during the task. They include, on a trial-by-trial basis, the prediction error δ (PE, the absolute difference between model belief and location of the coin), the probability that a change-point occurred (CPP, the likelihood that the sampling distribution of the coin's location has changed, thus that a change-point has occurred), and relative uncertainty (RU, the fraction of uncertainty about the generative mean that is not due to noise). RU was expressed as its inverse, termed model confidence (MC, the precision of the model's beliefs about the mean), to allow for a more direct comparison with confidence judgments from the task. For more detail on the model see Appendix F.

After fitting the model to the task data and obtaining the latent parameters for each subject, we assessed how these parameters related to participant behavior (action and confidence), and whether these relationships differed between the groups. Following previous studies, two separate mixed-effects models were assessed, where participant behavior (either action or confidence) was regressed against three model parameters: absolute PE, CPP and (1-CPP)(1-MC), and the categorical variable hit, indicating whether the particle was caught or not. Here, PE represents information regarding the most recent observation, while CPP and (1-CPP)(1-MC) represent the model's estimation that a change-point did or did not occur, given the sequence of past

observations, respectively. For the action model, the dependent variable was calculated as: LR * PE, which is equal to the bucket update, and the predictors were also interacted with PE, following previous work (McGuire et al., 2014; Nassar et al., 2019; Seow & Gillan, 2020; Vaghi et al., 2017). For both models, all fixed-effects were z-scored and interacted with group. Random intercepts and slopes of all predictors were also included in the models.

In the Bayesian model, the hazard rate is a constant of 0.125, which is equal to the hazard rate in the task. As additional sensitivity analyses we furthermore calculated the perceived hazard rate as a free parameter for each subject based on the best fit of the model on the participant's behavior (see Appendix F for more information).

Results

There were no differences in age (t_{49} = 0.42, p = 0.68), gender distribution (X^2 = 0.40, p = 0.52) or education level (t_{49} = -0.38, p = 0.71) between HC and GD groups. For details on demographics, clinical and task data, see Table 1.

Table 1: Demographic, clinical and task variables

	GD	HC
Age	36.8 (11.4)	35.6 (8.8)
Females (%)	4 (16.0%)	6 (23.1%)
Education Level	3.12 (0.9)	3.23 (1.2)
PGSI	15.1 (4.2)	
GBQ	56.4 (21.2)	
Accuracy (%)	60.17 (9.78)	62.28 (5.46)
Confidence	47.84 (25.16)	50.52 (21.45)
Confidence Update	15.12 (8.62)	13.35 (7.40)
Learning Rate	0.37 (0.14)	0.47 (0.21)
Action Update	18.59 (4.31)	19.62 (4.57)
Prediction Error	27.22 (10.99)	24.61 (2.29)

Abbreviations: GD = Gambling Disorder, HC = Healthy Controls, PGSI: Problem Gambling Severity Index, GBQ: Gamblers Belief Questionnaire. Data are reported as mean (standard deviation).

No group differences in action updating or confidence

Mixed-model analyses were conducted to test group differences in task behavior (i.e. action and confidence). No differences in the amount of action updating (β = -1.02 (1.24), t = -1.82, p=0.415), nor differences in confidence (β = -2.67 (6.54), t = -0.41, p=0.684) were found between groups. Accuracy was equal between the groups as well (t_{49}=-0.95, p=0.345). The proportion of trials in which no action update was performed was higher in GD, however (t_{49}=2.48, p=0.017; GD = 60.1%, HC = 50.5%).

Weaker action-confidence coupling in GD

Next, we evaluated whether the coupling between action update and confidence differed between the groups. As expected, a significant negative relationship between confidence and action update existed across groups, such that higher confidence was related to less action updating (i.e. action-confidence coupling) (β = -8.26 (1.14), t = -7.23, p<.001). Moreover, there was evidence for a distortion of this action-confidence coupling in GD, as a significant interaction between group and confidence was found (β = 3.28 (1.63), t = 2.01, p=0.045), indicating a weaker action-confidence coupling in GD (estimated marginal slope = -4.98 (1.17)) than in HC (estimated marginal slope = -8.26 (1.14)).

Within the GD group, no significant correlation was found between action-confidence coupling and PGSI score (r = -0.23, p=0.266), or GBQ score (r = -0.07, p=0.754).

Lower learning rates in GD

We also assessed differences in learning rates and the error sensitivity in terms of learning between the GD and HC groups. Across both groups, learning rates increased as a function of prediction error magnitude (β = 0.006 (0.0002), t = 39.51, p<.001), and thus learning rates were highest after large errors. Moreover, learning rates were found to be significantly lower overall in the GD group (β = -0.13 (0.05), t = -2.39, p=0.021), but no evidence was found for an interaction effect between PE and group.

To look at the group differences in cases of low, middle or high error magnitude, following previous research (Hoven, Mulder, et al., 2023; Vaghi et al., 2017), a mixed-model analysis binning the prediction error in 3 quantiles (i.e., low, medium or high error magnitude) was run. This indicated that patients with GD specifically had decreased learning rates when error magnitude was small (HC-GD estimate = 0.13 (0.05), Z-ratio: 2.60, p=0.009) and medium (HC-GD estimate = 0.19 (0.05), Z-ratio: 3.79, p<0.001). This indicates that when errors were of small or medium size, the influence of the most recent outcome on subsequent action (i.e. PE) was lower in the GD compared to the HC group, whilst this did not differ for larger PEs.

Stronger effect of uncertainty about the generative mean of the distribution on action in GD

Finally, we assessed whether task behavior (action and confidence) was differently predicted by the latent model parameters that represent different forms of uncertainty and feedback in the volatile environment. As expected, action was significantly predicted by all model-derived parameters and hit, such that increases in PE, CPP and (1-CPP)*(1-MC) predicted an increase in action, while a successful catch of the particle predicted a decrease in action. Moreover, a significant interaction between group and the (1-CPP)*(1-MC) parameter indicated a stronger effect of relative uncertainty of the belief about the mean of the distribution in the GD group compared to the HC group (β = 1.72 (0.72), t = 2.37, p=0.022).

Confidence was, as expected, significantly negatively predicted by CPP and (1-CPP)*(1-MC), but only marginally by PE, and significantly increased with a successful catch of the particle. We did not find any evidence for group differences in the strength of these effects (see Appendix F).

No group differences in perceived hazard rates

Sensitivity analyses using the subject-specific perceived hazard rate (see Appendix F) first of all showed no differences in hazard rate between groups (mean GD: 0.54, mean HC: 0.59: t_{49}=-0.61, p=0.542). Moreover, in sensitivity analyses we performed the same analyses as described above, but including the subject-specific hazard rate as a covariate. These analyses indicated that none of the significant group differences that were found were influenced by differences in perceived hazard rate. For more details, see Appendix F.

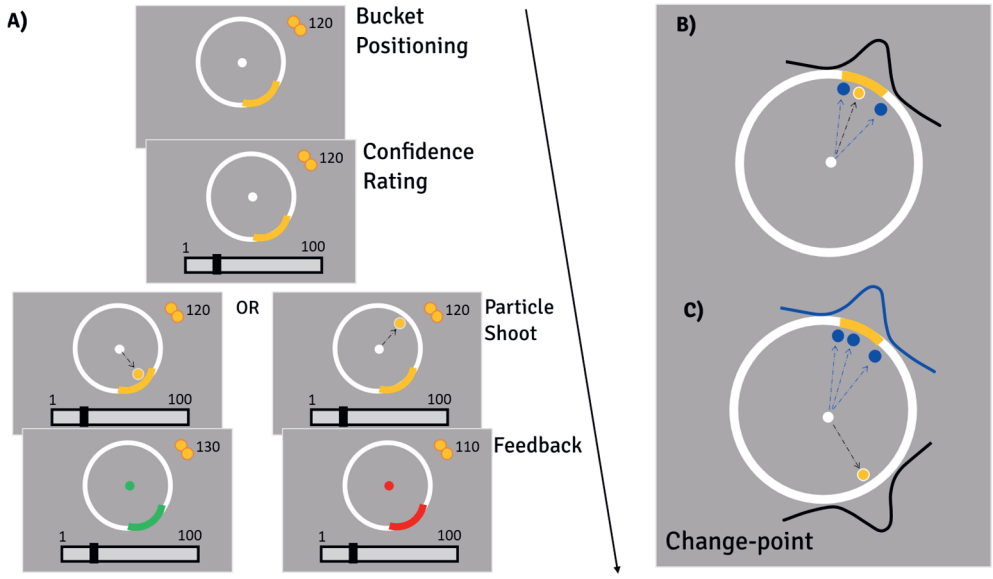


Figure 1: Predictive Inference Task. (A) Trial of the predictive inference task. Participants positioned their bucket (i.e. yellow bar) to catch a flying particle that was released from the center dot to the edge of the circle. After positioning their bucket, participants indicated their confidence in catching the particle. The particle was either caught (bar turned green) or missed (bar turned red), which resulted in gaining or losing points, respectively. The number of points obtained is shown in the right upper corner. **(B)** In every trial, the landing positions of the particles were sampled from a random Gaussian distribution with a standard deviation. This noise resulted in the particles to land close together with a small amount of noise. Current trial particle trajectory is marked in black, while previous trials particle trajectories are marked in blue. Over time participants learn about the Gaussian distribution from which the particle trajectories are drawn. **(C)** During a change-point the mean of the Gaussian distribution of the landing position changes. After a change point, the landing positions are again sampled using the new Gaussian distribution, until a new change-point occurs. Figure was adapted with permission from Seow et al., (2020).

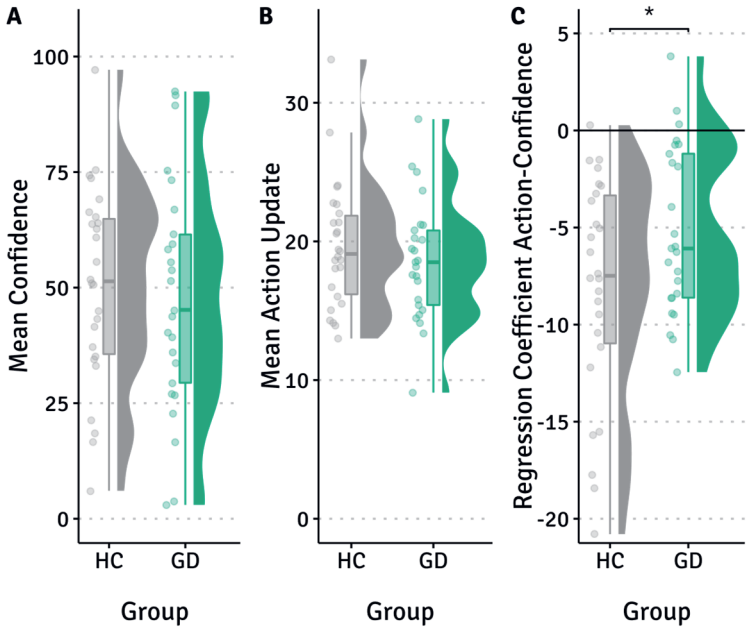


Figure 2: Task behavior across groups. Mean confidence **(A)** and action update **(B)** per group. **(C)** Regression coefficient from the relationship between action update and confidence. As expected, regression coefficients were negative indicating that lower confidence was associated with bigger action updates of the location of the bucket. Dots represent **(A)(B)** data from individual participants and **(C)** regression coefficients of individual subjects. Boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. Significance stars represent the main effects of group in the respective mixed-effects models. * p<.05, **p<.01, ***p<.001. HC = healthy control subjects, GD = gambling disorder patients.

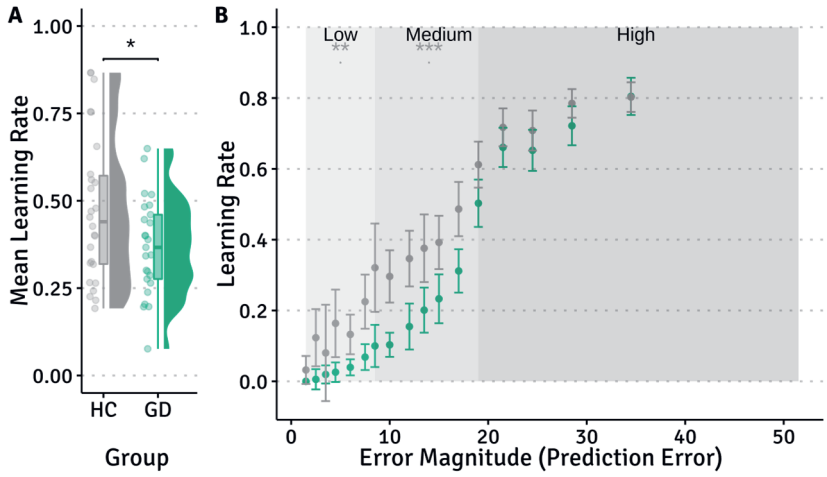


Figure 3: Learning rates and error sensitivity. (A) Mean learning rates per group ($\hat{\alpha}_t$). Patients had significantly decreased learning rates compared to the HC group. Dots represent learning rates of individual subjects, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. **(B)** The relationship between prediction error magnitude ($\hat{\delta}_t$) and learning rate for both group. Prediction errors were divided in 20 quantiles, of which 18 quantiles are shown here for visualization purposes. Dots represent mean learning rates per group, error bars represent the SEM. Overall, learning rates were higher when prediction errors were larger. Learning rates were lower in the GD group compared to the HC group at low and medium error magnitudes. * p<0.05, ** p<0.01, *** p<0.001

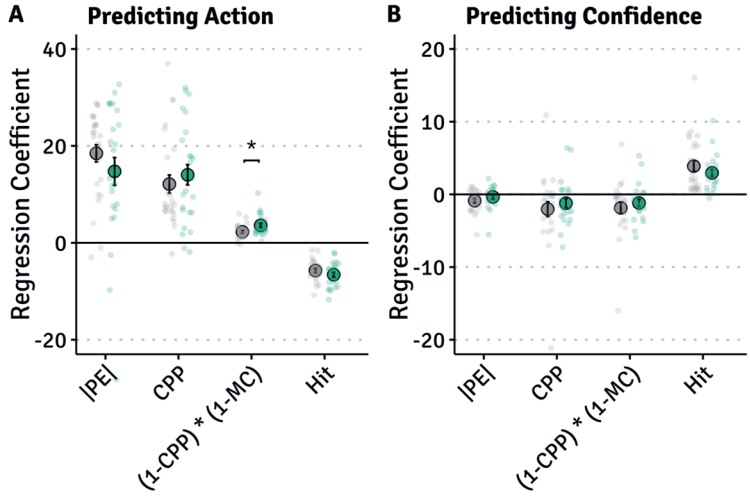


Figure 4: Model-based results on action and confidence. Regression coefficients of the regressions assessing the relationship between the parameters from the computational model and **(A)** human action (i.e. learning rate * absolute prediction error), or **(B)** human confidence. Small dots represent individual regression coefficients, big dots represent mean regression

coefficients per group, error bars denote SEM per group. Predictors included absolute prediction error (PE), change-point probability (CPP), model confidence (MC) and a categorical variable representing hits/misses. * p<0.05, ** p<0.01, *** p<0.001

Discussion

Drawing on previous observations of increased confidence and impaired reinforcement learning in GD, here we extended the literature by investigating the connection between confidence and action updating and subsequent learning in patients with GD. Our results showed that patients with GD demonstrated comparable levels of confidence, action updating, and performance, but had a weaker coupling between confidence and action. This indicates that patients with GD assign less significance to their confidence levels when performing actions under volatility. These findings support the hypothesis that GD is characterized by decreased confidence-action coupling.

This dissociation between action and confidence resembles the clinical presentation of GD, where patients often continue gambling despite knowing it is unwise. It suggests a disruption in metacognitive control, which might also be associated with a disruption of model-based action (Voon et al., 2015), as has been found in GD before (Bruder et al., 2021; Wyckmans et al., 2019). Though current models for compulsive gambling behavior do not incorporate the role of metacognition or confidence, we can draw on a recent model of obsessive-compulsive disorder (OCD) (Fradkin et al., 2020). This model describes that compulsive behavior can arise from overreliance on prior beliefs (e.g., overconfidence in those beliefs) at the expense of new evidence, leading to less learning, more stickiness, and habitual behavior. This kind of behavior was indeed observed in highly compulsive individuals from the general population, indicating lower learning rates and decreased action-confidence coupling (Seow & Gillan, 2020), although gambling symptoms were not explicitly assessed in this study.

The current findings indicate overall lower learning rates in GD, with a specific decrease in learning rates when the error magnitude was small or medium. GD patients overall seem to move their bucket position less frequently (i.e., significantly lower proportion of trials in which the bucket was moved), while there was no difference in the degree of movement (i.e., action update). This suggests that patients exhibit more sticky behaviour than healthy controls, which aligns with prior research (Perandrés-Gómez et al., 2021; van Timmeren et al., 2018; Wiehler et al., 2021). However, lower learning rates in GD were not always directly evident in experimental tasks (Hales et al., 2023). For example, a recent study employing a probabilistic instrumental learning task with three conditions (reward, avoidance, neutral) found no overall differences in the proportion

of correct choices between patients with GD and HCs in reward or avoidance trials. However, employing a computational model with two separate learning rates revealed that patients with GD exhibited relatively excessive sensitivity to positive prediction errors (PEs), but insensitivity to negative PEs (Suzuki et al., 2023). These findings underscore the notion that GD might be linked to subtle and specific differences in learning rates, which might not always be easily discernible without employing sensitive experiments and computational modeling (Hales et al., 2023).

Our study found no evidence of increased confidence judgements in patients with GD, a finding that aligns with previous research using a non-incentivized learning task (Brevers et al., 2014). This contrasts, however, with studies that have used monetary incentives, where GD patients have shown higher levels of confidence (Goodie, 2005; Hoven, de Boer, et al., 2022). As suggested (Hoven, Hirmas, et al., 2023), it appears that overconfidence in GD manifests mainly in disorder-relevant contexts, such as during gambling task or when gains or risk are involved. This raises important questions for future research: under what circumstances do distortions in confidence occur in GD, and how do these distortions impact learning and decision-making?

Recent investigations in healthy populations have begun to elucidate the relationship between learning biases and confidence biases (Lebreton, Bacily, et al., 2019; Salem-Garcia et al., 2023; C. Ting et al., 2023). These studies have shown that individuals tend to be more confident when learning to seek gains as opposed to avoiding losses. This 'valence-induced confidence bias' has been linked to reduced context-dependent learning, while a general overconfidence bias correlated with a confirmatory learning bias (Salem-Garcia et al., 2023; C. Ting et al., 2023). Applying this framework to GD, one could hypothesize that in an incentivized reinforcement learning task, GD patients would exhibit both elevated confidence and a more pronounced valence-induced confidence bias. This in turn could be associated with increased confirmatory learning and decreased context-dependent learning relative to HCs. This pattern could offer insights into rigid, disadvantageous decision-making in GD. Subsequent research should validate these hypotheses, potentially providing a more nuanced understanding of the cognitive biases at play in GD

Our current study comes with limitations. In line with prior research, we integrated model-based analyses for consistency. However, it's important to note that while recent findings suggested good internal consistency and test-retest reliability for the main measures of confidence and learning rate, the psychometric quality of the Bayesian model parameters was comparatively lower (Loosen et al., 2023). This implies that the utilization and interpretation of model-based metrics should be exercised cautiously, particularly when examining differences between individuals. Also, the

predictive inference task does not resemble a real-world gambling game. Hence, enhancing the ecological validity of our approach could involve using a task that simulates monetary involvement and enforces penalties for excessive action updating. Furthermore, our study population was drawn from therapy centers, encompassing individuals who had undergone cognitive-behavioral therapy (CBT) for their gambling disorder. Given that CBT targets the reduction of irrational gambling-related thoughts to mitigate the influence of outcome significance on decision-making (Sylvain et al., 1997; Toneatto, 1999), it's possible that CBT contributed to a reduction in overconfidence during the present task. It could be hypothesized that untreated GD patients might exhibit more pronounced overconfidence and/or a stronger disconnection between confidence and action.

In conclusion, our study investigated the connection between confidence and action in patients with GD in a volatile learning task. We found a weaker coupling between confidence and action, suggesting disrupted metacognitive control in GD, without a general positive confidence bias in GD. Additionally our findings indicated lower learning rates in GD, indicating differences in learning under volatile conditions. All in all, these findings suggest that GD is associated with disturbance in metacognitive control. Future research could advance by incorporating metacognitive ability as an important factor for comprehending disadvantageous decision-making in GD.

9

The role of attention in decision-making under risk in gambling disorder: an eye-tracking study

Hoven M

Hirmas A

Engelmann JB

van Holst RJ

Addictive Behaviors, 2023, 138, 1-10

Abstract

Gambling disorder (GD) is a behavioral addiction characterized by impairments in decision-making, favoring risk- and reward-prone choices. One explanatory factor for this behavior is a deviation in attentional processes, as increasing evidence indicates that GD patients show an attentional bias toward gambling stimuli. However, previous attentional studies have not directly investigated attention during risky decision-making. 26 patients with GD and 29 healthy matched controls (HC) completed a mixed gambles task combined with eye-tracking to investigate attentional biases for potential gains versus losses during decision-making under risk. Results indicate that compared to HC, GD patients gambled more and were less loss averse. GD patients did not show a direct attentional bias towards gains (or relative to losses). Using a recent (neuro)economics model that considers average attention and trial-wise deviations in average attention, we conducted fine-grained exploratory analyses of the attentional data. Results indicate that the average attention for gains in GD patients moderated the effect of gain value on gambling choices, whereas this was not the case for HC. GD patients with high average attention for gains started gambling at less high gain values. A similar trend-level effect was found for losses, where GD patients with high average attention for losses stopped gambling at lower loss values. This study gives more insight into how attentional processes in GD play a role in gambling behavior, which could have implications for the development of future treatments focusing on attentional training or for the development of interventions that increase the salience of losses.

Introduction

Gambling disorder (GD) is the first behavioral addiction to be classified in the DSM-5 (American Psychiatric Association, 2013) and is characterized by persistent and problematic gambling behavior despite the – often negative – consequences. One of the hallmarks of GD is an impairment in decision-making, which is biased toward risky choices with high pay-out (see for reviews (van Holst et al., 2010; Wiehler & Peters, 2015)). Indeed, relative to controls, patients with GD are found to be more risk-seeking in a number of tasks (Brand et al., 2005; Brevers et al., 2012; Goudriaan et al., 2005; Ligneul et al., 2013; Ochoa et al., 2013; Spurrier & Blaszczynski, 2014). Risk-seeking behavior in GD has also been associated with lower sensitivity to the expected value of choices (Limbrick-Oldfield et al., 2020), and diminished loss aversion (Gelskov et al., 2016; Giorgetta et al., 2014). Some research has also suggested that risky decision-making in GD is linked to abnormal processing of rewards (Goudriaan et al., 2006). Many neuroimaging studies have studied reactivity to monetary gains, showing mixed results. Increased, decreased and normal striatal responses to rewards have been found in GD, resulting in a complex picture on reward sensitivity (Limbrick-Oldfield et al., 2013; Luijten et al., 2017). An integrative model of addiction posits an explanation for the above-mentioned findings by virtue of the presence or absence of addiction-related cues (Leyton & Vezina, 2013). Finally, another factor that has been suggested to contribute to risky decision-making in GD is overconfidence, which could lead to overoptimistic behavior (Brevers et al., 2013, 2014; Goodie, 2005; Lakey et al., 2007). The mechanisms underlying these abnormal economic decisions across various contexts in GD are unclear to date. Recent advances in the fields of behavioral economics (Krajbich et al., 2010; Orquin & Mueller Loose, 2013) and neuroeconomics (Hare et al., 2011; Lim et al., 2011) indicate that attentional processes may underlie psychopathological distortions of (risky) decision-making. One intriguing possibility suggested by this body of work is that GD patients' attention deployment differs from healthy controls during risky decision-making.

Attentional processes play an active role in decision-making (Orquin & Mueller Loose, 2013). Choice options that we focus on longer and more often are more likely to be chosen (Krajbich et al., 2010; Lim et al., 2011; Pachur et al., 2018; Thomas et al., 2019), and higher valued choice options attract more attention (Anderson et al., 2011; Gluth et al., 2018, 2020). Indeed, studies in healthy subjects have indicated that attention directed towards reward cues predicted one's degree of risk-taking, with a larger attentional bias for high-value cues (San Martín et al., 2016). So far, studies investigating attention in GD have primarily analysed reaction times or fixation times when viewing addiction-relevant stimuli vs non-addiction-relevant stimuli, known as

attentional bias studies (Cox et al., 2014; Field & Cox, 2008; Hønsi et al., 2013). These studies have shown that increased attention to addiction-relevant cues plays an important role in the onset and the maintenance of addictive (and also gambling) behavior (Anselme & Robinson, 2020; Brevers et al., 2011; Ciccarelli et al., 2016, 2019; Grant & Bowling, 2015; McGrath et al., 2018; Sancho et al., 2021; Vizcaino et al., 2013), but also see (Anderson, 2016; Christiansen et al., 2015). These findings have led to the development of attentional bias modification training as interventions against addictive behavior (Heitmann et al., 2018), which is currently being tested in GD (Boffo et al., 2017; Hilgenstock et al., 2014). Eye-tracking provides a more direct measure of attentional processes than solely studying reaction times. Previous studies have revealed that people who feel emerged in gambling allocate their visual attention more to specific stimuli during gambling, such as 'amount won' messages (Rogers et al., 2017) or 'credit window' (Murch et al., 2020) than to more general stimuli. However, it is currently unknown whether GD is associated with abnormal attention towards gains or losses during decision-making and how it influences these decisions.

The current study aims to fill this knowledge gap by applying eye-tracking during a mixed gambles task in healthy controls and patients with GD. Moreover, by including confidence judgments in the current study, we can test how confidence relates to gambling propensity, since this has been postulated to be linked (Goodie, 2005; Hoven, de Boer, et al., 2022), but so far remains untested. We hypothesized that compared to controls, GD patients would show increased gambling propensity, higher confidence in their choices, increased reward sensitivity and a lower level of loss aversion. With respect to attentional processes, we hypothesized that GD patients would exhibit more attention to gains and less attention to losses and that attention would influence GD patients' choice behaviour more strongly.

Methods

Participants

A total of 27 GD patients and 30 HC subjects were included in this study (Table 1), matched on age, sex and education. Recruitment was performed via an online participation pool and patient clinics in the Netherlands. All GD patients had followed at least one treatment and had gambled regularly within the previous year. All HC subjects did not currently or in the previous 6 months suffer from any psychiatric disorder, including gambling disorder, and did not use medication.

All participants signed an informed consent form prior to the start of the experiment. The experiment was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki and was approved by the Ethics Board of the Behavioral Science Laboratory at the University of Amsterdam.

Experimental Task and Procedure

The mixed gambles task involved making risky decisions, choosing from two alternative options (Figure 1A). Mixed gambles were presented as a 50/50 chance of gaining or losing a specific value and participants were asked to decide between two options: rejecting (sure option) or accepting (gambling option) the gamble. The sure option always entailed opting for the initial endowment of €25 without the possibility of additional bonuses. The gambling option always entailed potentially gaining or losing additional bonuses, both with an equal probability of 50%. For more details on the task, see Appendix G.

Each trial started with a fixation cross, after which the gamble was shown for a maximum of 6 seconds or until the participant made their choice. Feedback indicated the chosen option, and after each choice subjects rated their confidence on a 7-point scale (Figure 1A). Each combination of gains and losses was shown twice to counterbalance the location of gains/losses. Gains or losses never appeared on the same side for more than three times in a row. All subjects performed a training session and the task consisted of 160 trials (in 4 blocks). For details on the set-up of the EyeLink eyetracker, see Appendix G.

Before the start of the experimental tasks, subjects filled in questionnaires measuring gambling problem severity (Problem Gamblers Severity Index, PGSI) (Ferris & Wynne, 2001), depressive symptoms (Hamilton Depression Rating Scale, HDRS) (Hamilton, 1960), behavioral inhibition and activation system (BIS/BAS) (Carver & White, 1994) and gambling beliefs (Gamblers Beliefs Questionnaire, GBQ, GD only) (Steenbergh et al., 2002).

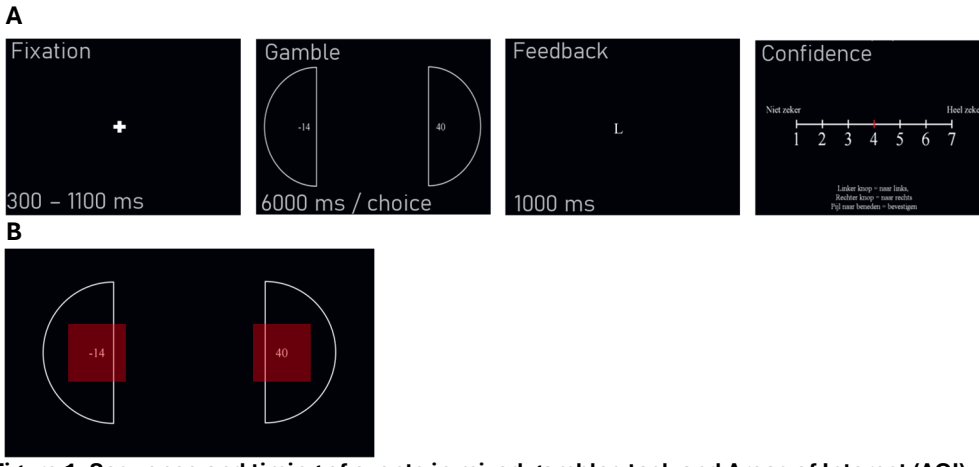


Figure 1: Sequence and timing of events in mixed gambles task and Areas of Interest (AOI). **(A)** At the beginning of each trial a fixation cross was shown, jittered between 300 and 1100 ms. Then the gamble was shown (i.e. gain and loss value stimuli; left stimulus centered on 480x540, right stimulus centered on 1440x540) for the duration of the decision with a maximum of 6000 ms. Subjects were asked to accept or reject the gamble using the up or down key, respectively, after which a brief feedback message was shown indicating and confirming their choice (1000ms. L = lottery option, X = safe option, 'Respond Faster' = if failed to respond within 6000 ms). After each choice participants rated their confidence on a scale from 1 (not sure) to 7 (very sure) (unlimited time). Subjects did not receive any feedback about the outcome of their choices (win or loss outcomes) until after completion of the experiment to avoid history and learning effects. **(B)** This figure represents the rectangular areas of interest centered on the gambling stimuli (in red) with margins of 150 pixels in each direction (approx 4 visual degrees). Since the areas of interest are widely separated in space and show no overlap, using a wide AOI margin is encouraged to minimalize false negatives.

Exclusions

Exclusions (see Appendix G for details on exclusion criteria) led to a final sample of 26 GD and 29 HC participants, with a total of 8295 trials. The quality of the eye-tracking data was checked and blocks or trials with poor data quality were excluded (see Appendix G).

Data Preparation

The EyeLink 1000 online parser was used to classify the raw eye movement data into events: saccades, fixations and blinks, using default settings. This parsed data was further analyzed using the *eyelinker* package in R (Barthelmé & Hurst, 2021). We constructed areas of interest (AOIs, Figure 1B) and excluded poor quality trials (see

Appendix G for details). Analyses focused on the period from gamble onset until choice and fixations that did not strictly fall within this period were trimmed.

Measures

Gamble propensity was measured as the percentage of accepted gambles per subject. For each trial we computed the total dwell time (i.e. the total fixation time on a specific AOI, Figure 1B) separately for the loss and gain AOIs. A relative dwell time on gains versus losses was calculated by subtracting the dwell time on losses from that on gains.

Analyses

For all analyses we used R (version 1.4.1106) in combination with the packages eyelinker (Barthelmé & Hurst, 2021), emmeans (Lenth et al., 2018), lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017).

Demographics

Age, sex, education level, gambling severity and depression levels were compared between groups using two-sample t-tests in case of continuous variables or chi-square tests in case of categorical variables.

Control Analyses

Due to exclusion of trials we could not analyze the full set of 160 trials for every subject. However, when comparing the average gain values, average loss values and expected values of the included trials between the groups, we found no differences (average gain values: $t^{53} = -1.01$, $p = 0.32$, average loss values: $t^{53} = 0.87$, $p = 0.39$, average EV: $t^{53} = -0.22$, $p = 0.83$).

Gambling Propensity, Loss Aversion and Choice Behavior

As a first step, we compared gambling propensity between the two groups using a two-sample t-test. Choice data were further analyzed by fitting mixed-effects models with a binomial family and logit link function on our trial-by-trial data (8295 observation in 55 subjects). All models used the binary choice to gamble as dependent variable (coded as a dummy with 0(1) for rejecting(accepting) the lottery). For this basic choice model we used the maximum possible random-effects structure (Barr et al., 2013). See Appendix G, Table G1A for the specification of the fixed and random effects. For further analyses, mixed-effect models were run using additional and/or different fixed and

random effects, described in detail in Appendix G, Table G1. For all mixed models continuous independent variables were z-scored (gain value, loss value, confidence, dwell times) and a deviation coding scheme was used for the categorical group variable (Singmann & Kellen, 2019). All models were run on a trial-by-trial basis and post-hoc tests were performed to quantify significant interaction.

Models of differing complexity (see Appendix G, Table G1 for full model specifications) were used to first evaluate the effect of value, confidence and their interaction with group on choice to gamble (Appendix G, Table G1B). Second, to test for group differences in the effects of value on confidence a linear mixed-effects model was run with confidence as dependent variable (Appendix G, Table G1C).

To compute loss aversion, we extracted random slopes for the effects of gains and losses per subject from the basic choice model (Appendix G, Table G1A) , in which the predictors were not z-scored in order to preserve the interpretation of the loss aversion parameter. Slopes represent the size of the contribution of the potential gain and loss values to each subject's choice. The ratio of the individual beta-weights was taken as a measure of loss aversion: $\lambda = |\beta loss / \beta gain|$. A λ of 1 indicates that losses and gains are weighted equally when choosing to gamble, whereas a $\lambda > 1$ indicates that losses are weighted stronger than gains, illustrating loss aversion. Group differences were assessed with a two-sample t-test. Moreover, to test whether GD and HC subjects showed loss aversion (i.e., $\lambda > 1$), we performed two separate one sample t-tests against 1. For analyses on expected value (EV), see Appendix G.

Attention

Dwell times reflect attentional focus, which we recorded separately for gains and losses (see AOIs in Figure 1B). To evaluate the effect of value, group and their interaction on dwell times, separate linear mixed-models with (1) dwell times on gains (Appendix G, Table G1D); (2) dwell times on losses (Appendix G, Table G1E) and (3) relative dwell time on gains versus losses (Appendix G, Table G1F) were estimated using linear mixed-models.

Influence of attention on choices

Finally, to explore the extent to which attention influences risky decision-making in GD compared to HC we explored the joint influence of attentional processes, gain/loss values, group and their interactions on choice to gamble. A logistic mixed-effects model was constructed with choice as dependent variable (Appendix G, Table G1G).

Exploratory Eye-tracking Analyses

As exploratory analyses we leveraged a newly developed attention-based decision-making model that discerns (1) average attention (i.e. average dwell time on gains/losses) and (2) trial-by-trial deviations around average attention (J. Engelmann et al., 2021). These measures, respectively, reflect a subject's average attentional strategy in solving a task (akin to goal-directed or 'top-down') and trial-based or contextual (akin to salience-based or 'bottom-up') attentional processes. Separating these two attentional processes has been shown to lead to significantly improved model fits in prior work (J. Engelmann et al., 2021), and enables further analyses of the underlying attentional differences between GD and HC during choice.

We aimed to explore the relative importance of these attentional processes, how they interact with values, and whether there are group differences in these interactions. Following our approach above, we used a logistic mixed model to predict choices (Appendix G, Table G1H).

Results

Demographics

No group differences were found for age, sex and education level. PGSI, HDRS and BIS scores were significantly higher in GD subjects (Table 1).

Table 1: Demographic and clinical variables

	Age	***Sex***	***Education level***	***PGSI***	***HDRS***	***BIS***	***BAS***	***GBQ***
GD	37.4 (12.1)	21 males, 5 females	3.08 (0.89)	15.3 (3.94)	2.5 (3.51)	19.3 (3.87)	32.0 (10.8)	59.3 (23.2)
HC	34.8 (8.61)	23 males, 6 females	3.31 (1.17)	0 (0)	16.9 (4.54)	16.9 (4.54)	32.6 (8.30)	NA
Test Statistic	t^{53} = 0.921	X^2 = $2.877 \cdot 10^{-31}$	t^{53} = 0.826	Welch's t^{25} = -19.866	Welch's t^{29} = -2.906	t^{53} = -2.099	t^{53} = 0.233	NA
p-value	0.361	1	0.413	**<0.001**	**0.007**	**0.041**	0.817	NA

Shown are descriptive statistics of demographic and clinical variables for both groups, together with test statistics and p-values reflecting differences between the groups.

Choice Behavior

Gambling propensity was higher for gamblers (60.7% ± 4.22%) than for HCs (42.7% ± 3.81%) (t^{53}=3.175, p=0.002) (Figure 2A). Loss aversion was significantly lower in GD (1.063 ± 0.069) compared to HC (1.728 ± 0.227) (t^{53}=-2.673, p=0.009) (Figure 2B). Note that a confirmatory analysis, removing statistical outliers in loss aversion, still showed lower loss aversion in GD compared to HC (t^{48}=-2.473, p=0.017). Moreover, one sample t-tests showed that loss aversion was significantly higher than 1 in HC (t^{28}=-3.205, p=0.003), but not in GD (t^{25}=0.913, p=0.37), indicating that GD patients did not show loss aversion.

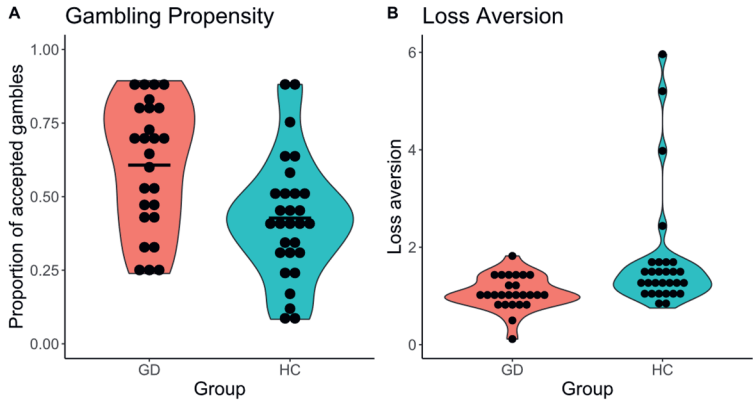


Figure 2: Choice behavior. A) Shown is the proportion of accepted gambles per group. Horizontal bars correspond to the means per group, dots represent individual gambling propensities. GD subjects showed a significantly higher gambling propensity than control subjects. B) Shown is the loss aversion (λ) value per group. Dots represent individual values. GD subjects have significantly lower loss aversion than control subjects.

We followed up on these initial results by conducting a trial-by-trial analysis investigating the influence of gain value, loss value and confidence on choice to gamble, and whether these effects differed between groups. Results showed a main effect of group, with higher gambling propensity in GD (Table 2). Moreover, strong significant main effects of both gain value and loss value were found; when there is more to gain and less to lose subjects gambled more (Figure 3A, 3B). No significant evidence was found for interactions with group or for effects of confidence on gambling choices.

When predicting confidence, a significant interaction effect between gain value and group was found (Table 2). Post-hoc tests indicated a significant positive slope of gain

value on confidence in GD (Z=2.24, p=0.025), whereas the gain slope did not differ from 0 in HC (Z=-0.62, p>0.5) (Figure 3C). This result agrees with our previous work, showing increased confidence in reaction to potential gain in patients with GD (Hoven, de Boer, et al., 2022).

Attention

First, we tested our hypothesis that attention to gains is enhanced in GD. To this end we regressed group and gain value on dwell times on gains and tested whether GD attended longer to gains of increasing size compared to HC. Interestingly, results revealed a significant main effect of gains, indicating higher dwell times with higher gain values. A significant interaction effect between gain value and group and post-hoc testing indicated, however, that the positive effect of gain value on dwell time was only significant in HC (post-hoc; HC: Z=3.33, p<0.001, GD: Z=0.31, p>0.75) (Table 2, Figure 4A). GD subjects thus did not increase their attention on gains when there was more to win. This was in contrast to dwell time on losses, where both groups showed an equal increase in dwell time on losses when there was more to lose, as confirmed by a significant main effect of loss value (Table 2, Figure 4B). In both models, we did not find evidence for a main effect of group, which indicates that groups did not differ in their average dwell times.

We also inspected the influence of values on the *relative* dwell time on gains versus losses. Results showed a significant effect of loss value, indicating that with increasing loss values, subjects' relative dwell time toward gains decreased in favor of dwell time on losses (Table 2). A positive trend effect of gain value was found, but no group differences or interactions.

Dwell times were strongly correlated to reaction times (RT). As a check, an additional mixed-model confirmed that there were no differences in reaction times (log-transformed due to skew) between groups (p>0.2) that could have impacted our results on dwell times.

Table 2: Results of mixed-effects models on choice behavior (Model B), confidence (Model C), dwell time on gains (Model D), dwell time on losses (Model E), relative dwell time on gains (Model F)

	Model B *Choice*		**Model C** *Confidence*		**Model D** *Dwell Time on Gains*		**Model E** *Dwell Time on Losses*		**Model F** *Relative Dwell Time on Gains*	
Parameter	**Estimate (SE)**	**p-value**	**Estimate (SE)**	**p-value**	**Estimate (SE)**	**p-value**	**Estimate (SE)**	**p-value**	**Estimate (SE)**	**p-value**
Intercept	0.45 (0.39)	0.252	5.47 (0.10)	**<0.001**	0.58 (0.03)	**<0.001**	0.53 (0.02)	**<0.001**	0.06 (0.01)	**0.002**
Gain Value	2.57 (0.20)	**<0.001**	0.04 (0.04)	0.235	0.02 (0.01)	**0.028**			0.01 (0.01)	*0.071*
Loss Value	-2.37 (0.18)	**<0.001**	-0.03 (0.04)	0.507			0.04 (0.01)	**<0.001**	-0.06 (0.02)	**0.003**
Group (GD)	2.25 (0.78)	**0.004**	0.14 (0.20)	0.473	-0.08 (0.05)	0.142	-0.07 (0.05)	0.179	-0.02 (0.03)	0.554
Confidence	-0.03 (0.12)	0.802								
Gain Value x Group (GD)	0.66 (0.39)	0.092	0.15 (0.07)	**0.046**	-0.03 (0.01)	**0.045**			-0.00 (0.01)	0.769
Loss Value x Group (GD)	0.09 (0.36)	0.796	-0.11 (0.08)	0.194			0.03 (0.02)	0.153	-0.01 (0.03)	0.649
Confidence x Group (GD)	-0.03 (0.24)	0.886								
AIC	5130.2		26149.35		8216.223		6402.445		12950.5	
R^2	0.885		0.345		0.219		0.232		0.080	

Results of mixed-models specified in detail in Appendix G, Table G1B, G1C, G1D, G1E, G1F, respectively. Shown are the estimates, their standard errors (SE) and 95% confidence intervals (CI), statistic and p-values. Loss values were entered as absolute values for easier interpretation. N = 55 subjects with a total of 8295 observations. *p<0.05, **p<0.01, ***p<0.001.

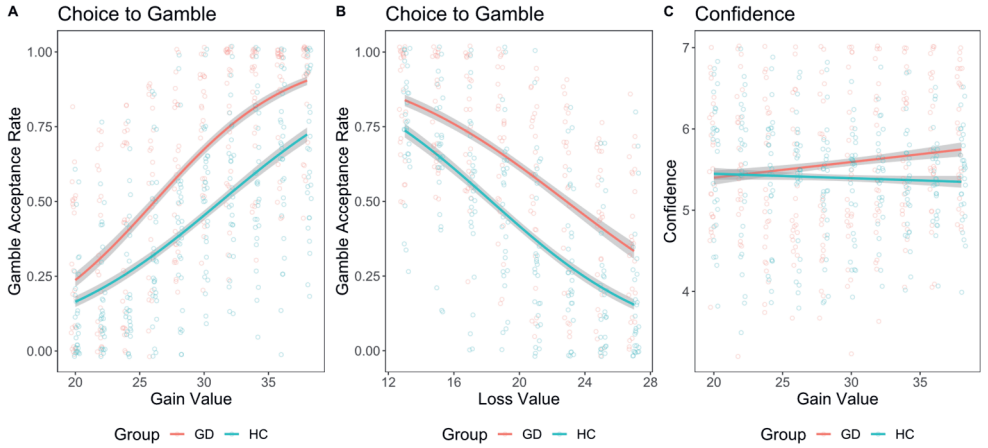


Figure 3: Behavioral results of gambling decision-making and confidence. All graphs show regression curves and 95% CI in grey, together with individual datapoints. Red lines represent model predictions for GD subjects, blue lines represent model predictions for HC subjects. Dots represent the average (A) acceptance rate per participant at each level of gain value, (B) acceptance rate per participant at each level of loss value, (C) confidence level at each level of gain value. Red dots represent GD subjects, whereas blue dots represent HC subjects. **A)** The logistic curve shows that participants from both groups accepted gambles more as the gain value increased, and shows an overal increased gamble acceptance rate in GD compared with control subjects. **B)** The logistic curve shows that participants from both groups accepted gambles less as the loss value increased, and shows an overal increased gamble acceptance rate in GD compared with control subjects. **C)** The linear regression line shows that an interaction between gain value and group on confidence level exists, where confidence increases as the gain value increased in GD subjects, but not in control subjects.

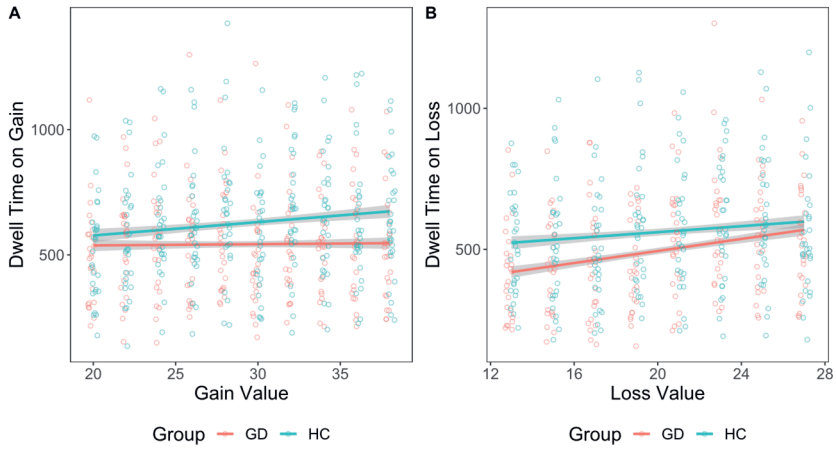


Figure 4: Effect of gain value, loss value and group on dwell times. All graphs show regression lines and 95% CI in grey, together with individual datapoints. Red lines represent models estimates for GD subjects, blue lines represent model estimates for HC subjects. Dots represent the average (A) dwell time on gains per participant at each level of gain value, (B) dwell time on losses per participant at each level of loss value. Red dots represent GD subjects, whereas blue dots represent HC subjects. **A)** The interaction effect between gain value and group on dwell time on gains shows that dwell time on gains increases with increases in gain value in control subjects, but not in GD subjects. **B)** The regression lines show that dwell time on losses increases as loss value increases, which did not differ between the groups.

Influence of attention on choices

Lastly, we turned to investigating the joint influence of attention and values on choice. We tested the hypothesis that GD subjects relative to HCs display an increased influence of their attention on gains and away from losses during decisions to gamble. We analyzed whether choices could be predicted by attentional measures and values, and whether these relationships differed between groups. Results replicate many of our previous results (Table 3A), including the strong influence of gain and loss values on choice, and increased gambling in GD. Results also showed a main effect of dwell time to losses such that all subjects gambled less when they had increased attention on losses, which effect was strongest when attending to losses relatively longer. No interactions between group and attention were found, however, indicating no evidence for an increased influence of attention on gambling choices in GD in this trial-by-trial analysis.

Exploratory analyses inspecting group effects on separate attention channels

One possibility for this null result is that different channels of attention are differentially affected in GD, which cannot be identified using the average. We address this in exploratory analyses that split the attentional data into average goal-directed and trial-based salience-based attentional processes. Results indicate that average attention largely drives the results reported in Table 3A, such that subjects gambled more when their *average* attention to gains was higher (at trend level) and gambled less when their *average* attention to losses was higher (Table 3B). Trial-by-trial increases in attention to losses (e.g. due to salience) also resulted in less gambling, while there was no effect of trial-by-trial increases in attention to gains. A trend level interaction effect between gain value and group was found, suggesting that the influence of gains on choices was stronger in GD compared to HC.

Interestingly, results showed that the relationship between gain value and average attention to gains on choices significantly differed per group. Post-hoc tests indicated that GD subjects with high average attention for gains showed a stronger effect of gain value on their choice to gamble than GD subjects with low average attention for gains (post-hoc: $Z=2.85$, $p=0.012$), whereas this was not the case in HCs (post-hoc: $Z=0.03$, $p>0.9$) (Figure 5A). This specifically resulted in significantly increased effects of gain on choice in GD compared to HC at average and high levels of dwell time to gains (average: $Z=1.95$, $p=0.051$, high: $Z=2.96$, $p=0.003$). In other words, GD subjects that had high attention for gains tended to accept gambles with lower gain values compared to GD subjects that had low attention for gains.

The three-way interaction between loss value, attention to losses and group showed a trend effect ($p=0.07$). Exploratory post-hoc tests indicated that GD subjects with high average attention for losses showed a marginally stronger effect of loss value on their choice to gamble than GD subjects with low average attention for losses (post-hoc: $Z=2.03$, $p=0.106$), whereas this was not the case in HCs (post-hoc: $Z=-0.26$, $p>0.9$) (Figure 5B). Thus, a marginal effect was found showing that GD subjects that had high attention for losses stopped gambling sooner when loss values increased, whereas those with relatively low attention continued gambling even when there was much to lose.

Table 3: Results of the mixed-effects model on the influence of attentional measures on choice

A)	Model 1G	
Parameter	**Estimate (SE)**	**p-value**
Intercept	0.41 (0.38)	0.283
Gain Value	2.64 (0.21)	**<0.001**
Loss Value	-2.42 (0.19)	**<0.001**
Group (GD)	2.24 (0.76)	**0.003**
Confidence	-0.07 (0.13)	0.568
Dwell Time on Gain	0.00 (0.08)	0.969
Dwell Time on Loss	-0.22 (0.08)	**0.005**
Gain Value x Group (GD)	0.61 (0.41)	0.136
Loss Value x Group (GD)	0.18 (0.37)	0.625
Confidence x Group (GD)	-0.06 (0.25)	0.819
Dwell Time on Gain x Group (GD)	-0.13 (0.15)	0.362
Dwell Time on Gain x Gain Value	-0.08 (0.05)	0.115
Dwell Time on Loss x Group (GD)	-0.13 (0.15)	0.380
Dwell Time on Loss x Loss Value	0.13 (0.05)	**0.006**
Dwell Time on Gain x Gain Value x Group (GD)	0.12 (0.10)	0.204
Dwell Time on Loss x Loss Value x Group (GD)	-0.08 (0.09)	0.377
AIC: 5050.9 R^2: 0.890		

B)	Model 1H	
Parameter	**Estimate (SE)**	**p-value**
Intercept	0.41 (0.38)	0.281
Gain Value	2.73 (0.21)	**<0.001**
Loss Value	-2.47 (0.20)	**<0.001**
Average Dwell Time Gain	1.13 (0.69)	*0.098*
Average Dwell Time Loss	-1.58 (0.69)	**0.021**
Trial-by-Trial Deviations Dwell Time Gain	-0.00 (0.07)	0.967
Trial-by-Trial Deviations Dwell Time Loss	-0.18 (0.07)	**0.007**
Group (GD)	2.20 (0.76)	**0.004**
Confidence	-0.07 (0.13)	0.561
Gain Value x Group (GD)	0.80 (0.41)	*0.051*
Gain Value x Average Dwell Time Gain	0.32 (0.16)	**0.045**
Gain Value x Trial-by-Trial Deviations Dwell Time Gain	-0.09 (0.05)	**0.042**
Group (GD) x Average Dwell Time Gain	2.51 (1.39)	*0.072*
Group (GD) x Trial-by-Trial Deviations Dwell Time Gain	-0.15 (0.14)	0.258
Loss Value x Group (GD)	0.05 (0.38)	0.890
Loss Value x Average Dwell Time Loss	-0.20 (0.13)	0.137
Loss Value x Trial-by-Trial Deviations Dwell Time Loss	0.10 (0.04)	**0.014**
Group (GD) x Average Dwell Time Loss	-2.89 (1.43)	**0.044**

Group (GD) x Trial-by-Trial Deviations Dwell Time Loss	-0.10 (0.13)	0.453
Confidence x Group (GD)	-0.05 (0.25)	0.835
Average Dwell Time Gain x Gain Value x Group (GD)	0.77 (0.31)	**0.014**
Trial-by-Trial Deviations Dwell Time Gain x Gain Value x Group (GD)	0.08 (0.09)	0.382
Average Dwell Time Loss x Loss Value x Group (GD)	-0.48 (0.27)	*0.074*
Trial-by-Trial Deviations Dwell Time Loss x Loss Value x Group (GD)	-0.08 (0.08)	0.323
AIC: 5056.9 R^2: 0.893		

A) Results of mixed-model model specified in detail in Appendix G Table G1G. **B)** Results of mixed-model model specified in detail in Appendix G Table G1H**.** Shown are the beta estimates, their standard error (SE) and 95% confidence intervals (CI), statistics and p-values. Loss values were entered as absolute values for easier interpretation. N = 55 subjects with a total of 8295 observations. **p<0.05, **p<0.01, ***p<0.001.

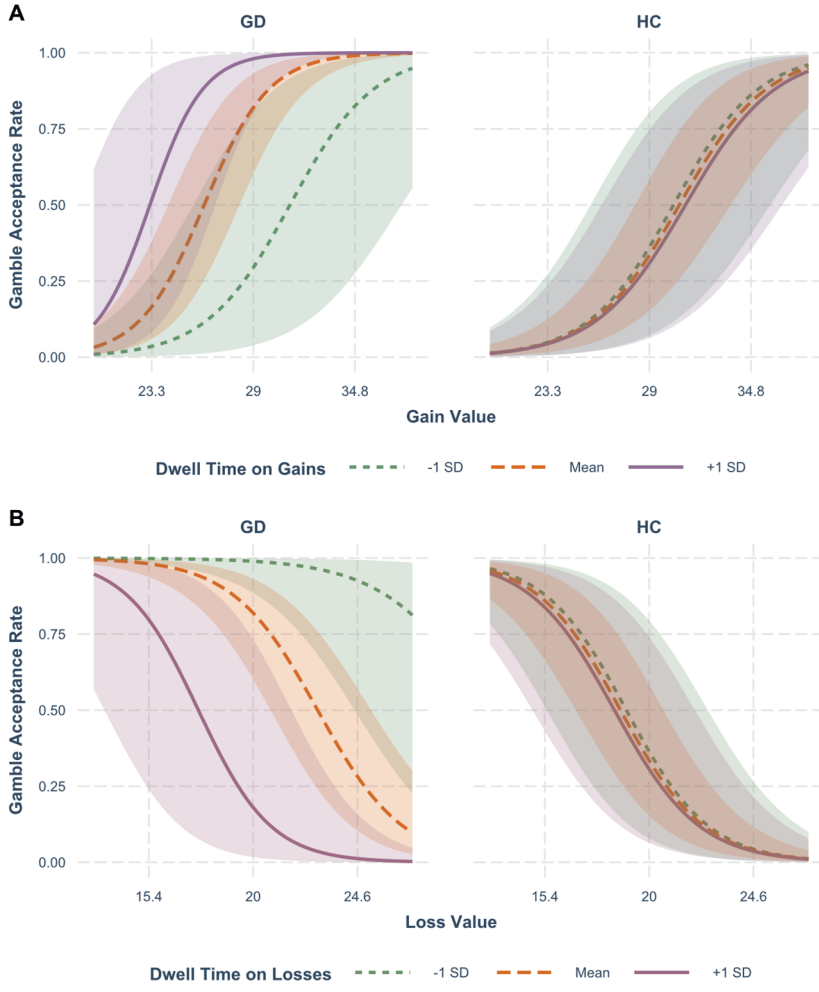


Figure 5: Interactions between group, values and attention on gambling choices. Plots of significant three-way interactions between group, average attention and values (either gains or losses) predicting gambling acceptance. Average dwell times on gains and losses were fixed at -1 standard deviation (based on subjects with relatively low average attention for gains/losses), the mean (based on subjects with relatively average attention for gains/losses) and +1 standard deviation (based on subjects with relatively high average attention for gains/losses). For purpose of illustration, z-scored gain and loss values were transformed to real numbers by calculating the mean ± 1 SD. **A)** The interaction effect between group, average attention toward gains and gain value. Regression curves show that when GD subjects have high average attention toward gains, they tend to accept gambles with lower gain values, compared to GD subjects with low average attention toward gains. This effect is not found in control subjects. **B)** The interaction effect between group, average attention toward losses and loss value. Regression curves show that when GD subjects have high average attention toward losses, they tend to reject gambles with lower loss values, compared to GD subjects with low average attention toward gains. This effect is not found in control subjects.

Discussion

The current study investigated how attention toward potential gains and losses during risky decision-making influences choices using a mixed-gambles task in GD patients and HCs. In line with previous findings, GD patients displayed higher gambling propensity (i.e., are more risk-taking (Brand et al., 2005; Brevers et al., 2012; Goudriaan et al., 2005; Ligneul et al., 2013; Ochoa et al., 2013; Spurrier & Blaszczynski, 2014)), lower loss aversion (Gelskov et al., 2016; Giorgetta et al., 2014), and increased influence of gains on confidence than HCs (Hoven, de Boer, et al., 2022). Also replicating earlier work, overall gambling propensity increased when there was more to gain and less to lose (J. Engelmann et al., 2021; J. B. Engelmann et al., 2015; J. B. Engelmann & Tamir, 2009). However, there were no group differences in the influence of gain value on gambling propensity, suggesting no increased reward sensitivity in GD.

We extend the current literature by investigating the role of attention in risky decision-making using eye-tracking, which has been underexplored in GD thus far. Consistent with previous studies in HCs, subjects' overall relative attention toward gains decreased in favor of attention toward losses when loss values increased (J. Engelmann et al., 2021; Gluth et al., 2018, 2020). We did not find group differences in attention to either gains or losses, suggesting no direct attentional biases in GD. However, while HCs increased their attention to gains with higher gain values, patients with GD did not. Moreover, while patients with GD displayed lower loss aversion, they did not show less attention to losses, rather, in both groups, increased trial-by-trial attention to losses resulted in less gambling.

The question arises whether attention modulates the effect of gains and losses on choice behavior differently in GD relative to controls. Our exploratory analyses that differentiated between two different channels of attention indeed indicated that the effect of gain value on gambling choices was modulated by the amount of *average* attention on gains in GD only. In other words, patients with GD who focused more on gains exhibited a greater gambling propensity at relatively low gain values. Notably, the strength of the effect of gain value on choice only significantly differed between groups at average and high levels of attention to gains, while patients with GD and HCs with relatively low levels of average attention to gains did not differ. Moreover, patients with GD who had relatively more average attention to losses showed a reduction in gambling propensity at relatively lower loss values, but note that this was at trend level. Since average attention relates to goal-directed or top-down attention, this measure likely reflects one's preferences and beliefs (Corbetta & Shulman, 2002; Pachur et al., 2018). Hence, the current results suggest that gambling choices in patients with GD, relative

to HCs are more influenced by their preferences for gains. Future studies are needed to verify if and how top-down attentional processes affect decision-making in GD.

Our study has limitations to address. In the current paradigm we weren't able to compute risk aversion (to compare against loss aversion), which could be an alternative explanation for increased gambling propensity (e.g.,(J. B. Engelmann et al., 2015; J. B. Engelmann & Tamir, 2009)). However, modeling both risk aversion and loss aversion in a mixed gamble task is difficult, as parameters are prohibitively correlated (Stewart et al., 2015). Moreover, it is possible that the relative unattractiveness of our task, relative to an actual gambling game, led to the relatively low reward sensitivity in GD, as suggested before (Leyton & Vezina, 2012, 2013; Van Holst, Veltman, Van Den Brink, et al., 2012). Eye-tracking during real-world gambling will likely be more sensitive and provide higher ecological validity and is needed to test whether similar cognitive processes are involved. Lastly, the current study population was recruited from therapy centers and thus included subjects who had received cognitive behavioral therapy (CBT) for their gambling disorder. Because CBT is partly focused on reducing irrational thoughts about gambling (Sylvain et al., 1997; Toneatto & Gunaratne, 2009) to dampen the influence of outcome salience on choice behavior, CBT might have reduced attentional biases towards gains, as well as gambling propensity in our sample. Such biases might be more pronounced in patients with untreated GD. Future studies on the role of attentional biases can address this by manipulating individuals' preferences and valuation processes, e.g. via priming (Cohn et al., 2015), and assessing how this affects attentional and gambling behavior. Finally, since real-life gamble products exploit outcome salience, i.e. via increasing (decreasing) the salience of wins (losses) (Yücel et al., 2018), it could be that in an untreated GD population the influence bottom-up attention towards presented gambling stimuli is enhanced.

Conclusions

In sum, the current study points toward nuanced effects of attention on the strength of the effect of rewards on (increased) risky decision-making in GD. GD patients who have more attention for gains are more strongly affected by gain values in their choice to gamble. Since this study is one of the first investigating attentional processes during a decision-making task in GD, more research is needed to establish a more detailed understanding of the relationship between choice preferences, attention and eventual gambling behavior. In time, this knowledge may help to improve the treatment of GD, for example by personalizing attentional bias modification training.

Data Availability

All data and code needed to evaluate or reproduce the figures and analyses described in the paper and supplementary materials are available online at 'https://osf.io/5v2j4/?view_only=5e56e989a84242bfa2f41c1292b77cdd'.

Acknowledgments

We are thankful for all the participants that participated in this study and for the funding we received from Amsterdam Brain and Cognition (ABC) to conduct this research.

Disclosure statement

None of the authors have any conflicts of interest to declare.

10

Confidence and risky decision-making in gambling disorder

Hoven M

Hirmas A

Engelmann JB

van Holst RJ

Journal of Behavioral Addictions, 2023, 1-7

Abstract

Background and aims

People with Gambling Disorder (GD) often make risky decisions and experience cognitive distortions about gambling. Moreover, people with GD have been shown to be overly confident in their decisions, especially when money can be won. Here we investigated if and how the act of making a risky choice with varying monetary stakes impacts confidence differently in patients with GD (n= 27) relative to healthy controls (HCs) (n=30).

Methods

We used data from our previous mixed-gamble study, in which participants were given the choice of a certain option or a 50/50 gamble with potential gains or losses, after which they rated their confidence.

Results

While HCs were more confident when making certain than risky choices, GD patients were specifically more confident when making risky choices than certain choices. Notably, relative to HCs, confidence of patients with GD decreased more strongly with higher gain values when making a certain choice, suggesting a stronger fear of missing out or "anticipated regret" of missing out on potential gains when rejecting the risky choice.

Discussion

The current findings highlight the potential relevance of confidence and "regret" as cognitive mechanisms feeding into excessive risk-taking as seen in GD. Moreover, this study adds to the limited previous work investigating how confidence is affected in value-based risky contexts.

Introduction

Gambling involves taking risks, typically with a high probability of loss against a smaller probability of gain. While for most people gambling is a leisure activity, for some people it develops into a gambling disorder (GD), described as the continuation or escalation of gambling despite the occurrence of negative consequences (American Psychiatric Association, 2013). It is often hard to grasp why people continue to show irrational gambling behavior when it is clear that, in the long term, "the house always wins".

Research on risk-taking and gambling-related cognition finds that people with GD make more risky decisions (Brand et al., 2005; Brevers et al., 2012; Ligneul et al., 2013; Ochoa et al., 2013; Spurrier & Blaszczynski, 2014), are less loss averse (Gelskov et al., 2016; Giorgetta et al., 2014; Hoven, Hirmas, et al., 2023) and exhibit higher levels of cognitive distortions about gambling than people without GD (Joukhador et al., 2003; Ledgerwood et al., 2020). Cognitive distortions about gambling often involve cognitions that minimalize the perceived risk of gambling and encourage gambling (Goodie & Fortune, 2013) (i.e., "the illusion of control" (Langer, 1975). Moreover, people with GD have been shown to be overly confident in their decisions (Brevers et al., 2013, 2014; Goodie, 2005; Lakey et al., 2007), especially when money can be won (Hoven, de Boer, et al., 2022). We recently replicated our findings using a mixed gamble task showing that relative to controls, patients with GD gambled more, and that increasing amounts of potential gains increased confidence more strongly in patients than in controls (Hoven, Hirmas, et al., 2023). Since accurate confidence is important for monitoring errors (Boldt & Yeung, 2015; Yeung & Summerfield, 2012), learning (Meyniel, Schlunegger, et al., 2015) and planning subsequent actions (Desender et al., 2018), having too much confidence in one's choices could contribute to risky decision-making (Hoven, de Boer, et al., 2022). While it has become clear that contextual cues, such as monetary incentives, can bias confidence, little is known about how the presence of risk and the act of making a risky choice impacts confidence and whether this interacts with incentive value.

One prior study conducted in healthy subjects investigated the impact of risky choices on confidence judgments (da Silva Castanheira et al., 2021). Their results indicated that confidence was significantly higher when selecting a certain prospect compared with a risky one - an intuitive finding that reflects the decision-makers feeling of uncertainty that comes with making a risky choice. Since there are little to no other studies that have investigated this in GD, it remains unknown whether risky choices and monetary incentives affect confidence of people with GD in the same way as healthy controls (HCs).

Based on previous findings (Brevers et al., 2013, 2014; Goodie, 2005; Lakey et al., 2007), we hypothesized that first, patients with GD relative to HCs are generally more confident in their choices made during an experiment where incentives can be won. Secondly, while HCs are relatively more confident in certain versus risky choices, patients with GD are relatively more confident in risky versus certain choices because of their experience with gambling. Finally, we hypothesized that confidence judgments of patients with GD compared to HCs are more sensitive to increases in potential gains, in line with the suggestion that gamblers might be more overconfident with greater potential gains (Hoven, de Boer, et al., 2022) such that increasing gain value increases(/decreases) confidence more strongly when making a risky(/certain) choice. We tested these hypotheses by utilizing data from our previous mixed-gamble study (Hoven, Hirmas, et al., 2023).

Methods

Participants

As the current study used data from our previous mixed-gamble study (Hoven, Hirmas, et al., 2023), the description of the participants are the same as in our previous paper. 27 patients with GD and 30 HCs were included, matched on age, sex and education, recruited online and via patient clinics in the Netherlands. All patients with GD had been in treatment and gambled regularly within the previous year and were diagnosed by a certified medical professional for gambling disorder using the DSM-5 criteria. All subjects did not currently or in the previous 6 months suffer from any psychiatric disorder, except for gambling disorder for the GD group, and did not use medication.

Experimental Task and Procedure

All subjects performed a mixed-gamble task including 160 trials (Figure 1A). Gambles were presented with an equal (50/50) chance of either gaining or losing a specific value and subjects chose between two options: rejecting (certain option) or accepting (risky option) the gamble. The certain option entailed opting for the initial endowment of €25 without the possibility of bonuses. The risky option entailed potentially gaining or losing additional bonuses as presented by the gamble. After each choice, feedback indicated the chosen option (but no feedback about wins and losses was provided until the end of the experiment to prevent learning) and subjects were asked to rate their confidence on a 7-point scale. Each combination of gains and losses was shown twice for counterbalancing, and gains or losses never appeared on the same side for more than

three times in a row. All subjects performed a training session. For more details see Figure 1A and Hoven, Hirmas, et al., (2023). The same exclusion criteria as in our previous study were applied, leading to a final sample of 26 patients GD and 29 HCs (Hoven, Hirmas, et al., 2023).

Analyses

For all analyses we used R (version 1.4.1106) with the packages emmeans (Lenth et al., 2018) , lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017). Age, sex, education level, gambling severity and percentage gambling choices were compared between groups using two-sample t-tests or chi-square tests.

To test for group differences in the effects of choice type, value and their interaction on confidence, we fit two mixed-effects models on our trial-by-trial data. In the first model, the effects of choice, group and expected value (0.5*loss value + 0.5*gain value), and their three-way interaction on confidence were investigated. Moreover, a covariate of the log of the reaction time (logRT, due to skewness), random intercepts and random slopes of EV and choice were included. In the second model, instead of using expected value, we investigated the separate effects of gain and loss value and their interactions with choice and group (choice*gain*group and choice*loss*group) on confidence. In both models, LogRT, EV, gain and loss values were z-scored and an effects coding scheme was used for the categorical group and choice variables. Post-hoc tests were performed to quantify significant interactions.

Ethics

The experiment was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Board of the University of Amsterdam.

Results

No group differences were found in age (GD: 37.4 ± 12.1; HC: 34.8 ± 8.61; $t^{53} = 0.92$, $p = 0.36$), sex (GD: 21 males, 5 females; HC: 23 males, 6 females; $X2 = 2.8*10^{-31}$, $p = 1$) or education level (GD: 3.08 ± 0.89; HC: 3.31 ± 1.17; $t^{53}=0.83$, $p = 0.41$). Problem Gambling severity index (PGSI (Ferris & Wynne, 2001)) scores were significantly higher in patients with GD (15.3 ± 3.94) than in HCs (0) (Welch's $t^{25}=-19.87$, $p<.001$). Patients with GD

scored 59.3 ± 23.2 on the Gamblers Beliefs Questionnaire (GBQ (Steenbergh et al., 2002)), a self-report measure of gamblers' cognitive distortions, where a higher score indicates more cognitive distortions. In general, patients with GD made more risky choices (60.7% ± 4.22%) than HCs (42.7% ± 3.81%) (t_{53}=3.175, p=0.002), and previous work using this dataset indicated less loss aversion in patients with GD compared to HCs (Hoven, Hirmas, et al., 2023).

The first mixed-effects model showed a significant main effect of reaction time on confidence, indicating increased confidence for choices with faster reaction times (**Table 1A**). A significant interaction between choice and group showed that GD patients were more confident in risky than certain choices (post-hoc: Z=2.781, p=.005), and a trend effect for the opposite pattern for HCs (post-hoc: Z-ratio=-1.772, p=.076) (Figure 1B). The significant interaction between choice and EV indicated that confidence increased with increasing EV for risky choices (slope = 0.583), but decreased for certain choices (slope = -0.587). The significant three-way interaction between choice, EV and group indicated that GD patients, compared to HC, showed even lower confidence rates when rejecting high EV gambles (Figure 1C). Post-hoc analyses confirmed that the negative effect of EV on confidence when making a certain choice was stronger in GD (slope: -0.734) than in HC (slope -0.439; Z=-3.573, p<.001)

Model 2 (Table 1B) separated the effects of gain and loss value, which were orthogonalized, allowing us to inspect whether the interaction effects observed in model 1 are driven by either gain or loss values. We find a similar three-way interaction effect between choice, *gain* value and group (Figure 1D), but not with *loss* value. Indeed, confidence of the GD group declined more strongly with increasing gain value when chosing the certain option, relative to HCs (GD slope:-0.513, HC slope:-0.276, post-hoc Z=-3.628 p<.001). Moreover, a significant interaction between loss value and choice indicated that with increasing loss value, all subjects became more confident when they chose the certain option (slope: 0.448), but became less confident when they chose the risky option (slope: -0.438). A significant interaction between loss value and group showed that the GD group was less sensitive to loss value (slope: 0.08) than the HC group (slope: -0.07) in terms of decreasing their confidence. Finally, neither PGSI nor GBQ score within the GD group correlated significantly with mean confidence (both p>.25) or confidence for risky choices (both p>.5).

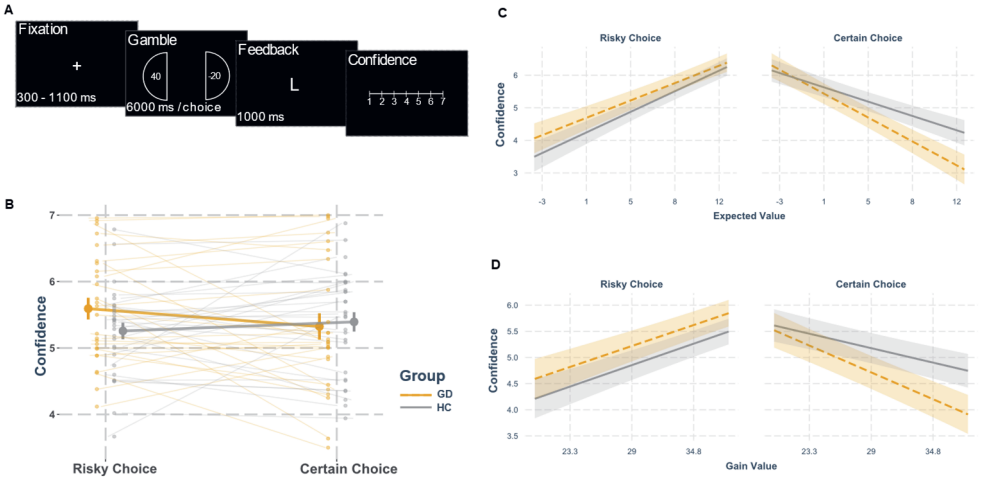


Figure 1: Mixed gambles task and results. (A) At the beginning of each trial a fixation cross was shown, jittered between 300 and 1100 ms. Then the gamble was shown (i.e., gain and loss value stimuli; left stimulus centered on 480x540, right stimulus centered on 1440x540) for the duration of the decision with a maximum of 6000 ms. Subjects were asked to accept or reject the gamble using the up or down key, respectively, after which a brief feedback message was shown indicating and confirming their choice (1000ms. L = lottery (risky) option, X = certain option, 'Respond Faster' = if failed to respond within 6000 ms). After each choice participants rated their confidence on a scale from 1 (not sure) to 7 (very sure) (unlimited time). Subjects did not receive any feedback about the outcome of their choices (win or loss outcomes) until after completion of the experiment to avoid history and learning effects. **(B)** The significant interaction effect between choice type and group shows that the GD group were more confident while making risky choices, while the HC group were more confidenct when making certain choices. Large dots and error bars signify means and standard errors, smaller dots represent individual subject data points (**Table 1A**). **(C)** A significant three-way interaction between expected value, choice type and group shows that increasing expected value of the gamble has a stronger negative effect on confidence in the GD group when making certain choices (**Table 1A**). **(D)** A significant three-way interaction between gain value, choice type and group shows that increasing gain value has a stronger negative effect on confidence in the GD group when making certain choices (see **Table 1B**). Yellow color indicates GD, grey color indicates HC.

Table 1: Results of mixed-effects models on confidence

A) **Model 1: expected value**

Parameter	**Estimate (SE)**	**t-value**	**p-value**
Intercept	5.00 (0.09)	58.49	**<0.001**
Choice (Certain Option)	-0.05 (0.06)	-0.81	0.419
Group (GD)	-0.03 (0.09)	-0.40	0.687
Expected Value	-0.001 (0.04)	-0.05	0.960
Reaction Time	-0.37 (0.01)	-25.86	**<0.001**
Choice (Certain Option) x Group (GD)	-0.21 (0.06)	-3.24	**0.002**
Choice (Certain Option) x Expected Value	-0.58 (0.02)	-32.62	**<0.001**
Group (GD) x Expected Value	-0.10 (0.04)	-2.70	**0.009**
Choice (Certain Option) x Group (GD) x Expected Value	-0.05 (0.02)	-2.80	**0.005**
AIC: 24001.46 R^2: 0.506 # observations: 8295 trials (of 55 subjects)			

B) **Model 2: gain and loss value**

Parameter	**Estimate (SE)**	**t-value**	**p-value**
Intercept	4.99 (0.08)	58.81	**<0.001**
Choice (Certain Option)	-0.04 (0.06)	-0.69	0.491
Group (GD)	-0.02 (0.08)	-0.29	0.774
Gain Value	0.01 (0.03)	0.19	0.849
Loss Value	0.01 (0.03)	0.17	0.862
Reaction Time	-0.36 (0.01)	-25.65	**<0.001**
Choice (Certain Option) x Group (GD)	-0.21 (0.06)	-3.20	**0.002**
Choice (Certain Option) x Gain Value	-0.40 (0.02)	-25.62	**<0.001**
Group (GD) x Gain Value	-0.06 (0.03)	-2.19	**0.033**
Choice (Certain Option) x Loss Value	0.44 (0.01)	29.65	**<0.001**
Group (GD) x Loss Value	0.08 (0.03)	2.57	**0.013**
Choice (Certain Option) x Group (GD) x Gain Value	-0.06 (0.02)	-3.72	**<0.001**
Choice (Certain Option) x Group (GD) x Loss Value	-0.001 (0.01)	-0.04	0.972
AIC: 23912.11 R^2: 0.514 # observations: 8295 trials (of 55 subjects)			

Shown are the estimates, their standard errors (SE), t-values and p-values. Loss values were entered as absolute values for easier interpretation. The value of the choice options was modeled as expected value in model 1 and, separately as (experimentally orthogonalized) gain and loss value in model 2.

As a sensitivity analyses to verify whether fatigue or time on task affected our results, we included trial number as a covariate and also tested for the trial*group interaction in the two described models. The results indicated no significant effect of trial number nor trail*group interactions on any of the results. Moreover, including this variable did not change any of our previously reported results.

Discussion

Why do patients with GD continue to gamble regardless of all the negative consequences? One answer may lie in overconfidence in their actions, specifically when risk and monetary incentives are involved. This study investigated how making a risky choice with varying monetary stakes impacts confidence and whether patients with GD are affected differently than HCs.

There was some evidence for our first hypothesis of general increased confidence in GD compared with HCs. There was convincing support for our second hypothesis: relative to HCs (who have higher confidence when making a certain versus risky choice), patients with GD are more confident when making risky choices than certain choices. Notably, relative to HCs, confidence of patients with GD decreased more strongly with higher gain values when making a certain choice. Hence, these findings also partly support our third hypothesis of increased sensitivity to gain values in GD patients and point to our measure of confidence capturing a stronger fear of missing out or "anticipated regret" of missing out on potential gains when rejecting gambles.

This fear of missing out on potential gains is recognized in the clinical presentation of GD (Ladouceur, 2004) and may be reflected in the higher willingness to gamble. Patients often describe that their only solution to solving their financial problems is taking excessive risks in the hope of obtaining high gains. Indeed, research has shown that scarcity creates "bandwidth taxes" that reduce mental resources, impairing cognitive ability and causing counterproductive behavior, such as risk-taking (Liang et al., 2021), which perpetuates poverty (Haushofer & Fehr, 2014; Ong et al., 2019). In the current study, as expected, more patients with GD (n=18) experienced debts than HCs (n=5). In that light, this stronger fear of missing out on potential gains may not only speak to patients with GD but also to people who experience financial debts. The lower confidence when selecting the certain option found in the GD group aligns well with recent findings by Wu et al. (2021). They computationally assessed an anticipatory regret parameter that captured the difference between the worst outcome in one gamble versus the best outcome in the other gamble and found that people with GD experience increased anticipatory regret relative to controls. While our findings and

those of Wu et al. (2021) should be considered preliminary, they highlight the relevance of regret and confidence as cognitive mechanisms in disordered gambling. Future studies on this topic are needed and can draw on extensive mathematical and experimental methods developed by behavioral economics.

The current study adds to the limited previous work investigating how confidence is affected in value-based risky contexts. Da Silva Castanheira et al. (2021) found that healthy people are more confident when selecting certain options. Moreover, consistent with previous findings (De Martino et al., 2012; Folke et al., 2017), in the absence of risk, higher subjective values and faster RTs were associated with higher confidence ratings (da Silva Castanheira et al., 2021). We replicated these findings in our HCs and observed the weakening of these well-documented relations with risky decisions relative to certain decisions (see Appendix H). These findings fit the notion that risky choices are accompanied by an inherent uncertainty about the option's value and that RTs are slower under greater uncertainty (D. G. Lee & Daunizeau, 2021; D. G. Lee & Hare, 2023).

Our results should be interpreted with some limitations in mind. First, all included patients received treatment, and the task's relative unattractiveness and artificial nature may have attenuated natural risk-taking behavior in our patient sample. Future studies should assess whether the current findings generalize to more realistic gambling situations and to untreated patients. Additionally, longitudinal studies are needed to dissect whether alterations in confidence under risk are a cause or consequence of GD. Furthermore, the influence of financial debts on confidence and anticipated regret needs to be established. Finally, the current results were secondary to our previous work (Hoven, Hirmas, et al., 2023) and can be considered exploratory. Nonetheless, these results provide an important initial demonstration of how subjective confidence during risky decision-making is differently affected in patients with GD relative to HCs.

In sum, the current study points out that compared to HCs, patients with GD are generally more confident when taking risks versus playing it safe. Importantly, they become less confident about playing it safe when the potential winnings increase. This behavioral pattern matches anticipatory regret of missing out on potential gains, which may contribute to excessive risk-taking in GD patients.

Acknowledgments

We are thankful for the funding we received from Amsterdam Brain and Cognition (ABC) to conduct this research.

Disclosure statement

None of the authors have any conflicts of interest to declare.

Part IV

Discussion

11

Summary and General Discussion

In this thesis I investigated confidence judgments as a metacognitive construct, exploring its neurobiological foundations, biases, and its relationship with psychiatric symptoms and disorders. Through a series of comprehensive studies encompassing clinical samples of patients with OCD and/or GD, healthy controls and general population samples, I have investigated the disruptions in metacognitive abilities across different contexts. By employing a range of methodologies such as fMRI, eye-tracking, cognitive computer tasks, questionnaires and computational modeling, we have gained a multifaceted understanding of confidence in psychiatry. Overall, our findings indicate that disturbances in confidence, and in a broader sense, metacognition, are a central aspect of mental health.

Summary of the main findings

Part I: Confidence and its Biases in Psychiatry

Chapter 2 presents a literature review on confidence abnormalities across psychiatric disorders and symptoms in both clinical and subclinical populations. There was compelling evidence for an association between obsessive-compulsive symptoms/behavior and reduced confidence in (sub)clinical populations. In schizophrenia, we consistently found overconfidence, specifically in errors, leading to an impaired confidence discrimination in (sub)clinical samples. In GD, subclinical studies showed overconfidence, while studies in clinical substance-use dependency showed worsened metacognitive sensitivity. (Sub)clinical studies indicated lower confidence levels related to depression and anxiety. Our review highlighted confidence abnormalities across various (sub)clinical psychiatric conditions, with specific directions for different symptom presentations.

In **Chapter 3**, we investigated the neural correlates of confidence and the biasing effect of incentives on confidence signals in healthy participants. We replicated the earlier found incentive confidence bias, where gains(/losses) increase(/decrease) confidence without affecting performance. The fMRI results showed that vmPFC activity was related to an early certainty signal and to incentive value, but not strongly to confidence during rating. Activity in the vmPFC correlated with confidence solely in the gain context, but not in the neutral or loss contexts. Overall, this study demonstrated that motivational processes can influence confidence and revealed that confidence signals in the vmPFC can undergo modulation by motivational signals.

Chapter 4 reports an fMRI study investigated the neural correlates of confidence and the biasing effect of incentives in two clinical populations: OCD and GD. The fMRI

results showed positive relationships between vmPFC activity and confidence in all groups, and replicated that this relationship was strongest in the gain context. No group differences were found for confidence or motivational BOLD signals. Behaviorally, the incentive confidence bias was replicated. Moreover, confidence was significantly(/marginally) higher in GD patients compared to OCD patients(/controls). No differences in confidence were found between OCD and controls. A trend interaction effect indicated that GD patients were specifically more confident than OCD patients and controls in gain context. No evidence was found for lower confidence specifically in loss context in OCD patients.

In **Chapter 5** we explored confidence on multiple hierarchical levels (local confidence, global confidence and higher-order self-beliefs) across a wide range of psychopathology using a transdiagnostic approach. There were significant positive relationships between the hierarchical levels of confidence. Moreover, subjects scoring high on anxious-depressive (AD) symptoms showed significantly lower confidence (*under*confidence), while subjects scoring high on compulsive-behavior-and-intrusive-thought (CIT) symptoms showed significantly more *over*confidence and more distorted coupling between the hierarchical confidence levels. Low self-beliefs were the strongest predictor of all transdiagnostic symptom dimensions, while higher local confidence positively predicted the severity of CIT symptoms. No relationships were found between transdiagnostic symptom dimensions and metacognitive efficiency.

Part II: Confidence in OCD

In **Chapter 6** we tested confidence in various groups: clinical medication-free OCD patients, a healthy, and a highly compulsive general population sample. We tested the assumption that highly compulsive individuals from the general population (HComp, matched on the severity of OCD symptoms) resemble clinical OCD patients in terms of disturbances in (meta)cognitive processes. The results indicated that OCD patients exhibited significantly lower local and global confidence levels compared to healthy controls, indicating underconfidence relative to the control group. Conversely, the HComp group demonstrated significant overconfidence and poorer metacognitive sensitivity relative to the clinical patient group. Thus, clinical OCD patients have distinct metacognitive patterns compared to a highly compulsive group from the general population.

In **Chapter 7** we focused on exploring the relationship between learning and confidence under volatility in OCD, comparing medication-free OCD patients, healthy controls, and highly compulsive subjects from the general population. The findings

demonstrated that OCD patients had lower confidence and higher error-sensitivity (i.e., higher learning rates for small prediction errors) compared to both healthy and highly compulsive individuals. No evidence was found to support a decoupling between action and confidence in the OCD group. These results suggest that in unmedicated OCD, underconfidence, rather than a decoupling between action and confidence may underlie compulsive behaviors. Overall, these findings give way to the idea that obsessive-compulsive symptoms can go together with different (meta)cognitive and behavioral profiles, depending on the sample.

Part III: Confidence in GD

In **Chapter 8**, we explored the association between confidence and learning under volatility in GD compared to HC. We found no significant group differences in confidence or action updating. However, the coupling between confidence and action was weaker in the GD group compared to the control group, which might suggest that patients with GD may give less weight to their feelings of confidence when taking actions. Additionally, the GD group exhibited lower overall learning rates, indicating a reduced influence of the most recent outcome on subsequent action in GD. Since we did not find evidence for higher confidence ratings on this task, this may point to the notion of context-dependent (i.e., disorder-relevant context) increases of confidence in GD. Overall, the weaker coupling between confidence and action suggests that confidence judgments guide decisions less in GD.

Chapter 9 reports on the study of the role of attention, value and confidence in risky decision-making within GD using eye-tracking. The GD group gambled more, was less loss averse and more confident in their gambles when gain value increased compared to the control group. No group differences in average attention to either gains or losses were found. Overall, participants gambled more(/less) when their average attention towards gains(/losses) was higher. A specific pattern emerged in the GD group, where GD patients with higher average attention towards gains were more strongly influenced by gain values in their decision to gamble compared to GD patients with lower average attention towards gains.

Finally, **Chapter 10** extended Chapter 9 by further testing the effects of risk-taking on confidence during gambling in GD patients. While the control group was more confident when making certain choices compared to risky choices, the GD group was more confident when making risky choices. While all participants demonstrated decreases in confidence when opting to forgo a gamble with increasing potential gain value, this effect was more pronounced in the GD group. The findings suggest that GD

is associated with a stronger anticipatory regret, we postulate that this strong anticipatory regret may fuel excessive gambling.

General Discussion

OCD patients generally have lower confidence, but intact metacognitive sensitivity

Deficits in metacognitive ability are a central aspect of the phenomenology of OCD, including excessive doubting (Dar, 2004; Dar et al., 2000; Hermans et al., 2008), intolerance of uncertainty (Gentes & Ruscio, 2011; Jacoby et al., 2013; Pinciotti et al., 2021) and, as we show, decreased confidence. Our findings of general decreases in confidence in OCD (**Chapter 2**) were substantiated in a recent meta-analysis (Dar et al., 2022) and across the chapters of this thesis. Specifically, our research demonstrated that OCD patients exhibited lower local and global confidence in a perceptual decision-making task (**Chapter 6**) and a predictive inference learning task (**Chapter 7**) compared to controls. In the incentivized confidence task (**Chapter 4**), however, neither evidence of decreased confidence nor specific sensitivity to loss was found. Additionally, across **Chapters 4**, **6** and **7**, the ability to integrate local confidence into a global feeling of confidence was found to be intact, and no deviations in metacognitive sensitivity, as measured by discrimination, meta-d' or meta-d'/d', were found. Together, this points to a specific negative bias of confidence judgments in OCD, while patients have an intact ability to use their confidence to inform and discriminate their decisions.

Several factors should be considered to interpret the lack of finding lower confidence in OCD in **Chapter 4** compared to the other chapters. Importantly, **Chapter 4** included a different clinical OCD sample compared to **Chapters 6** and **7**. Although the demographic and clinical descriptions indicated similar age ranges, symptom severity and sex distributions, many patients in **Chapter 4** used medication for their OCD symptoms, while those in **Chapter 6** and **7** did not. This may have influenced the results, as studies have suggested improvements in confidence abnormalities among medicated versus unmedicated patients (Marzuki et al., 2022). This could additionally partly explain our null results regarding loss sensitivity in confidence, as previous research demonstrated that medicated OCD patients exhibit less loss aversion compared to unmedicated OCD patients (Sip et al., 2018). Moreover, the incentivized confidence task (**Chapter 4**) was conducted inside an MRI scanner, which may influence neurocognitive task outcomes, particularly in clinical samples (Kolodny et al., 2022; van Maanen et al., 2016). Finally, the sample size was smaller in the

incentivized confidence study (N = 28) compared to the other studies (N~40), leading to reduced statistical power to detect group differences. Since we did find a trend effect of lower confidence in OCD compared to controls, a higher sample size could have resulted in significant differences. Overall, this thesis provides evidence for a negative confidence bias in OCD across various contexts, extending beyond local confidence in memory or perception, without abnormalities in metacognitive sensitivity or efficiency.

GD patients have increased confidence in gambling-related context

Metacognition has received less attention in GD, while clinical characteristics do suggest that patients with GD may struggle to critically evaluate their gambling beliefs and exhibit increased confidence in those beliefs (Armstrong et al., 2020). This thesis provides compelling evidence supporting these notions. Patients with GD, compared to controls, demonstrated increased levels of confidence, particularly in contexts involving potential gains (**Chapter 4**), high gain values (**Chapter 9**), and risk (**Chapter 10**), but not in a neutral learning task (**Chapter 8**). Moreover, the coupling between confidence and action was distorted in GD patients, suggesting that their confidence informed their actions to a lesser extent (**Chapter 8**). These findings suggest that increased confidence specifically manifests within gambling-related contexts, tied to potential gains or risks. Research on global confidence measures in GD is scarce, although studies have consistently reported lower levels of self-esteem and self-efficacy in this population (Casey et al., 2008; Choi & Kim, 2021; Hawker et al., 2021; Kaare et al., 2009; Park et al., 2019). Future studies are needed to clarify how increased local confidence judgements in gambling contexts go together with lower levels of higher-order self-beliefs, a pattern that we also observed in people scoring high on CIT symptoms (**Chapter 5**).

The observation of context-specific overconfidence is supported by earlier research that also did not find confidence differences between problematic gamblers and controls in a neutral grammar task (Brevers et al., 2014). Additionally, it fits with work showing gambling-specific reward sensitivity over other natural rewards in GD, indicating a motivational hierarchy that may contribute to gambling-related overconfidence (Sescousse et al., 2013). The influential model by Leyton & Vezina (2013) underscores the role of addiction-related cues in addiction behaviors. For example, gambling acceptance in a gambling task has been found to be dependent on and sensitive to gambling cues (Genauck et al., 2020). Domain-specific compulsivity is considered a central mechanism in behavioral addictions (Perales et al., 2020), which could relate to gambling-specific overconfidence. However, testing context

dependency is challenging, and studies have not always found support for the hypothesis that gambling cues impact cognitive function in GD (Van Timmeren et al., 2023).

In our studies, all patients with GD were either currently in treatment or had recently received treatment, mostly through cognitive behavioral therapy (CBT). This therapy aims to tackle cognitive biases, which could have led to reduced general confidence biases and diminished our ability to find differences using neutral tasks that do not trigger gambling-related biases, such as overconfidence in winning a gamble. Directly comparing confidence and metacognitive ability between gambling irrelevant and gambling relevant environments will help us better understand these processes.

Transdiagnostic and hierarchical approaches to study confidence in psychiatry

An impactful shift in the field is the use of a transdiagnostic approach to the study of (meta)cognition in psychiatry (Dalgleish et al., 2020; Gillan et al., 2016). Transdiagnostic research especially seems to hold promise for understanding OCD, where obsessive-compulsive and anxiety symptoms often coexist. Rouault, Seow, et al. (2018) were among the first to demonstrate that transdiagnostic symptom dimensions better captured abnormalities in confidence than disorder-specific symptoms. Transdiagnostic anxious-depressive (AD) symptoms related negatively to local confidence (mirroring the effects shown in clinical OCD samples), while a dimension of compulsive behavior and intrusive thoughts (CIT) positively related to local confidence. These findings have been replicated in this thesis (**Chapter 5**), and in the general knowledge and predictive inference domain (Benwell et al., 2022; Seow & Gillan, 2020).

Another recent shift has expanded the study of confidence beyond the local, trial-by-trial assessments to incorporate higher-order levels within an interconnected hierarchical framework comprising global confidence and higher-order self-beliefs (Seow et al., 2021). Our findings are one of the first empirical confirmations of this theoretical framework, confirming the positive relationships between the different levels of the confidence hierarchy (**Chapter 5**), which has been replicated recently (Katyal et al., 2023). The field of metacognition research in psychiatry has for a long time solely focused on local confidence, with the idea that abnormalities in local confidence would relate to higher-order feelings about the self and pathological behavior in daily life (Rouault et al., 2022). Actually testing these relationships is crucial to bridge the gap between experimental studies and their implications for patients' daily lives. In this thesis, we made an important step forward by showing that the same (negative)

abnormalities of local confidence indeed extend to higher-order levels of confidence, such as in symptoms of depression and apathy (**Chapter 5**), and clinical OCD patients (**Chapter 6**). This is not the case for every symptom dimension, however. In **Chapter 5,** we further provided crucial insights by merging the transdiagnostic and hierarchical approach to demonstrate that the hierarchical levels of confidence related differently to the symptom dimensions. While individuals with high AD symptoms showed negative biases across the hierarchy, individuals with high CIT symptoms showed a dissociation between the levels of the hierarchy. These findings highlight the importance of merging these approaches and call for a deeper investigation into underlying mechanisms, dynamics between levels of the hierarchy, and potential targets for treatment.

Indeed, recent computational work has shown that distinct mechanisms may underlie the confidence abnormalities found for the different symptom dimensions (Katyal et al., 2023). The authors showed that individuals with high AD symptoms were overly sensitive to trials with low local confidence when forming their global confidence estimates that stretched longer periods of time, indicating that global underconfidence in AD, in part, arises from the bottom-up influence of low local confidence. Interestingly, individuals with high CIT symptoms demonstrated an even stronger weighting of trials with low versus high local confidence when forming their global confidence. This finding implies that while individuals with high CIT symptoms are generally quite confident about their trial-by-trial choices, when they are asked to reflect on their global task performance they mostly rely on those instances in which they were not so confident. This hypersensitivity to low local confidence could in turn be fueled by top-down influences of overly negative (prior) self-beliefs, both in individuals with high CIT and AD symptoms. This also fits with our finding of the coexistence of heightened local confidence and lowered global confidence in individuals scoring high on CIT symptoms (**Chapter 5**). Since we showed that self-beliefs were more strongly affected in individuals with high AD compared to CIT symptoms (**Chapter 5**), it might be that these negative self-beliefs have a more direct and pronounced effect on the lower levels of the confidence hierarchy in individuals with high AD scores. On the other hand, in individuals with high CIT scores, negative self-beliefs may have an indirect influence, specifically affecting the integration of low local confidence to global confidence.

While Katyal et al. (2023) focused on the bottom-up influence of local on global confidence, recent work has indeed indicated that manipulating top-down prior self-beliefs about task ability causally and persistently influences local confidence measures, inducing biases of under- and overconfidence (Van Marcke et al., 2022). In

their model, prior beliefs shaped the mapping between accumulated evidence and confidence, where subjects with a very negative prior self-belief would never feel highly confident about their choices regardless of how much evidence they accumulated. Future work could benefit from applying these models to clinical data, exploring if the mapping between evidence and confidence is more tuned towards low confidence in individuals with high AD symptoms due to their stronger negative self-belief priors, than people with high CIT symptoms. Models incorporating both bottom-up and top-down influences across the hierarchy, where prior self-beliefs concurrently influence lower levels of confidence and are updated by these feelings of confidence could offer a more complete picture of the mechanisms underlying metacognitive dysfunction in psychiatric disorders. This could, in the long run, offer possible therapeutic entry points to restore accurate self-evaluation, that could have strong impact given the significance of self-beliefs for daily life. One potential therapeutic entry point is inspired by both studies (Katyal et al., 2023; Van Marcke et al., 2022), showing that positive trial-by-trial feedback positively impacted local confidence persistently over blocks and cognitive domains, which was mediated by changes in global confidence (Katyal et al., 2023). Strikingly, positive feedback also improved self-beliefs in the form of endorsement of positive vs. negative words (Katyal et al., 2023). This might be specifically promising for individuals suffering from underconfidence, such as patients with OCD or anxious and depressive symptoms.

Next to deviations of confidence, another finding that has been consistently reported is that of lower task performance in individuals with high CIT symptoms (Benwell et al., 2022; J. K. Lee et al., 2023). It has been suggested that high CIT symptoms are associated with an altered decision formation process, potentially linked to differences in the evidence accumulation process (Banca et al., 2015), or the formation of a higher-order model of task structure (Voon et al., 2015). Convincingly, a recent study revealed a negative relationship between a variable called ‘decision acuity’ (a measure of decision-making ability across a wide variety of decision-making paradigms, independent of IQ) and a dimension encompassing symptoms of schizotypy, compulsivity and obsessionality, similar to our CIT symptom dimension (Moutoussis et al., 2021). This suggests that individuals with high CIT symptoms not only show distortions in the ability to reflect on their decisions, but also in the decision process itself. These processes are interlinked, as difficulty in evidence accumulation could bias confidence in these decisions. Importantly, we did not find evidence for lower task performance in clinical OCD patients, implying differences in cognitive decision-making processes between clinical and sub-clinical groups suffering from obsessive-compulsive symptoms. Our findings in GD, however, also point to less decision acuity, indicated by more risk taking (**Chapter 9**) and slower learning rates (**Chapter 8**), and

thus resemble individuals scoring high on symptoms of schizotypy, compulsivity and obsessionality (Moutoussis et al., 2021). However, currently, GD symptoms have not often been taken into account within transdiagnostic approaches, so it remains to be tested how symptoms of GD would relate to transdiagnostic dimensions.

Generalizability of findings from the general population samples to clinical samples

Neurocognitive research in psychiatry usually involves studying either clinical samples or analog samples from the general population with high scores on the symptom under consideration, with the assumption that the associations between symptoms and the neurocognitive process of interest are similar in clinical and non-clinical general population samples (Abramowitz et al., 2014). In this thesis we showed that clinical OCD patients generally have a negative confidence bias (**Chapter 6,7**), while individuals from the general population scoring high on CIT symptoms have increased local confidence (**Chapter 5**). Meanwhile, local confidence did not relate specifically to OCD symptoms in the general population (**Chapter 5**). Indeed, while directly comparing OCD patient groups to highly compulsive general population groups, we found evidence for different metacognitive profiles. The OCD patients exhibited local and global underconfidence relative to the healthy and highly compulsive groups (**Chapter 6, 7**), along with higher learning rates and higher error sensitivity (**Chapter 7**).

These findings suggest that similar levels of obsessive-compulsive symptom severity relate to distinct behavioral profiles dependent on the nature of the sample. A recent model of OCD proposes that difficulty with processing state transitions may underlie obsessive-compulsive behaviors (Fradkin et al., 2020). On the one hand, OCD patients may have excessively low confidence in their actions because they have difficulty learning from past experiences to form an internal model of the world that guides their future actions. This could result in overreliance on immediate feedback (or obsessions) at the expense of accumulated knowledge, whilst the latter in theory is a more precise estimate of the state of the world. Due to the lack of confidence in their model of the world, patients' behavior can become compulsive and habitual, constantly reacting to incoming feedback (or obsessions). This notion aligns with our findings of lower local and global confidence and increased learning rates in OCD patients (**Chapter 6,7**). On the other hand, rigid prior beliefs about the model of the world that are resistant to new sensory feedback can be held with high confidence. This can make behavior compulsive, habitual and inflexible, aligning with our findings of increased confidence and lower learning rates in the highly compulsive group (**Chapter 7**) and GD patients

(**Chapter 8**). In this way, different behavioral profiles may be associated with compulsivity. It is important to recognize that our findings come from a group-level analysis, and individual variations exist within the groups, likely lying along a behavioral spectrum rather than fitting precisely into one profile or the other. Additionally, our samples of OCD patients that were compared with high compulsive individuals, consisted of individuals without medication and comorbidities. It is possible that OCD patients with more severe symptoms and/or those using medication might show a distinct behavioral pattern.

An important consideration that could separate clinical from highly compulsive general population samples is the extent to which symptoms affect daily life. In our papers, groups were matched on OCI-R score. The OCI-R assesses distress related to specific and select types of obsessions and compulsions, using three items per symptom type, which can confound severity with the type (and number) of symptoms that are present (Abramovitch et al., 2020). The Y-BOCS is likely more sensitive to heterogeneity of symptoms, as it first identifies the primary obsessions and compulsions before rating them on severity, frequency, duration and functional interference (i.e., impact on daily life). Therefore, it might be that functional impairments in daily life differ between the groups. Unfortunately, we do not have data for the non-clinical samples to directly test this, but we did find that our clinical group had a decreased feeling of autonomy in daily life compared to our highly compulsive group (**Chapter 6**). Corroborating this idea, recent studies have shown that students from the general population scoring high on OCD symptoms, compared to clinical OCD patients, indeed have identical OCI-R scores but lower Y-BOCS scores, indicating similar levels of distress but less impact on daily life in non-clinical samples (Abramovitch et al., 2023). To get a better understanding of the nature and impact of the obsessive-compulsive symptoms in these samples, future research should conduct a more comprehensive assessment of OCD symptoms using various different instruments in which severity is not influenced by the number of type of symptoms.

Another consideration is that, even though our samples showed comparable severity of obsessive-compulsive, depressive and anxiety symptoms, it remains possible that the highly compulsive groups differ from the clinical OCD groups in terms of other (transdiagnostic) symptoms. The pattern of results observed in the highly compulsive general population samples more closely resembles that of individuals with high CIT symptoms rather than AD symptoms. Conversely, the pattern of our clinical OCD group more closely resembles the AD pattern. Indeed within the highly compulsive group, CIT symptoms were higher than the overall average of the entire general population sample from which it was drawn, while the AD symptoms were lower than the overall average

of the entire general population sample. This might suggest that symptoms of impulsivity, schizotypy, addiction or eating disorders are more prominent in the general population high compulsive groups than in the clinical groups, partially driving these results. Indeed, previous research has demonstrated that OCD symptoms can be mediated by different subgroups of compulsive-impulsive phenotypes (Prochazkova et al., 2018), and that specific compulsive symptom clusters are associated with either schizotypal and body dysmorphia symptoms, while other clusters are associated with mood-related symptoms (Fontenelle et al., 2022). It is important to disentangle these effects in future research by including a more diverse set of questionnaires in the clinical population.

In light of our results, I would not directly advise against using analog samples altogether. I would, however, advise caution in bluntly generalizing these findings to clinical populations. This advice might be less urgent for some cognitive processes within OCD (e.g., attentional biases (Bar-Haim et al., 2007) or attachment problems (van Leeuwen et al., 2020)) or for disorders other than OCD (e.g., depression (**Chapter 2**, Rouault, Seow, et al., 2018; Sax et al., 2023; Seow & Gillan, 2020)). An important step forward is to apply transdiagnostic approaches not only to analog samples, but also to clinical samples to investigate if transdiagnostic dimensions can better explain (meta)cognitive deficits in actual patients compared to diagnostic categories. Recent studies have indeed demonstrated that transdiagnostic compulsivity symptoms had better predictive power than a binary OCD (or GD) diagnosis in explaining goal-directed behavior and neurobiological connectivity (Gillan et al., 2020; Parkes et al., 2019). In general, I believe more effort should go into directly comparing clinical and highly symptomatic general population samples, not only to study generalizability of (meta)cognitive ability, but also to get a better idea of the differences between these groups that could relate to the transition to treatment seeking or significant functional impairment.

Learning processes and confidence in OCD and GD

Confidence is tightly coupled to our actions and decisions (Schulz et al., 2023). We observed a weaker action-confidence coupling for patients with GD, but not for patients with OCD (respectively **Chapter 8, 7**). Additionally, GD patients and highly compulsive individuals from the general population showed decreased learning rates compared to controls (i.e., more ‘sticky’ behavior), while OCD patients showed increased learning rates (i.e., more ‘volatile’ behavior).

These findings align with earlier work indicating that patients with OCD have higher levels of switching between choices, or reduced stickiness, in probabilistic learning tasks (Hauser, Iannaccone, et al., 2017; Kanen et al., 2019; Marzuki et al., 2021). Furthermore, OCD patients more often changed their responses even when receiving positive feedback, which was positively related to checking symptoms (Benzina et al., 2021). Our findings in GD align with studies showing that patients with GD generally display greater behavioral inflexibility and more perseveration (Perandrés-Gómez et al., 2021; van Timmeren et al., 2018). Moreover, they tend to engage less in directed exploration of the environment (Wiehler et al., 2021) and learn slower in stable conditions compared to controls (Perandrés-Gómez et al., 2021). These findings suggests that different learning mechanism may be at play in these disorders, that might also be related to confidence in a different manner.

Our work indicates that patients with GD have difficulty using their confidence to inform their actions (**Chapter 8**). However, none of the discussed earlier studies have investigated the interplay between learning biases and confidence biases in GD. Recently, studies have started exploring the relationship between reinforcement learning and confidence in more detail, revealing a valence-induced confidence bias indicating higher confidence when learning to seek gains than to avoid losses, despite equal performance (Lebreton, Bacily, et al., 2019; C. C. Ting et al., 2020). A new computational model proposes that these confidence biases in reinforcement learning actually stem from learning biases (Salem-Garcia et al., 2023). In our ongoing work, not included in this thesis, we aim to explore the relationship between confidence and reinforcement learning biases in patients with GD. Based on computational work (Salem-Garcia et al., 2023), we expect to find increased confidence and a stronger value-induced confidence bias in GD patients compared to controls, which go together with more pronounced learning biases in GD patients.

Compulsivity and confidence in OCD and GD

Although our primary objective was not to directly compare confidence and metacognitive ability between OCD and GD patients, intriguing patterns worth discussing emerged. These disorders share a phenotype of compulsivity (Fineberg et al., 2016), but exhibit substantial symptomatic differences. As anticipated, patients with OCD and GD reside at opposite ends of the confidence spectrum (**Chapter 4**). This suggests that compulsivity can go together with both heightened confidence and reduced confidence, with the direction of the bias dependent on the type of compulsive behavior. A recent research line has focused on the transdiagnostic study of the

concept of compulsivity. Using psychological, cognitive and neurobiological data, Den Ouden et al. (2022) uncovered three unique profiles relating to compulsivity (albeit with a small sample of patients, warranting replication). A compulsive non-avoidant profile was characterized by low-to-mild compulsivity without obvious emotional processing problems, a compulsive reactive profile was characterized by moderate compulsivity, a stronger reward bias and a tendency to avoid negative emotions, and a compulsive stressed profile related to high compulsivity, poor ability to manage stress and maladaptive emotion regulation. In addition, the latter profile showed significantly higher anxiety and depression levels than the former two profiles. Given these findings, I would expect that the compulsive reactive type resembles GD patients and individuals with high CIT symptoms, and would thus relate to overconfidence, while the compulsive stressed type would resemble OCD patients and individuals with high AD symptoms and would relate to underconfidence. Moving beyond OCD symptoms alone and specifically decomposing transdiagnostic compulsive behavior within diverse patient groups, shows promise in cognitive neuroscience (Albertella et al., 2020; Parkes et al., 2019), and could give more insight into how compulsivity and metacognition are related.

Additionally, the findings of a stronger confidence bias in disorder-related context in GD raises the question of whether confidence biases are also magnified in a disorder-related context in OCD, such as during symptom provocation, as also discussed in our review (**Chapter 2**). In our ongoing research, not presented in this thesis, we test this hypothesis. by investigating confidence in OCD patients during symptom provocation.

Clinical implications

Abnormalities in confidence have significant functional consequences for individuals. Negative confidence biases can be detrimental for self-esteem, motivation, and learning, while positive confidence biases could relate to risky behavior, rigid beliefs and dogmatism with harmful consequences.

In recent years, metacognition has gained attention as a therapeutic target (Philipp et al., 2020). Metacognitive therapy focuses on addressing broader inflexible metacognitive beliefs and thoughts that patients may have (Wells, 2019), while metacognitive training aims to enhance awareness of disorder-specific (meta)cognitive biases to improve the ability to think about thinking (Philipp et al., 2020). For OCD specifically, the inference-based approach (IBA) therapy targets the distrust of sensory information, doubts and obsessive reasoning processes (Julien et al., 2016). In addition to these broader metacognitive interventions, our research provides evidence of a

disturbance in the general ability to construct an accurate reflection of reality across various psychiatric symptoms.

While our research is primarily experimental, we have identified some potential avenues for improving therapies, which I will discuss below. However, first, I will address a few critical points regarding the clinical implications of our work. Importantly, while our work points to interesting windows of therapeutic opportunity, the field is still in its infancy in terms of direct clinical implications. First, our studies are of experimental nature. Whilst this allows for rigor of measurement, the tasks used remain distant from the day-to-day reality that patients encounter. I believe context is a very important variable for bridging this gap, and it is important to extend our type of experimental work to situations that are more ecologically valid and tap into disorder-specific situations. Methods such as virtual reality and ecological momentary assessment could assist in these directions (Porras-Segovia et al., 2020; Shamay-Tsoory & Mendelsohn, 2019). Second, there is quite some individual variability in metacognitive ability and confidence, also within patient groups, making it challenging to provide a one-fits-all therapy focusing on confidence. Computational modeling advances and transdiagnostic therapy might be valuable assets in addressing this challenge (Katyal et al., 2023; Vujanovic et al., 2017). Relatedly, I believe that the extent to which confidence abnormalities are a central characteristic differs between disorders and symptoms. While (local) confidence seem a central aspect for OCD and GD, this might be less so for schizotypy (Rouy et al., 2021), while for depression disturbances of higher-order levels of confidence might be more central to the disorder. As we have shown that different levels of confidence relate differently to symptom profiles, this has important implications for the focus of interventions. Additionally, the awareness of deficits in confidence about cognitive abilities might differ between patient groups, where patients with depressive symptoms might be very aware of their feeling of being incompetent, while patients with gambling disorder might be less aware of their miscalibrations in confidence. Finally, it is important to acknowledge that metacognitive ability is just one variable within the complex structure of psychiatric disorders, and multivariate approaches are likely necessary to fully understand and treat these conditions.

The work in this thesis has also brought about insights regarding improving metacognitive ability in clinical populations. A potential avenue for therapy that I feel holds promise could involve the different hierarchical levels of metacognition. Recent insights have shown that positive feedback interventions can increase global confidence and improve affective self-beliefs in individuals with AD symptoms, which, in turn, may help ameliorate local confidence biases and symptom severity (Katyal et

al., 2023). Computational modeling approaches could assist by simulating expected improvements of metacognition under different feedback schemes, conditional on the symptomatology, which holds promise for personalized interventions. There are, however, other ways to break the self-perpetuating cycle of negative self-beliefs, which might also have an impact on confidence biases and learning biases (Müller-Pinzler et al., 2019). For example, CBT for OCD also addresses the tendency of patients to focus on their shortcomings or errors by deliberately cultivating global confidence and self-esteem. Patients have to reflect on successful exposure experiences, redirecting their attention towards positive achievements. Research has shown that CBT indeed improves self-esteem in OCD patients (Toledano et al., 2020), and that positive self-esteem and mastery experiences predicted decreases of Y-BOCS scores (Schwartz et al., 2017). It thus seems that improving global confidence in patients' ability to perform certain (disorder-specific) tasks also bolsters self-esteem. Studying the temporal precedence of metacognition at various levels and their interaction with symptoms is an area of active research, still in its early stages. Large longitudinal studies incorporating continuous measures of metacognitive ability and confidence at different levels, together with over-time assessment of symptoms are needed to get a better understanding of these processes and will hopefully inspire more fruitful ways to improve therapies.

Supplementing existing therapies (e.g., CBT or psycho-education) with interventions specifically centered on improving patients' confidence calibration and metacognitive ability in disorder-relevant contexts could be effective in improving the ability to construct an accurate reflection of reality. Recent studies already indicated that CBT concurrently improves transdiagnostic AD symptoms and increases confidence, with the greatest improvements in symptoms observed in individuals who experienced the largest increase in confidence (Fox et al., 2023). Possibly, including a special focus on improving confidence could be even more effective in improving symptoms and confidence biases in the long term.

This thesis also brings about new ideas for treatment of GD. First, we showed that patients with GD were more confident when taking a risk versus playing it safe (**Chapter 10**), together with anticipatory regret, as confidence decreased more strongly when turning down gambles with potential increasing gain values. Recent computational work demonstrating heightened sensitivity to the anticipation of regret in GD patients chimes with these findings (Wu et al., 2021). Regret often stems from counterfactual thinking (i.e., 'what could have been') (Zeelenberg & Pieters, 2004). In the context of gambling, Petrocelli & Crysel (2009) revealed that counterfactual thinking leads to overconfidence in future gambles, subsequently increasing the likelihood of continued

gambling with larger bets (Petrocelli & Sherman, 2010). These findings speak to the pragmatic relevance of regret sensitivity in GD. Currently, CBT for GD focuses on challenging cognitive biases and erroneous beliefs (Bodor et al., 2021; Fortune & Goodie, 2012). Regret differs from typical cognitive gambling biases, however, as it deals with emotions of the outcome rather than with the probability of the outcome. A valuable approach could be to focus more directly on faulty perceptions about (anticipatory) regret within the CBT framework by emphasizing the emotional consequences (e.g., regret) of taking risks and counterfactual thinking (Tochkov, 2008). Second, patients with GD who exhibited higher average attention toward the potential gains of a gamble were more likely to engage in gambling with lower potential gains (**Chapter 9**). On the other hand, gamblers with higher average attention toward the potential loss tended to already stop gambling with lower potential losses and engage in gambling less frequently. These findings suggest that training GD patients to actively direct their attention away from gains and toward losses during gambling could help reduce gambling behavior. Currently, attentional bias training focuses on remedying attentional biases towards gambling stimuli (Boffo et al., 2017). Future attentional therapies could go beyond this, and instead focus on directly steering the attentional process during actual gambling.

Strengths and limitations

The main strength of this thesis lies in its comprehensive exploration. We employed a wide range of cognitive tasks, including perceptual decision-making tasks, with and without an incentivized element, inferential learning tasks and gambling tasks. Moreover, we used multiple methodologies, including behavioral measures, fMRI, computational modeling, and eye-tracking. Next, we incorporated hierarchal and transdiagnostic approaches to study confidence. Finally, we included different samples, among which healthy (control) samples, general population samples, highly compulsive samples from the general population, patients with GD and patients with OCD both taking medication and medication free. Overall, this broad perspective has yielded a deeper understanding of the intricate nature of confidence and its deviations in psychiatric disorders.

However, this thesis and the research field are not without limitations, some of which I touched upon in the preceding discussion. First, as mentioned before, the tasks used are of artificial nature and it is not yet clear how metacognitive ability measured in these tasks translates to day-to-day functioning. Second, all our studies are cross-sectional, which limits our ability to establish causal relationships and to study temporal

precedence. Third, there are methodological limitations to measuring confidence. Over the past years, substantial efforts have been made to develop 'pure' metacognitive metrics and methods that offer better control over confounding factors, such as the development of the (hierarchical) meta-d' framework (Fleming, 2017). A recent comprehensive examination of metacognitive measures showed that all measures are valid, show similar levels of precision, and high split-half reliabilities when using 100 or more trials (Rahnev, 2023). However, metacognitive sensitivity and efficiency as measured using meta-d' and meta-d'/d' exhibit lower stability with fewer than a couple of hundred trials (Rouault, McWilliams, et al., 2018). Additionally, many measures of metacognitive ability are not entirely independent of performance or confidence biases (Guggenmos, 2021; Rahnev, 2023). Numerous studies have used a staircase design to equalize performance across participants to specifically isolate changes in confidence. However, research has shown that using a continuous staircase inflates measures of metacognition, and the authors instead recommend using a single difficulty level to enhance precision (Rahnev & Fleming, 2019). As a result, it is likely that studies using a continuous staircase have underestimated their effect sizes (Rahnev & Fleming, 2019). Moreover, important for studies investigating individual differences and longitudinal designs, test-retest reliability of confidence judgments has been found to be relatively poor even with large numbers of trials (Rahnev, 2023). However, these test-retest reliability findings are based on a single dataset and require replication and cautious interpretation. Contrarily, test-retest reliability was found to be stronger for metacognitive efficiency or metacognitive bias (i.e., calibration) (Ais et al., 2016). Overall, these studies emphasize the importance of using a large number of trials and awareness of the limitations that these measures hold. Fourth, relatedly, even though the replication of findings using the extensively used transdiagnostic factors (AD, CIT, SW) is meaningful, the questionnaires used for this approach do not exhaustively cover all psychiatric symptoms and should not be considered the definitive standard. For example, they do not include symptoms of GD. Although GD symptoms share similarities with the CIT dimension and also relate to increased confidence, these symptoms should be formally incorporated in a transdiagnostic framework. Lastly, while significant progress has been made in developing rigorous methodologies to investigate metacognition, this progress has come at the cost of ignoring the broader scope of metacognitive processes (Katyal & Fleming, 2023). An important step forward has been made by appreciating confidence at various hierarchical levels. As suggested by Katyal & Fleming (2023), further improvements can be made by studying the social aspect of metacognition, and broadening the field to explore metacognition in domains without an objective ground truth, such as affective states and value-based decision-making.

Future directions and ongoing research

Looking ahead, I propose several key points for future research. First, there is a need for larger clinical samples with longitudinal approaches to establish a clearer understanding of the cause-and-effect structure and temporal dynamics of metacognition and symptoms, as well as to investigate the effects of treatment on metacognitive ability. Larger clinical samples would also facilitate transdiagnostic approaches that can be compared to findings from general population samples. Smartphone-based research methodologies hold great potential in achieving these goals (Gillan & Rutledge, 2021). Second, greater emphasis should be placed on examining the impact of contextual factors on metacognitive ability, especially in the light of disorder-related context, as those are likely the type of situations that provoke the strongest symptoms. Additionally, contextual factors might be especially potent modulators of higher-order levels of confidence, since these form over longer timescales. Indeed, these two important key points align with the recent compelling call for integrating time and context into (computational) psychiatry research (Hitchcock et al., 2022).

Third, it is crucial to incorporate the broader scope of metacognition in our studies. A first step has been made by looking beyond local confidence, but most paradigms usually still focus on the metacognitive abilities of an individual in isolation, while all of us are embedded in a social context. Next to intrapersonal functions, metacognition also serves interpersonal functions simply because we communicate and share our private states of confidence (Pescetelli et al., 2016). In other words, 'metacognition is tuned for social interaction by social interaction' (Heyes et al., 2020). An intriguing path for future studies involves investigating whether intrapersonal confidence biases spill over into social communication. This idea gets support from a recent study showing that patients with schizophrenia displayed overconfidence in their advice to others when compared to a control group (Hertz et al., 2020). Moreover, the field could focus on further exploring how social information in turn impacts intrapersonal metacognition. Fourth, the development of more extensive computational models of confidence formation and biases will contribute to a better understanding of the underlying mechanisms and ways to treat them. These models could incorporate multiple levels of metacognition, contextual influences, and even biological information. Finally, it is important to acknowledge and consider the substantial variation in the manifestation, incidence, underlying mechanisms, and treatment response of psychiatric symptoms between sexes and gender (Bangasser & Cuarenta, 2021; Holingue et al., 2020; LeGates et al., 2018; Mallorquí-Bagué et al., 2021; Raines et al., 2018), and their relation to metacognition, as research has revealed sex

differences in metacognitive ability (**Chapter 4**, Ariel et al., 2018; Moses-Payne et al., 2021).

Within our ongoing research not included in this thesis, we aim to build on some of these proposals. We are currently investigating the effects of disorder-specific context by way of symptom provocation on confidence judgments in OCD and the integration of post-decision evidence into confidence, particularly in relation to changes of mind. Moreover, we are studying the interplay between confidence biases and learning biases in reversal learning and contextual reinforcement learning in GD. On a neurobiological level, we are currently delving deeper into the relationship between confidence and the dopamine system in GD using ^{18}F-DOPA PET imaging to assess dopamine synthesis capacity. With our research, I hope to have paved, and continue to pave the way for future studies further exploring the fascinating '*mind's mirror*' in health and disease.

Part V

Appendices

Supplementary materials

Appendix A

Supplement to Chapter 3

Full behavioral models

To assess whether our main behavioral results on confidence still hold in a full model, considering various other factors, we performed a model selection procedure of various linear mixed effect models. We used linear mixed-effects models (LMEM) as implemented in the lmer function from the lme4 package in R (Version 1.1-12; (Bates et al., 2015)).

We iteratively built several LMEMs (Table A1), and the final one was selected by model comparison, assessing model fit by using chi-square tests on the log-likelihood values, as well as comparison of the AIC and BIC model values. Model predictors were added whenever model fit was significantly improved.

The final model included fixed effects of incentive value (gain (1), neutral (0) or loss (-1)), evidence, accuracy (correct (1) or incorrect (0)), the interaction of accuracy end evidence, reaction time, and difficulty level (easy (1), medium (2), difficult (3)), as well as a random intercept and slope for the effect of incentive on confidence (model 9, see Table A1). Satterthwaite approximations (Schaalje et al., 2002) were used to calculate degrees of freedom and p-value estimates for the fixed effects' regression coefficients by using the 'lmerTest' package (Version 2.0-36 (Kuznetsova et al., 2017)). Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality.

Final model results revealed that the significant effect of net incentive value on confidence still holds, while considering all other factors ($\beta = 0.88 \pm 0.30$, $t_{32} = 2.94$, $P = 0.006$) (Table A2). Moreover, we found a significant effect of RT on confidence, showing that quicker choices lead to higher confidence levels ($\beta = -5.24 \pm 0.21$, $t_{4305} = -25.34$, $P < 2e\text{-}16$) (Table A2). We also replicated that the link between confidence ratings and evidence is positive for correct and negative for incorrect responses.

Table A1: Model descriptions and comparison

Model	Model notation	AIC	BIC	Model comp.	χ^2	P-value	Winning model
1	Confidence ~ Incentive + (1|Subject)	35083	34109				
2	Confidence ~ Incentive + (1+Incentive|Subject)	34077	34115	1 vs. 2	10.64	0.005	2
3	Confidence ~ Incentive + Accuracy + (1+Incentive|Subject)	33920	33964	2 vs. 3	158.78	<0.001	3
4	Confidence ~ Incentive + Accuracy + Evidence + (1+Incentive|Subject)	33817	33868	3 vs. 4	104.68	<0.001	4
5	Confidence ~ Incentive + Accuracy*Evidence + (1+Incentive|Subject)	33767	33824	4 vs. 5	51.92	<0.001	5
6	Confidence ~ Incentive + Accuracy*Evidence + Gender + (1+Incentive|Subject)	33769	33833	5 vs. 6	0.006	0.936	5
7	Confidence ~ Incentive + Accuracy*Evidence + Age + (1+Incentive|Subject)	33769	33833	5 vs. 7	0.16	0.687	5
8	Confidence ~ Incentive + Accuracy*Evidence + Difficulty + (1+Incentive|Subject)	33788	33808	5 vs. 8	33.11	<0.001	8
9	Confidence ~ Incentive + RT + Accuracy*Evidence + Difficulty + (1+Incentive|Subject)	33142	33218	8 vs. 9	598.25	<0.001	9

Shown here are the model notations of all nine models with their respective AIC and BIC values, as well as model comparisons with corresponding χ^2 and P-values, resulting in the winning model 9.

Table A2: Results of linear mixed-effects model

Full Behavioral Results *Confidence ~ Incentive + RT + Accuracy*Evidence + Difficulty + (1+Incentive|Subject)*	
Intercept (B0)	β = 76.56 ± 1.27 t_{45} = 60.36 P < 2e-16
Incentive	β = 0.88 ± 0.30 t_{32} = 2.94 P = 0.006
RT	β = -5.24 ± 0.21 t_{4305} = -25.34 P < 2e-16
Accuracy	β = 3.30 ± 0.42 t_{4290} = 7.86 P = 4.71e-15
Accuracy * Evidence	β = 2.83 ± 0.50 t_{4275} = 5.69 P = 1.38e-08
Difficulty hard	β = -2.22 ± 0.43 t_{4258} = -5.20 P = 2.07e-07
Difficulty medium	β = -1.53 ± 0.41 t_{4256} = -3.71 P = 0.0002

Shown here are the results of the full linear mixed-effects model of the winning model. β: estimated regression coefficients for fixed effects ± estimated standard error of the regression coefficients, with corresponding t- and P-values.

Explorative analyses VS

We also applied our ROI analytical strategy to the VS. Like for the VMPFC and dACC analyses we built an independent anatomical ROI of the VS from the Brainnetome Atlas (Figure A1A) (Fan et al., 2016).

We compared early certainty, incentive and confidence-related activations during both time-points in all available GLMs within the VS ROI (see Figure 4 in main text for comparable analysis in VMPFC). Thus, we extracted individual standardized regression coefficients (t-values) from the VS, corresponding to these respective activations and statistically compared them using repeated measure ANOVAs and post-hoc paired t-tests (Figure A1, Table A3). Activations for early certainty during choice moment were similar for all GLMs (ANOVA $F(4,29)= 0.43$, $P=0.787$; Figure A1B), and was only positively related to early certainty in GLM2b (but marginally positively related in all other GLMs) (GLM1: $t_{29}= 2.01$, $P = 0.0541$; GLM2a: $t_{29}= 2.01$, $P = 0.0536$; GLM2b: $t_{29}= 2.12$, $P = 0.0428$; GLM3: $t_{29}= 2.00$, $P = 0.0531$; GLM4: $t_{29}= 2.00$, $P = 0.0547$). GLM specification had an impact on the incentive activation (ANOVA, main effect of GLM; $F(3,29) = 9.28$, $P <$

0.001; Figure A1C), but not on the confidence activations (ANOVA, main effect of GLM; F(3,29) = 1.37, P = 0.2561; Figure A1D) during incentive/rating moment. In the incentive case, post-hoc t-tests showed that T-values extracted from the GLM3 that related to the EV regressor were significantly higher than from other GLMs with a different coding of incentives (GLM1 versus GLM3: t_{29}= -3.39, P = 0.002; GLM2b versus GLM3: t_{29}= -3.62, P = 0.001; GLM4 versus GLM3: t_{29} = -3.75, P<0.001), but activity related to EV and confidence or certainty during rating moment were found to be similarly strong.

Finally, we repeated the qualitative falsification exercise (see Figure 5 in the main text) for the VS ROI. We extracted the VS activations for all regressors in GLM5 using our ROI, and compared them with the theorized qualitative patterns (Figure A2, Table A4-5). At the stimulus/choice moment, we found no effect of incentive conditions on dACC baseline activity, nor on its correlation with confidence – "slope" (ANOVA baseline: P = 0.9616; ANOVA slope: P = 0.2595). At rating moment, incentive conditions had an effect on VS baseline activity (ANOVA F(2,29)= 6.40, P= 0.0031). Post-hoc testing revealed that VS baseline activity was significantly positive in all incentive conditions (Loss: t_{29} = 5. 26, P <0.001 ; Neutral: t_{29} = 3.37, P = 0.022; Gain: t_{29} = 6.17, P<0.001), but larger in gain versus loss (t_{29} = -2.20, P = 0.036) and in gain vs neutral conditions (t_{29} = -2.93, P = 0.006), but not in loss vs neutral condition (t_{29} = 1.87, P = 0.072) (see Table A4-5). Incentive conditions had a significant effect (ANOVA F(2,29)= 5.94 P = 0.005) on the slope of the correlation of VS activity with confidence, where only in the gain condition the slope was positive (t_{29} = 2.79, P = 0.009). Post-hoc testing showed that the correlation with confidence was significantly higher in gain versus loss (t_{29} = -3.16, P = 0.0036), and higher for gain versus neutral conditions (t_{29} = -2.72, P = 0.0109), whereas no difference was found for neutral versus loss condition (t_{29} = 0.41, P = 0.688). Again, similar to the results in the VMPFC and dACC, the observed pattern of VS activity was not featured in the EV model, nor in the confidence model, or any other model prediction, and thus points to a more complex picture of disruption of metacognitive signals due to motivational signals.

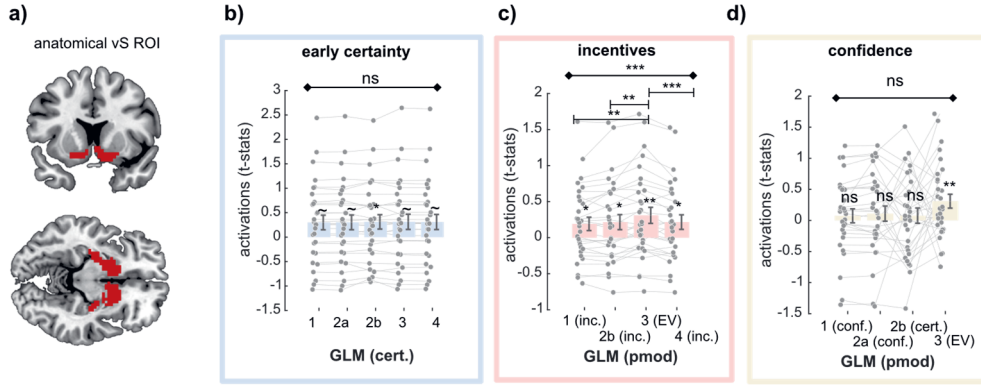


Figure A1: Activation in Ventral Striatum Across Mode. a) Anatomical VS region of interest (ROI). **b-d)** Comparison of dACC activations to different specifications of early certainty during choice moment (B), incentives during incentive/rating moment (C) and confidence during incentive/rating moment (D), as implemented in the different GLMs. Dots represent individual activations (N=30); bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different specifications. Cert: early certainty; Inc.: incentives; conf: confidence; EV: expected value; Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired t-tests. ~ P < 0.10; * P<0.05; ** P<0.01; *** P < 0.001. For repeated-measure ANOVA results: ns P>0.05, for one-sample t-tests: ns P>0.1.

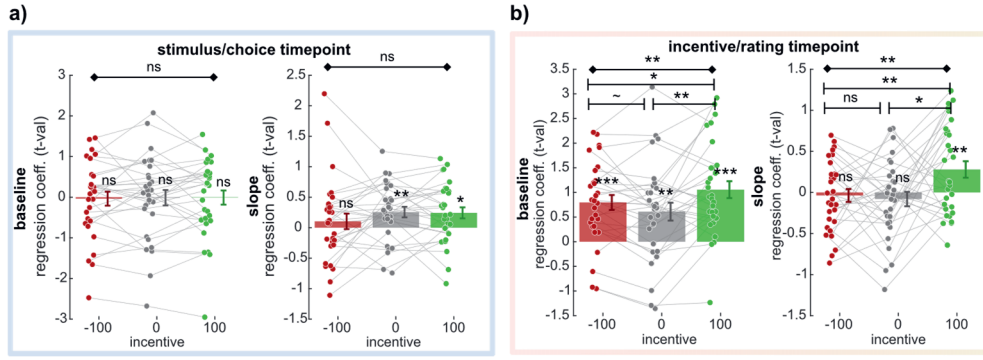


Figure A2: Activation in Ventral Striatum across Incentives and Timepoints. a-b) VS ROI analysis. T-values corresponding to baseline and regression slope were extracted in the three incentive conditions, and at the two time-points of interest (A: stimulus/choice; B: incentive/rating). Dots represent individual activations (N=30); bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different incentive conditions. Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired t-tests. ~ P < 0.10; * P<0.05; ** P<0.01; *** P < 0.001. For repeated-measure ANOVA results: ns P>0.05, for one-sample t-tests: ns P>0.1.

Table A3: Comparison of VS parametric activity (t-values) as a function of model specification (GLMs)

Early certainty	**GLM1**	**GLM2a**	**GLM2b**	**GLM3**	**GLM4**
	0.30 ± 0.15 t_{29} = 2.0075 P = 0.0541	0.30 ± 0.15 t_{29} = 2.0120 P = 0.0536	0.31 ± 0.15 t_{29} = 2.1190 P = 0.0428	0.31 ± 0.16 t_{29} = 2.0162 P = 0.0531	0.31 ± 0.15 t_{29} = 2.0025 P = 0.0547
	ANOVA (Main effect of GLM)				
	F(4,29)=0.43 P=0.7870				
Incentive		**GLM1**	**GLM2b**	**GLM3**	**GLM4**
		0.19 ± 0.09 t_{29} = 2.0684 P = 0.0476	0.22 ± 0.10 t_{29} = 2.0750 P = 0.0470	0.31 ± 0.11 t_{29} = 2.7793 P = 0.0095	0.22 ± 0.10 t_{29} = 2.1448 P = 0.0405
	ANOVA (Main effect of GLM)	**T-Test (3 vs 1)**	**T-Test (3 vs 2b)**		**T-Test (3 vs 4)**
	F(3,29)= 9.28 P=2.17206e-05	-0.12 ± 0.03 t_{29} = -3.3930 P = 0.0020	-0.09 ± 0.03 t_{29} = -3.6234 P = 0.0011		-0.09 ± 0.02 t_{29} = -3.7459 P = 7.9381e-04
Confidence		**GLM1**	**GLM2a**	**GLM2b**	**GLM3**
		0.07 ± 0.11 t_{29} = 0. 6626 P = 0. 5128	0.11 ± 0.12 t_{29} = 0. 8974 P = 0. 3769	0.08 ± 0.12 t_{29} = 0. 6265 P = 0. 5359	0.31 ± 0.11 t_{29} = 2. 7793 P = 0. 0095
	ANOVA (Main effect of GLM)				
	F(3,29) = 1.37 P = 0.2561				

The table reports descriptive and inferential statistics on VS ROI parametric activations with three different variables of interest: early certainty effects at choice moment, incentive effects at rating moment and confidence effects at rating moment (see Figure A5). Per effect of interest, results of one-sample t-tests against zero, repeated-measure (RM) ANOVAs on the main effect of GLMS, and post-hoc t-test results are shown.

Table A4: Comparison of VS activity at choice moment (t-values), as a function of incentive condition

Choice/Stim	**baseline**	**Inc. -100**	**Inc. 0**	**Inc. +100**	**ANOVA**
		-0.04 ± 0.17 t_{29} = -0.2118 P = 0. 8337	-0.01 ± 0.19 t_{29} = -0.0692 P = 0. 9453	-0.01 ± 0.17 t_{29} = -0.0481 P = 0. 9620	F(2,29) = 0.04 P = 0.9616
	slope	**Inc. -100**	**Inc 0**	**Inc. +100**	**ANOVA**
		0.10 ± 0.13 t_{29} = 0.8188 P = 0.4196	0.26 ± 0.09 t_{29} = 2.9434 P = 0.0063	0.24 ± 0.09 t_{29} = 2.6902 P = 0.0117	F(2,29) = 1.38 P = 0.2595

The table reports descriptive and inferential statistics on VS ROI parametric activations in our three incentive conditions during choice moment, for both baseline activity as well as the correlation with early certainty (i.e., slope) (see Figure A6). Results of RM ANOVAs and one-sample t-tests against 0 are shown.

Table A5: Comparison of VS activity at rating moment (t-values), as a function of incentive condition

Incentive/rating	**baseline**	**Inc -100**	**Inc 0**	**Inc +100**	**ANOVA**
		0.80 ± 0.15 t_{29} = 5.2603 P = 1.2305e-05	0.61 ± 0.18 t_{29} = 3.3655 P = 0.0022	1.06 ± 0.17 t_{29} = 6.1747 P = 9.8752e-07	F(2,29) = 6.40 P = 0.0031
		T-Test [-100 vs 0]	**T-Test** [0 vs 100]	**T-Test** [-100 vs 100]	
		0.19 ± 0.10 t_{29} = 1.8707 P = 0.0715	-0.45 ± 0.15 t_{29} = -2.9268 P = 0.0066	-0.26 ± 0.12 t_{29} = -2.1995 P = 0.0360	
	slope	**Inc -100**	**Inc 0**	**Inc +100**	**ANOVA**
		-0.04 ± 0.08 t_{29} = -0.4695 P = 0.6422	-0.08 ± 0.09 t_{29} = -0.9138 P = 0.3684	0.28 ± 0.10 t_{29} = 2.7922 P = 0.0092	F(2,29) = 5.94 P = 0.0045
		T-Test [-100 vs 0]	**T-Test** [0 vs 100]	**T-Test** [-100 vs 100]	
		0.04 ± 0.11 t_{29} = 0.41 P = 0.6877	-0.36 ± 0.13 t_{29} = -2.7197 P = 0.0109	-0.32 ± 0.10 t_{29} = -3.1642 P = 0.0036	

The table reports descriptive and inferential statistics on VS ROI parametric activations in our three incentive conditions during rating moment, for both baseline activity as well as the correlation with confidence (i.e., slope) (see Figure A6). Results of one-sample t-tests against 0, RM ANOVAs and post-hoc t-tests are shown.

Explorative analyses dACC

Explorative analysis of dACC results show overlap between confidence and EV signal

While we did not find clear evidence for VMPFC activity correlating with confidence at our pre-specified statistical threshold, we did find a cluster of dACC activity positively correlating with both confidence (Figure 2A main text) and EV (Figure 2B main text). We therefore applied our ROI analytical strategy – originally designed for the VMPFC – to the dACC. Like for the VMPFC analyses we built an independent anatomical ROI of the dACC from the Brainnetome Atlas (Fan et al., 2016) (Figure A2A).

We compared early certainty, incentive and confidence-related activations during both time-points in all available GLMs within the dACC ROI (see Figure 4 in main text for comparable analysis in VMPFC). Thus, we extracted individual standardized regression coefficients (t-values) from the dACC, corresponding to these respective activations and statistically compared them using repeated measure ANOVAs and post-hoc paired t-tests (Figure A2, Table A6). Activations for early certainty during choice moment were similar for all GLMs (ANOVA $F(4,29)= 1.75$, $P=0.149$; Figure A2B), and all were significantly negatively related to early certainty (GLM1: $t_{29}= -2.48$, $P = 0.019$; GLM2a: $t_{29}= -2.48$, $P = 0.019$; GLM2b: $t_{29}= -2.39$, $P = 0.024$; GLM3: $t_{29}= -2.48$, $P = 0.019$; GLM4: $t_{29}= -2.51$, $P = 0.018$). GLM specification had an impact on the incentive activation (ANOVA, main effect of GLM; $F(3,29) = 19.13$, $P < 0.001$; Figure A3C), but not on the confidence activations (ANOVA, main effect of GLM; $F(3,29) = 1.95$, $P = 0.127$; Figure A3D) during incentive/rating moment. In the incentive case, post-hoc t-tests showed that T-values extracted from the GLM3 that related to the EV regressor were significantly higher than from other GLMs with a different coding of incentives (GLM1 versus GLM3: $t_{29}= -5.22$, $P < 0.001$; GLM2b versus GLM3: $t_{29}= -4.45$, $P<0.001$; GLM4 versus GLM3: $t_{29} = -4.31$, $P<0.001$), but activity related to EV and confidence or certainty during rating moment were found to be similarly strong.

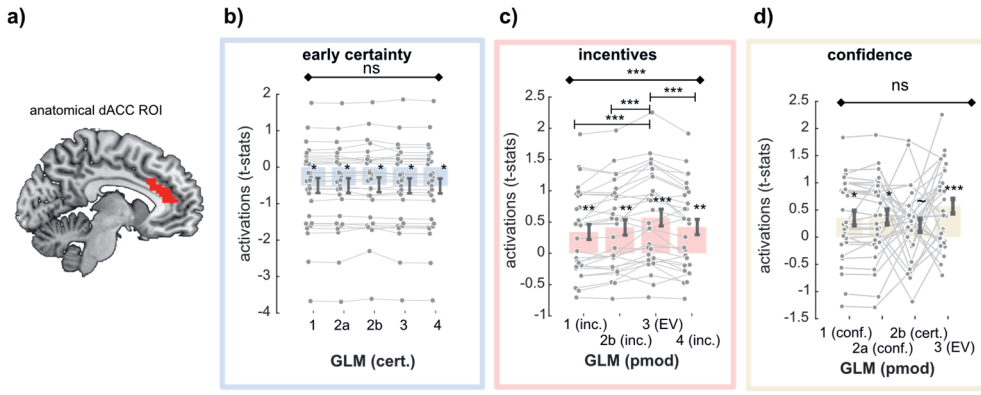


Figure A3: Activation in Dorsal Anterior Cingulate Cortex Across Models. a) Anatomical dACC region of interest (ROI). **b-d)** Comparison of dACC activations to different specifications of early certainty during choice moment (B), incentives during incentive/rating moment (C) and confidence during incentive/rating moment (D), as implemented in the different GLMs. Dots represent individual activations (N=30); bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different specifications. Cert: early certainty; Inc.: incentives; conf: confidence; EV: expected value; Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired t-tests. ~ P < 0.10; * P<0.05; ** P<0.01; *** P < 0.001. For repeated-measure ANOVA results: ns P>0.05, for one-sample t-tests: ns P>0.1.

Finally, we repeated the qualitative falsification exercise (see Figure 5 in the main text) for the dACC ROI. We extracted the dACC activations for all regressors in GLM5 using our ROI, and compared them with the theorized qualitative patterns (Figure A3, Table A7-8). At the stimulus/choice moment, we found no effect of incentive conditions on dACC baseline activity, nor on its correlation with confidence – "slope" (ANOVA baseline: P = 0.952; ANOVA slope: P = 0.534). At rating moment, incentive conditions had an effect on dACC baseline activity (ANOVA F(2,29)= 12.30, P<0.001). Post-hoc testing revealed that dACC baseline activity was significantly positive in all incentive conditions (Loss: t_{29} = 3.96, P <0.001 ; Neutral: t_{29} = 2.69, P = 0.011; Gain: t_{29} = 6.31, P<0.001), but larger in gain versus loss (t_{29} = -3.63, P = 0.001) and in gain vs neutral conditions (t_{29} = -4.10, P < 0.001), but not in loss vs neutral condition (t_{29} = 1.71, P = 0.098) (see Table A7-8). Incentive conditions had a marginally significant effect (ANOVA F(2,29)= 3.12 P = 0.052) on the slope of the correlation of dACC activity with confidence, where only in the gain condition the slope was positive (t_{29} = 3.35, P = 0.002). Post-hoc testing showed that the correlation with confidence was only significantly higher in gain versus loss (t_{29} = -2.37, P = 0.025), and marginally higher for gain versus neutral conditions (t_{29} = -1.95, P = 0.060), whereas no difference was found for neutral versus

loss condition (t_{29} = -0.18, P = 0.860). Again, similar to the results in the VMPFC, the observed pattern of dACC activity was not featured in the EV model, nor in the confidence model, or any other model prediction, and thus points to a more complex picture of disruption of metacognitive signals due to motivational signals.

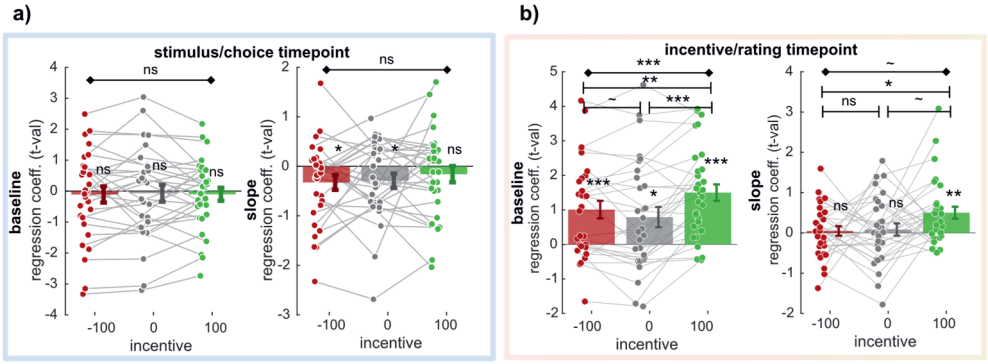


Figure A4: Activation in Dorsal Anterior Cingulate Cortex across Incentives and Timepoints. **a-b)** dACC ROI analysis. T-values corresponding to baseline and regression slope were extracted in the three incentive conditions, and at the two time-points of interest (A: stimulus/choice; B: incentive/rating). Dots represent individual activations (N=30); bar and error bars indicate sample mean ± standard error of the mean. Grey lines highlight within subject variation across the different incentive conditions. Diamond-ended horizontal bars indicate the results of repeated-measure ANOVAs. Dash-ended horizontal bars indicate the result of post-hoc paired t-tests. ~ P < 0.10; * P<0.05; ** P<0.01; *** P < 0.001. For repeated-measure ANOVA results: ns P>0.05, for one-sample or two sample t-tests: ns P>0.1.

Table A6: Comparison of ACC parametric activity (t-values) as a function of model specification (GLMs)

Early certainty	**GLM1**	**GLM2a**	**GLM2b**	**GLM3**	**GLM4**
	-0.5 ± 0.2 t_{29} = -2.48 P = 0.019	-0.51 ± 0.2 t_{29} = -2.48 P = 0.019	-0.48 ± 0.2 t_{29} = -2.39 P = 0.024	-0.51 ± 0.21 t_{29} = -2.48 P = 0.019	-0.52 ± 0.21 t_{29} = -2.51 P = 0.018
	ANOVA (Main effect of GLM)				
	F(4,29)=1.75 P=0.1439				
Incentive		**GLM1**	**GLM2b**	**GLM3**	**GLM4**
		0.34 ± 0.12 t_{29} = 2. 82 P = 0. 0085	0.41 ± 0.12 t_{29} = 3. 34 P = 0. 0023	0.57 ± 0.13 t_{29} = 4. 25 P = 0. 0002	0.42 ± 0.12 t_{29} = 3. 43 P = 0. 0018
	ANOVA (Main effect of GLM)	**T-Test (3 vs 1)**	**T-Test (3 vs 2b)**		**T-Test (3 vs 4)**
	F(3,29)=19.13 P= 1.0e-08	-0.23 ± 0.09 t_{29} = -5.22 P = 1.378e-05	-0.16 ± 0.07 t_{29} = -4.45 P = 1.16e-04		-0.15 ± 0.07 t_{29} = -4.3 P = 1.71e-04
Confidence		**GLM1**	**GLM2a**	**GLM2b**	**GLM3**
		0. 35 ± 0.13 t_{29} = 2. 65 P = 0. 0128	0.37 ± 0.14 t_{29} = 2. 75 P = 0. 0102	0.22 ± 0.12 t_{29} = 1.82 P = 0. 0795	0.57 ± 0.13 t_{29} = 4. 25 P = 0. 0002
	ANOVA (Main effect of GLM)	**T-Test (3 vs 1)**	**T-Test (3 vs 2a)**	**T-Test (3 vs 2b)**	
	F(3,29) = 1.95 P = 0.1272	-0.22 ± 0.36 t_{29} = -1. 24 P = 0. 2257	-0.20 ± 0.35 t_{29} = -1.15 P =0.2583	-0.35 ± 0.33 t_{29} = -2.20 P = 0. 036	

The table reports descriptive and inferential statistics on ACC ROI parametric activations with three different variables of interest: early certainty effects at choice moment, incentive effects at rating moment and confidence effects at rating moment (see Figure A3). Per effect of interest, results of one-sample t-tests against zero, repeated-measure (RM) ANOVAs on the main effect of GLMS, and post-hoc t-test results are shown.

Table A7: Comparison of ACC activity at choice moment (t-values), as a function of incentive condition

Choice/Stim	**baseline**	**Inc. -100**	**Inc. 0**	**Inc. +100**	**ANOVA**
		-0.11 ± 0.26 t_{29} = -0.43 P = 0.67	-0.07 ± 0.27 t_{29} = -0.26 P = 0.80	-0.10 ± 0.21 t_{29} = -0.49 P = 0.63	F(2,28) = 0.05 P = 0.95
	slope	**Inc. -100**	**Inc 0**	**Inc. +100**	**ANOVA**
		-0.33 ± 0.15 t_{29} = -2.17 P = 0.04	-0.29 ± 0.15 t_{29} = -1.98 P = 0.06	-0.16 ± 0.16 t_{29} = -0.98 P = 0.34	F(2,28) = 0.63 P = 0.53

The table reports descriptive and inferential statistics on ACC ROI parametric activations in our three incentive conditions during choice moment, for both baseline activity as well as the correlation with early certainty (i.e., slope) (see Figure A4). Results of RM ANOVAs and one-sample t-tests against 0 are shown.

Table A8: Comparison of ACC activity at rating moment (t-values), as a function of incentive condition

Incentive/rating	**baseline**	**Inc -100**	**Inc 0**	**Inc +100**	**ANOVA**
		1.01 ± 0.25 t_{29} = 3.96 P = 0.0004	0.79 ± 0.29 t_{29} = 2.69 P = 0.0117	1.50 ± 0.24 t_{29} = 6.31 P = 6.83×10^{-7}	F(2,28) = 12.30 P = 3.52×10^{-5}
		T-Test [-100 vs 0]	**T-Test** [0 vs 100]	**T-Test** [-100 vs 100]	
		0.22 ± 0.13 t_{29} = 1.71 P = 0.0984	-0.71 ± 0.17 t_{29} = -4.10 P = 3.01×10^{-4}	-0.49 ± 0.14 t_{29} = -3.63 P = 0.0011	
	slope	**Inc -100**	**Inc 0**	**Inc +100**	**ANOVA**
		0.05 ± 0.12 t_{29} = 0.41 P = 0.68	0.08 ± 0.15 t_{29} = 0.22 P = 0.58	0.50 ± 0.15 t_{29} = 3.35 P = 0.0022	F(2,28) = 3.12 P = 0.0517
		T-Test [-100 vs 0]	**T-Test** [0 vs 100]	**T-Test** [-100 vs 100]	
		-0.04 ± 0.20 t_{29} = -0.18 P = 0.86	-0.42 ± 0.21 t_{29} = -1.95 P = 0.06	-0.45 ± 0.19 t_{29} = -2.37 P = 0.0246	

The table reports descriptive and inferential statistics on ACC ROI parametric activations in our three incentive conditions during rating moment, for both baseline activity as well as the correlation with confidence (i.e., slope) (see Figure A4). Results of one-sample t-tests against 0, RM ANOVAs and post-hoc t-tests are shown.

Additional behavioral analyses: properties of confidence judgments

Similarly to Lebreton et al. (2018), we performed additional behavioral analyses to confirm three main properties of confidence judgements, as theorized in a recent paper by Sanders and colleagues (Sanders et al., 2016). There, the authors outlined three main properties of confidence judgments, which should be observed if participants compute the probability of a choice being correct given some level of noisy evidence: (1) confidence ratings correlate with the probability of being correct; (2) the link between confidence ratings and evidence is positive for correct and negative for incorrect responses; (3) the link between evidence and performance differs between high and low confidence trials.

To assess the first property, we sorted trials according to the confidence ratings at the individual level. Then, we averaged trials over 8 bins per participant, and computed the frequency of correct choices in each bin. Finally, the correlation between the bins' confidence and performance was computed at the individual level. These measures were positively correlated ($R = 0.59 \pm 0.05$; Figure A4A).

To assess the second property, the following linear regression was estimated at the individual level, using all trials from the confidence elicitation task (Model 1):

(1) $Conf = \beta0 + \beta1 \times Correct \times Evidence + \beta2 \times Incorrect \times Evidence$,

where Incorrect is a dummy variable coding for incorrect answers, and Correct is a dummy variable coding for correct answers. Then, we tested the parameters of this model at the population level using one-sample t-tests. The results (Figure A4B), summarized in the table below (Table A9), demonstrate that confidence judgments are indeed positively associated with evidence for correct trials, and negatively for incorrect trials.

To assess the third property, we proceeded similarly to the second: the following logistic regression was estimated at the individual level, using all trials (Model 2):

(2) $Correct = \beta0 + \beta1 \times High \times Evidence + \beta2 \times Low \times Evidence$,

where High is a dummy variable coding for high confidence trials (i.e., confidence > median(confidence)), and Low is a dummy variable coding for low confidence trials (i.e., confidence ≤ median(confidence)). Then, the parameters of this model were tested at the population level, using one-sample t-tests. The results (Figure A4C), summarized in the table below (Table A9), indeed demonstrate that the curve has a steeper slope in the high than in the low confidence trials, as was expected.

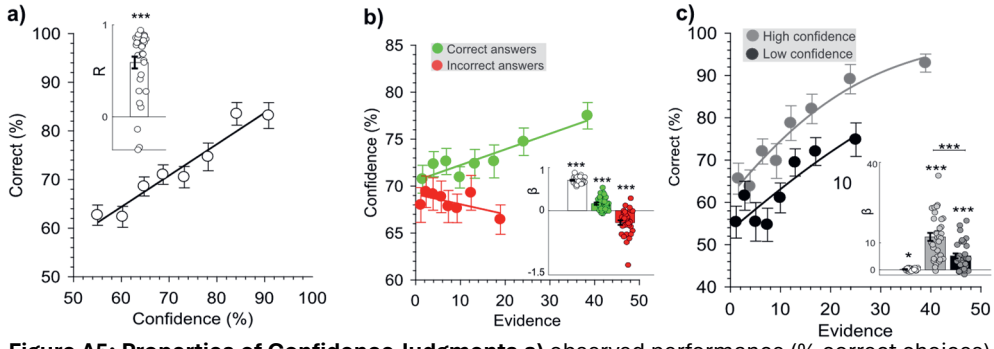


Figure A5: Properties of Confidence Judgments a) observed performance (% correct choices) as a function of reported confidence. **b)** reported confidence as function of evidence for correct (green) and incorrect (red) choices. **c)** observed performance (% correct choices) as a function of evidence, for high (gray) and low (black) confidence trials. The insets presented on the side of each graph depict the results of the population-level analyses on the correlation coefficients (a) or on the regression coefficients (b and c). Error bars indicate inter-subject standard errors of the mean. N = 32. *: P<.05; **: P<.01; ***P<.001

Table A9: Results of linear mixed-effects models for properties of confidence judgments

Model 1 (Figure A5B)	
Intercept (β_0)	$\beta = 0.71 \pm 0.01$ $t_{31} = 53.06$ P = 5.3528e-32
Confidence/Evidence Correct Answers (β_1)	$\beta = 0.16 \pm 0.03$ $t_{31} = 5.86$ P = 1.8326e-06
Confidence/Evidence Incorrect Answers (β_2)	$\beta = -0.28 \pm 0.05$ $t_{31} = -5.36$ P = 7.7032e-06
Model 2 (Figure A5C)	
Intercept (β_0)	$\beta = 0.14 \pm 0.07$ $t_{31} = 2.04$ P = 0.0495
Performance/Evidence High confidence (β_1)	$\beta = 12.23 \pm 1.49$ $t_{31} = 8.21$ P = 2.8097e-09
Performance/Evidence Low confidence (β_2)	$\beta = 5.14 \pm 0.93$ $t_{31} = 5.50$ P = 5.1270e-06
Difference (β_1 - β_2)	$t_{31} = 5.45$ P = 5.8544e-06

Early certainty

In this section, we provide further details about the computation and properties of the early certainty variable. To verify that our model of early certainty is an appropriate proxy of confidence judgments, we performed similar behavioral analyses to confirm the three main properties of confidence judgments still hold for our early certainty variable. We performed identical analyses, substituting subjective confidence judgments for early certainty values.

Our results show that the measures of early certainty and performance are highly correlated (R = 0.67 ± 0.07; Figure A5A, Table A10). Early certainty is also positively associated with evidence for correct trials, and negatively for incorrect trials (Figure A5B, Table A10). Finally, the relationship between performance and evidence is indeed higher in trials with high early certainty versus low early certainty (Figure A5C, Table A10).

When inspecting the beta values for the second model (Figure A5C, Table A10), we observed three statistical outliers (i.e., >1.5 times the interquartile range away from the 75th percentile) in the effect of evidence on performance in trials with high early certainty (β_1). These outliers were caused by the median-split of the early certainty trials into high and low variants, as these subjects performed (almost) perfectly in the high early certainty trials, causing the betas to inflate. Importantly, when excluding these subjects from the analyses, we found identical results, albeit stronger (β_0 = 0.09 ± 0.07, t_{28} = 1.26, P = 0.217; β_1 = 17.99 ± 2.40, t_{28} = 7.49, P < 0.001; β_2 = 3.83 ± 0.94, t_{28} = 4.06, P < 0.001; Difference (β_1 - β_2): t_{28} = 5.87, P<0.001).

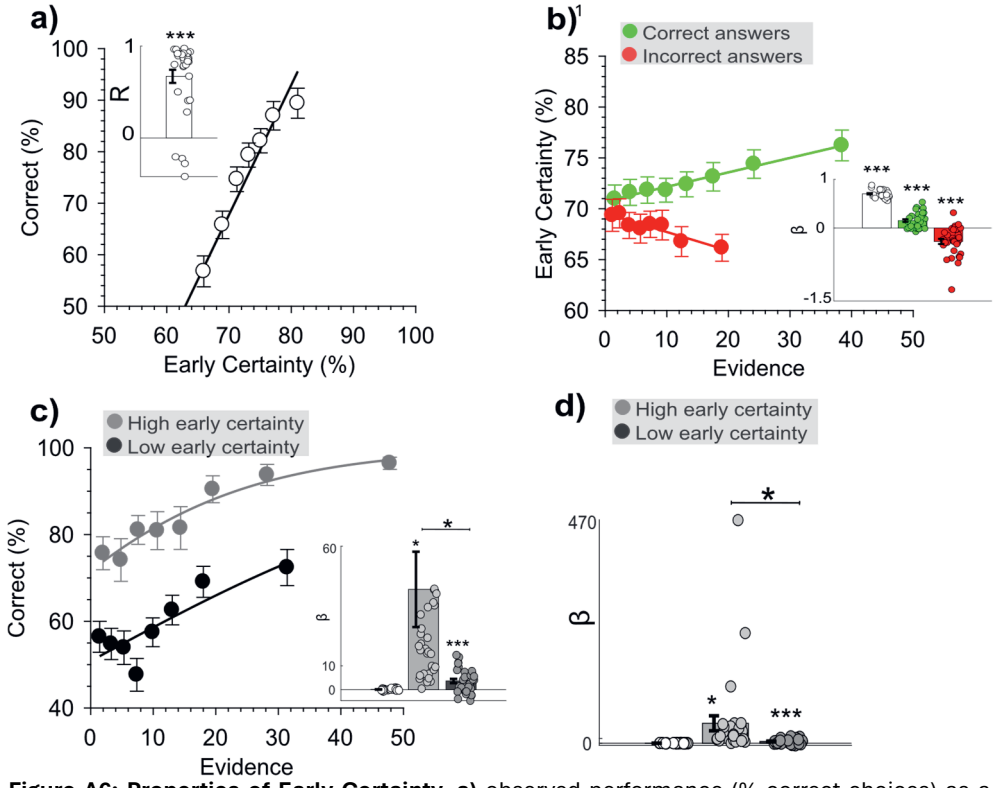


Figure A6: Properties of Early Certainty. a) observed performance (% correct choices) as a function of early certainty. **b)** early certainty as function of evidence for correct (green) and incorrect (red) choices. **c)** observed performance (% correct choices) as a function of evidence, for high (gray) and low (black) early certainty trials. The insets presented on the side of each graph depict the results of the population-level analyses on the correlation coefficients (a) or on the regression coefficients (b and c), where dots represent individual correlation coefficients (a), or regression coefficients (b and c) (N=32); bar and error bars indicate sample mean ± inter-subject standard error of the mean. Main plots and insets in plot a and b include the three statistical outliers. **d)**: Shown here are the three statistical outliers for the individual regression coefficients for the high early certainty trials. For visibility we excluded those three outliers in the inset in plot c. *: P<.05; **: P<.01; ***P<.001

Table A10: Results of linear mixed-effects models for properties of early certainty

Model 1 (Figure A6B)	
Intercept (β_0)	$\beta = 0.70 \pm .01$ $t_{31} = 53.91$ P = 3.3040e-32
Confidence/Evidence Correct Answers (β_1)	$\beta = 0.15 \pm .03$ $t_{31} = 5.45$ P = 5.9907e-06
Confidence/Evidence Incorrect Answers (β_2)	$\beta = -0.27 \pm .05$ $t_{31} = -5.16$ P = 1.3702e-05
Model 2 (Figure A6C)	
Intercept (β_0)	$\beta = 0.11 \pm 0.07$ $t_{31} = 1.51$ P = 0.14
Performance/Evidence High confidence (β_1)	$\beta = 41.95 \pm 15.75$ $t_{31} = 2.67$ P = 0.0122
Performance/Evidence Low confidence (β_2)	$\beta = 3.64 \pm 0.90$ $t_{31} = 4.06$ P = 0.0003
Difference ($\beta_1 - \beta_2$)	$t_{31} = 2.40$ P = 0.0226

Moreover, to validate that our model of early certainty correlates highly with subjective confidence and choice and stimulus features, but does not show a statistical relationship with incentives, we built a linear mixed-effects model using the lme4 package in R. We used early certainty as dependent variable and added RT, accuracy, evidence and the interaction between evidence and accuracy as predictors. Indeed, the results showed that RT, accuracy and the accuracy * evidence interaction all significantly contributed to early certainty, while no effect of incentive value on early certainty was found (Table A11).

Table A11: Results of general linear mixed-effects model

Early Certainty GLMER Results *Early Certainty ~ Incentive + RT + Accuracy*Evidence + (1|Subject)*	
Intercept (B0)	β = 75.52 ± 1.15 t_{33} = 65.63 P=<2e-16
Incentive	β = 0.07 ± 0.08 t_{4288} = 0.84 P = 0.404
RT	β = -5.69 ± 0.08 t_{4292} = -71.90 P < 2e-16
Accuracy	β = 3.61 ± 0.16 t_{4288} = 22.75 P = <2e-16
Accuracy * Evidence	β = 2.50 ± 0.19 t_{4288} = 13.16 P = <2e-16

Shown here are the results of the full linear mixed-effects model. β: estimated regression coefficients for fixed effects ± estimated standard error of the regression coefficients, with corresponding t- and P-values.

Table A12: Whole-brain activation GLM1 and GLM3

GLM1								
Effect	**Brain Region**	**k**	**Peak z-score**	**P (cluster FWE corrected)**	**Peak voxel MNI coordinates**			
Early certainty +	VMPFC	95	3.97	0.004	3	29	-7	LR
					-9	65	-4	LR
					-3	38	-7	LR
	PPC	59	3.79	0.03	-3	-43	32	LR
					6	-52	32	LR
					-3	-58	29	LR
Early certainty -	Insula Inferior frontal gyrus	873	5.92	<0.001	33	20	8	R
	RLPFC / DLPFC				45	14	2	R
	Putamen				45	38	-4	R
	DLPFC	176	4.55	<0.001	-42	26	35	L
	RLPFC				-30	47	11	L
					-36	29	26	L
	Supplementary motor area	583	5.51	<0.001	3	11	47	LR
	Mid cingulate cortex				6	23	35	LR
	Anterior cingulate cortex				9	17	41	LR
	Supramarginal gyrus	76	4.45	0.011	51	-19	20	R
					42	-22	23	R
	Posterior Insula				36	-13	20	R
	Inferior Occipital Gyrus	57	4.33	0.034	-48	-70	-1	L
					-42	-61	-7	L
	Inferior parietal lobe	332	4.99	<0.001	-48	-34	41	L
					-30	-52	41	L
	Postcentral gyrus				-57	-22	35	L
	Anterior insula	170	4.95	<0.001	-33	17	5	L
					-33	26	-7	L
	Cerebellum	121	4.98	0.001	-36	-55	-31	L
					-15	-43	-25	L
Incentive +	VMPFC	46	4.32	0.011	3	47	-4	LR
	Anterior medial prefrontal cortex / DMPFC	36	4.14	0.032	0	56	11	LR
	DLPFC	39	4.01	0.023	-27	38	55	L
					-6	35	32	L
Incentive -	Angular gyrus	54	4.06	0.005	39	-55	41	R

					39	-58	50	R
	Superior occipital lobe				27	-61	38	R
Confidence +	Cerebellum	997	6.19	<0.001	12	-52	-16	R
	Lingual gyrus (visual cortex)				18	-70	-13	R
					21	-70	-4	R
	Putamen	328	4.83	<0.001	-33	-10	-1	L
					-30	2	5	L
					-42	-8	11	L
	Primary motor cortex	244	4.88	<0.001	-33	-28	59	L
					-42	-22	41	L
					-42	-19	56	L
	Anterior cingulate cortex	90	4.55	0.001	-6	23	29	L
					3	20	26	R
	Mid cingulate cortex				-6	2	38	LR
	Parahippocampal gyrus	64	4.01	0.008	-24	-37	-13	L
	Fusiform gyrus				-24	-46	-10	L
	Middle temporal gyrus	56	3.97	0.016	48	-70	2	R
					48	-52	17	R
					51	-64	17	R
	Precuneus	75	4.29	0.004	-15	-49	8	L
					-6	-58	23	L
					-9	-58	14	L
Confidence -	Lingual gyrus (visual cortex)	302	5.82	<0.001	-15	-82	-4	L
	Cerebellum				-15	-55	-16	L
	Primary motor cortex	55	4.26	0.018	39	-19	59	R
					36	-19	44	R
					57	-16	44	R

GLM 3								
Effect	**Brain Region**	**k**	**Peak z-score**	**P (cluster FWE corrected)**	**Peak voxel MNI coordinates**			
Expected Value +	VMPFC	336	4.93	<0.001	0	47	-4	LR
	Anterior medial prefrontal cortex / DMPFC				0	56	11	LR
	Anterior cingulate cortex (dorsal + ventral)				0	32	11	LR
	Insula	37	4	0.038	-36	5	1	L
					-36	-1	5	L

Brain activations (whole brain analyses) of GLM1 and GLM3, showing activity related to early certainty at choice moment, as well as activity related to incentive and confidence at incentive/rating moment, and expected value at incentive/rating moment. All whole-brain activation maps were thresholded using family-wise error correction for multiple correction (FWE) at cluster level (P FWE_clu < 0.05), with a voxel cluster-defining threshold of P<0.001 uncorrected. Activity that positively correlates to given variable is denoted by '+', whereas negative correlations are denoted by '-'.

Table A13: GLM1 activation with exclusive motor mask

GLM1								
Effect	**Brain Region**	**k**	**Peak z-score**	**P (cluster FWE corrected)**	**Peak voxel MNI coordinates**			
Early certainty +	VMPFC	95	3.97	0.004	3	29	-7	LR
					-9	65	-4	LR
					-3	38	-7	LR
	PPC	59	3.79	0.03	-3	-43	32	LR
					6	-52	32	LR
					-3	-58	29	LR
Early certainty -	Insula Inferior frontal gyrus	718	5.92	<0.001	33	20	8	R
	RLPFC / DLPFC				45	14	-1	R
	Putamen				45	38	-4	R
	DLPFC	174	4.55	<0.001	-42	26	35	L
	RLPFC				-30	47	11	L
					-36	29	26	L
	Mid cingulate cortex	334	5.51	<0.001	6	23	35	LR
	Anterior cingulate cortex				9	17	41	LR
					-6	17	41	LR
	Supramarginal gyrus	62	4.44	0.025	42	-22	23	R
					-30	47	11	R
	Posterior Insula				36	29	26	R
	Superior parietal lobe	87	4.42	0.006	-30	-52	41	L
					-30	-40	38	L
	Inferior parietal lobe				-39	-40	38	L
	Angular gyrus				54	-46	35	R
					33	-43	41	R
					45	-43	44	R
	Anterior insula	167	4.95	<0.001	-33	17	5	L
					-33	26	-7	L
Incentive +	VMPFC	46	4.32	0.011	3	47	-4	LR
					-3	53	-7	LR

	Anterior medial prefrontal cortex / DMPFC	36	4.14	0.032	0	56	11	LR
	DLPFC	39	4.01	0.023	-27	38	55	L
					-6	35	32	L
Incentive -	Angular gyrus	54	4.06	0.005	39	-55	41	R
					39	-58	50	R
	Superior occipital lobe				27	-61	38	R
Confidence +	Lingual gyrus (visual cortex)	689	6.07	<0.001	18	-70	-13	R
					21	-70	-4	R
					24	-61	-10	R
	Putamen	91	4.83	0.001	-33	-10	-1	L
					-30	-22	5	L
					-30	-22	14	L
	Anterior cingulate cortex	62	4.55	0.01	-6	23	29	L
					3	20	26	R
	Parahippocampal gyrus	64	4.01	0.008	-24	-37	-13	L
	Fusiform gyrus				-24	-46	-10	L
	Middle temporal gyrus	56	3.97	0.016	48	-52	17	R
					48	-70	2	R
					51	-64	17	R
	Precuneus	75	4.29	0.004	-15	-49	8	L
					-6	-58	23	L
					-9	-58	14	L
Confidence -	Lingual gyrus (visual cortex)	217	5.82	<0.001	-15	-82	-4	L
					-15	-55	-16	L

Brain activations of GLM1 activation table with exclusive motor mask showing activity related to early certainty at choice moment, as well as activity related to incentive and confidence at incentive/rating moment, exclusively masked for motor-related activity patterns using a Neurosynth mask. All whole-brain activation maps were thresholded using family-wise error correction for multiple correction (FWE) at cluster level (P FWE_clu < 0.05), with a voxel cluster-defining threshold of P<0.001 uncorrected. Activity that positively correlates to given variable is denoted by '+', whereas negative correlations are denoted by '-'.

Appendix B

Supplement to Chapter 4

Supplemental Methods

Participants

All subjects underwent screening with the MINI structured psychiatric interview to confirm the absence of any other psychiatric disorder (Sheehan et al., 1998). OCD symptom severity was measured using the Yale-Brown Obsessive Compulsive Scale (YBOCS) (Goodman et al., 1989), and GD symptom severity was measured using the Problem Gambling Severity Index (PGSI) (Ferris & Wynne, 2001). Anxiety symptoms were assessed using the Hamilton Anxiety Rating Scale (HAMA) (Hamilton, 1959) and depression symptoms using the Hamilton Rating Scale for Depression (HDRS) (Hamilton, 1960). We analyzed whether age, sex, IQ, Y-BOCS, PGSI, HAMA and HDRS score differed between the three groups using ANOVAs for all variables but sex, which was assessed using a Chi-square test. When appropriate, two-sample t-tests were executed post-hoc.

Exclusion Criteria

The exclusion criteria included having a diagnosis of major depressive disorder, bipolar disorder, psychotic disorders, substance-use disorders, using tricyclic antidepressants or antipsychotics, having any contraindications for MRI, and having a history of or current treatment for neurological disorders, major physical disorders or brain trauma.

Moreover, session-level behavioral and fMRI data were excluded when task accuracy was below 50%, which would signal below chance level performance and would hinder our interpretation of the underlying cognitive processes. Also, when subjects did not show sufficient variation in their confidence reports (standard deviation of <5 confidence points), which indicates careless responding, those data were excluded. Session-level fMRI data was additionally excluded when participants displayed more than 3.5 mm head movement in any direction. Overall, for the behavioral analyses, this led to the full exclusion of four GD patients and one OCD patient, as well as one out of two session exclusions for four GD patients, two OCD patients and two HCs. For the fMRI analyses, three additional GD patients, one OCD patient and two HCs were fully

excluded, as well as one out of two session exclusions for three additional GD patients and one OCD patient.

Experimental Procedure

After demographic and clinical interviews, all participants performed an initial calibration session (consisting of 144 trials) to tailor the difficulty level of the task to each individual. This was done to keep average performance similar across individuals. Following, all subjects performed two fMRI sessions, each consisting of 72 trials (24 per incentive condition), presented in a random order. After the fMRI task, six random trials were drawn (i.e., two of each incentive condition) on which the final payment was based, and the total amount of points was converted to money.

Behavioral Analyses: Properties of Confidence Judgments

Similarly to Lebreton et al. (2018), we performed additional behavioral analyses to confirm three main properties of confidence judgements, as theorized in a recent paper by Sanders and colleagues (Sanders et al., 2016). There, the authors outlined three main properties of confidence judgments, which should be observed if participants compute the probability of a choice being correct given some level of noisy evidence: (1) confidence ratings correlate with the probability of being correct; (2) the link between confidence ratings and evidence is positive for correct and negative for incorrect responses; (3) the link between evidence and performance differs between high and low confidence trials.

- To assess the first property, we sorted trials according to the confidence ratings at the individual level. Then, we averaged trials over 8 bins per participant, and computed the frequency of correct choices in each bin. Finally, the correlation between the bins' confidence and performance was computed at the individual level. These measures were positively correlated ($R = 0.59 \pm 0.03$; Figure B1A).
- To assess the second property, the following linear regression was estimated at the individual level, using all trials from the confidence elicitation task (Model 1):
  $Conf = \beta 0 + \beta_1 \times Correct \times Evidence + \beta_2 \times Incorrect \times Evidence,$
  where **Incorrect** is a dummy variable coding for incorrect answers, and **Correct** is a dummy variable coding for correct answers. Then, we tested the

parameters of this model at the population level using one-sample t-tests. The results (Figure B1B), summarized in the table below (Table B1), demonstrate that confidence judgments are indeed positively associated with evidence for correct trials, and negatively for incorrect trials.

- To assess the third property, we proceeded similarly to the second: the following logistic regression was estimated at the individual level, using all trials (Model 2).

  Correct = $\beta 0 + \beta_1 \times High \times Evidence + \beta_2 \times Low \times Evidence$, where **High** is a dummy variable coding for high confidence trials (i.e., confidence > median(confidence)), and **Low** is a dummy variable coding for low confidence trials (i.e., confidence ≤ median(confidence)). Then, the parameters of this model were tested at the population level, using one-sample t-tests. The results (Figure B1C), summarized in the table below (Table B1), indeed demonstrate that the curve has a steeper slope in the high than in the low confidence trials, as was expected. Nota bene: when inspecting the data for the last model, we observed that the regression model was inestimable for two subjects. This was due to the median-split of the early certainty trials into high and low variants, since for both subjects the amount of low confidence trials was not sufficient (i.e., <nbins) to estimate the model, since the distribution of their confidence judgments was very skewed. This resulted in $\boldsymbol{\beta_2}$ being inestimable. Therefore, we excluded those two subjects (only) from the analyses of the last model when testing at the population level, whilst including them for the other models.

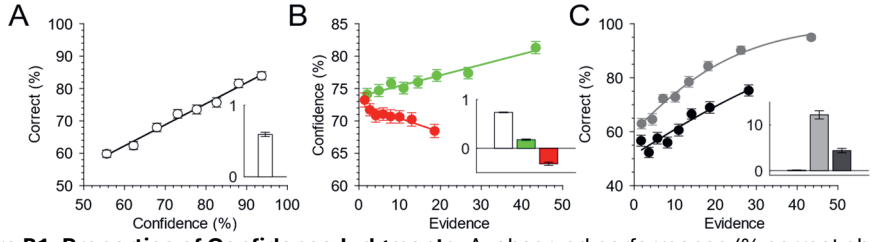


Figure B1: Properties of Confidence Judgments. A: observed performance (% correct choices) as a function of reported confidence. B: reported confidence as function of evidence for correct (green) and incorrect (red) choices. C: observed performance (% correct choices) as a function of evidence, for high (gray) and low (black) confidence trials. The insets presented on the side of each graph depict the results of the population-level analyses on the correlation coefficients (A) or on the regression coefficients (B and C). Error bars indicate inter-subject standard errors of the mean. *: P<.05; **: P<.01; ***P<.001

Table B1: Results of properties of confidence judgments

Model 1 (Figure B1B)	
Intercept (β_0)	β = 0.7366 ± 0.0085 t_{109} = 86.9948 P = 1.5197e-102
Confidence/Evidence Correct Answers (β_1)	β = 0.1736 ± 0.0174 t_{109} = 9.9870 P = 4.5659e-17
Confidence/Evidence Incorrect Answers (β_2)	β = -0.3184 ± 0.0335 t_{109} = -9.5124 P = 5.5456e-16
Model 2 (Figure B1C)	
Intercept (β_0)	β = 0.0987 ± 0. 0446 t_{107} = 2.2134 P = 0.0290
Performance/Evidence High confidence (β_1)	β = 12.2197 ± 0. 9019 t_{107} = 13.5494 P = 4.0743e-25
Performance/Evidence Low confidence (β_2)	β = 4.3919 ± 0. 4810 t_{107} = 9.1305 P = 4.1081e-15
Difference (β_1 - β_2)	t_{107} = 10.7705 P = 7.3768e-19

Behavioral Analyses: Properties of Early Certainty

Here we provide further details about the computation and properties of the early certainty variable. To verify that our model of early certainty is an appropriate proxy of confidence judgments, we performed similar behavioral analyses to confirm the three main properties of confidence judgments still hold for our early certainty variable. We performed identical analyses, substituting subjective confidence judgments for early certainty values.

Our results show that the measures of early certainty and performance are highly correlated (R = 0.73 ± 0.0362; Figure B2A, Table B2). Early certainty is also positively associated with evidence for correct trials, and negatively for incorrect trials (Figure B2B, Table B2). Finally, the relationship between performance and evidence is indeed higher in trials with high early certainty versus low early certainty (Figure B2C, Table B2). Nota bene: when inspecting the data for the last model, we observed that the regression model was inestimable for four subjects. This was due to the median-split of the early certainty trials into high and low variants, where these four subjects had an average performance of 100% in the *high confidence* trials, making $\boldsymbol{\beta_1}$ inestimable. Therefore, we excluded those four subjects from the analyses of the last model when testing at the population level, whilst including them for the other models.

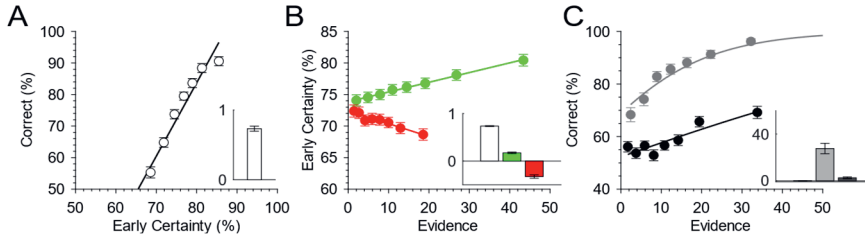


Figure B2: Properties of Early Certainty. A: observed performance (% correct choices) as a function of early certainty. B: early certainty as function of evidence for correct (green) and incorrect (red) choices. C: observed performance (% correct choices) as a function of evidence, for high (gray) and low (black) early certainty trials. The insets presented on the side of each graph depict the results of the population-level analyses on the correlation coefficients (A) or on the regression coefficients (B and C). Error bars indicate inter-subject standard errors of the mean. *: P<.05; **: P<.01; ***P<.001

Table B2: Results of properties of early certainty

Model 1 (Figure B2B)	
Intercept (β_0)	β = 0. 7364 ± .0084 t_{109} = 87.1574 P = 1.2434e-102
Confidence/Evidence Correct Answers (β_1)	β = 0.1714 ± .00177 t_{109} = 9.7058 P = 2.0064e-16
Confidence/Evidence Incorrect Answers (β_2)	β = -0.3266 ± .0355 t_{109} = - 9.2108 P = 2.6975e-15
Model 2 (Figure B2C)	
Intercept (β_0)	β = 0. 0858 ± 0. 0484 t_{105} = 1.7401 P = 0.0848
Performance/Evidence High confidence (β_1)	β = 27.5829 ± 4. 4499 t_{105} = 6.0848 P = 1.9264e-08
Performance/Evidence Low confidence (β_2)	β = 2.5038± 0. 7493 t_{105} = 3.2801 P = 0. 0014
Difference (β_1 - β_2)	t_{105} = 5.5184 P = 2.4830e-07

Moreover, to validate that our model of early certainty correlates highly with subjective confidence and choice and stimulus features, but does not show a statistical relationship with incentives, we built a linear mixed-effects model using the afex package in R. We used early certainty as dependent variable and added RT, accuracy,

evidence and the interaction between evidence and accuracy as predictors. Indeed, the results showed that RT ($F_{1,\ 15169}$ = 11622.7231, p<.001), accuracy ($F_{1,15157}$ = 1855.0251, p<.001) and the accuracy * evidence interaction ($F_{1,15156}$ = 626.7032, p<.001) all significantly contributed to early certainty, while no effect of incentive value on early certainty was found ($F_{1,15154}$ = 1.9232, p=.1462).

Behavioral Analyses: Model Comparisons

We iteratively built linear mixed effects models (LMEMs) and compared those by assessing model fit by using Chi-square tests on the log-likelihood values, and by comparing of the AIC and BIC model values. We started with a basic model with fixed effects of incentive, group and their interaction on confidence, together with a random subject intercept and slope of incentive per subject. Model predictors of accuracy and evidence, together with their interaction and the interaction with group were added whenever it significantly improved model fit. See Table 1 in the main text for the model comparison results. The final model (Model 1) consisted of fixed effects of incentive, group, accuracy and evidence (z-scored) and interactions between incentive and group, as well as two-way and three-way interactions between evidence, accuracy and group. All models included trial-by-trial data, and a random subject intercept as well as a random slope of incentive per subject.

Behavioral Analyses: Integration of Evidence in Confidence Judgments

Theoretical models of confidence formation suggest that confidence builds – at least partly - on the integration of noisy perceptual evidence used for decision-making (Fleming & Daw, 2017; Sanders et al., 2016). A resulting signature of confidence is its statistical dependence on an interaction of accuracy and perceptual evidence, which is typically illustrated as an 'X-pattern' where confidence increases/decreases with increasing evidence for correct/incorrect decisions, respectively. To study if GD and OCD patients show aberrant integration of evidence in confidence signals, we have included a three-way interaction term between evidence, accuracy and group in Model 1. Post-hoc testing was performed by comparing the groups on the slopes of evidence integration in confidence separately for correct and incorrect trials using the emtrends() function.

Behavioral Analyses: Confidence Calibration

Confidence calibration – also known as confidence bias – is the difference between average confidence and average performance per subject. If this measure is positive, this indicates overconfidence, whereas negative numbers indicate underconfidence. We calculated confidence calibration for each subject per incentive condition. We then performed a mixed ANOVA implemented in the afex package, to test for main effects of incentive conditions, groups, and their interaction. When a main effect was found significant, we performed post-hoc testing using the emmeans package, correcting for multiple comparisons using Tukey's method.

Behavioral Analyses: Metacognitive Sensitivity

Metacognitive sensitivity is a measure that indicates how well one's confidence judgments discriminate between one's correct and incorrect answers. One of the metrics used to express metacognitive sensitivity is *discrimination*. Discrimination is calculated as the difference between one's average confidence in their correct answers and their incorrect answers. The higher this metric, the more sensitive one's metacognitive abilities are. Another metric for sensitivity is *meta-d'*, which represents how much information in signal-to-noise units is available for the formation of confidence judgments (Maniscalco & Lau, 2012). The higher meta-d', the higher the metacognitive sensitivity.

We calculated discrimination for each subject per incentive condition. Moreover, we computed meta-d' per incentive and group using a hierarchical Bayesian framework (Fleming, 2017). We then performed two mixed ANOVA implemented in the afex package, to test for main effects of incentive conditions, groups, and their interaction on discrimination and meta-d' separately.

fMRI Analyses: Acquisition & Preprocessing

All our analyses were performed using MATLAB with SPM12 software (Wellcome Department of Cognitive Neurology, London, UK). Raw multi-echo functional scans were weighed and combined into 570 single volumes per scan session, using the first 30 dummy scans to calculate the optimal weighting of echo times for each voxel by applying a PAID-weight algorithm. During the combining process, realignment was performed on the functional data by using linear interpolation to the first volume. Subsequently, the functional images were co-registered, segmented for normalization

to MNI space and smoothed. To reduce motion-related artifacts, the Art-Repair toolbox (Mazaika et al., 2007) was used to detect large volume-to-volume movement and repair outlier volumes. Outliers were detected using a threshold for the variation of the mean intensity of the BOLD signal and a volume-to-volume motion threshold. A threshold of 1.5% variation from the mean intensity was used to detect and repair volume outliers by interpolating from the adjacent volumes.

fMRI Analyses: General Linear Models

GLM 1 consisted of three regressors for each timepoint: 'choice', 'incentive/rating' and 'feedback', to which parametric modulators (pmods) were added. All regressors were specified as stick functions time-locked to the onset of the respective events. The choice regressor was modulated by two pmods: early certainty (z-scored on subject level) and button press (left or right) to control for motor-related activation. The incentive/rating regressor was modulated by two pmods: incentive value ([-1,0,1]) and confidence rating (z-scored on subject level). The feedback regressor was additionally modulated by a pmod representing choice accuracy.

GLM 2 consisted of regressors for each of two time points (choice moment and incentive/rating moment) and three incentive conditions, as well as a single regressor at feedback moment, resulting in a total of seven regressors. All regressors at choice moment were modulated by a pmod of button press (left/right) and signed evidence: a variable that signifies the interaction between evidence and accuracy. Signed evidence was calculated as the absolute value of evidence in case of correct answers and the negative absolute evidence (i.e. –abs(evidence)) in case of incorrect answers. All regressors at rating moment were modulated by a pmod of confidence, and the feedback regressor was modulated by a pmod of accuracy. Thus, for all these events we could examine both baseline activity and regression slopes relating to their respective pmod.

For both GLMs pmods were not orthogonalized and thus competed to explain variance. We included six motion parameters as nuisance regressors. Regressors were modeled separately for each scanning session and constants were included to account for between-session differences in mean activation. All events were modeled by convolving a series of delta functions with the canonical hemodynamic response function (HRF) at the onset of each event and were linearly regressed onto the functional BOLD-response signal. Low frequency noise was filtered with a high pass filter with a cut off of 128 seconds. We controlled for the number of sessions while making the first-level contrasts. All contrasts were computed at subject level and then

taken to group level analyses. For GLM 1 we assessed group differences by performing a one-way ANOVA to our contrasts of interest, using an F-contrast test to test for any group differences (i.e. [1 -1 0; 0 1 -1]). In addition, to gain a complete picture of areas involved in our contrasts of interest, we grouped all subjects together and performed one-sample t-tests against 0.

Supplemental Results

Demographics

Age was not significantly different between the three groups ($F_{2,107} = 0.253$, $p > 0.75$), but IQ was, ($F_{2,107} = 3.222$, $p = 0.0438$). Post-hoc t-tests showed that HC subjects had a significantly higher IQ score than GD patients ($t = 2.53$, $p=0.014$). As expected, Y-BOCS scores and PGSI scores differed significantly between groups ($F_{2,107} = 322.2$, $p<.001$;$F_{2,107} = 380.5$, $p<.001$, respectively), with OCD patients having higher Y-BOCS scores than HCs ($t = -16.97$, $p<.001$) and GD patients ($t = -36.67$, $p<.001$), and GD patients having higher PGSI scores than HCs ($t = -15.99$, $p<.001$) and OCD patients ($t = -14.32$, $p<.001$). HAMA scores were significantly different between groups ($F_{2,107} = 48.02$, $p<.001$), post-hoc tests revealed higher HAMA scores for OCD patients than HCs ($t = -8.50$, $p<.001$) and GD patients ($t = 4.58$, $p<.001$), and higher HAMA scores for GD patients compared to HCs ($t = -2.44$, $p=.002$). HDRS scores were significantly different between groups ($F_{2,107} = 24.97$, $p<.001$), with higher scores for OCD versus HC ($t = -7.76$, $p<.001$), and higher scores for GD versus HC ($t=-3.03$, $p=.005$). Lastly, using a Chi-square test we found a significant difference in sex distribution between the groups ($X = 14.483$, $df = 2$, $p<.001$),

Behavioral Descriptive Results

Here we show the descriptive results that are depicted in Figure 2 main text.

Table B3: Descriptive behavioral results

Group	Incentive	Confidence	Accuracy	RT	Evidence
GD	Loss	76.31 ± 1.91	71.22 ± 1.66	1175.57 ± 81.19	15.05 ± 1.00
GD	Neutral	78.56 ±1.64	73.07 ± 1.61	1135.56 ± 71.15	17.60 ± 1.14
GD	Gain	81.12 ±1.84	71.37 ± 1.69	1132.38 ± 79.91	16.97 ± 1.26
OCD	Loss	71.30 ± 1.85	72.62 ± 1.42	1215.41 ± 74.70	13.84 ± 0.79
OCD	Neutral	73.27 ± 1.70	73.14 ± 1.39	1219.32 ± 65.71	15.56 ± 0.94
OCD	Gain	73.70 ± 1.65	75.07 ± 1.78	1224.67 ± 68.70	14.69 ± 0.84
HC	Loss	73.02 ± 1.03	70.87 ± 1.27	1130.16 ± 44.63	14.87 ± 0.63
HC	Neutral	75.05 ± 1.01	71.86 ± 1.19	1142.39 ± 48.26	16.90 ± 0.79
HC	Gain	75.68 ± 1.07	72.35 ± 1.27	1149.16 ± 46.71	16.05 ± 0.76

Shown here are the descriptive results of confidence, accuracy, reaction times (RT) and evidence per group and incentive condition. Shown are means ± sems.

Behavioral Analyses: Integration of Evidence in Confidence Judgments

The evidence integration effect differed per group, as signaled by a significant three-way interaction between accuracy, evidence and group ($F_{2,15094} = 3.0533$, $p=0.04723$) (Figure B3, Table B3). Post-hoc, we compared the groups on the slopes of evidence integration in confidence separately for correct and incorrect trials using the emtrends() function, and found that the slope for evidence integration into confidence was less steep for correct answers in GD patients compared to both HCs (GD - HC = -1.712 ± 0.283, Z-ratio = -6.057, $p<0.001$) and OCD patients (GD - OCD = -2.110 ± 0.357, Z-ratio = -5.912, $p<0.001$). This indicates that GD patients' confidence ratings were less influenced by the perceptual evidence when they made a correct choice. No differences between OCD patients and HC were found regarding evidence integration effects.

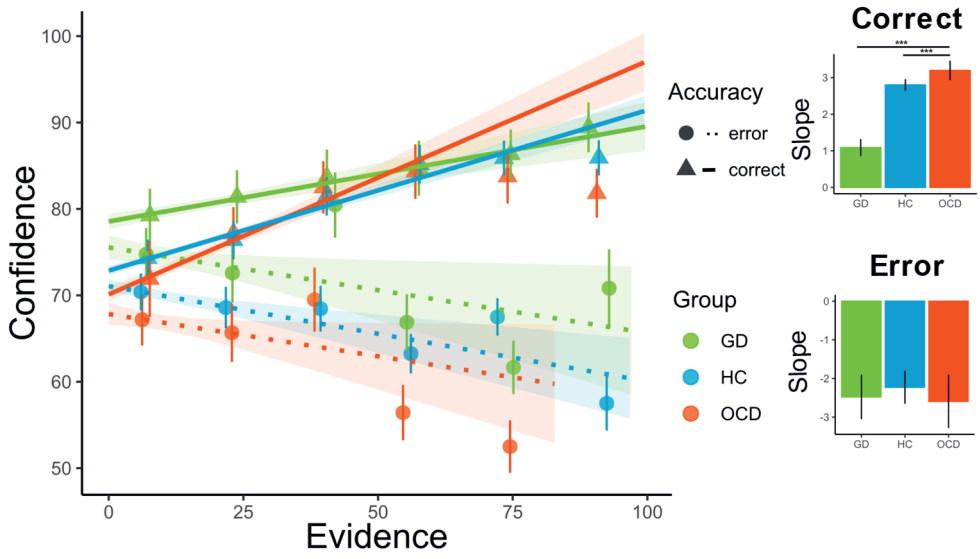


Figure B3: Evidence integration per group and accuracy level in confidence. Linking evidence, accuracy and group. Triangles represent mean reported confidence as a function of evidence for correct answers, and dots for incorrect answers, with different colors for the three groups. The solid lines represent the best linear regression fit for each group separately at the population level for correct answers, and the dotted lines for incorrect answers. Error bars represent SEM per group, shaded areas represent 95% confidence interval. Insets represent slopes (estimated marginal means of trends, taken from emtrends() function, error bars represent SEM) of correct and incorrect answers per group. Results from post-hoc testing are shown, where the slope for correct answer is significantly lower for gambling disorder (GD) versus both healthy controls (HC) and obsessive-compulsive disorder (OCD) (* p<0.05, ** p<0.01, *** p<0.001).

Behavioral Analyses: Confidence Calibration

We found a significant main effect of group ($F(2,107)=4.40$, $p = 0.015$), but no effect of incentive, nor an interaction effect between group and incentive. Post-hoc tests showed that GD patients showed increased calibration compared to OCD patients (t_{107}: -2.967, $p = 0.0103$), but no differences between GD or OCD patients and HC subjects. This indicated that GD patients are more overconfident

Behavioral Analyses: Metacognitive Sensitivity

We did not find a significant main effect of group or incentive, nor an interaction effect between group and incentive, both for discrimination and meta-d'. Average discrimination values were positive and average meta-d' was close to 1, indicating sensitive metacognition.

Behavioral Analyses: Clinical Correlations

We performed additional correlational analyses to explore whether subject's mean confidence level correlates with various clinical questionnaires of interest, separately for OCD and GD patients. In OCD patients there were no significant correlations with severity of OCD symptoms as measured with the YBOCS ($p>.5$) or with obsessive beliefs measured with the OBQ-44 ($p>.5$). In GD patients there was also no significant correlation with symptom severity measured using PGSI ($p>.4$), but there was a significant positive correlation between confidence level and BAS (Behavioral Approach System) scores ($r = 0.4608$, $p = 0.01784$).

Behavioral Analyses: Evidence Across Conditions

Due to a technical bug, perceptual evidence was not equal across incentive conditions. We performed a mixed ANOVA with within-subject factor incentive and between-subject factor group, which showed that evidence differed significantly over incentive conditions ($F_{2,205}=39.94$ $p<.001$), but not over groups ($F_{2,107}=0.94$ $p>.3$), and no interaction between incentive and group was found ($F_{3.83,205}=0.82$ $p>.5$). Post-hoc testing using t-tests revealed that evidence was highest in neutral, followed by gain, followed by loss condition (neutral versus loss: t-ratio= 7.844, $p<.001$; neutral versus gain: t-ratio=3.306, $p=0.001$; gain versus loss: t-ratio: 5.537, $p<.001$). Since evidence did not differ between the groups, it cannot account for any group differences we find in our data. Importantly, there are no effects of incentive on performance. Moreover, the difference in evidence over incentive conditions does not drive our incentive-induced confidence bias, since we do find a parametrically increase in confidence over incentive value, with a significant difference between all pairs. This means that confidence is higher in gain versus neutral conditions, even though evidence was significantly higher in neutral versus gain conditions. This shows that even though trials were easier in the neutral condition, participants were still more confident when they could gain points.

Behavioral Analyses: Clinical Groups With Their Own Control Group

In order to explore whether the behavioral analyses as in the main results with better matched control groups to the demographics of the two clinical groups would reveal similar results, we selected two subsets from our bigger sample of HCs (OCD control group N = 31, GD control group N = 32, with a slight overlap of N=8) of control groups to compare them with the two clinical groups.

Even though the groups were better matched and did not significantly differ from the clinical groups on terms of age, sex or IQ, we found similar results. When comparing OCD patients to a matched HC group, using Model 4 as described in the methods, we only found a significant main effect of incentive ($F_{2,111} = 9.5665$, $p<.001$), accuracy ($F_{1,8217} = 337.5033$, $p<.001$), evidence ($F_{1,8214} = 6.7033$, $p=.009$), an interaction effect between accuracy and evidence ($F_{1,8224} = 118.2244$, $p<.001$) and an interaction effect between group and accuracy ($F_{1,8227} = 7.9859$, $p=.004726$). No group effects were found.

When comparing GD patients to a matched HC group, using Model 4 from the main methods, we found significant main effects of incentive ($F_{2,60} = 15.6065$, $p<.001$), accuracy ($F_{1,8033} = 365.6563$, $p<.001$), evidence ($F_{1,8029} = 7.7733$, $p=.0053$), an interaction between accuracy and evidence ($F_{1,8027} = 117.9345$, $p<.001$), and between accuracy evidence and group ($F_{1,8027} = 6.5439$, $p=.0105$). No main group effect was found. Post-hoc analyses showed that that the slope for evidence integration into confidence is less steep for correct answers in GD patients compared to HCs (GD - HC = -1.329 ± 0.326, Z-ratio = -4.071, $p<0.001$).

These additional analyses thus show that even when using a better matched control group, we find no evidence for abnormalities in confidence level for OCD nor GD patients. For GD patients, we do, however, replicate that GD patients have a lower slope of evidence integration in confidence for correct answers compared to HCs.

fMRI: Interaction Between Metacognition and Incentives in VS (GLM 2)

We performed an ROI analysis by leveraging our factorial design. We extracted VS activations for both time points (choice and rating), all incentives (loss, neutral and gain), all groups (HC, OCD and GD), for both baseline activity and a regression slope with (1) signed evidence and (2) confidence judgments for all these events.

First, one-sample t-tests showed that, overall, VS baseline activations did not differ from 0 at choice moment ($t100 = -0.317$, $p >0.75$), while it was positive for baseline activations at rating moment ($t100 = 8.238$, $p < 0.001$). The correlations between VS activity and signed evidence at choice moment was significantly positive ($t100 = 4.985$, $p < 0.001$). However, the correlation between VS activity and confidence at rating moment did not differ from 0 ($t100 = 1.664$, $p = 0.099$) (Figure B4). This implies that activity in VS is related to incentive presentation, but also that it is related to signed evidence (i.e., the interaction between accuracy and evidence, showing that VS activity was lowest when one had high levels of evidence but was incorrect, and highest when

one had a lot of evidence and was in fact correct). Then, we turned to see whether there were effects of incentive condition and group around this general signal. As expected, at choice moment there were no effects of incentive condition on VS baseline activity, nor on its correlation with the signed evidence signal (i.e., slope) (Figure B4, Table B4). Moreover, we did not find a group nor an interaction effect on both baseline VS activity and the correlation with signed evidence at choice moment. At rating moment, however, incentive condition had a significant effect on both the baseline VS activity, as well as its correlation with confidence. Post-hoc testing showed that the baseline VS activity was highest during gain, followed by loss, and lowest during neutral (loss versus gain: t196: -4.590, $p < 0.001$, neutral versus gain: t196: -7.710, $p < 0.001$, loss versus neutral: t196 = 3.119, $p = 0.006$). The correlation of VS activity with confidence was significantly higher (i.e., increased slope) in gain versus neutral (t196 = -2.607, $p = 0.0265$), while no differences between gain and loss, or between neutral and loss were found. Moreover, there was a significant group effect on VS baseline activity during rating moment. This effect did not remain significant in the post-hoc tests, however, which showed that GD subjects had subthreshold decreased activity compared with HCs, averaged over incentive conditions (t98 = -2.272, $p = 0.0646$). No interaction effects between group and incentive were found on baseline activity or its correlation with confidence at rating moment.

Table B4: Results of VS ROI analysis

	Incentive	**Group**	**Incentive:Group**
Choice Baseline	F(1.92, 188.46) = 0.16, p = 0.846	F(2, 98) = 1.32, p = 0.271	F(3.85, 188.46) = 0.66, p = 0.615
Choice Slope 'Signed Evidence'	F(1.99, 195.35) = 1.63, p = 0.198	F(2, 98) = 0.44, p=0.647	F(3.98, 195.35) = 0.49, p = 0.741
Rating Baseline	F(1.85, 181.48) = 30.08, p<**0.001**	F(2, 98) = 3.48, p = **0.035**	F(3.70, 181.48) = 0.90, p = 0.460
Rating Slope 'Confidence Judgment'	F(1.92, 188.55) = 3.41, p = **0.037**	F(2, 98) = 1.68, p = 0.192	F(3.83, 188.55) = 0.69, p = 0.593

Shown here are the results of the mixed ANOVAs of t-statistics in the ventral striatum (VS) region of interest (ROI). Shown are the main effects of incentive condition, group and their interaction effect on the choice and rating time points, focusing on both the baseline activity as well as the slope of signed evidence and confidence judgments, respectively. F-values, with corresponding degrees of freedom and p-values are reported.

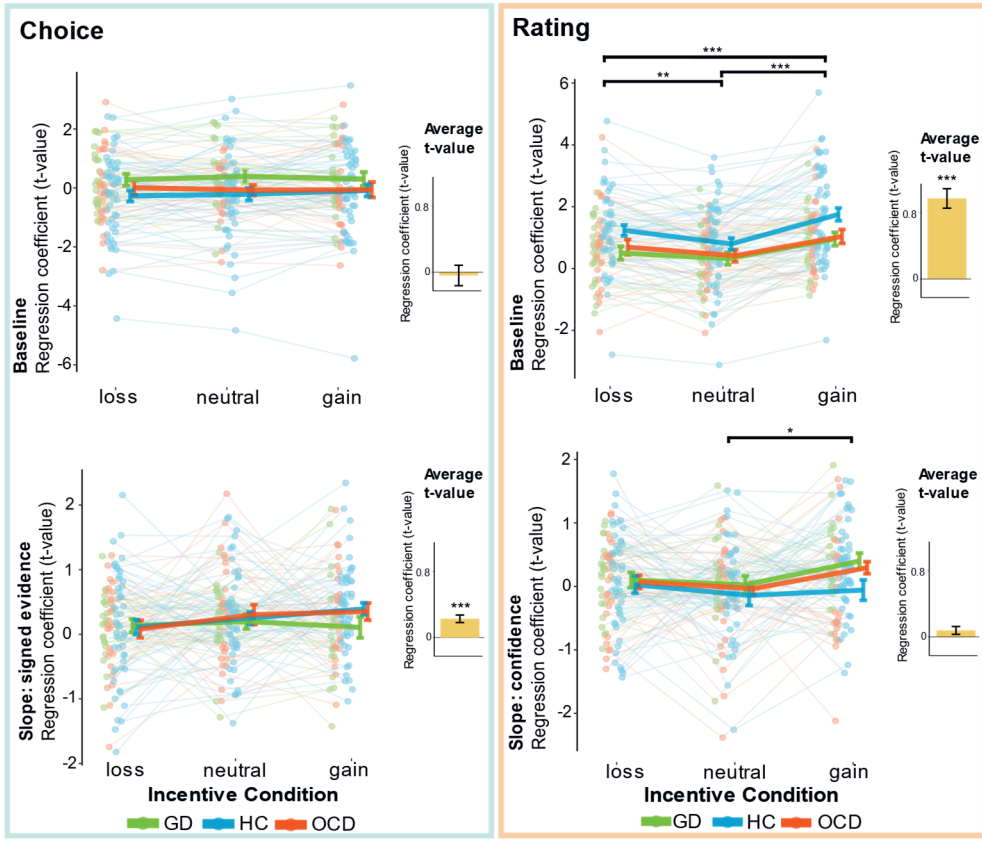


Figure B4: Ventral striatum cortex region of interest (ROI) analysis. T-values corresponding to baseline and regression slopes were extracted for all three groups and three incentive conditions, at two time points of interest: choice and incentive/rating moment. Green dots and lines represent gambling disorder patients, blue dots and lines represent healthy controls and red dots and lines represent obsessive-compulsive disorder patients. Dots represent individual t-statistics, and error bars represent sample mean ± SEM per group. Black bars represent significant post-hoc tests. Yellow bars represent average t-values, with corresponding significance level of one-sample t-tests against 0. (* p<0.05, ** p<0.01, *** p<0.001). GD = gambling disorder, HC = healthy control, OCD = obsessive-compulsive disorder.

Appendix C

Supplement to Chapter 5

Supplementary Methods

Local and Global Confidence Task

Participants performed short blocks with two randomly interleaved perceptual games indicated by two arbitrary color cues (Figure 1 main text). All games involved perceptual discrimination judgments. Each block had 12 trials, 6 trials from each game in pseudo-randomized order. Participants were instructed and incentivized to learn about their performance on both games. In each trial, two black boxes with white dots were briefly shown, and participants indicated which box contained a higher number of dots using the Z (left) and M (right) keys. The perceptual evidence was governed by the difference in dot number between the boxes. Two task features were varied between games: games could either be easy or difficult (i.e., difficulty feature), and delivered veridical feedback or no feedback (i.e., feedback feature). These features resulted in six possible pairings of games in the learning blocks. Each possible pairing was repeated twice, in a randomized order.

On all trials without feedback, participants were asked to provide a *local confidence* rating about the probability of their perceptual judgment being correct on a continuous scale from '50% correct (chance level)' to '100% correct (perfect)', with intermediate options of 60, 70, 80 and 90% correct. Confidence judgments were self-paced.

After each block, participants were asked to choose the game for which they believed they performed best (*global choice*). Subjects were instructed that their payment bonus depended on their average performance in the chosen game, incentivizing subjects to truthfully pick the game they believed to have performed best at. To indicate their choice, participants responded with two keys that differed from the keys used in the perceptual decisions, to avoid carry-over effects. After providing their global choice, participants were asked to rate their overall performance on each of the two games in the block (*global confidence*). The scale ranged from '50% correct (chance level)' to '100% correct (perfect)', with intermediate options of 60, 70, 80 and 90% correct, and these ratings were self-paced. Afterwards, participants received a break after which a new block started with two new games, indicated by two new color cues. Both these measures (global choice and global confidence) are metrics of *global metacognition*.

Each block consisted of two games, and each trial started with the presentation of a central color cue (1200 ms), indicating which of the two games will be presented in the current trial. Following the cue, the two black boxes were presented (300 ms). The dot difference was constant for each difficulty condition, but the spatial configuration of the dots within the boxes varied across trials. One box was always filled halfway (313 dots), the other box either contained 313 + 24 dots (difficult condition) or 313 + 60 dots (easy condition), based on earlier studies and targeting performance levels of around 70% and 85% correct, respectively (Rouault et al., 2019). The location of the box (left/right) was pseudo-randomized so that half of the trials in each block had the box with the most dots on the left. After a choice was made, the chosen box was highlighted (300 ms), and, in the feedback condition, a colored rectangle with the corresponding cued color of the current trial was presented, showing feedback (Correct / Incorrect) (1500 ms). In the no-feedback condition, the confidence rating scale was presented, with the color cue on top of the screen. The ITI was 600 ms. All participants first completed a practice block with longer stimulus presentation times for one game at a time.

Questionnaires

The symptoms assessed included alcoholism (Alcohol Use Disorders Identification Test (AUDIT)) (Saunders et al., 1993), apathy (Apathy Evaluation Scale (AES)) (Marin et al., 1991), impulsivity (Barratt Impulsiveness Scale (BIS-11)) (Patton et al., 1995), eating disorders (Eating Attitudes Test (EAT-26)) (Garner et al., 1982), social anxiety (Liebowitz Social Phobia Scale) (Liebowitz, 1987), obsessive-compulsive disorder (Obsessive-Compulsive Inventory Revised (OCI-R)) (Foa et al., 2002), schizotypy (Short Scales for Measuring Schizotypy) (Mason et al., 2005), depression (Zung Self-Rating Depression Scale) (Zung, 1965) and generalized anxiety (Generalized Anxiety Disorder 7-item scale (GAD-7)) (Spitzer et al., 2006).

The self-belief constructs assessed were: autonomy (Amsterdam Autonomy Scale (AAS)) (Bergamin et al., 2023), self-efficacy (Generalized Self-Efficacy Scale) (Schwarzer & Jerusalem, 1995), mastery (Sense of Mastery Scale) (Pearlin & Schooler, 1978) and self-esteem (Rosenberg Self-esteem Scale (RSE)) (Rosenberg, 1965) and Self-esteem Rating Scale Short Form (Lecomte et al., 2006). Self-esteem is a global construct concerned with one's self-worth that spans many personal domains, and low self-esteem has been related to the development of depression and anxiety disorders (Quiles et al., 2015; Sowislo & Orth, 2013). Self-efficacy is defined as "people's beliefs in their ability to influence events that affect their lives", and is strongly related to

emotional wellbeing, motivation and mastery (Bandura, 1977). Autonomy is seen as one's ability to live a meaningful life of their own making, which is undermined in many psychiatric disorders (Bergamin et al., 2022). Mastery is the degree to which one believes they can control various influences in their life, which is closely related to quality of life (Eklund et al., 2003; O'Kearney et al., 2020). Self-esteem is the broadest concept concerned with overall self-worth across all life domains (physical, academic, and social abilities, among others). All the above constructs are typically considered as trait characteristics, relatively stable for a given individual and only (currently) measurable by interview or questionnaire.

Exclusion Criteria

First, we excluded all participants who failed both of the two catch questions interspersed within the questionnaires (5 participants), who did not enter similar demographics details when asked twice (7 participants), and who failed comprehension tests about the usage of the confidence scale (91 participants). The comprehension test was passed when subjects rated *perfect* performance at least 10% higher than *chance* performance. This criterium is slightly different than what we pre-registered (pass when *perfect* performance $\geq$ 60 and *guess* performance between 40% and 60%), because we wanted to adhere to the same exclusion criteria the original authors of the task had set (Rouault et al., 2019). 16 participants were excluded for responding at or below chance level (50%), and 60 participants were excluded for having too little variation in their confidence judgment (a standard deviation < 5%), signaling that they hardly changed their confidence from the default setting. Some subjects did not meet multiple of our criteria. In total, we excluded 135 subjects, an exclusion rate of ~21%, consistent with a meta-analysis showing that web-based experiments typically exclude between 3% and 37% of their sample (Chandler et al., 2014). Our final sample consisted of 489 subjects with an average age of 27.2 years (± 8.5 years), of which 318 were male.

Moreover, for each participant, we excluded single trials when the choice reaction time (RT) was either > 10 seconds, < 200 ms or deviated more than 3 standard deviations from the participant average (median percentage of trials removed: 4.86%). We also reproduced all our analyses with either all RT < 100 ms removed, or with all subjects who had more than 50% of trials removed, for details see section 'Behavioral analyses with different RT exclusion criteria'. We also originally planned to exclude subjects who showed an average reading time of the primary instruction page of <5 seconds; this could unfortunately not be traced back from our data.

Factor Analysis

Even though our sample size is smaller than the original study (N=1413 versus N=489), we performed a *de-novo* factor analysis. This was for two reasons. First, recent work by Rouault, Seow, et al. (2018) using a comparable sample size (N=497) indicated that a de-novo factor analysis recovered similar symptom dimensions with high correlations between factor loadings. Second, our set of questionnaires did not fully match the original set of Gillan et al. (2016), since we included the GAD-7 instead of the STAI to measure general anxiety symptoms. Therefore, it was not possible to use the loadings derived from Gillan et al. (2016) to calculate factor scores for the current set of participants. However, for the overlapping items, our *de-novo* item loadings were very strongly correlated (all $r > 0.85$ and $p<0.001$) with the item loadings from (Rouault, Seow, et al., 2018), ensuring that our factor solution replicates previous factor solutions using this set of questionnaires, validating these results.

Following Rouault, Seow, et al. (2018), we performed a factor analysis with Maximum Likelihood estimation using the fa() function within the *psych* package in R-studio. We used all 197 unique individual questionnaire items as variables for the factor analysis (Figure C1A). Oblique rotations were used since factors were expected to correlate (and indeed correlations between factors were > 0.3). The social anxiety questionnaire individual item scores were calculated as the average of the avoidance and fear subitems. Since some responses were binary (schizotypy scale), a heterogeneous correlation matrix was computed using the *hetcor*() function within the *polycor* package in R, which allowed for the calculation of Pearson correlation between numerical variables and polyserial correlations between numeric and binary variables and polychoric correlations between binary variables. Factor selection was based on earlier work (Gillan et al., 2016; Rouault, Seow, et al., 2018), and on Cattell's criterion which states that an 'elbow' in the screeplot signifies the number of factors to retain. We used the Cattell-Nelson-Gorsuch test using the *nCng()* function within the *nFactors* package, which indicated that a 3-factor structure explained the item-level responses best and most parsimoniously, replicating previous studies.

For consistency with earlier studies, factors were given the same labels according to the items that loaded the most strongly (even though these labels are somewhat arbitrary). The loadings of the items on factor 1 were dominated by apathy ($M = 0.49 \pm 0.14$) and depression ($M=0.30 \pm 0.23$), whereas no other questionnaire on average reached a threshold of 0.25 (threshold taken from (Gillan et al., 2016)), even though general anxiety ($M = 0.23 \pm 0.07$) was close. The highest loading items of the GAD-7 questionnaire describe 'not being able to stop worrying', 'feeling nervous, anxious or on edge', and 'having trouble relaxing'. Thus, factor 1 was labelled 'Anxious-Depression'

(AD). For factor 2, items pertaining to OCD (M=0.46 ±0.09), general anxiety (M=0.43 ± 0.07) and alcoholism (M = 0.38 ± 0.10) showed the highest loadings. The top loading items of factor 2 pertain mostly to obsessive thoughts (i.e. obsessions) and compulsive behaviors for almost all items of the OCD and alcohol addiction questionnaires, which are both characterized as compulsive disorders (Figee et al., 2016). The top loading GAD-7 items interrogated feelings of 'being so restless it is hard to sit still', 'becoming easily annoyed or irritable' and 'feeling afraid as if something awful might happen'. Therefore, factor 2 was labeled 'Compulsive Behavior and Intrusive Thoughts' (CIT). In factor 3, items of social anxiety dominated, showing the highest loadings (M = 0.54 ± 0.13) without loading too strongly on general anxiety (M = 0.20 ±0.09). Therefore, factor 3 was labeled as 'Social Withdrawal' (SW). See Figure C1 for the factor loadings.

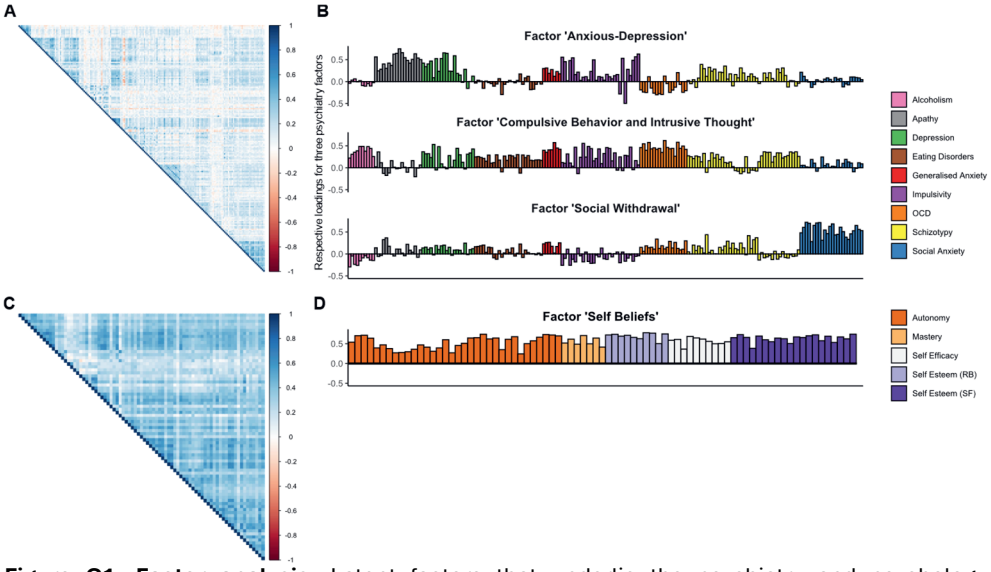


Figure C1: Factor analysis. Latent factors that underlie the psychiatry and psychology questionnaire items. **(A)** Correlation matrix of 197 questionnaire items pertaining to psychiatry, showing correlation coefficients between scores on all individual items across subjects. **(B)** Individual item loadings on each of the three factors, named 'anxious-depression' (AD), 'compulsive behavior and intrusive thought' (CIT) and 'social withdrawal' (SW), according to content of the strongest item loadings and in concordance with earlier work. **(C)** Correlation matrix of the questionnaire items pertaining to psychological constructs, showing correlation coefficients between individual item scores across subjects. **(D)** Individual item loadings on the 'Self-Beliefs' factor.

Statistical Analysis of the Local and Global Confidence Task

First, we completed several control analyses to examine whether the effects of the task features (feedback and difficulty) on performance and metacognitive measures replicated (Rouault et al., 2019). We performed a 2x2 ANOVA with feedback and difficulty as factors on performance, reaction times, global choice and global confidence judgments. Since global choice are proportions, they were transformed with a classic arcsine square root transformation. To confirm that both global confidence and global choice consistently captured global metacognition, a paired t-test was performed to compare global confidence for chosen and unchosen tasks.

Moreover, to explore how participants compute local confidence, it was compared between easy and hard trials, and correct and incorrect trials using paired t-tests. Finally, to investigate whether internal fluctuations in local confidence predict global choices over and above differences in accuracy or reaction times, we performed a logistic regression in blocks where both tasks did not provide feedback and instead asked for local confidence judgments. Differences in local confidence level, accuracy and RT between the two tasks in a block were used as predictors:

Global Choice ~ Δaccuracy + ΔRT + Δlocal confidence

Similarly, using a linear regression, we sought to predict the difference in global confidence between tasks using the fluctuations in local confidence, accuracy and RT between tasks:

Δ Global Confidence ~ Δaccuracy + ΔRT + Δlocal confidence

In both models, predictors were standardized (z-scored).

Relating Questionnaire Scores to Task Variables

We studied the relationship between our task variables pertaining to both performance and metacognition (including performance, global confidence, local confidence, local calibration, global calibration, metacognitive efficiency and the correlation between local and global confidence) and scores on (1) psychiatric symptom questionnaires and (2) self-belief construct questionnaires, as predictors, using multiple linear regressions.

All regressors were z-scored before entering the models to obtain standardized (i.e., comparable) regression coefficients. All questionnaire scores and metacognitive efficiency values were log-transformed and total scores were entered for each

questionnaire. Following Rouault, Seow, et al. (2018), due to high correlations between questionnaires, we ran a separate regression for each dependent variable and each symptom while controlling for demographic variables, as follows:

Behavioral Variable ~ Questionnaire Score + Age + IQ + Gender

To correct for multiple comparisons, a Bonferroni correction was applied that took the number of dependent variables into account, following Rouault, Seow, et al. (2018).

Additional Analyses Comparing the Effects of the Confidence Levels on Symptoms

Local confidence and global confidence correlate strongly together. In order to examine whether self-beliefs remained the strongest predictor of all three psychiatry factors when using a single predictor that was the average of local and global confidence, we performed additional regression analyses:

AD/CIT/SW ~ Average of Local and Global Confidence + Self-Beliefs + Age + IQ + Gender

Moreover, we also constructed separate analyses where only either local confidence or global confidence was used as a predictor alongside the predictor of self-beliefs.

Additional Analyses on Behavioral Patterns

In the main analyses, it was found that high scoring CIT individuals performed significantly worse, while high scoring AD individuals performed significantly better on the task. Please note that these analyses were not pre-registered and have been added as exploratory analyses to examine these behavioral patterns more in depth.

First, we computed parameters that give information about careless responding to questionnaires using the *careless* package in R (Richard & Wilhelm, 2022). The parameters 'long string index' (i.e., the longest consecutive string of identical responses), and 'intra-individual response variability' (IRV, i.e., the standard deviation of responses across a set of consecutive item responses) were calculated. A higher long string index and a lower IRV indicated more careless responding. We calculated the average long string index and IRV per subject and regressed them against their AD, CIT and SW scores.

Second, we performed analyses to examine whether AD, CIT and SW dimension scores related to the average RT on the task using a regression analysis, assuming that shorter RTs could be reflective of more impulsive response styles.

Third, we performed analyses to investigate whether there were differences in the relationships between the psychiatric dimensions and the manipulation of feedback on confidence/performance. To do so, we computed the difference in performance, global confidence and global calibration between feedback and no-feedback conditions for each subject, and regressed them separately against the AD, CIT and SW dimension scores.

Supplementary Results

Association Between Task Measures and Transdiagnostic Dimension Scores

Regression analyses were performed to investigate the relationships between various task measures and the transdiagnostic dimension scores (Figure 2A main text), of which the exact statistical outcomes are shown in Table C1.

Association Between Task Measures and Self-Beliefs

Regression analyses were performed to investigate the relationships between various task measures and the scores on the Self-Beliefs dimension (Figure 2B main text), of which the exact statistical outcomes are shown in Table C2.

Predicting Psychopathology with Levels of Confidence

Regression analyses were performed to assess the influence and relative importance of the hierarchical levels in predicting transdiagnostic psychiatric symptom dimensions (Figure 3 main text), of which the exact statistical outcomes are shown in Table C3. Outcomes of post-hoc testing, comparing the influence of the various levels of confidence on the transdiagnostic symptoms is shown in Table C4.

Table C1: Associations between task measures and transdiagnostic dimension scores.

	Predictors			
	AD	**CIT**	**SW**	***Post-hoc AD vs CIT***
Performance	β = 0.180 SE = 0.046 t = 3.948 p_{uncor} < 0.001 p_{cor} < 0.001	β = -0.213 SE = 0.046 t = -4.588 p_{uncor} < 0.001 p_{cor} < 0.001	β = 0.018 SE = 0.047 t = 0.369 p_{uncor} > 0.7 p_{cor} = 1	t = -5.626 p_{uncor} < 0.001 p_{cor} < 0.001
Local Confidence	β = -0.192 SE = 0.046 t = -4.165 p_{uncor} < 0.001 p_{cor} < 0.001	β = 0.095 SE = 0.047 t = 2.034 p_{uncor} < 0.05 p_{cor} > 0.2	β = -0.034 SE = 0.048 t = -0.719 p_{uncor} > 0.4 p_{cor} = 1	t = 4.071 p_{uncor} < 0.001 p_{cor} < 0.001
Global Confidence	β = -0.125 SE = 0.046 t = -2.712 p_{uncor} < 0.01 p_{cor} < 0.05	β = 0.013 SE = 0.047 t = 0.287 p_{uncor} > 0.7 p_{cor} = 1	β = -0.048 SE = 0.048 t = -1.001 p_{uncor} > 0.3 p_{cor} = 1	t = 1.960 p_{uncor} = 0.051 p_{cor} > 0.3
Local Calibration	β = -0.297 SE = 0.046 t = -6.491 p_{uncor} < 0.001 p_{cor} < 0.001	β = 0.231 SE = 0.047 t = 4.950 p_{uncor} < 0.001 p_{cor} < 0.001	β = -0.040 SE = 0.048 t = -0.835 p_{uncor} > 0.4 p_{cor} = 1	t = 7.527 p_{uncor} < 0.001 p_{cor} < 0.001
Global Calibration	β = -0.309 SE = 0.045 t = -6.802 p_{uncor} < 0.001 p_{cor} < 0.001	β = 0.232 SE = 0.046 t = 5.027 p_{uncor} < 0.001 p_{cor} < 0.001	β = -0.066 SE = 0.047 t = - 1.395 p_{uncor} > 0.1 p_{cor} = 1	t = 7.781 p_{uncor} < 0.001 p_{cor} < 0.001
Metacognitive Efficiency	β = 0.060 SE = 0.050 t = 1.188 p_{uncor} > 0.2 p_{cor} = 1	β = 0.097 SE = 0.050 t = 1.894 p_{uncor} = 0.059 p_{cor} > 0.4	β = -0.005 SE = 0.052 t = -0.091 p_{uncor} > 0.9 p_{cor} = 1	t = 0.483 p_{uncor} > 0.6 p_{cor} = 1
Correlation Local & Global Confidence	β = 0.005 SE = 0.047 t = 0.095 p_{uncor} > 0.9 p_{cor} = 1	β = -0.199 SE = 0.048 t = -4.128 p_{uncor} < 0.001 p_{cor} < 0.001	β = 0.067 SE = 0.049 t = 1.370 p_{uncor} > 0.1 p_{cor} = 1	t = -2.807 p_{uncor} < 0.01 p_{cor} < 0.05

Results of two-sided regression models testing the associations between various task measures and psychiatric factor scores (AD = 'Anxious-Depression', CIT = 'Compulsive Behavior and Intrusive Thoughts', SW = 'Social Withdrawal'. Reported are the corresponding β-values, standard errors (SE), t-values and p-values. Results of post-hoc tests comparing the associations between the AD and CIT factors with each task variable are shown. The gray shaded squares represent significant effects. P_{uncor} = uncorrected p-value, p_{cor} = corrected p-value ($p\text{-value}_{uncor}$ * 7).

Table C2: Associations between task variables and Self-Beliefs.

	Predictor: Self-Beliefs
Performance	β = -0.062 SE = 0.044 t = -1.407 $p_{uncor} > 0.1$ $p_{cor} = 1$
Local Confidence	β = 0.173 SE = 0.043 t = 3.977 $p_{uncor} < 0.001$ $p_{cor} < 0.001$
Global Confidence	β = 0.149 SE = 0.043 t = 3.452 $p_{uncor} < 0.001$ $p_{cor} < 0.01$
Local Calibration	β = 0.185 SE = 0.045 t = 4.133 $p_{uncor} < 0.001$ $p_{cor} < 0.001$
Global Calibration	β = 0.211 SE = 0.044 t = 4.756 $p_{uncor} < 0.001$ $p_{cor} < 0.001$
Metacognitive Efficiency	β = -0.086 SE = 0.047 t = -1.810 $p_{uncor} = 0.07$ $p_{cor} > 0.4$
Correlation Local & Global Confidence	β = 0.080 SE = 0.045 t = 1.782 $p_{uncor} > 0.05$ $p_{cor} > 0.5$

Results of two-sided regression models testing the associations between various task measures and the self-beliefs factor. Reported are the corresponding β-values, standard errors (SE), t-values and p-values. The gray shaded squares represent significant effects. P_{uncor} = uncorrected p-value, p_{cor} = corrected p-value ($p\text{-value}_{uncor} * 7$).

Table C3: Predicting transdiagnostic dimensions with levels of confidence

		Predictors		
		Local Confidence	**Global Confidence**	**Self-Beliefs**
Dependent Variables	**AD**	β = -0.082 SE = 0.045 t = -1.840 p = 0.066	β = 0.051 SE = 0.045 t = 1.121 p > 0.25	β = -0.827 SE = 0.026 t = -31.889 p < 0.001
	CIT	β = 0.252 SE = 0.071 t = 3.555 p < 0.001	β = -0.174 SE = 0.072 t = -2.428 p < 0.05	β = -0.410 SE = 0.042 t = -9.941 p < 0.001
	SW	β = 0.088 SE = 0.067 t = 1.316 p > 0.1	β = -0.076 SE = 0.067 t = -1.134 p > 0.25	β = -0.532 SE = 0.039 t = -13.749 p < 0.001

Results of two-sided regression models testing the associations between psychiatric factor scores and various hierarchical levels of confidence (i.e., local confidence, global confidence and Self-Beliefs (SB) factor score). Reported are the corresponding β-values, standard errors (SE), t-values and p-values. The gray shaded squares represent significant effects.

Table C4: Post-hoc analyses of predicting transdiagnostic dimensions with levels of confidence

Dependent Variable	**Post-hoc comparison**	**Statistics**
AD	Local confidence vs. global confidence	t = -1.558 p_{cor} > 0.3
AD	Local confidence vs. SB	t= 13.882 p_{cor} < 0.001
AD	Global confidence vs. SB	t = 16.736 p_{cor} < 0.001
CIT	Local confidence vs. global confidence	t = 3.149 p_{cor} < 0.01
CIT	Local confidence vs. SB	t = 7.766 p_{cor} < 0.001
CIT	Global confidence vs. SB	t = 2.830 p_{cor} < 0.05
SW	Local confidence vs. global confidence	t = 1.290 p_{cor} > 0.5
SW	Local confidence vs. SB	t = 7.742 p_{cor} < 0.001
SW	Global confidence vs. SB	t = 5.826 p_{cor} < 0.001

Two-sided post-hoc tests comparing the associations between psychiatric factor scores and various hierarchical levels of confidence (i.e., local confidence, global confidence and Self-Beliefs (SB) factor score) using the esticon() function in R, with Bonferroni correction applied (i.e., p-value*3). Reported are the corresponding t-values and p-values for one sample t-tests of each regression coefficient. For all three psychiatry factors, SB factor score (the highest hierarchical level), was the strongest predictor.

Behavior on the Local and Global Confidence Task

We studied how our task features of feedback and difficulty affected subjects' performance. Replicating Rouault et al. (2019), a 2x2 ANOVA showed a main effect of difficulty ($F(1,488) = 1986.37$, $p = 3.75 \cdot 10^{-174}$), but no main effect of feedback ($F(1,488) = 0.14$, $p = 0.71$), nor an interaction effect ($F(1,488) = 0.41$, $p = 0.52$) on performance. Thus, performance was higher in the easy (85.4% correct ± 1.6%) versus the difficult (67.7% correct ± 2.1%) condition, while it was not different for no feedback (76.4% ± 1.9%) or feedback (76.5% ± 1.9%) tasks (Figure C2A). This allowed us to study the influence of feedback on global confidence measures, irrespective of performance differences between feedback conditions. For reaction times (RTs), we similarly found a significant main effect of difficulty ($F(1,488) = 246.85$, $p = 2.6 \cdot 10^{-45}$), but also a significant main effect of feedback ($F(1,488) = 94.39$, $p = 1.62 \cdot 10^{-20}$) and no interaction ($F(1,488) = 1.09$, $p = 0.30$). RTs were significantly faster in easy (555.42 ms± 22.43 ms) versus difficult (618.54 ms ± 23.44 ms) tasks, and in tasks with feedback (563.62 ms ± 21.41 ms) than without feedback (610.77 ms ± 24.40 ms) (Figure C2B).

We then investigated how task features influenced global and local confidence, and examined whether subjects formed metacognitive judgments that accurately matched their performance. We first examined how our task features influenced (1) global choices, and (2) global confidence judgments. Consistent with previous findings (Rouault et al., 2019), 2x2 ANOVAs showed a significant main effect of both feedback ($F(1,488) = 687.98$, $p = 2.95 \cdot 10^{-95}$) and difficulty ($F(1,488) = 857.59$, $p = 1.49 \cdot 10^{-109}$) on global confidence judgments, as well as a significant interaction effect ($F(1,488) = 57.15$, $p = 2.01 \cdot 10^{-13}$). These results showed that tasks providing feedback were rated with higher global confidence than no-feedback tasks, and even more so for easy tasks (Figure C2D). This was mirrored in global choice (main effect of difficulty: ($F(1,488) = 638.74$, $p = 1.02 \cdot 10^{-90}$), main effect of feedback: ($F(1,488) = 639.73$, $p = 8.24 \cdot 10^{-91}$), interaction effect: ($F(1,488) = 36.94$, $p = 2.46 \cdot 10^{-9}$). Participants more often chose easy than difficult tasks, and more so in case they received feedback (Figure C2C). This result signals that, even though subjects' performance is equal in the presence or absence of feedback, subjects rated their performance worse when no feedback was present. Notably, we also found consistency between the two measures of global metacognition, with global confidence ratings higher for tasks that were chosen versus those that were not ($t_{488} = -45.158$, $p = 2.15 \cdot 10^{-176}$).

Replicating Rouault et al. (2019)[1], we showed that subjects gave higher local confidence ratings for easy (83.16 ± 0.67) versus hard (76.72 ± 0.69) tasks ($t_{488} = -27.21$, $p = 7.32 \cdot 10^{-100}$), as well as for correct (81.90 ± 0.68) versus incorrect (73.49 ± 0.67) trials ($t_{488} = -33.42$, $p = 2.97 \cdot 10^{-128}$), which signals metacognitive sensitivity (Figure C3A). We

also calculated M-Ratio as a measure of metacognitive efficiency, which indexes how well a subject can discriminate between correct and incorrect answers using their confidence. We found a mean posterior of 1.01, which is close to the ideal value of 1, and in line with previous studies on perceptual metacognition (e.g., (Faivre et al., 2018)).

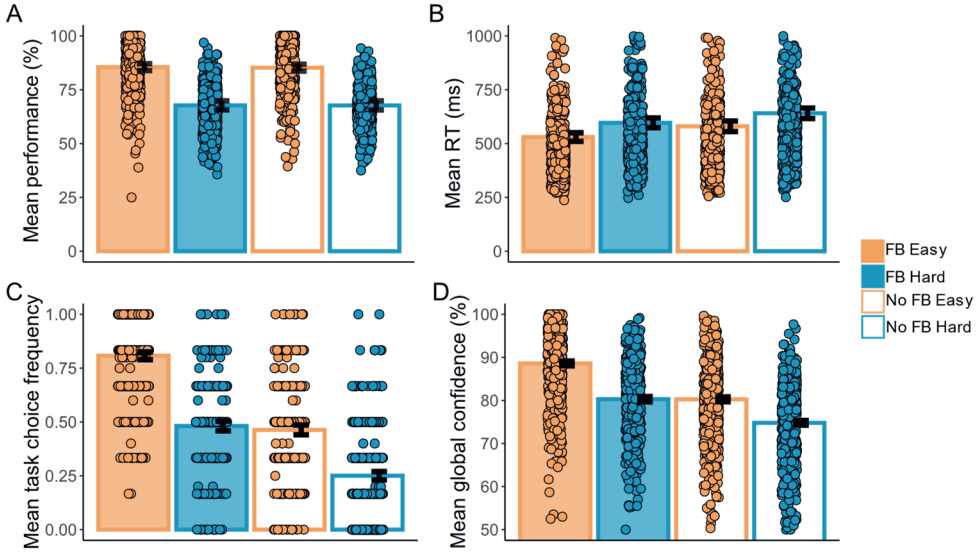


Figure C2. Behavioral results on objective performance and global confidence measures. A) Accuracy (mean % correct) was higher for easy than hard tasks, but was similar for tasks with ("FB") or without ("No FB") feedback. **B)** Reaction times were faster for easy than hard tasks, and for tasks with feedback than without feedback. These patterns were dissociated from the metacognitive patterns, where both global choices **(C)** and global confidence judgments **(D)** were higher in the presence than in the absence of feedback. This difference was even more pronounced when tasks were easy, despite objective performance being equal between feedback conditions. Bars represent mean across subjects. Black error bars represent SEM across subjects (N=489). Dots represent individual data points; for task choice frequency the individual data points take discrete values due to the finite number of blocks per participant.

After confirming that global and local confidence were indeed sensitive to our task features, we sought to examine whether fluctuations in local confidence affected both global confidence measures, over and above fluctuations in accuracy or RTs. A logistic regression showed that differences in local confidence ratings between tasks significantly predicted global choice ($\beta = 0.12$ ($\pm$ 0.01), $p < 0.0001$) over and above differences in accuracy ($\beta = 0.32$ ($\pm$ 0.31), $p = 0.29$) and RTs ($\beta = -0.0002$ ($\pm$ 0.0004), $p = 0.69$) (Figure C3B). This regression model was a better fit to subjects' global choices compared to a reduced model without local confidence fluctuations (BIC = 1011 for the model with local confidence, BIC = 1144 for the reduced model), and a model with only

local confidence fluctuations as predictor was also a better fit (BIC = 999). Moreover, a linear regression demonstrated that differences in local confidence between tasks also significantly explained differences in global confidence between those tasks (β = 0.69 (± 0.04), p < 0.0001), over and above differences in accuracy (β = 2.20 (± 1.28), p = 0.09) and RTs (β = 0.0008 (± 0.002), p = 0.62) (Figure C3C).

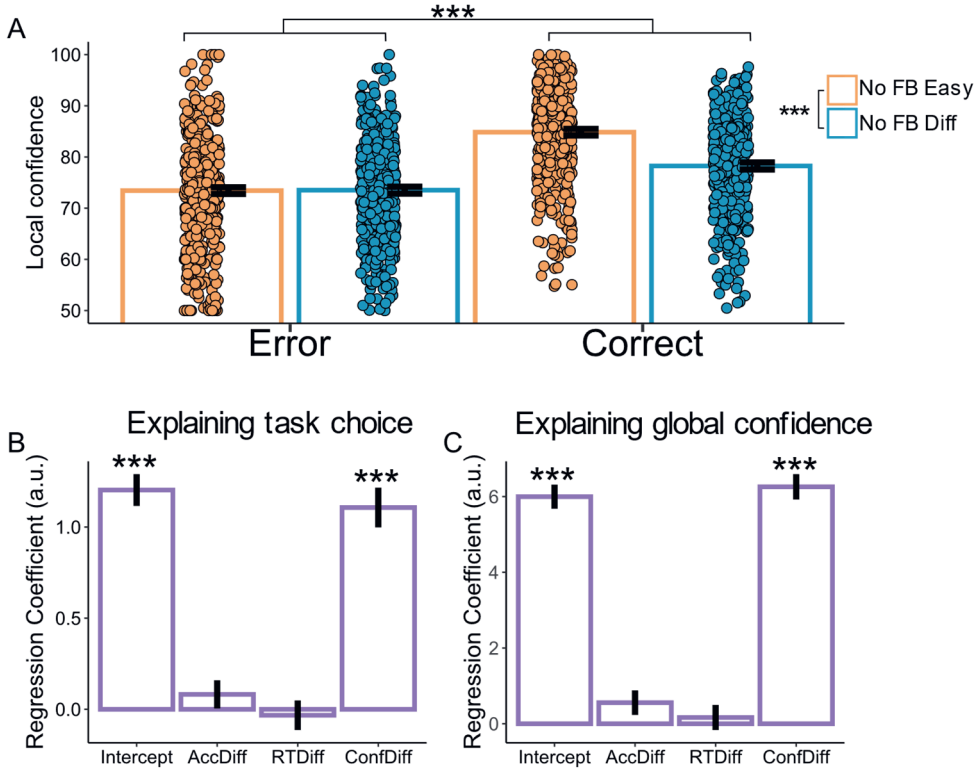


Figure C3. The association between local confidence and global confidence measures. A) Local confidence was rated higher in correct than incorrect trials and even more so for easy tasks, thus showing that local confidence judgments are affected by objective performance. N = 489 independent subjects. Bars and error bars represent mean and SEM over subjects. Dots represent individual data points. Significance stars represent two-sided ANOVA main effects of accuracy and feedback. Two-sided regressions on both **(B)** global choice and **(C)** global confidence judgments demonstrated that the difference in local confidence level between tasks (ConfDiff) explained subjects' task choices and global confidence over and above differences in accuracy (AccDiff) and reaction times (RTDiff) between tasks. N = 489 independent subjects. Error bars represent SEM. *** p < 0.001 indicates the statistical significance of the regression coefficient.

Relating Questionnaire Scores to Task Variables

Psychiatry Scores Regressions

We investigated how self-reported symptom scores relate not only to performance and local metacognition, but also to global confidence measures, metacognitive sensitivity and the correlation between local and global confidence, using regression analyses (Supplementary Methods and Figure C4).

Our regression analyses showed that self-reported OCD and alcoholism scores were significantly negatively related to performance (Alcoholism: $\beta = -0.133 \pm 0.045$, $t = -2.959$, $p < 0.05$; OCD: $\beta = -0.146 \pm 0.044$, $t = -3.299$, $p < 0.01$). In non-standardized terms, every 1 standard deviation increase in OCD symptom score, respectively alcoholism symptom score resulted in a 1.194%, respectively 1.085% decrease in performance. Regarding our metacognitive measures, we found that local confidence was negatively related to apathy scores ($\beta = -0.143 \pm 0.043$, $t = -3.302$, $p<0.01$), while global confidence was negatively related to both apathy and depression scores (apathy: $\beta = -0.117 \pm 0.043$, $t = -2.713$, $p < 0.05$; depression: $\beta = -0.117 \pm 0.043$, $t = -2.720$, $p < 0.05$). Local calibration was significantly higher (i.e., larger discrepancy between local confidence and actual performance) for subjects with higher OCD scores ($\beta = 0.131 \pm 0.046$, $t = 2.858$, $p < 0.05$), which was presumably driven by lower performance levels because of the lack of association between OCD scores and local confidence. Contrarily, local calibration was significantly lower (i.e., more underconfident) for subjects with higher apathy scores ($\beta = -0.171 \pm 0.045$, $t = -3.837$, $p = 0.001$). Global calibration was lower (i.e., more underconfident) for subjects with higher depression ($\beta = -0.136 \pm 0.045$, $t = -3.036$, $p < 0.05$) and apathy scores ($\beta = -0.187 \pm 0.044$, $t = -4.227$, $p < 0.001$). These findings of underconfidence in apathy and depression were presumably driven by decreases in local and global confidence rather than performance changes.

Interestingly, we found a significant negative relationship between the correlation of local with global confidence and both impulsivity ($\beta = -0.152 \pm 0.045$, $t = -3.351$, $p < 0.01$) and alcoholism ($\beta = -0.141 \pm 0.046$, $t = -3.081$, $p < 0.05$) scores, showing that the coupling between local and global confidence is more distorted in subjects with higher impulsivity and alcoholism scores. There were no significant associations between questionnaires and metacognitive efficiency that resisted correction for multiple comparisons, except for a small positive association between depression symptoms and metacognitive efficiency.

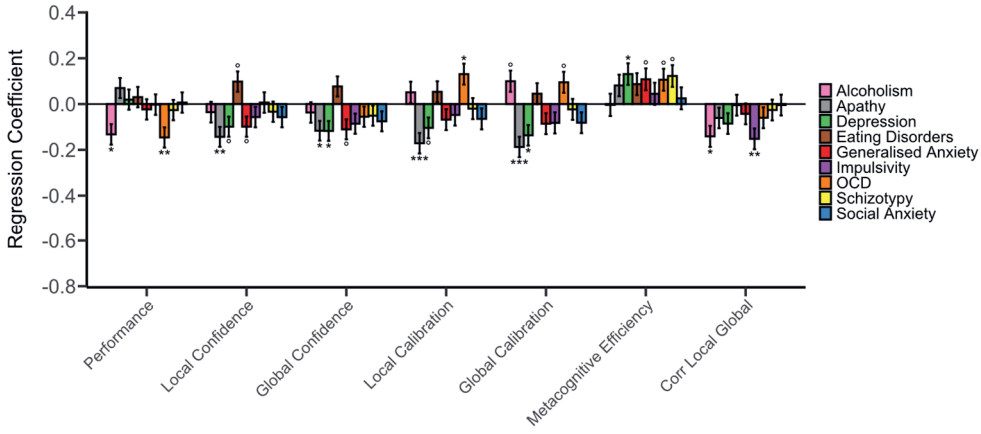


Figure C4: The association between task variables and symptoms. Regression coefficients of the two-sided regressions of all self-reported psychiatric symptom scores and various dependent variables measuring aspects of performance and metacognition. Each psychiatry symptom score was assessed in a separate regression model whilst controlling for age, IQ and gender. Since all variables were z-scored, the y-axis corresponds to the change in the dependent variable for each change of 1 standard deviation of that particular symptom score. Results are corrected for multiple testing. N = 489 independent subjects. Alcoholism and OCD symptoms were related to lower performance. Apathy symptoms were related to lower local confidence and calibration, as well as global confidence and calibration. Depression symptoms were associated with lower global confidence and calibration. The correlation between local and global confidence was diminished with high impulsivity and alcoholism symptoms. Error bars represent SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons over the number of dependent variables tested, ° $p < 0.05$ uncorrected. OCD = obsessive compulsive disorder, Corr Local Global = correlation between local confidence and global confidence.

Psychological Construct Regressions

We examined whether the higher-level psychological constructs were also related to performance and/or metacognitive measures (Figure C5). Our regressions revealed significant positive relationships between local confidence and autonomy ($\beta = 0.156 \pm 0.043$, $t = 3.585$, $p < 0.01$), mastery ($\beta = 0.122 \pm 0.044$, $t = 2.804$, $p < 0.05$) and self-efficacy ($\beta = 0.139 \pm 0.044$, $t = 3.193$, $p < 0.05$) scores, while global confidence ratings were positively related to autonomy ($\beta = 0.135 \pm 0.043$, $t = 3.141$, $p < 0.05$) and mastery ($\beta = 0.125 \pm 0.043$, $t = 2.897$, $p < 0.05$) scores. Moreover, both local and global calibration were found to be significantly higher (i.e. more overconfident) for subjects with higher autonomy (local: $\beta = 0.157 \pm 0.045$, $t = 3.492$, $p < 0.01$; global: $\beta = 0.168 \pm 0.045$, $t = 3.757$, $p = 0.001$), self-esteem (Short-Form Questionnaire: local: $\beta = 0.131 \pm 0.045$, $t = 2.889$, $p < 0.05$; global: $\beta = 0.163 \pm 0.045$, $t = 3.617$, $p < 0.01$ & Rosenberg's Questionnaire: local: $\beta = 0.141 \pm 0.045$, $t = 3.124$, $p < 0.05$; global: $\beta = 0.174 \pm 0.045$, $t =$

3.885, p = 0.001) and self-efficacy (local: β = 0.131 ± 0.045, t = 2.894, p < 0.05; global: β = 0.138 ± 0.045, t = 3.082, p < 0.05) scores. We found no significant associations between psychological questionnaire scores and metacognitive efficiency (M-Ratio), nor with the correlation between local and global confidence.

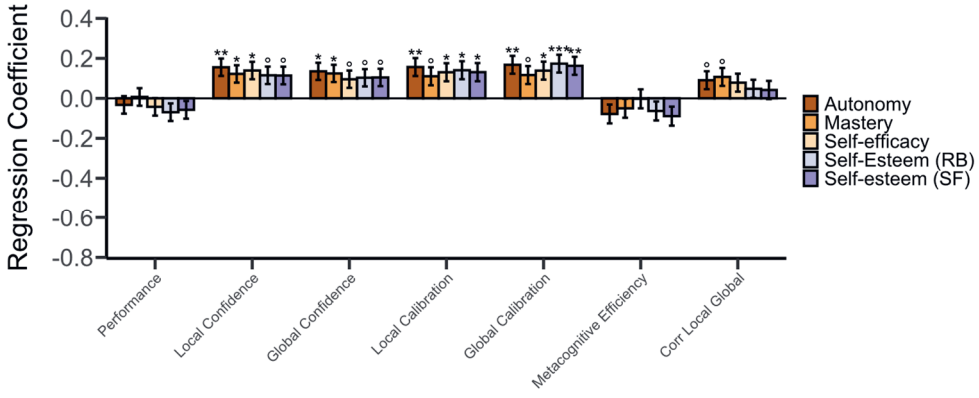


Figure C5: The association between task variables and psychology questionnaires. Regression coefficients of the two-sided regressions of all self-reported psychology questionnaire scores and various dependent behavioral variables pertaining to performance and metacognitive variables. Each questionnaire score was assessed in a separate regression model whilst controlling for age, IQ and gender. Since all variables were z-scored, the y-axis corresponds to the change in the dependent variable for each change of 1 standard deviation of that particular questionnaire score. Results are corrected for multiple testing. N = 489 independent subjects. While no relationships with accuracy were found, these results indicate that all five psychology questionnaires were positively related to local and global confidence judgments and calibration levels (although not all findings survived correction for multiple comparisons and should thus be interpreted with caution). Error bars represent SEM. * p < 0.05, ** p < 0.01, *** p < 0.001, corrected for multiple comparisons over the number of dependent variables tested, ° p < 0.05 uncorrected. RB = Rosenberg's, SF = short-form, Corr Local Global = correlation between local confidence and global confidence.

Additional Analyses Comparing the Effects of the Confidence Levels on Symptoms

Results from regression analyses with a predictor representing the average local and global confidence together with a predictor of self-beliefs showed similar results as reported in the main text. Here, we again found that self-beliefs remain the strongest predictor for AD (Effect of SB: β = -0.829 ± 0.026, t = -31.916, p < .001 & Effect of average local/global confidence: β = -0.031 ± 0.027, t = -1.162, p = 0.246), CIT (Effect of SB: β = -0.405± 0.042, t = - 9.750, p < .001 & Effect of average local/global confidence: β = 0.079±

0.043, t = 1.819, p = 0.070) and SW (Effect of SB: β = - 0.530± 0.039, t = - 13.703, p < .001 & Effect of average local/global confidence: β = 0.012± 0.040, t = 0.308, p = 0.758).

Moreover, when we constructed separate regression models where only either local or global confidence was a predictor alongside self-beliefs, we again get very similar results (Table C5).

Table C5: Results from two-sided regression analyses using either local confidence or global confidence as predictor for AD, CIT or SW scores alongside the predictor of Self-Beliefs

AD / CIT / SW ~ *Local* Confidence + Self-Beliefs + Age + IQ + Gender			
	AD	**CIT**	**SW**
Predictor			
Self-Beliefs	β = -0.827 ± 0.026 t = -31.864 p < .001	β = -0.412 ± 0.041 t = -9.940 p < .001	β = -0.533± 0.039 t = -13.770 p < .001
Local Confidence	β =-0.042 ± 0.027 t = -1.575 p = 0.116	β =0.114 ± 0.043 t = 2.679 p = 0.008	β =0.027 ± 0.040 t = 0.681 p = 0.496
AD / CIT / SW ~ *Global* Confidence + Self-Beliefs + Age + IQ + Gender			
	AD	**CIT**	**SW**
Predictor			
Self-Beliefs	β = -0.832 ± 0.026 t = -32.108 p < .001	β = -0.396 ± 0.042 t = -9.544 p < .001	β = -0.528± 0.039 t = -13.676 p < .001
Global Confidence	β =-0.016 ± 0.027 t = -0.592 p = 0.554	β =0.030 ± 0.043 t = 0.701 p = 0.483	β =-0.005 ± 0.040 t = -0.130 p = 0.896

Additional Analyses on Behavioral Patterns

First, we showed strong negative relationships between the long string index and all three symptom dimensions (AD: β = -0.208 ± 0.052, t = -4.013. p<.001; CIT: β = -0.648 ± 0.051, t = -12.615. p<.001; SW: β = -0.254 ± 0.053, t = -4.8195. p<.001), indicating that subjects scoring higher on all symptom dimensions had lower scores on the long string index, and thus less careless responding. Post-hoc tests showed that this negative relationship was stronger with CIT than with AD scores (t = 5.614, corrected p < .001).

IRV was negatively related to both CIT (β = -0.138 ± 0.012, t = -11.710. p<.001) and SW (β = -0.039 ± 0.012, t = -3.211. p=0.001), but not to AD (β = -0.017 ± 0.011, t = -1.412. p = 0.159). This contrarily points to more careless responding in high scoring CIT participants. However, since it is likely that subjects with high scores on CIT or SW

symptom dimensions have less variation in their answers, as most answers that will be selected will be on the high end of the spectrum.

Critically, both IRV and long string index did not correlate significantly with mean performance on the task (long string index: r = -0.04, p = 0.349; IRV: r = -0.03, p = 0.450). Overall, there was no clear pattern of increased or decreased careless responses in high scoring CIT participants that could have impacted their task performance.

With respect to reaction times, a weak positive relationship between CIT scores and RT was found (β = 0.102 ± 0.049, t = 2.092. p = 0.037), such that subjects with higher CIT scores were on average a bit *slower*, while we found a negative relationship between AD scores and RT (β = -0.094 ± 0.048, t = -1.966. p = 0.049), such that subjects with higher AD scores were on average a bit faster. These effects disappear when controlling for multiple testing. In sum, from the RTs on the task there are no evidence that the high scoring CIT subjects responded more impulsively.

In addition, no significant relationships were found between any of the factor scores and the difference in performance between the feedback and no-feedback conditions (AD: p = 0.599; CIT: p = 0.713; SW: p = 0.919). We did find a significant positive relationship between AD scores and the difference in global confidence between feedback conditions (β = 0.700 ± 0.279, t = 2.668. p = 0.008), indicating that subjects with higher AD scores have a larger increase in global confidence when there is feedback versus no feedback (so a stronger effect of the feedback manipulation). This relationship was significantly negative for CIT scores (β = -0.745 ± 0.277, t = -2.527. p = 0.012), indicating a weaker effect of the feedback manipulation. Next, looking at global calibration, we only found a significantly positive effect of AD score, such that higher scoring AD subjects had a stronger increase in calibration in the feedback versus no-feedback condition (β = 0.936 ± 0.392, t = 2.389. p = 0.017). Overall, we only find that CIT subjects have a weaker effect of the feedback manipulation on their global confidence, but not on their performance, and thus it is unlikely that this has negatively influenced their overall performance. These results thus do not give reason to think that the lower performance in high scoring CIT individuals is due to a different reaction to the feedback manipulation.

Distributions of Questionnaire Scores

Total scores of all questionnaires across subjects showed considerable spread, indicating that we had successfully sampled a large variability of symptoms in our general population sample (Figure C6A).

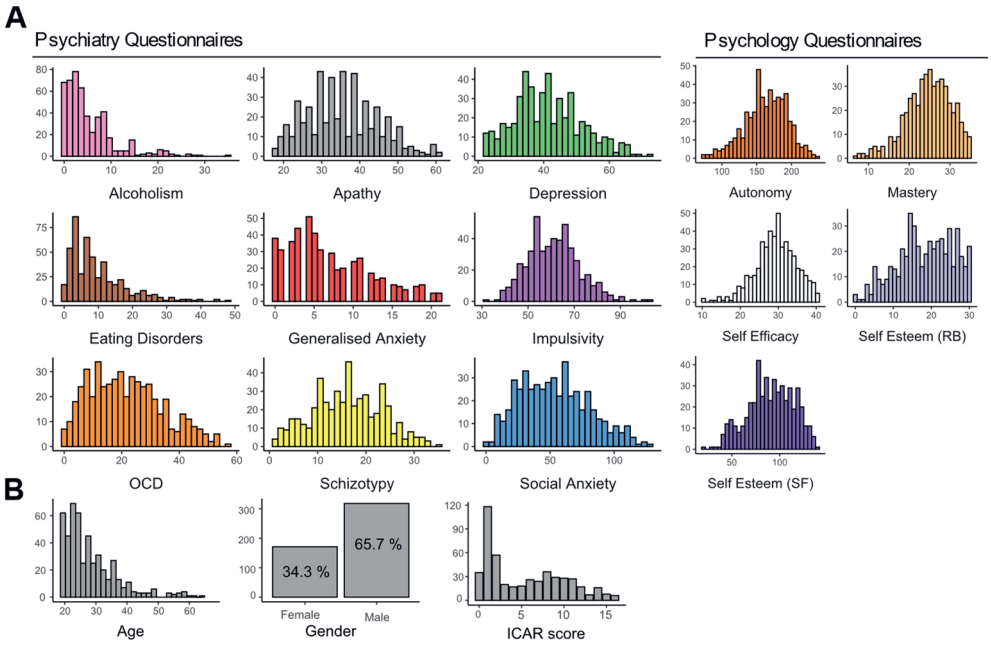


Figure C6: Questionnaire scores. Distributions of scores on **(A)** psychiatry (left panels) and psychology (right panels) questionnaires and **(B)** demographics across subjects (N=489). The list of the questionnaires employed is provided in the Supplementary Methods section.

Distributions of Variables

The distribution of all the task variables and transdiagnostic dimension scores were plotted (Figure C7), showing considerable spread and virtually normal distributions.

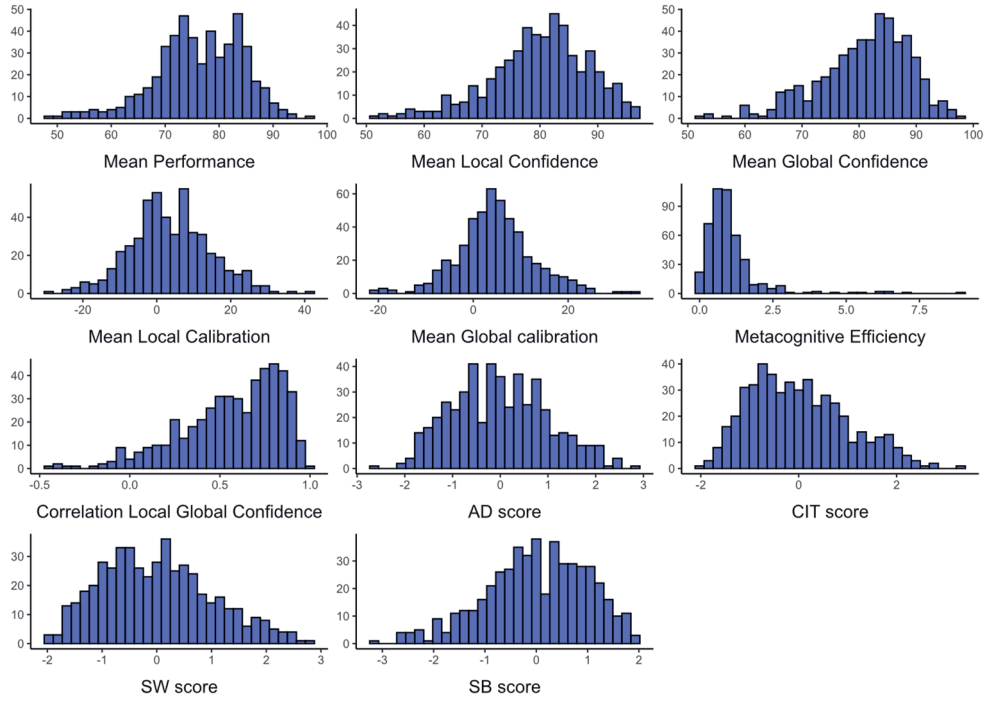


Figure C7: Distribution of variables. Distributions of task variables and transdiagnostic symptom scores (N=489).

Correlations Between Item Loadings

We calculated correlations between the item loadings from the present study and the item loadings from Rouault, Seow, et al. (2018) for the three psychiatry factors. We excluded all questionnaire items from the GAD-7 and STAI, since those were only used in one of the two studies and then correlated the remaining item loadings. Loadings were strongly positively correlated, indicating a satisfactory recovery of similar latent factors in the current study (Figure C8).

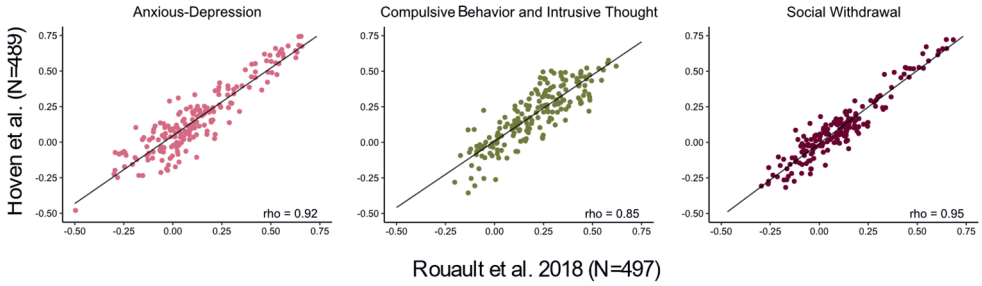


Figure C8: Correlations between item loadings of transdiagnostic factors. Correlations between questionnaire items from the current study (N=489 participants) and Rouault et al. (2018) study (N=497 participants). Item loadings for all three transdiagnostic factors were strongly correlated (all rhos > 0.85, all two-sided p < 0.001).

Item Loadings on Psychiatry and Psychology Factors

We calculated correlations between total scores on all psychiatry (Figure C9A) and psychology (Figure C9C) questionnaires and illustrated the mean loadings of each questionnaire on the three psychiatry dimensions (Figure C9B) and on the Self-Beliefs factor (Figure C9D).

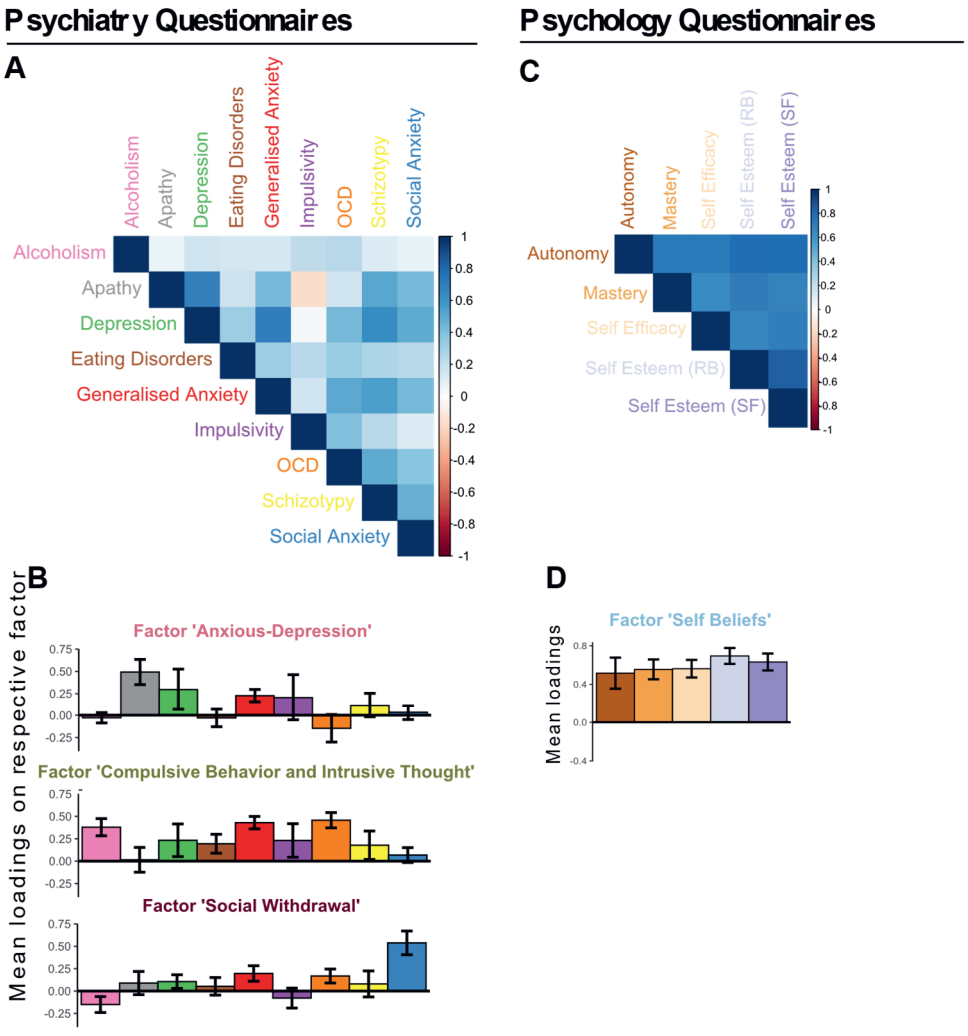


Figure C9: Factor analysis. A: Correlation matrix of the scores on the psychiatry questionnaires. **B:** Mean loadings of psychiatry questionnaire scores on the three latent psychiatry factors. **C:** Correlation matrix of the average scores on the psychology questionnaires. **D:** Mean loadings of the psychology questionnaire scores on the latent Self Beliefs factor. In B and D, error bars represent standard deviations over items within each questionnaire.

Relationships Between Demographics, Performance and Metacognition

We evaluated the relationships between demographics and all dependent variables without entering questionnaire scores in the regression (Figure C10). In this first set of regressions, with only birthyear, IQ and gender as predictors, we found that younger subjects had a better performance score ($\beta = 0.174 \pm 0.045$, $t = 3.899$, $p < 0.001$), as well as higher local and global confidence ratings (local: $\beta = 0.211 \pm 0.045$, $t = 4.723$, $p < 0.001$; global: $\beta = 0.267 \pm 0.044$, $t = 6.062$, $p < 0.001$). Moreover, higher performance scores ($\beta = 0.323 \pm 0.094$, $t = 3.443$, $p < 0.01$) as well as higher local confidence ratings ($\beta = 0.278 \pm 0.094$, $t = 2.975$, $p < 0.05$) were found for males versus females (Fig C10).

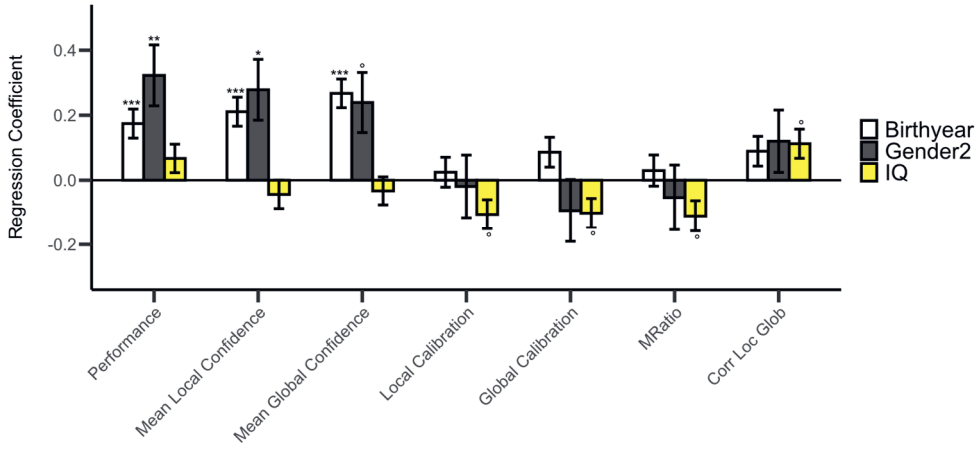


Figure C10: Associations between task variables and demographics. Associations between performance, metacognitive variables, and demographics (birthyear, gender and IQ) assessed with two-sided regression analyses. Y-axis indicate the change in each dependent variable for 1 standard deviation increase in birthyear or IQ. The reference group for gender is females, against which males are compared. N = 489 independent subjects. Results are corrected for multiple testing. Error bars represent SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, corrected for multiple comparisons over the number of dependent variables tested, ° $p < 0.05$ uncorrected. Corr Loc Glob means the correlation between local confidence and global confidence.

Behavioral Analyses with Different RT Exclusion Criteria

We reproduced all behavioral analyses with a sample in which we excluded all RTs < 100 ms. Since subjects are already presented with the stimuli for 300 ms, after which they can start making a response, there are many quick button presses. Almost all results remained, except for some results that were significant, but did not survive Bonferroni correction, including: the negative relationship between global confidence

and both depression and apathy symptom scores, the positive relationship between local calibration and OCD symptom scores, the negative relationship between global calibration and depression symptom scores, the positive relationship between mastery and both local and global confidence and the negative relationship between AD dimension scores and global confidence.

Moreover, we reproduced all behavioral analyses with a sample in which we excluded all subjects who had more than 50% of their trials removed when removing trials with an RT < 200 ms, in an attempt to remove outliers. Again, most results remained similar, except for some results that were significant, but did not survive Bonferroni correction, including: the negative relationship between alcoholism and performance, the negative relationship between global confidence and both apathy and depression symptom scores, the negative relationship between the correlation of local and global confidence and both alcoholism and impulsivity symptom scores, and the negative relationship between AD dimension scores and global confidence. In summary, our approach to treating RT outliers did not impact our conclusions.

Appendix D

Supplement to Chapter 6

Correlations between questionnaire scores

Table D1: Correlations between questionnaire scores

	RSES	OCI-R	MCQ-30	DASS	ZUNGDEP	GAD-7
ASA	.77***	-.47***	-.65***	-.75***	-.55***	-.58***
RSES		-.54***	-.66***	-.71***	-.60***	-.68***
OCI-R			.72***	.76***	.53***	.35*
MCQ-30				.80***	-	-
DASS					-	-
ZUNGDEP						.66***

Spearman correlation coefficients between questionnaire scores. For the correlations between rSES, OCI-R and ASA, all 120 subjects were included. For the correlations including MCQ-30 and DASS, only the 80 participants in the OCD and HC group were included. For the correlations including the ZungDEP and GAD-7 questionnaires only the HComp group was included. Abbreviations: OCD = Obsessive-Compulsive Disorder, HCs = Healthy Controls, HComp = High-Compulsive subjects, OCI-R: Obsessive-Compulsive Inventory-Revised, ASA: Autonomy Scale Amsterdam, rSES: Rosenberg Self-Esteem Scale, DASS: Depression Anxiety and Stress Scale, GAD-7: Generalized Anxiety Disorder-7 Questionnaire, ZungDEP: Zung's Depression scale, * = p<.05, ** = p<.01, *** = p<.001.

Comparing highly compulsive subjects to healthy controls

The healthy control (HC) group performed significantly better on the task compared to the highly compulsive (HComp) group, and were also significantly more confident at both local and global levels (Table D2). Even though the HComp group had higher local and global calibration values, and both groups showed overconfidence, this did not differ significantly between the groups. Also, no group differences in discrimination or the correlation between local and global confidence were found. Together, this shows that the HComp group is just as overconfident in their abilities as the HC group, even though the HComp group performs worse and is less confident overall.

Table D2: Differences in demographics, clinical data and task performance between HC and HComp groups.

	HCs (N = 40)	HComp (N=40)	HC vs. HComp
Age in years	38.58 (11.11)	36.53 (12.73)	T = 0.77 P = 0.45
Females (%)	27 (67.5%)	28 (70%)	X2 = 0.06 P = 0.81
Years of education	10.20 (3.13)	10.35 (2.64)	T = -0.23 P = 0.82
OCI-R	2.90 (2.48)	23.35 (13.18)	T = -9.65 P < .001
ASA	168.13 (19.18)	160.35 (33.99)	T = 1.26 P = 0.21
rSES	23.48 (3.94)	18.53 (7.56)	T = 3.67 P < .001
Accuracy (percent correct)	76.49 (7.76)	69.90 (8.64)	F = 13.15 P < 0.001
Local Confidence (on 50-100 scale)	81.14 (8.11)	76.82 (9.58)	F = 4.76 P = 0.032
Global Confidence	80.69 (7.27)	76.21 (8.83)	F = 6.08 P = 0.016
Local Calibration	4.82 (8.92)	6.63 (11.22)	T = 0.80 P = 0.429
Global Calibration	4.20 (6.98)	6.31 (9.62)	T = 1.13 P = 0.264
Correlation Local & Global Confidence	0.56	0.52	T = -0.60 P = 0.552
Discrimination	8.34 (4.77)	6.73 (4.66)	T = -1.53 P = 0.130

Abbreviations: HCs = Healthy Controls, HComp = High-Compulsive subjects, OCI-R: Obsessive-Compulsive Inventory-Revised, ASA: Autonomy Scale Amsterdam, rSES: Rosenberg Self-Esteem Scale, T = T-value from two-sample t-test, F = F-value from ANOVA, P = P-value. Data are reported as mean (standard deviation).

Comparing OCD and HC groups while controlling for anxiety and depression symptoms

Table D3: Regression results from models comparing OCD and HC groups while controlling for anxiety and depression symptoms.

Dependent Variable	Intercept	DASS score	Group (OCD)
Local confidence	β = 82.193 SE = 1.659 T = 49.533 P < .001	β = 1.391 SE = 1.401 T = 0.993 P = 0.324	β = -8.508 SE = 2.785 T = -3.055 P = 0.003
Global confidence	β = 81.480 SE = 1.505 T = 54.141 P < .001	β = 1.041 SE = 1.271 T = 0.819 P = 0.415	β = -6.027 SE = 2.526 T = -2.386 P = 0.020
Local calibration	β = 7.872 SE = 2.098 T = 3.753 P < .001	β = 4.030 SE = 1.772 T = 2.275 P = 0.026	β = -11.091 SE = 3.521 T = -3.150 P = 0.002
Global calibration	β = 6.315 SE = 1.604 T = 3.938 P < .001	β = 2.798 SE = 1.354 T = 2.066 P = 0.042	β = -7.234 SE = 2.691 T = -2.688 P = 0.009

Abbreviations: HC = Healthy Controls, OCD = Obsessive compulsive disorder, DASS: Depression Anxiety Stress Scale, SE = Standard Error, T = T-value, P = P-value.

Comparing the effect of OCI-R score on local confidence between OCD and HComp groups

Table D4: Regression results from model comparing the effects of OCI-R score on local confidence between OCD and HComp groups.

Local Confidence ~	Intercept	OCI-R score	Group (HComp)	Group (HComp) x OCI-R score
β	74.724	-2.508	2.086	4.034
SE	1.388	1.698	1.963	2.087
T	53.830	- 1.477	1.063	1.933
P	<.001	0.144	0.291	0.057

Abbreviations: HComp = Highly compulsive subjects, OCD = Obsessive compulsive disorder, OCI-R: Obsessive-Compulsive Inventory-Revised, SE = Standard Error, T = T-value, P = P-value.

Comparing M-Ratio (metacognitive efficiency) between groups

For the sake of completeness, we calculated metacognitive efficiency for each participant. The signal detection theory framework assumes constant signal strength, and therefore metacognitive efficiency (i.e., M-Ratio) was calculated separately for the easy and hard trials (36 trials per subject per M-Ratio calculation). The M-Ratio was taken as the average M-Ratio over the easy and hard condition, and compared between groups using two-sample t-tests. Some subjects (8 OCD, 2 HC, 6 HComp) with a negative M-Ratio likely due to the low number of trials to estimate M-Ratio, were excluded for these analyses.

The average M-Ratio for OCD patients was 0.859, for HC it was 0.927, and for Hcomp it was 1.11. There were no differences in M-Ratio between the OCD and HC groups (t_{68} = 0.487, p = 0.628), and neither between the OCD and HComp groups (t_{64} = 1.136, p = 0.260).

Appendix E

Supplement to Chapter 7

Supplementary Methods

Quasi-optimal Bayesian Observer Model

The quasi-optimal Bayesian observer model uses trial-by-trial feedback to form a point-estimate of its belief about the mean of the Gaussian distribution that generates the particle's landing locations. This current belief B_t is updated on each trial in proportion to the prediction error δ_t, using a delta-rule:

$$\mathrm{B}_{t+1} = \mathrm{B}_t + \alpha_t \times \delta_t$$

The degree to which prediction errors drive learning depends on the trial-by-trial learning rate α_t. Model prediction error is the difference between the estimated mean of the distribution B and the actual landing location of the particle X on each trial t.

$$\delta_t = \mathrm{X}_t - \mathrm{B}_t$$

In contrast to the fixed learning rates commonly used in reinforcement learning models, the strength of Bayesian belief updating is its dynamic updating. For the reduced Bayesian model, this means that the model learning rate α_t is updated on each trial in proportion to the surprise induced by the new evidence (Ω_t) and the confidence the model has in its own belief estimate (υ_t).

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

Surprise is quantified by change-point probability (CPP, Ω_t), and refers to the estimated probability that the mean of the sampling distribution has changed, given the current evidence. Model confidence (MC, υ_t), in turn, is a measure of the estimated reliability of the model's beliefs about the mean. When model confidence is low, even small prediction errors will be strongly weighted. The values of CPP and MC interact in such a way that when new evidence is surprising (a change-point), the learning rate will increase even when model confidence is high.

CPP is calculated as the relative likelihood that the current evidence X is generated by a uniform generative distribution over all possible locations $U(1{,}360)$ (i.e., during a change-point), as opposed to being drawn from the Gaussian (N) that generated the current belief B_t:

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

To calculate the CPP, the model takes into account three variables. First, the current evidence (i.e., landing location) X_t. Second, the hazard rate (H), referring to the degree of unexpected (transition) uncertainty intrinsic to the task. When H is maximal (H~ 1), each new observation is completely independent from previous observations. CPP is necessarily increased when H is maintained at a higher value. In our task, H is a constant at a value of 0.125. In contrast to the model, participants have no prior knowledge about the value of H and must infer it from the data. Third, the variance of the predicted distribution σ_t^2. This term describes the estimated noise in the generative Gaussian distribution, by modulating known noise constant σ_N by the current MC υ_t. The noise constant is a measure of expected uncertainty, and is inserted into the model as the true value of the generative variance (σ_N = 12). Participants must also infer this from the data.

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

Model confidence (υ) interacts in a precision-weighting manner with the variance of the generative Gaussian σ_N. When model confidence is low (e.g., right after a change-point), estimated noise over the predictive distribution σ_t^2 will be higher, and as the model gains more evidence and becomes more confident of its predictions, the estimated noise σ_t^2 will decrease and approach the true value of σ_N. Because the value of σ_t^2 is used in the calculation of the CPP, the model will be less likely to attribute new evidence to the occurrence of a change-point when model confidence is low. Unlike the other model variables, model confidence does not depend on the current location of the particle X_t and is calculated at the end of trial t for the subsequent trial. υ_{t+1} represents the inverse of the fraction of total uncertainty about the next location of the particle that is due to imprecise estimation of the mean, relative to the known uncertainty due to noise ${\sigma_N}^2$.

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

The numerator consists of three terms. The first term represents the uncertainty of the generative distribution ${\sigma_N}^2$ weighted by the estimated probability that a change-point has occurred. In the second term, this uncertainty ${\sigma_N}^2$ is weighted by the estimated probability that no change-point occurred. The third term reflects the model's uncertainty about whether or not a change-point occurred. These terms are repeated in the denominator, so that the value of this uncertainty can be calculated relative to

the uncertainty that is due to noise. If a change-point occurred on the current trial (Ω_t = 1), model confidence is automatically set to 0.5 (its minimum value) for trial t+1. Due to interdependencies intrinsic to the calculation of the parameters, they are expected to correlate.

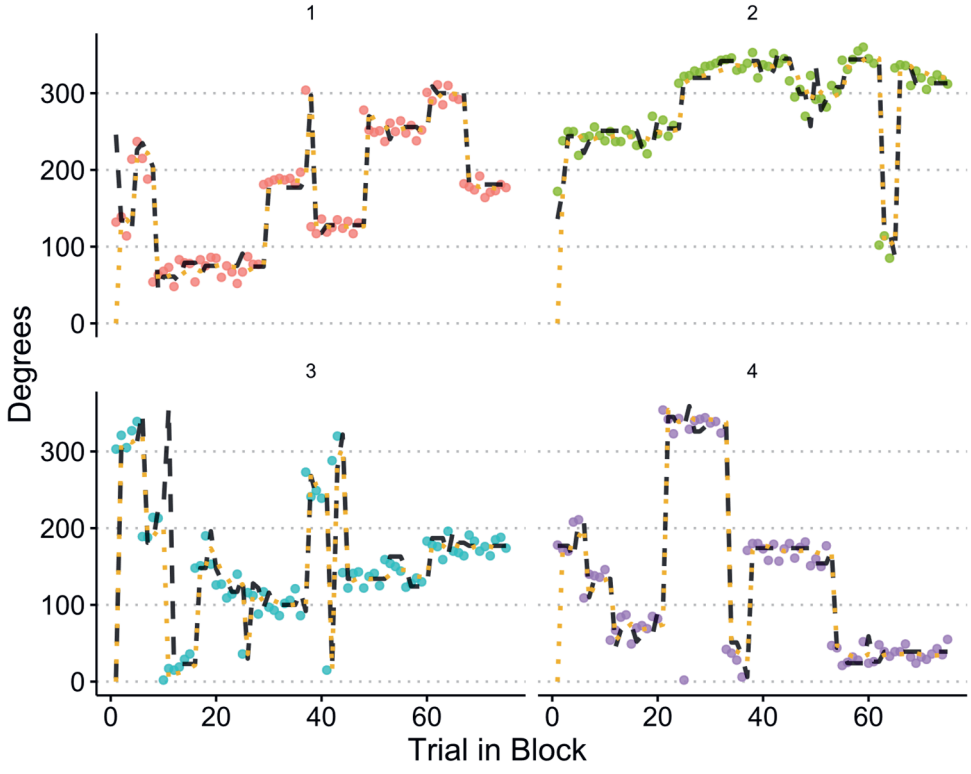


Figure E1: Human and model behavior. Colored dots represent the landing positions of the particle, per block (1-4).The black line represents the participants bucket position per trial, and the orange line represents the prediction of the quasi-optimal Bayesian model (B_t). Data are shown for a sample HC subject.

Control Analyses

To compare various control variables between our groups, we used two sample t-tests. When variances were not equal between groups, Welch t-tests were used to approximate the degrees of freedom.

First, we checked whether accuracy (i.e., % hits) was equal between the groups. Indeed, accuracy was equal between the OCD and HC group ($t_{71.4}$=-0.08, p=0.94) and between the OCD and HComp group (t_{112}=1.57, p=0.12) (Table E1). The percentage of excluded trials relative to the total number of trials per subject did not differ between OCD and HC groups (t_{73}=0.11, p=0.92), and neither between OCD and HComp groups (t_{73}=0.09, p=0.93). Next, the percentage of trials per subject in which no action update was performed did not differ between OCD and HC groups (t_{73}=1.51, p=0.14). It was higher, however, in the HComp group compared to the OCD group (t_{112}=3.96, p<0.001; HComp = 43.9 %, OCD = 27.2 %). The percentage of trials in which no confidence update was performed was equal between OCD and HC groups (t_{73}=0.34, p=0.73), and between OCD and HComp groups (t_{112}=0.37, p=0.71). Similarly, the percentage of trials in which confidence was kept at the default rating did not differ between OCD and HC groups (t_{73}=-0.10, p=0.92), and neither between OCD and HComp groups (t_{112}=0.70, p=0.49).

Then, we calculated the percentage of change-point trials per subject relative to their total number of trials and compared between groups. This did not differ between the OCD and HC groups (t_{73}=-0.18, p =0.86), but there was a small difference with slightly lower percentage of change-points for the HComp group (t_{112}=-2.59, p =0.01; HComp = 12.7 %, OCD = 13.1 %). Moreover, we calculated the number of trials between consecutive change-points, since change-points could occur at any moment in the task. This did not differ between the OCD and HC groups ($t_{40.9}$=1.10, p =0.28), and neither between the OCD and HComp groups ($t_{37.6}$=-1.61, p=0.012). Since the landing location of the particle was drawn from a uniform distribution during a change-point, we calculated the average difference between the position of the particle preceding a change-point and at change-point per subject. This did not differ between the OCD and HC groups (t_{73}=-1.55, p=0.13), and neither between the OCD and HComp groups (t_{73}=-0.34, p=0.73).

Hazard Rate Analyses

The hazard rate H in our task is a constant of 0.125. However, it is possible that participants inferred a different value of H from the evidence, since they have no prior knowledge about the value of H. Our main analyses were carried out using a constant H of 0.125. In addition, for each participant (OCD, HC and HComp), we carried out an exhaustive search for the best fitting H parameter between 0 and 1. The best fitting value was determined by using a minimum least squares fit between the bucket positions (i.e., participant behavior) and model belief B_t about the landing positions. As per the model, higher hazard rates result in higher values of CPP (Ω) and learning rate (α_t). When hazard rate equals 1, learning rate also equals 1, such that each new observation is independent from previous observations and consequently, the model belief is completely determined by the prediction error. When hazard rate become smaller, the model belief gradually is less influences by the prediction error, and mostly influenced by the belief in the previous trial.

Overall, the perceived hazard rates were higher than the constant of 0.125 (Figure E2). No significant differences were found between the OCD (0.79 ± 0.22) and HC (0.70 ± 0.27) groups (t_{73} = -1.61, p = 0.112), but hazard rate was higher for the OCD group than the HComp group (0.50 ± 0.30, $t_{96.5}$ = -5.74, p < .001). Hazard rate was strongly positively correlated with action update (clinical sample: r = 0.65, p<.001, analogue sample: r = 0.74, p<.001) and learning rate (clinical sample: r = 0.68, p<.001, analogue sample: r = 0.77, p<.001), and negatively correlated with accuracy (clinical sample: r = -0.42, p<.001, analogue sample: r = -0.35, p<.001). No significant correlations were found within the OCD and HComp groups between hazard rate and OCI-R score.

Sensitivity analyses were performed using the individually fitted hazard rate parameter as covariate to control for potential effects of hazard rate on our group differences. We included hazard rate as a fixed effect in our mixed-effects models investigating (1) group differences on confidence, (2) group differences on action update, (3) interaction between group and prediction error bin on learning rate, (4) group differences in the effects of the Bayesian parameters on action, and (5) on confidence.

In the clinical and analogue sample, the group effect of lower confidence in OCD compared with HC or HComp remained significant (clinical: β = -16.87 ± 4.89, t = -3.45, p < .001, analogue: β = -12.53 ± 5.02, t = -2.49, p = 0.014). This implies that the difference in confidence between the groups was not driven by a difference in hazard rates between the groups. Moreover, the main effect of group on action update remained non-significant in the clinical sample (β = -0.73 ± 1.13, t = -0.64, p = 0.523), with a strong main effect of hazard rate on action update (β = 17.25 ± 2.31, t = 7.48, p < .001). The

significant group effect of higher action update in OCD compared with HComp disappeared (β = 0.76 ± 0.90, t = 0.84, p = 0.401), and a strong main effect of hazard rate on action update was found (β = 14.07 ± 1.45, t = 9.71, p < .001). This implies that the difference in action update between the OCD and HComp groups was driven by a difference in perceived hazard rate between the groups. Just as our main analyses, our sensitivity analyses did not show a group difference in action-confidence coupling in both samples (clinical: β = 1.06 ± 1.19, t = 0.89, p = 0.379, analogue: β = 0.65 ± 1.25, t = 0.52, p = 0.603). Next, the effect of a higher learning rate in OCD compared to HC specifically for small prediction errors remained significant (Z-ratio: -1.99, p = 0.046). Overall, the main effect of higher learning rate in the OCD group compared to the HComp group remained significant (β = 0.20 ± 0.04, t = 5.73, p<.001). When zooming in on different prediction error bins, this effect only remained significant for low and medium prediction errors (low: Z-ratio: -5.729, p < .001, medium: Z-ratio: -4.193, p<.001), but not for high prediction errors (Z-ratio: 0.12, p = 0.905).

In terms of the model-based analyses, when adding hazard rate as a covariate in the clinical sample all of the original results remained. In the analogue sample, all of the original results remained, except that the main effect of group on action disappeared (β = 1.17 ± 0.83, t = 1.40, p = 0.164).

Taken together, these sensitivity analyses indicate that all group differences between OCD patients and HCs were not influenced by differences in perceived hazard rate between groups. While the difference in perceived hazard rate between OCD patients and HComp participants explained their difference in action updating, but not their differences in confidence and learning rate.

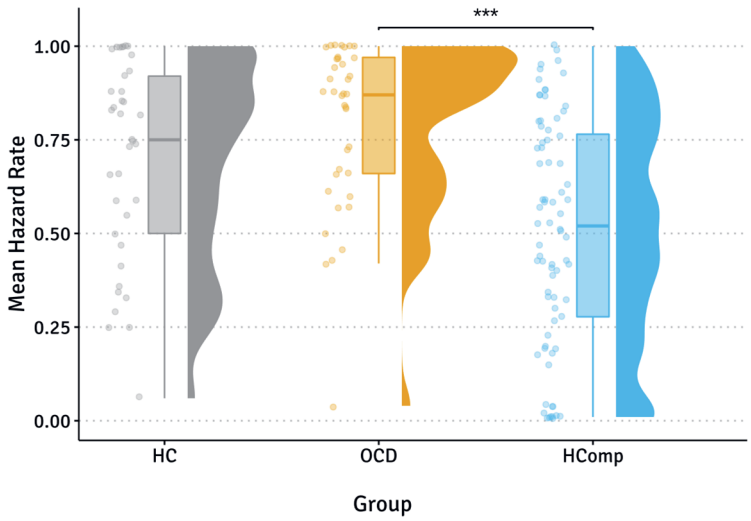


Figure E2: Hazard rate per group. Dots show data from individual participants, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. Significance starts represent the main effects of group in the respective mixed-effects models. ***p<.001. HC = healthy control subjects, OCD = obsessive-compulsive disorder patients, HComp= highly compulsive subjects from the general population.

Table E1: Mean and standard deviation of task variables per group.

	OCD	HC	HComp
Accuracy (%)	56.39 (8.57)	56.23 (9.66)	59.04 (8.51)
Confidence	40.64 (20.93)	59.56 (21.85)	55.07 (23.51)
Confidence Update	13.95 (6.65)	14.50 (6.21)	15.45 (7.17)
Learning Rate	0.73 (0.27)	0.65 (0.29)	0.45 (0.21)
Action Update	23.30 (6.54)	22.49 (6.30)	18.52 (4.96)

Table E2: Results of linear mixed-effects models predicting the effects of computational variables and group on action and confidence.

OCD versus HC						
	Dependent Variable: Action			**Dependent Variable: Confidence**		
Predictors	***Beta (SE)***	***t***	***p***	***Beta (SE)***	***t***	***p***
PE	20.01 (2.00)	9.99	**<.001**	-0.96 (0.61)	-1.58	0.117
CPP	9.44 (2.02)	4.67	**<.001**	-3.22 (0.80)	-4.04	**<.001**
(1-MC)*(1-CPP)	1.72 (0.44)	3.92	**<.001**	-3.15 (0.61)	-5.12	**<.001**
Hit	-6.36 (0.48)	-13.28	**<.001**	5.84 (0.69)	8.45	**<.001**
PE * Group (OCD)	0.60 (2.82)	0.21	0.831	0.53 (0.85)	0.62	0.534
CPP * Group (OCD)	-0.52 (2.84)	-0.18	0.831	-1.04 (1.11)	-0.93	0.356
(1-MC)*(1-CPP) * Group (OCD)	0.10 (0.62)	0.17	0.867	-0.26 (0.86)	-0.31	0.761
Hit * Group (OCD)	0.09 (0.67)	0.14	0.890	-1.14 (0.97)	-1.17	0.245
OCD versus HComp						
	Dependent Variable: Action			**Dependent Variable: Confidence**		
Predictors	***Beta (SE)***	***t***	***p***	***Beta (SE)***	***t***	***p***
PE	6.42 (1.96)	3.28	**0.001**	-1.06 (0.57)	-1.88	0.062
CPP	21.08 (1.79)	11.77	**<.001**	-4.82 (0.80)	-6.02	**<.001**
(1-MC)*(1-CPP)	4.70 (0.49)	9.59	**<.001**	-4.42 (0.57)	-7.74	**<.001**
Hit	-5.47 (0.32)	-17.21	**<.001**	4.59 (0.51)	8.93	**<.001**
PE * Group (OCD)	14.35 (3.13)	4.59	**<.001**	0.59 (0.84)	0.70	0.484
CPP * Group (OCD)	-12.32 (2.79)	-4.42	**<.001**	0.56 (1.26)	0.45	0.657
(1-MC)*(1-CPP) * Group (OCD)	-2.91 (0.79)	-3.68	**<.001**	0.98 (0.95)	1.04	0.302
Hit * Group (OCD)	-0.79 (0.52)	-1.53	0.129	0.08 (0.85)	0.10	0.923

Appendix F

Supplement to Chapter 8

Quasi-optimal Bayesian Observer Model

The same model as used in previous studies (Hoven, Mulder, et al., 2023; Seow & Gillan, 2020; Vaghi et al., 2017)has been employed for the current study. The quasi-optimal Bayesian observer model uses trial-by-trial feedback to form a point-estimate of its belief about the mean of the Gaussian distribution that generates the particle's landing locations. This current belief B_t is updated on each trial in proportion to the prediction error δ_t, using a delta-rule:

$$\mathrm{B}_{t+1} = \mathrm{B}_t + \alpha_t \times \delta_t$$

The degree to which prediction errors drive learning depends on the trial-by-trial learning rate α_t. Model prediction error is the difference between the estimated mean of the distribution B and the actual landing location of the particle X on each trial t.

$$\delta_t = \mathrm{X}_t - \mathrm{B}_t$$

In contrast to the fixed learning rates commonly used in reinforcement learning models, the strength of Bayesian belief updating is its dynamic updating. For the reduced Bayesian model, this means that the model learning rate α_t is updated on each trial in proportion to the surprise induced by the new evidence (Ω_t) and the confidence the model has in its own belief estimate (υ_t).

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

Surprise is quantified by change-point probability (CPP, Ω_t), and refers to the estimated probability that the mean of the sampling distribution has changed, given the current evidence. Model confidence (MC, υ_t), in turn, is a measure of the estimated reliability of the model's beliefs about the mean. When model confidence is low, even small prediction errors will be strongly weighted. The values of CPP and MC interact in such a way that when new evidence is surprising (a change-point), the learning rate will increase even when model confidence is high.

CPP is calculated as the relative likelihood that the current evidence X is generated by a uniform generative distribution over all possible locations $U(1{,}360)$ (i.e. during a change-point), as opposed to being drawn from the Gaussian (N) that generated the current belief B_t:

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

To calculate the CPP, the model takes into account three variables. First, the current evidence (i.e. landing location) X_t. Second, the hazard rate (H), referring to the degree of unexpected (transition) uncertainty intrinsic to the task. When H is maximal (H~ 1), each new observation is completely independent from previous observations. CPP is necessarily increased when H is maintained at a higher value. In our task, H is a constant at a value of 0.125. In contrast to the model, participants have no prior knowledge about the value of H and must infer it from the data. Third, the variance of the predicted distribution σ_t^2. This term describes the estimated noise in the generative Gaussian distribution, by modulating known noise constant σ_N by the current MC υ_t. The noise constant is a measure of expected uncertainty, and is inserted into the model as the true value of the generative variance (σ_N = 12). Participants must also infer this from the data.

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

Model confidence (υ) interacts in a precision-weighting manner with the variance of the generative Gaussian σ_N. When model confidence is low (e.g. right after a change-point), estimated noise over the predictive distribution σ_t^2 will be higher, and as the model gains more evidence and becomes more confident of its predictions, the estimated noise σ_t^2 will decrease and approach the true value of σ_N. Because the value of σ_t^2 is used in the calculation of the CPP, the model will be less likely to attribute new evidence to the occurrence of a change-point when model confidence is low. Unlike the other model variables, model confidence does not depend on the current location of the particle X_t and is calculated at the end of trial t for the subsequent trial. υ_{t+1} represents the inverse of the fraction of total uncertainty about the next location of the particle that is due to imprecise estimation of the mean, relative to the known uncertainty due to noise ${\sigma_N}^2$.

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

The numerator consists of three terms. The first term represents the uncertainty of the generative distribution ${\sigma_N}^2$ weighted by the estimated probability that a change-point has occurred. In the second term, this uncertainty ${\sigma_N}^2$ is weighted by the estimated probability that no change-point occurred. The third term reflects the model's uncertainty about whether or not a change-point occurred. These terms are repeated in the denominator, so that the value of this uncertainty can be calculated relative to

the uncertainty that is due to noise. If a change-point occurred on the current trial (Ω_t = 1), model confidence is automatically set to 0.5 (its minimum value) for trial t+1. Due to interdependencies intrinsic to the calculation of the parameters, they are expected to correlate.

Control Analyses

To compare various control variables between our groups, we used two sample t-tests. When variances were not equal between groups, Welch t-tests were used to approximate the degrees of freedom.

Accuracy (i.e. % hits) was equal between the groups (t_{49}=-0.95, p=0.345), as well as the average prediction error (t_{49}=-1.18, p=0.242). The percentage of excluded trials relative to the total number of trials per subject did not differ between groups (t_{49}=0.02, p=0.981). The percentage of trials per subject in which no action update was performed was higher in GD (t_{49}=2.48, p=0.017; GD = 60.1 %, HC = 50.5 %). The percentage of trials in which no confidence update was performed was equal between groups (t_{49}=1.29, p=0.202). The percentage of change-point trials per subject relative to their total number of trials did not differ between the groups (t_{49}=-0.58, p =0.566). Moreover, we calculated the number of trials between consecutive change-points, since change-points could occur at any moment in the task, which did differ between the groups, with a slightly higher number of trials in GD compared to HC ($t_{27.07}$=5.66, p<0.01; GD = 6.9 trials, HC = 6.7 trials). Since the landing location of the particle was drawn from a uniform distribution during a change-point, we calculated the average difference between the position of the particle preceding a change-point and at change-point per subject. This did not differ between the groups (t_{49}=0.60, p=0.55).

Results Excluding Outlier Participant Based on Task Accuracy

We did not use accuracy-based exclusion criteria in our analyses. However, the average accuracy of one GD participant was characterized as an outlier, being 18.2%. When removing this participant from our main analyses all results remained similar, however, the effect of a weaker action-confidence coupling in GD turned non-significant, but still indicated a trend effect (β = 3.07 +-1.65, t = 1.86, p = 0.063).

Hazard Rate Analyses

The hazard rate H in our task is a constant of 0.125. However, it is possible that participants inferred a different value of H from the evidence, since they have no prior knowledge about the value of H. Our main analyses were carried out using a constant H of 0.125. In addition, for each participant, we carried out an exhaustive search for the best fitting H parameter between 0 and 1. The best fitting value was determined by using a minimum least squares fit between the bucket positions (i.e. participant behavior) and model belief B_t about the landing positions. As per the model, higher hazard rates result in higher values of CPP (Ω) and learning rate (α_t). When hazard rate equals 1, learning rate also equals 1, such that each new observation is independent from previous observations and consequently, the model belief is completely determined by the prediction error. When hazard rate become smaller, the model belief gradually is less influences by the prediction error, and mostly influenced by the belief in the previous trial.

Overall, the perceived hazard rates were higher than the constant of 0.125 (Figure F1). No significant differences were found between the GD (mean: 0.54, sd: 0.30) and HC (mean: 0.59, sd: 0.31) groups (t_{49} = -0.61, p=0.542). Hazard rate was strongly positively correlated with action update (r = 0.72, p<.001) and learning rate (r = 0.74, p<.001), but not with accuracy (r = -0.03, p=0.85). No significant correlations were found within the GD group between hazard rate and PGSI or GBQ score.

Sensitivity analyses were performed using the individually fitted hazard rate parameter as covariate to control for potential effects of hazard rate on our group differences. We included hazard rate as a fixed effect in our mixed-effects models investigating (1) group differences on confidence, (2) group differences on action update, (3) interaction between group and prediction error bin on learning rate, (4) group differences in the effects of the Bayesian parameters on action and (5) on confidence. This did not change any of the effects reported in the main manuscript. In fact, including hazard rate as a covariate in the learning rate analyses strengthened the group effect of lower learning rates in GD compared to HC (general mixed-model: β = -0.10 (0.04), t = -2.89, p = 0.0056). When zooming in on different prediction error bins, the effect of lower learning rates for low and medium prediction errors strengthened as well (low: Z-ratio: 3.143, p = 0.002, medium: Z-ratio: 4.879, p<.001).

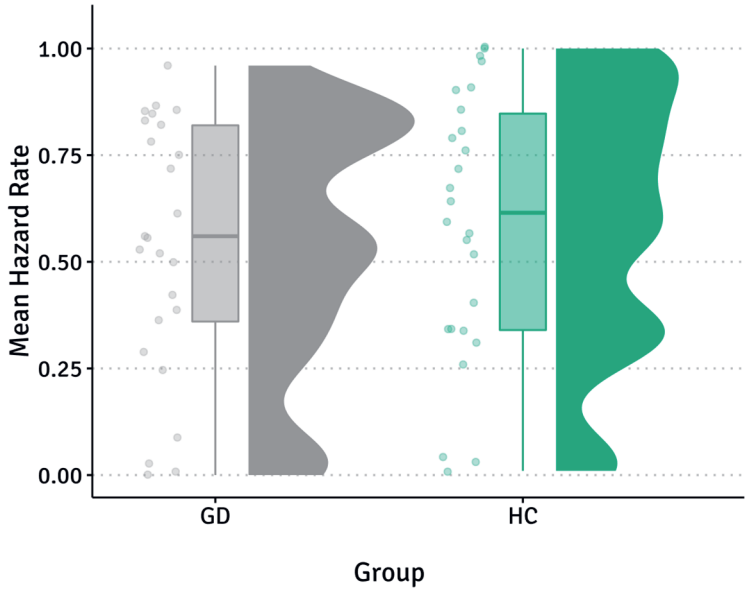


Figure F1: Hazard rate per group. Dots show data from individual participants, boxplots show median and upper/lower quantile with whiskers indicating the 1.5 interquartile range, distributions show the probability density function of all data points per group. HC = healthy control subjects, GD = gambling disorder.

High Volatility Analyses

A subset of our sample additionally performed the predictive inference task using a higher hazard rate H of 0.20, for which we performed similar exclusion criteria and analyses as were developed in the main manuscript. The final sample consisted of 24 GD and 15 HC participants.

First, no group differences were found in the perceived hazard rate (Mean HC: 0.68, mean GD: 0.71, $t_{37} = 0.33$, p=0.741). Again, no effects of group on confidence (β = -1.86 (6.13), t = -0.30, p=0.764) or action update (β = 0.44 (1.96), t = 0.23, p=0.822) were found. Here, no evidence was found of a weaker action-confidence coupling in GD (interaction effect: β = 2.85 (1.84), t = 1.55, p=0.122), and again no correlations were found between action-confidence coupling and PGSI score (r = 0.14, p = 0.503) or GBQ score (r = 0.14, p=0.507) in the GD sample.

No evidence was found for an overall group difference in learning rate (β = -0.10 (0.09), t = -1.10, p=0.278), however in post-hoc analyses it was shown that GD patients have higher learning rate particularly when prediction error was low (HC-GD Z-ratio = 2.307,

p = 0.021). Regarding the Bayesian model-based analyses, no group differences were found on the effects of the latent parameters on either action or confidence.

Comparing High to Low Volatility

Using the subset of our sample that completed the original predictive inference task (termed here: low volatility task), as well as the high volatility task, we aimed to compare behavior between the groups on the two different variations of the task. After applying similar exclusion criteria, the final sample consisted of 23 GD and 14 HC participants.

First of all we compared perceived hazard rates between groups and between task type using a mixed ANOVA. As expected, perceived hazard rate was higher was higher in the high volatility versus low volatility task ($F_{1,35}$ = 18.66, p<.001), but this did not differ between groups ($F_{1,35}$ = 0.0008, p = 0.978), and no interaction between group and task type was found ($F_{1,35}$ = 1.75, p = 0.195).

Then, using mixed-models, the effects of group and task type and their interaction on confidence, action update and learning rate were tested. The models showed significant effects of task type on confidence (β = 1.33 (0.48), t = 2.79, p=0.005),on action update (β = -8.51 (0.89), t = -9.56, p<.001), and on learning rate (β = -0.16 (0.01), t = -12.34, p<.001) with lower confidence, higher action update and higher learning rates in the high volatility task. However, no interactions between group and task type were found on confidence (β = 0.26 (0.60), t = 0.43, p=0.671), on action update (β = -1.31 (1.13), t = -1.16, p=0.246), or on learning rate (β = -0.03 (0.02), t = -1.58, p=0.115).

Using a mixed-model with action update as dependent variable and fixed effects of confidence, group, task type and their interactions, we again showed a significant main effect of confidence (β = -9.20 (1.29), t = -7.15, p<.001), indicating the c0upling of action and confidence. Moreover, a significant main effect of task type was found, with more action updating in the high versus low volatility task (β = -8.51 (0.87), t = -9.77, p<.001). A significant interaction between group and confidence showed a weaker confidence coupling in GD (β = 3.29 (1.63), t = 2.01, p=0.044), however there was no evidence for a three-way interaction between group, confidence and task type, indicating that the weaker coupling of action and confidence in GD was not exaggerated further in a high volatility context (β = -0.05 (1.10), t = -0.04, p=0.965).

For the model-based analyses we used mixed-models to test the interaction between group, task type and the various latent parameters from the model on action and confidence. For the action model, no significant interactions were found. For the

confidence model, a significant interaction between group, task type and the hit parameter was found (β = -3.39 (0.74), t = -4.56, p<.001), which indicated that the effect of a hit on the previous trial on confidence was stronger in HC compared to GD when task volatility was low, but contrarily was stronger in GD compared to HC when the task volatility was high.

Overall, while increasing the volatility of the task directly impacts the amount of action updating, confidence and learning rates, there is little evidence that higher volatility of the task had different effects on task behavior in the GD versus the HC group.

Table F1: Results of linear mixed-effects models predicting the effects of computational variables and group on action and confidence in the original predictive inference task.

GD versus HC						
	Dependent Variable: Action			**Dependent Variable: Confidence**		
Predictors	***Beta (SE)***	***t***	***p***	***Beta (SE)***	***t***	***p***
PE	19.27 (2.82)	6.83	**<.001**	-0.75 (0.55)	-1.38	0.167
CPP	11.31 (2.58)	4.38	**<.001**	-2.21 (1.06)	-2.09	**0.042**
(1-CPP)* (1-MC)	2.09 (0.51)	4.14	**<.001**	-1.81 (0.74)	-2.43	**0.019**
Hit	-5.57 (0.50)	-11.13	**<.001**	3.91 (0.73)	5.32	**<.001**
PE * Group (GD)	-5.52 (4.04)	-1.37	0.178	0.40 (0.80)	0.50	0.619
CPP * Group (GD)	3.74 (3.69)	1.01	0.316	1.07 (1.52)	0.70	0.484
((1-CPP)* (1-MC)) * Group (GD)	1.72 (0.72)	2.37	**0.022**	0.59 (1.06)	0.56	0.581
Hit * Group (GD)	-1.16 (0.72)	-1.62	0.111	-0.96 (1.05)	-0.92	0.364

Appendix G

Supplement to Chapter 9

Supplementary Methods

Participants

HCs were recruited from our participant database that exists of HCs that participated in previous studies conducted in our lab. Participants were recruited online and tested in the lab. In line with the medical ethics committee approval and procedures, all invited participants received an information sheet describing what to expect from participating and explaining that the research concerned eye-tracking during a gambling game. Please note that all participants were naive to the mixed-gamble task before entering this study.

Task design

The mixed-gamble task consisted of mixed gambles that were always presented as a 50/50 chance of gaining or losing a specific value. Subjects were asked to decide between two options: rejecting (sure option) or accepting (gambling option) the gamble. The sure option always entailed opting for the initial endowment of €25 without the possibility of gaining or losing additional bonuses. The gambling option always entailed potentially gaining or losing additional bonuses, both with an equal probability of 50%. The amounts that participants chose over varied as follows: gain values ranged from 20 to 38 credits in steps of 2, and losses ranged from -13 to -27 credits in steps of 2. Each gamble constituted a combination of one gain value and one loss value. Gains and losses were orthogonalized, as all possible combinations were presented, resulting in independency between gains and losses. To incentivize participants, one of 160 trials would be chosen at random at the end of the experiment. If the gamble was rejected in this trial, subjects received the initial endowment. If the gamble was accepted on this trial, the outcome of the lottery was determined via a virtual coin flip. The outcome was converted to euros (1 credit = €0.185), and consequently added to the initial endowment if it was a gain, or subtracted if it was a loss.

Eye tracker set-up

The task was run using Presentation software (Version 18.0, Neurobehavioral Systems Inc, Berkeley, CA). Eye movement data was collected using the EyeLink 1000 desk-mounted eye-gaze tracking system (SR Research Ltd., Ottowa, Ontario), which uses infrared and corneal reflection techniques. The sampling rate was 500 Hz, and head position was kept stable using a head-chin rest. Participants' eyes were positioned 60 cm away from the monitor. The stimuli were presented on a Iiyama monitor (1920x1080). At the start of the experiment and before the start of each new block a 9-point calibration and validation was performed to ensure proper validity of the eye-tracking data throughout the experiment.

Data preparation and exclusions

Behavioral data

Data preparation and analyses were performed using Rstudio (RStudio version 1.4.1106). Subjects who did not show any variation in their choices were excluded (N=2). Moreover, all trials in which participants failed to make a decision within the time limit were excluded (41 trials in total). Three subjects experienced some technical issues, as a result of which their data was collected in two separate runs of 80 trials each.

Eye-tracking data

First, the quality of the validation was checked per block by inspecting the visual degrees in the central (960x540), left (115x540) and right (1804x540) calibration points, since the gamble stimuli only covered the central horizontal area of the monitor. Note that our experimental design separated gain and loss stimuli widely on the screen (by ca. 1700 pixels), allowing for a higher degree of error tolerance than with closely spaced areas of interest (AOIs) commonly used (Dalrymple et al., 2018). We therefore excluded blocks only when their average validation accuracy of a single calibration point exceeded 2°, or when the average of the validation accuracy of the three calibration points exceeded 1.5° (10 blocks in total, 8 subjects for 1 block and 1 subject for 2 blocks) .

The wide separation of stimuli together with high calibration accuracy allowed us to select large margins for our AOIs, reducing false negatives without the risk of increasing false positives since the AOIs by no means overlap (Kennedy, 2016; Orquin et al., 2016). Two rectangular AOIs with a margin of 150 pixels in each direction (approx. 4° visual

angle) were established around the gamble stimuli of 100x100 pixels. The AOIs were centered on (480,540) and (1440,540) for the left and right stimulus, respectively (Figure 1B). Moreover, an AOI was created for the fixation cross, centered on (960x540) with a margin of 100 pixels in each direction.

Finally, trials in which subjects made no fixations to the gain or loss AOIs were removed from further analyses (52 trials). Moreover, some trials did not contain eye-tracking data due to temporary data loss or closed eyes (12 trials).

Mixed-model specifications

For the analyses explained in the main text, we have built several mixed-models, of which detailed specifications can be found in Table G1.

Table G1: Specification of all mixed-effect models.

	Dependent Variable	Fixed Effects	Random Effects (per subject)
(A) Basic Choice Model	Choice	• Gain Value * Group • Loss Value * Group	• Intercept • Slope of Gain Value • Slope of Loss Value
(B) Extended Choice Model	Choice	• Gain Value * Group • Loss Value * Group • Confidence * Group	• Intercept • Slope of Gain Value • Slope of Loss Value • Slope of Confidence
(C) Confidence Model	Confidence	• Gain Value * Group • Loss Value * Group	• Intercept • Slope of Gain Value • Slope of Loss Value
(D) Dwell Time Gain Model	Dwell time on gains	• Gain Value * Group	• Intercept • Slope of Gain Value
(E) Dwell Time Loss Model	Dwell time on losses	• Loss Value * Group	• Intercept • Slope of Loss Value

(F)	**Relative Dwell Time Model**	Relative dwell time on gains versus losses	• Gain Value * Group • Loss Value * Group	• Intercept • Slope of Gain Value • Slope of Loss Value
(G)	**Choice x Attention Model**	Choice	• Dwell Time Gain * Gain Value * Group • Dwell Time Loss * Loss Value * Group • Confidence * Group	• Intercept • Slope of dwell time gain • Slope of Gain Value • Slope of dwell time loss • Slope of Loss Value • Slope of Confidence
(H)	**Choice x Attention Channels Model**	Choice	• Average Dwell Time Gain * Gain Value * Group • Trial-by-Trial Deviations Dwell Time Gain * Gain Value * Group • Average Dwell Time Loss * Loss Value * Group • Trial-by-Trial Deviations Dwell Time Loss * Loss Value * Group • Confidence * Group	• Intercept • Slope of Gain Value • Slope of Loss Value • Slope of Trial-by-Trial Deviations Dwell Time Gain • Slope of Trial-by-Trial Deviations Dwell Time Loss • Slope of Confidence

Presented are detailed specifications of all mixed-effect models used in the analyses. Reported are the dependent variable, fixed effects including interaction terms (all main effects of the variables in an interaction term were included as well) and random effects per subject that were specified in each model.

Correlational analyses

Exporatory correlations

Pearson's correlation tests were performed between various variables of interest. Gambling propensity and dwell time on gains/losses were correlated to gambling severity (in GD only), GBQ score (in GD only), BIS score and BAS score, and subscores where relevant.

Expected value

Expected value (EV) was calculated for each trial:

EV = gv*p1 + lv*p2

Where gv and lv represent the size of the potential gain and loss values, respectively, and p1 and p2 represent the probabilities of gain and loss (0.5 for both).

First, we investigated whether the EV of a gamble, taking in both loss and gain value and their 50% probability, would predict gambling choices, and whether this relationship differed between the groups. We used mixed-effects models were fit. For all further mixed models continuous independent variables were z-scored (EV, confidence, dwell times).

Second, to test for group differences in the effects of expected value on confidence a mixed-effects model was run with confidence as dependent variable, and fixed factors of expected value in interaction with group, random intercepts and random slopes of EV.

Third, to test for group differences in the effects of expected value on the relative dwell time towards gains versus losses, a mixed-effects model was run with relative dwell times as dependent variable, and fixed factors of expected value in interaction with group, random intercepts and random slopes of EV.

Finally, to test the influence of relative dwell times and EV on choice to gamble, and to test whether these effects differed per group, a final mixed-effects model was run with choice as dependent variable, and fixed factors of relative dwell time toward gains, EV and group and all interactions, together with random slopes and random effects of relative dwell times and EV per subject.

Supplementary Results

Correlational analyses

Gambling severity or gambling beliefs did not significantly correlate with gambling propensity or dwell times in GD. Across the whole sample, there was a significant negative correlation between BAS scores and top-down dwell time on gains ($r = -0.34$, $p = 0.01$) and between BAS scores and top-down dwell time on losses ($r = -0.36$, $p = 0.007$). When splitting up the groups, only correlations between BAS scores and dwell

time on gains (r = -0.46, p = 0.01) and dwell time on losses (r = -0.60 , p = 0.0008) stayed significant for the HC group, but not for the GD group.

Expected value analyses

First, results showed a strong significant effect of EV on choice to gamble, with higher EV leading to more gambling (Table G2A).

Our second model did not show a main effect of EV on confidence level (Table G2B). A trend-level interaction effect was found, which showed that patients with GD showed a more positive relationship between EV and confidence.

Table G2: Results of the mixed-effects model on choice behavior (A) and confidence (B).

A)		
Parameter	**Estimate (SE)**	**p-value**
Intercept	0.55 (0.37)	0.141
Expected Value	3.26 (0.24)	**<0.001**
Group (GD)	2.05 (0.74)	**0.005**
Expected Value x Group (GD)	0.67 (0.47)	0.151
B)		
Parameter	**Estimate (SE)**	**p-value**
Intercept	5.47 (0.10)	**<0.001**
Expected Value	0.05 (0.05)	0.315
Group (GD)	0.14 (0.20)	0.474
Expected Value x Group (GD)	0.18 (0.10)	*0.068*

A) Results of mixed-model: Choice ~ Expected Value x Group + (1 + Expected Value | Subject). **B)** Results of mixed-model: Confidence ~ Expected Value x Group + (1 + Expected Value | Subject). Shown are the estimates, their standard errors (SE) and 95% confidence intervals (CI), statistic and p-values. *p<0.05, **p<0.01, ***p<0.001.

With regards to relative dwell times to gains, we found a significant main effect of EV showing that subjects in general focus more attention towards gains versus losses during gambles with a higher EV (Table G3). No interaction effect with group was found.

Table G3: Results of the mixed-effects model on relative dwell times.

Parameter	**Estimate (SE)**	**p-value**
Intercept	0.06 (0.01)	**0.002**
Expected Value	0.04 (0.01)	**0.001**
Group (GD)	-0.02 (0.03)	0.572
Expected Value x Group (GD)	0.01 (0.02)	0.750

Results of model: Relative Dwell Time To Gains Versus Losses ~ Expected Value x Group + (1 + Expected Value | Subject). Shown are the beta estimates, their standard error (SE) and 95% confidence intervals (CI), statistic and p-values. *p<0.05, **p<0.01, ***p<0.001.

Finally, looking in to the effects of expected value, relative dwell time to gains, group and their interactions on choice behavior, results showed only main effects of expected value and group, but no interaction effects (Table G4).

Table G4: Results of the mixed-effects model on the influence of attentional measures on choice.

Parameter	Estimate (SE)	p-value
Intercept	0.50 (0.36)	0.168
Expected Value	3.26 (0.23)	**<0.001**
Group (GD)	2.04 (0.72)	**0.005**
Expected Value x Group (GD)	0.56 (0.46)	0.223
Relative Dwell Time x Expected Value	-0.10 (0.06)	0.092
Relative Dwell Time x Group (GD)	-0.10 (0.10)	0.339
Relative Dwell Time x Expected Value x Group (GD)	0.06 (0.11)	0.582

Results of model: Choice ~ Expected Value x Relative Dwell Time x Group + (1 + Expected Value + Relative Dwell Time | Subject). Shown are the beta estimates, their standard error (SE) and 95% confidence intervals (CI), statistic and p-values. *p<0.05, **p<0.01, ***p<0.001.

Appendix H

Supplement to Chapter 10

Supplementary Analyses

Previous work on the influence of value and risk-taking on confidence indicated that the well-documented negative relationship between confidence and reaction times was less strong when healthy subjects made risky versus certain choices (da Silva Castanheira et al., 2021). In our current work we tested these same relationships using mixed-effects models.

First, we tested these relationships within our sample of healthy controls. We adjusted the first model of our main paper to test whether the relationship between reaction times and confidence was different for the two choice types. To do so, we included an interaction between log reaction time and choice type:

Confidence ~ Choice Type * Expected Value + Choice Type * Reaction Time + (1 + Choice Type + Expected Value + Reaction Time | Subject)

Here we found trend level evidence for an interaction effect between reaction times and choice type (β = -0.04 +- 0.02, t = -1.89, p = 0.058), which indicated that the negative relationship between reaction time and confidence was less strong during risky choices (β = - 0.335 +- 0.077) than during certain choices (β = -0.408 +- 0.075).

Second, we also adjusted the second model of our main paper to include an interaction between reaction time and choice type within our sample of healthy controls:

Confidence ~ Choice Type * Gain Value + Choice Type * Loss Value + Choice Type * Reaction Time + (1 + Choice Type + Gain Value + Loss Value + Reaction Time | Subject)

We found evidence for a significant interaction effect between reaction times and choice type (β = -0.06 +- 0.02, t = -3.01, p = 0.003), indicating that the negative relationship between reaction time and confidence was less strong during risky choices (β = - 0.300 +- 0.077) than during certain choices (β = -0.413 +- 0.075).

Finally, we included the GD sample to explore whether the difference between choice types in the confidence-reaction time relationship was expressed differently in the groups by adding three-way interactions between choice type, reaction time and group. We did not find evidence for a significant three-way interaction.

Nederlandse samenvatting

De Innerlijke Spiegel: Een Neurocognitieve Benadering van Zekerheid en Metacognitie in de Psychiatrie

Introductie

Stel je voor dat je aan de kant van een donkere, mistige weg staat en je naar de overkant moet. Je vraagt je af of er een auto aankomt of niet. Komt het licht dat je ziet van de straatlantaarns, of toch van een naderende auto? Bij het maken van deze beslissing zul je af gaan op je gevoel van *zekerheid* over je overtuiging dat het licht van de straatlantaarns komt. Dit gevoel van zekerheid (ook wel '*confidence*') ontstaat bij vrijwel elke keuze die je maakt. Het is erg belangrijk dat de mate van zekerheid die je voelt in je beslissingen of overtuigingen overeenkomt met wat er daadwerkelijk gebeurt. Als je namelijk té zeker bent en de weg op stapt terwijl er eigenlijk een auto nadert, kunnen de gevolgen fataal zijn. Maar als je blijft aarzelen terwijl de kust veilig is, kom je nooit aan de overkant. Dit proces waarbij mensen nadenken over en reflecteren op hun eigen beslissingen, overtuigingen, acties en ideeën wordt *metacognitie* genoemd.

Metacognitie en zekerheid

Metacognitie wordt gedefinieerd als 'het denken over je eigen denken' (Flavell, 1979). Dit is iets wat we voortdurend doen (zowel bewust als onbewust) en vervolgens gebruiken om ons gedrag bij te sturen. Een vorm van metacognitie is '*confidence*', wat wordt gedefinieerd als het subjectieve gevoel van zekerheid dat je een correcte keuze hebt gemaakt of dat je idee juist is (Pouget et al., 2016). Hierbij wordt onderscheid gemaakt tussen '*metacognitieve monitoring',* wat het vermogen is om je gedrag te monitoren m.b.v. zekerheidsinschattingen, en '*metacognitieve controle*', wat het vermogen is om je zekerheid vervolgens te gebruiken om je gedrag te sturen.

Een goed werkend metacognitief systeem is van groot belang om je gedrag en keuzes zo goed mogelijk aan te passen in een wereld die vaak onvoorspelbaar is en voortdurend verandert. Als je ergens minder zeker over bent, zul je geneigd zijn om meer informatie te verzamelen, sta je open voor nieuwe ideeën, leer je sneller en verander je misschien zelfs van strategie en gedachten. Aan de andere kant, als je erg zeker bent van je zaak, sta je minder open voor feedback, stop je met informatie vergaren, en houd je vast aan je initiële keuze. Het gevoel van zekerheid fungeert dus eigenlijk als een innerlijke maatstaf die je gedrag beïnvloedt wanneer er geen externe feedback is, wat

vaak het geval is in het echte leven. Op deze manier speelt het een essentiële rol bij het sturen van optimaal gedrag.

Zekerheid in de psychiatrie

Wanneer je gevoel van zekerheid over de juistheid van je beslissingen niet overeenkomt met wat daadwerkelijk juist is, kan dit in extreme gevallen leiden tot problematisch gedrag zoals dwangmatig (ook wel: compulsief) gedrag. Compulsief gedrag kan verschillende vormen aannemen en wordt omschreven als het uitvoeren van herhaalde handelingen waarbij iemand het gevoel heeft dat ze 'moeten' worden uitgevoerd, terwijl ze zich er tegelijkertijd van bewust zijn dat deze handelingen niet in lijn zijn met hun overkoepelende doelen (Luigjes et al., 2019). Compulsiviteit is een karakteristiek van verschillende psychiatrische stoornissen, waaronder obsessieve-compulsieve stoornis (OCS) en verslaving (Figee et al., 2016). Je kan je voorstellen dat iemand die te weinig zekerheid (‘*underconfidence*’) heeft in het goed op slot doen van de deur, de deur obsessief blijft controleren, zoals voorkomt bij patiënten met OCS. Aan de andere kant kan te veel vertrouwen hebben (‘*overconfidence*’), bijvoorbeeld denken dat je veel vaker zult winnen bij roulette dan je eigenlijk zal, gepaard gaan met dwangmatig gokgedrag bij mensen met een gokverslaving.

Dit proefschrift onderzoekt de vraag of en hoe het gevoel van zekerheid verstoord kan zijn in relatie tot verschillende psychiatrische symptomen en aandoeningen. Hiervoor heb ik gebruik gemaakt van een neurocognitieve benadering. In essentie houdt deze benadering in dat ik kwantitatief experimenteel onderzoek heb uitgevoerd met verschillende groepen mensen, zowel psychiatrische patiënten als controle proefpersonen zonder psychiatrische symptomen. Hierbij heb ik zekerheid bestudeerd met behulp van cognitieve gedragstaken die metacognitieve vaardigheden meten, 'eye-tracking'-technologie om aandacht te meten, en medische beeldvormingstechnieken zoals functionele MRI om het functioneren van de hersenen in kaart te brengen.

Het meten van zekerheid

Er bestaan verschillende methodes om zekerheid te meten. Meestal wordt de deelnemer gevraagd om een keuze te maken, bijvoorbeeld: ‘in welk van deze twee vierkanten zag je de meeste stippen, links of rechts?’, waarna de deelnemer op een schaal moet aangeven hoe zeker ze is dat die specifieke keuze goed is. Dit wordt ‘*lokale* zekerheid’ genoemd. Hiermee kunnen we zien hoe zeker iemand gemiddeld is over haar keuzes, en of die zekerheid overeenkomt met hoe juist haar keuzes eigenlijk waren. ‘Overconfidence’ (te zeker zijn) treedt op wanneer iemand gemiddeld veel zekerder is van haar keuzes dan dat haar werkelijke prestaties rechtvaardigt. ‘Underconfidence’ (te onzeker zijn) daarentegen gebeurt wanneer iemand juist minder zekerheid toont in haar

keuzes, terwijl ze eigenlijk beter presteert dan ze denkt. We kunnen ook 'metacognitieve sensitiviteit' meten, wat aangeeft hoe goed iemand kan beoordelen of ze goed of fout zaten op basis van hun zekerheidsinschattingen. Normaal gesproken zouden mensen minder zeker moeten zijn wanneer ze een fout maken en zekerder wanneer ze correcte keuzes maken. Hoe groter het verschil is tussen de zekerheid die je hebt bij foute en correcte keuzes, des te beter is de metacognitieve sensitiviteit.

Naast 'lokale' zekerheid over specifieke keuzes, bestaat zekerheid ook op hogere niveaus in een hiërarchie. Zo kun je ook zekerheid voelen over hoe goed je bent in een bepaalde taak, of bijvoorbeeld over hoe sterk je geheugen is – dit noemen we '*globale zekerheid*'. Daarnaast is er een nog hogere vorm van zekerheid die zich uitstrekt over verschillende aspecten van je leven en betrekking heeft op jezelf, zoals zelfvertrouwen, zelfwaardering en autonomie. Tezamen noemen we dit hoogste niveau '*self-beliefs*'.

Resultaten

Zekerheidsafwijkingen in de psychiatrie

Toen ik aan mijn promotieonderzoek begon, heb ik eerst systematisch al het onderzoek in kaart gebracht dat zich bezighield met zekerheid in de psychiatrie, beschreven in **Hoofdstuk 2**. Ik vond overtuigend bewijs voor verbanden tussen specifieke psychiatrische symptomen en afwijkingen in zekerheid. Er werd een verband gevonden tussen obsessieve-compulsieve symptomen en verminderde zekerheid. Daarentegen bleek bij schizofrenie een neiging naar overmatige zekerheid te bestaan, vooral wanneer patiënten eigenlijk fouten maakten. Mensen met een gokverslaving bleken een te hoge mate van zekerheid te hebben, terwijl bij depressie en angststoornissen een lager niveau van zekerheid werd waargenomen.

De neurobiologische basis van zekerheid en de rol van financiële prikkels

Er is veel onderzoek gedaan naar de neurobiologische basis van zekerheid en er is een sterke consensus dat twee gebieden van de hersenen, de ventromediale prefrontale cortex (vmPFC) en het ventrale striatum (VS), een belangrijke rol spelen (Fleming & Dolan, 2012; Vaccaro & Fleming, 2018). Precies dezelfde gebieden zijn ook betrokken bij het verwerken van beloningen en de waarde die dingen hebben (Bartra et al., 2013; Lebreton et al., 2009). De mate van zekerheid die je ervaart kan ook worden beïnvloed door externe factoren, zoals het vooruitzicht om geld te winnen or verliezen (Lebreton et al., 2018). In **Hoofdstuk 3** bouwde ik voort op deze eerdere bevindingen en toonde ik aan dat mensen doorgaans zekerder zijn als ze iets kunnen winnen dan als ze iets

kunnen verliezen, terwijl hun daadwerkelijke prestaties hetzelfde zijn. Verder ontdekte ik dat de activiteit in de vmPFC alleen gerelateerd was aan het gevoel van zekerheid in situaties waarin mogelijk iets gewonnen kon worden, maar niet wanneer er mogelijk iets te verliezen was of als er niks op het spel stond. Hiermee toonde ik aan dat motivatieprocessen, zoals de kans om geld te winnen/verliezen, invloed hebben op het gevoel van zekerheid, en op de hersensignalen die hiermee te maken hebben.

Vervolgens heb ik deze processen onderzocht bij patiënten met OCS en patiënten met een gokverslaving. OCS wordt gekenmerkt door verontrustende dwanggedachten (obsessies) en repetitieve handelingen (compulsies), en treft ongeveer 2-3% van de bevolking (American Psychiatric Association, 2013). Gokverslaving wordt gekenmerkt door dwangmatig gokgedrag wat tot ernstige problemen leidt, en treft tot wel 5.8% van de wereldbevolking (American Psychiatric Association, 2013). Hoewel beide stoornissen gekenmerkt worden door compulsief gedrag, zijn er ook grote verschillen. Terwijl OCS gepaard gaat met gevoeligheid voor verlies en risicomijdend gedrag, wordt gokverslaving juist gekenmerkt door gevoeligheid voor winst en risico zoekend gedrag (Clark et al., 2019; Shephard et al., 2021). In **Hoofdstuk 4** testte ik onze hypothese dat patiënten met gokverslaving extra zeker zouden zijn, maar vooral wanneer er iets te winnen viel, terwijl patiënten met OCS juist extra onzeker zouden zijn, en vooral wanneer er iets te verliezen viel. Ik vond dat zekerheid significant hoger was bij patiënten met een gokverslaving vergeleken met patiënten met OCS en gezonde controles, vooral wanneer er iets te winnen was. Ik vond geen bewijs voor lagere zekerheid in patiënten met OCS vergeleken met gezonde controles. Ook waren er geen verschillen tussen de groepen in hersenactiviteit gerelateerd aan zekerheid.

Transdiagnostische benadering van zekerheid

Twee patiënten met dezelfde diagnose kunnen heel verschillende symptomen vertonen doordat er veel variabiliteit is in symptomen, en co-morbiditeit (het hebben van meer dan één diagnose) veel voorkomt. Er wordt hierdoor steeds meer gepleit voor een ‘transdiagnostische’ benadering. Bij deze benadering worden neurocognitieve processen gekoppeld aan symptomen die over diagnosegrenzen heen gaan, in plaats van te focussen op specifieke stoornissen. In **Hoofdstuk 5** heb ik deze benadering gebruikt om de verschillende niveaus van zekerheid (lokale zekerheid, globale zekerheid en ‘self-beliefs’) te onderzoeken. Ik heb gevonden dat de verschillende niveaus van zekerheid onderling positief samenhangen. Ook hadden degenen die hoog scoren op een transdiagnostische symptoomdimensie van angst en depressie een significant lagere zekerheid (‘underconfidence’). Degenen die hoog scoorden op transdiagnostische symptomen van compulsief gedrag en indringende gedachten toonden juist te hoge zekerheid (‘overconfidence’). Deze patronen waren minder

duidelijk wanneer ik specifieke diagnoses in plaats van transdiagnostische symptomen gebruikte. Dit wijst erop dat verschillende transdiagnostische symptoomdimensies samen gaan met specifieke afwijkingen in zekerheid.

Het vergelijken van zekerheid tussen patiënten en de algemene populatie

In de psychiatrie wordt vaak onderzoek gedaan door ofwel patiëntengroepen te vergelijken met gezonde mensen, ofwel met behulp van grootschalig onderzoek met grote groepen mensen uit de algemene bevolking die een hoge score hebben op bepaalde symptomen. Bij dit soort 'algemene populatie' onderzoeken wordt aangenomen dat mensen in de algemene bevolking met hoge scores op bijvoorbeeld OCS-symptomen vergelijkbaar zijn met klinische OCS-patiënten, zowel qua symptomen als cognitieve processen (Abramowitz et al., 2014). Maar in feite worden deze groepen bijna nooit direct met elkaar vergeleken. In **Hoofdstuk 6** heb ik daarom onderzocht hoe zekerheid op verschillende niveaus verschilt tussen klinische OCS-patiënten, een hoog-compulsieve groep uit de algemene bevolking (met even hoge OCS-scores als de patiënten maar zonder officiële diagnose), en mensen zonder psychische symptomen. Hierbij zou je op basis van de aanname verwachten dat de zekerheid van de patiëntengroep met een officiële diagnose en de hoog-compulsieve groep erg op elkaar lijkt. Ik heb daarentegen gevonden dat klinische OCS patiënten en de hoog-compulsieve groep juist verschillende patronen van zekerheid vertonen. Patiënten hadden significant lagere lokale en globale zekerheid dan de controlepersonen, terwijl de hoog-compulsieve groep juist significant méér zekerheid toonde vergeleken met de patiëntengroep. Dit benadrukt dat voorzichtigheid is gebonden bij het generaliseren van conclusies over (meta)cognitie en zekerheid van de hoog-compulsieve groep naar klinische patiënten, omdat ze niet één-op-één vergelijkbaar zijn.

De koppeling tussen zekerheid en leren

In **Hoofdstuk 7** heb ik onderzocht hoe zekerheid invloed heeft op leergedrag in een veranderlijke omgeving bij klinische OCS-patiënten, een hoog-compulsieve groep uit de algemene bevolking, en mensen zonder psychische symptomen. Gewoonlijk is iemand die erg zeker is van haar keuze minder ontvankelijk voor het leren van nieuwe informatie en minder geneigd zich aan te passen. Dit wijst op een sterke koppeling tussen zekerheid en acties die iemand onderneemt. Opnieuw toonden de resultaten dat de patiëntengroep en de hoog-compulsieve groep verschillende zekerheidspatronen laten zien. OCS patiënten hadden lagere zekerheid en waren gevoeliger voor fouten in vergelijking met zowel de controlegroep als de hoog-compulsieve groep. Er was geen bewijs voor een verstoring in de koppeling tussen

zekerheid en acties in OCS. Deze bevindingen tonen aan dat OCS symptomen samen kunnen gaan met diverse patronen van zekerheid en leergedrag, afhankelijk van de onderzochte groep mensen (klinisch of algemene bevolking).

In **Hoofdstuk 8** heb ik opnieuw de relatie tussen zekerheid en leren onderzocht, maar nu bij patiënten met een gokverslaving vergeleken met controlepersonen. De zekerheid verschilde niet tussen de groepen, maar er was wel een verzwakte koppeling tussen zekerheid en acties in patiënten met een gokverslaving. Dit wijst erop dat de patiëntengroep hun gevoel van zekerheid minder in beschouwing nemen bij het uitvoeren van hun acties. Bovendien toonde de patiëntengroep een lager leertempo, wat suggereert dat ze minder vatbaar zijn voor nieuwe informatie vergeleken met de controlegroep.

Aandacht en zekerheid

Bij het nemen van beslissingen speelt naast metacognitie ook aandacht een rol (Orquin & Mueller Loose, 2013). Patiënten met gokverslaving vertonen vaak een 'aandachts*bias*', waarbij hun aandacht sterker getrokken wordt door gokgerelateerde zaken. In **Hoofdstuk 9** heb ik de rol van aandacht en zekerheid bij beslissingen om te gokken onderzocht met behulp van eye-tracking. Hiermee kunnen we nauwkeurig volgen waar mensen naar kijken, en dus waar hun aandacht naartoe gaat. Mijn bevindingen toonden aan dat de patiënten met gokverslaving vaker gokken, minder gevoelig zijn voor mogelijke verliezen en meer zekerheid hebben in hun gokkeuzes naarmate er meer geld te winnen valt. Hoewel ik geen bewijs heb gevonden voor een specifieke aandachts*bias* gericht op mogelijke winst of verlies bij gokkeuzes in de gokgroep, ontdekte ik wel dat specifiek diegenen in de gokgroep die een hogere aandacht voor potentiële winst hadden ook sterker beïnvloed werden door de hoogte van die winst en daardoor eerder geneigd waren om te gaan gokken.

Risico en zekerheid

Risico en zekerheid werden in **Hoofdstuk 10** onderzocht bij patiënten met gokverslaving. De controlegroep toonde meer zekerheid bij het maken van veilige keuzes dan bij risicovolle gokkeuzes, terwijl dit bij de gokgroep andersom was. Terwijl iedereen minder zeker werd over hun keuze niet te gokken naarmate de hoogte van de potentiële winst toenam, was dit effect extra sterk bij de gokgroep. Dit suggereert dat de gokgroep een sterker gevoel heeft van 'spijt' dat ze een kans om te gokken voorbij hebben laten gaan naarmate er meer te winnen viel, doordat hun zekerheid in deze keuzes sterker achteruit ging. Dit zou de neiging tot excessief gokken kunnen versterken.

Discussie

In **Hoofdstuk 11** heb ik de bevindingen van dit proefschrift in de bredere context van de literatuur beschreven en een aantal kanttekeningen, limitaties en klinische implicaties benoemd. Samenvattend heb ik in dit proefschrift bewijs gevonden dat verstoringen in zekerheid een centraal aspect zijn binnen de geestelijke gezondheid.

Overkoepelende bevindingen

Specifiek bleek uit mijn onderzoek dat patiënten met OCS een lagere zekerheid en 'underconfidence' ervaren, zowel op lokaal als globaal niveau, vooral bij patiënten die geen medicatie gebruikten. Tegelijkertijd heb ik bewijs gevonden voor verhoogde zekerheid en 'overconfidence' in patiënten met gokverslaving, vooral in situaties gerelateerd aan gokken, winstmogelijkheden en risico. Deze bevinding sluit aan bij theorieën die benadrukken dat context en prikkels die met verslaving te maken hebben een grote invloed hebben op pathologisch gedrag bij gokverslaving (Genauck et al., 2020; Leyton & Vezina, 2012; Perales et al., 2020)

De transdiagnostische en hiërarchische benadering van zekerheid

Ik heb ook de relevantie van zowel de transdiagnostische als de hiërarchische benadering van zekerheid besproken. Mijn resultaat van de positieve verbanden tussen de verschillende niveaus van zekerheid bekrachtigen recente theorieën (Seow et al., 2021). De hiërarchische benadering van zekerheid is belangrijk aangezien het de kloof tussen experimenteel onderzoek (dat vooral lokale zekerheid m.b.v. cognitieve taken bestudeert) en het dagelijks leven van patiënten (waar hogere zekerheidsniveaus relevanter zijn) kan overbruggen. Door de hiërarchische en transdiagnostische benadering te combineren, heb ik laten zien dat de zekerheidsniveaus verschillend gerelateerd zijn aan de transdiagnostische symptoomdimensies. Dit heeft belangrijke implicaties voor mogelijke therapieën, aangezien recent onderzoek heeft aangetoond dat verschillende onderliggende mechanismen ten grondslag liggen aan de specifieke verstoringen van zekerheid die we zien bij de transdiagnostische symptomen (Katyal et al., 2023).

Zekerheid in klinische patiënten versus compulsiviteit in de algemene bevolking

Ook heb ik de vergelijkbaarheid en generaliseerbaarheid van zekerheidsverstoringen tussen patiëntgroepen met OCS en hoog-compulsieve groepen in de algemene bevolking besproken. Dit proefschrift heeft verschillende gedragsprofielen onthuld wat betreft zekerheid en leergedrag in deze groepen, ondanks de gelijke ernst van OCS symptomen tussen de groepen. Ik heb deze resultaten in het licht van een recent

theoretisch model besproken (Fradkin et al., 2020), dat suggereert dat obsessief-compulsief gedrag enerzijds kan ontstaan door een gebrek aan zekerheid, leidend tot overactieve reacties op feedback en compulsief inflexibel gedrag, en anderzijds door een teveel aan zekerheid in rigide ideeën en te weinig invloed van feedback, resulterend in habitueel inflexibel gedrag.

Het is erg aannemelijk dat deze groepen, ondanks de vergelijkbare psychologische onrust als gevolg van de symptomen, verschillen in de mate waarin de symptomen functionele beperkingen veroorzaken en impact hebben op het dagelijks leven, wat waarschijnlijk ernstiger is in klinische patiënten (Abramovitch et al., 2023). Daarnaast kunnen de groepen verschillen in de mate waarin ze andere symptomen buiten OCS ervaren. De hoog-compulsieve groep kan bijvoorbeeld meer symptomen van impulsiviteit, schizotypie en verslaving ervaren, terwijl de OCS groep wellicht meer last heeft van angst en depressie symptomen.

Klinische implicaties

Afwijkingen in zekerheid hebben aanzienlijke implicaties. Als je onterecht te weinig zekerheid hebt in je eigen vermogen, kan dat ook je zelfvertrouwen, motivatie en leervermogen schaden. Aan de andere kant, onterecht hoge zekerheid in je eigen vermogen kan samen gaan met risicovol gedrag, rigide overtuigingen en dogmatisme.

Hoewel mijn onderzoek interessante mogelijkheden voor therapie oplevert, staan de directe klinische toepassingen binnen dit veld nog in de kinderschoenen. Dit komt onder andere door de experimentele aard van het onderzoek dat nog ver verwijderd is van de dagelijkse realiteit, maar ook door grote individuele variabiliteit in metacognitieve vaardigheden wat een universele therapie lastig maakt. Daarnaast zijn er verschillen tussen stoornissen in de mate waarin afwijkingen in zekerheid centraal staan. Bovenal is zekerheid slechts één puzzelstukje in de multifactoriële complexe structuur van psychiatrische stoornissen. Ook is mijn onderzoek niet zonder limitaties. De gebruikte taken zijn kunstmatig en mijn onderzoek is cross-sectioneel (dat wil zeggen, op één moment in de tijd gemeten) waardoor ik weinig kan zeggen over oorzaak-gevolg relaties over de tijd heen. Ook zijn er methodologische limitaties wat betreft het meten van zekerheid, en heeft dit onderzoek ook een beperkte focus.

Potentiële klinische implicaties worden besproken, zoals positieve feedback interventies gefocust op hogere niveaus van zekerheid (Katyal et al., 2023; Van Marcke et al., 2022), en interventies die specifiek gericht zijn op het kalibreren van zekerheid in stoornis-specifieke context. Voor gokverslaving worden implicaties besproken die zich focussen op spijtgevoeligheid en aandachtstraining.

De toekomst van het veld

Ook heb ik diverse suggesties en aandachtspunten voor toekomstig onderzoek voorgesteld. Er is behoefte aan grootschaligere klinische studies met een longitudinale aanpak (over de tijd heen) om de temporele dynamiek van zekerheid en (transdiagnostische) symptomen beter te begrijpen. Daarnaast moeten we zekerheid beter bestuderen in specifieke symptoomcontext. Ook is het cruciaal om dit onderzoek in te bedden in het bredere perspectief van metacognitie, zoals de sociale en interpersoonlijke functies van zekerheid. Verder zouden uitgebreidere computationele modellen ons begrip van de onderliggende processen van zekerheidsvorming en verstoringen verfijnen.

In ons lopende onderzoek, dat niet in dit proefschrift is opgenomen, bouw ik voort op enkele van deze ideeën. Ik doe momenteel onderzoek naar het effect van stoornis-specifieke context op zekerheid en naar het proces van ‘changes of mind’ in OCS. Daarnaast onderzoek ik hoe afwijkingen in leerprocessen samenhangen met afwijkingen in zekerheid bij gokverslaving, en duik ik dieper in de relatie tussen zekerheid en dopamine in het brein met behulp van PET (positron emissie topografie) scans. Met mijn werk hoop ik een waardevolle bijdrage te hebben geleverd aan een dieper begrip van onze 'innerlijke spiegel', zowel in de context van gezondheid als in de psychiatrie.

Bibliography

Abbott, M. W. (2020). The changing epidemiology of gambling disorder and gambling-related harm: public health implications. *Public Health*, *184*, 41–45.

Abitbol, R., Lebreton, M., Hollard, G., Richmond, B. J., Bouret, S., & Pessiglione, M. (2015). Neural mechanisms underlying contextual dependency of subjective values: Converging evidence from monkeys and humans. *Journal of Neuroscience*, *35*(5), 2308–2320.

Abramovitch, A., Abramowitz, J. S., Riemann, B. C., & McKay, D. (2020). Severity benchmarks and contemporary clinical norms for the Obsessive-Compulsive Inventory-Revised (OCI-R). *Journal of Obsessive-Compulsive and Related Disorders*, *27*(7), 1–8.

Abramovitch, A., Robinson, A., Buckley, M. J., Çek, D., de Putter, L., & Timpano, K. R. (2023). Are student cohorts with psychopathology representative of general clinical populations? The case for OCD. *Journal of Obsessive-Compulsive and Related Disorders*, *37*, 1–7.

Abramowitz, J. S., Fabricant, L. E., Taylor, S., Deacon, B. J., McKay, D., & Storch, E. A. (2014). The relevance of analogue studies for understanding obsessions and compulsions. *Clinical Psychology Review*, *34*(3), 206–217.

Admon, R., Bleich-Cohen, M., Weizmant, R., Poyurovsky, M., Faragian, S., & Hendler, T. (2012). Functional and structural neural indices of risk aversion in obsessive-compulsive disorder (OCD). *Psychiatry Research - Neuroimaging*, *203*(2–3), 207–213.

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386.

Albertella, L., Chamberlain, S. R., Le Pelley, M. E., Greenwood, L. M., Lee, R. S. C., Den Ouden, L., Segrave, R. A., Grant, J. E., & Yücel, M. (2020). Compulsivity is measurable across distinct psychiatric symptom domains and is associated with familial risk and reward-related attentional capture. *CNS Spectrums*, *25*(4), 519–526.

Allen, M., Frank, D., Samuel Schwarzkopf, D., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *ELife*, *5*, 1–17.

Allen, M., Glen, J. C., Müllensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., Callaghan, M. F., & Rees, G. (2017). Metacognitive ability correlates with hippocampal and prefrontal microstructure. *NeuroImage*, *149*, 415–423.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

Anderson, B. A. (2016). What is abnormal about addiction-related attentional biases? *Drug and Alcohol Dependence*, *167*, 8–14.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371.

Anselme, P., & Robinson, M. J. F. (2020). From sign-tracking to attentional bias: Implications for gambling and substance use disorders. In *Progress in Neuro-Psychopharmacology and Biological Psychiatry* (Vol. 99, pp. 1–10). Elsevier.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407.

Ariel, R., Lembeck, N. A., Moffat, S., & Hertzog, C. (2018). Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence*, *70*, 42–51.

Armstrong, T., Rockloff, M., & Browne, M. (2020). Gamble with Your Head and Not Your Heart: A Conceptual Model for How Thinking-Style Promotes Irrational Gambling Beliefs. *Journal of Gambling Studies*, *36*(1), 183–206.

Ashbaugh, A. R., & Radomsky, A. S. (2007). Attentional focus during repeated checking influences memory but not metamemory. *Cognitive Therapy and Research*, *31*(3), 291–306.

Bachrach, N., Bekker, M. H. J., & Croon, M. A. (2013). Autonomy-Connectedness and Internalizing-Externalizing Personality Psychopathology, Among Outpatients. *Journal of Clinical Psychology*, *69*(7), 718–726.

Bacon, E., Danion, J. M., Kauffmann-Muller, F., & Bruant, A. (2001). Consciousness in schizophrenia: A metacognitive approach to semantic memory. *Consciousness and Cognition*, *10*(4), 473–484.

Bacon, E., & Izaute, M. (2009). Metacognition in Schizophrenia: Processes Underlying Patients' Reflections

on Their Own Episodic Memory. *Biological Psychiatry*, *66*(11), 1031–1037.
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. *Journal of Neuroscience*, *33*(42), 16657–16665.
Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1–11.
Banca, P., Vestergaard, M. D., Rankov, V., Baek, K., Mitchell, S., Lapa, T., Castelo-Branco, M., & Voon, V. (2015). Evidence Accumulation in Obsessive-Compulsive Disorder: The Role of Uncertainty and Monetary Reward on Perceptual Decision-Making Thresholds. *Neuropsychopharmacology*, *40*, 1192–1202.
Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.
Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087.
Bangasser, D. A., & Cuarenta, A. (2021). Sex differences in anxiety and depression: circuits and mechanisms. *Nature Reviews Neuroscience*, *22*(11), 674–684.
Baptista, A., Maheu, M., Mallet, L., & N'Diaye, K. (2021). Joint contributions of metacognition and self-beliefs to uncertainty-guided checking behavior. *Scientific Reports*, *11*(1), 1–10.
Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, *133*(1), 1–24.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.
Barthelmé, S., & Hurst, A. (2021). *Getting Started With eyelinker*. https://cran.r-project.org/web/packages/eyelinker/vignettes/basics.html
Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48.
Baxter, A. J., Vos, T., Scott, K. M., Ferrari, A. J., & Whiteford, H. A. (2014). The global burden of anxiety disorders in 2010. *Psychological Medicine*, *44*(11), 2363–2374.
Beck, A. (2003). Cognitive models of depression. *Clinical Advances in Cognitive Psychotherapy: Theory and Application*, *14*(1), 29–61.
Bekker, M. H. J., & Belt, U. (2006). The role of autonomy–connectedness in depression and anxiety. *Depression and Anxiety*, *23*(5), 274–280.
Bekker, M. H. J., & Croon, M. A. (2010). The roles of autonomy-connectedness and attachment styles in depression and anxiety. *Journal of Social and Personal Relationships*, *27*(7), 908–923.
Ben Shachar, A., Lazarov, A., Goldsmith, M., Moran, R., & Dar, R. (2013). Exploring metacognitive components of confidence and control in individuals with obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 255–261.
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), 1–17.
Benzina, N., N'Diaye, K., Pelissolo, A., Mallet, L., & Burguière, E. (2021). A cross-species assessment of behavioral flexibility in compulsive disorders. *Communications Biology*, *4*(1), 1–12.
Bergamin, J., Hoven, M., Nevick, B., van Holst, R., Denys, D., Bockting, C., & Luigjes, J. (2023). Development and Validation of the Amsterdam Autonomy Scale. *Submitted*.
Bergamin, J., Luigjes, J., Kiverstein, J., Bockting, C. L., & Denys, D. (2022). Defining Autonomy in Psychiatry. *Frontiers in Psychiatry*, *13*, 1–11.
Berner, E. S., & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5), 2–23.
Bey, K., Lennertz, L., Riesel, A., Klawohn, J., Kaufmann, C., Heinzel, S., Grützmann, R., Kathmann, N., & Wagner, M. (2017). Harm avoidance and childhood adversities in patients with obsessive-compulsive disorder and their unaffected first-degree relatives. *Acta Psychiatrica Scandinavica*, *135*(4), 328–338.
Bey, K., Weinhold, L., Grützmann, R., Heinzel, S., Kaufmann, C., Klawohn, J., Riesel, A., Lennertz, L., Schmid, M., Ramirez, A., Kathmann, N., & Wagner, M. (2020). The polygenic risk for obsessive-

compulsive disorder is associated with the personality trait harm avoidance. *Acta Psychiatrica Scandinavica*, *142*(4), 326–336.
Bhatt, R., Laws, K. R., & McKenna, P. J. (2010). False memory in schizophrenia patients with and without delusions. *Psychiatry Research*, *178*(2), 260–265.
Black, D. W., & Shaw, M. (2019). The epidemiology of gambling disorder. *Gambling Disorder*, 29–48.
Bodor, D., Ricijaš, N., & Filipčić, I. (2021). Treatment of gambling disorder: review of evidence-based aspects for best practice. *Current Opinion in Psychiatry*, *34*(5), 508–513.
Boffo, M., Willemen, R., Pronk, T., Wiers, R. W., & Dom, G. (2017). Effectiveness of two web-based cognitive bias modification interventions targeting approach and attentional bias in gambling problems: Study protocol for a pilot randomised controlled trial. *Trials*, *18*(452), 1–13.
Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *5*(1), 1–12.
Boldt, A., & Yeung, N. (2015). Shared Neural Markers of Decision Confidence and Error Detection. *Journal of Neuroscience*, *35*(8), 3478–3484.
Boog, M., Höppener, P., Ben, B. J. M., Goudriaan, A. E., Boog, M. C., & Franken, I. H. A. (2014). Cognitive inflexibility in gamblers is primarily present in reward-related decision making. *Frontiers in Human Neuroscience*, *8*(569), 1–6.
Bora, E., Yücel, M., & Pantelis, C. (2010). Cognitive impairment in schizophrenia and affective psychoses: implications for DSM-V criteria and beyond. *Schizophrenia Bulletin*, *36*(1), 36–42.
Borkowski, J. G., Carr, M., Rellinger, E., & Pressley, M. (1990). Self-regulated cognition: Interdependence of metacognition, attributions, and self-esteem. *Dimensions of Thinking and Cognitive Instruction*, 53–92.
Boschen, M. J., & Vuksanovic, D. (2007). Deteriorating memory confidence, responsibility perceptions and repeated checking: Comparisons in OCD and control samples. *Behaviour Research and Therapy*, *45*(9), 2098–2109.
Bowie, C. R., & Harvey, P. D. (2006). Cognitive deficits and functional outcome in schizophrenia Profile of cognitive impairments in schizophrenia. *Neuropsychiatric Disease and Treatment*, *2*(4), 531–536.
Brand, M., Kalbe, E., Labudda, K., Fujiwara, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-making impairments in patients with pathological gambling. *Psychiatry Research*, *133*(1), 91–99.
Brevers, D., Cleeremans, A., Bechara, A., Greisen, M., Kornreich, C., Verbanck, P., & Noël, X. (2013). Impaired Self-Awareness in Pathological Gamblers. *Journal of Gambling Studies*, *29*, 119–129.
Brevers, D., Cleeremans, A., Bechara, A., Greisen, M., Kornreich, C., Verbanck, P., & Noël, X. (2014). Impaired Metacognitive Capacities in Individuals with Problem Gambling. *Journal of Gambling Studies*, *30*(1), 141–152.
Brevers, D., Cleeremans, A., Bechara, A., Laloyaux, C., Kornreich, C., Verbanck, P., & Noël, X. (2011). Time Course of Attentional Bias for Gambling Information in Problem Gambling. *Psychology of Addictive Behaviors*, *25*(4), 675–682.
Brevers, D., Cleeremans, A., Goudriaan, A. E., Bechara, A., Kornreich, C., Verbanck, P., & Noël, X. (2012). Decision making under ambiguity but not under risk is related to problem gambling severity. *Psychiatry Research*, *200*(2–3), 568–574.
Broihanne, M. H., Merli, M., & Roger, P. (2014). Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Finance Research Letters*, *11*(2), 64–73.
Bruder, L. R., Wagner, B., Mathar, D., & Peters, J. (2021). Increased temporal discounting and reduced model-based control in problem gambling are not substantially modulated by exposure to virtual gambling environments. *BioRxiv*, 1–48.
Bruno, N., Sachs, N., Demily, C., Franck, N., & Pacherie, E. (2012). Delusions and metacognition in patients with schizophrenia. *Cognitive Neuropsychiatry*, *17*(1), 1–18.
Bucarelli, B., & Purdon, C. (2016). Stove checking behaviour in people with OCD vs. anxious controls. *Journal of Behavior Therapy and Experimental Psychiatry*, *53*, 17–24.
Cahill, L. (2006). Why sex matters for neuroscience. *Nature Reviews Neuroscience*, *7*(6), 477–484.
Calado, F., & Griffiths, M. D. (2016). Problem gambling worldwide: An update and systematic review of empirical research (2000–2015). *Journal of Behavioral Addictions*, *5*(4), 592–613.
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.
Carver, C. S., & White, T. L. (1994). Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales. *Journal of Personality and Social*

Psychology, *67*(2), 319–333.
Casey, L. M., Oei, T. P. S., Melville, K. M., Bourke, E., & Newcombe, P. A. (2008). Measuring self-efficacy in gambling: The gambling refusal self-efficacy questionnaire. *Journal of Gambling Studies*, *24*(2), 229–246.
Chai, X. J., Whitfield-Gabrieli, S., Shinn, A. K., Gabrieli, J. D. E., Nieto Castañón, A., McCarthy, J. M., Cohen, B. M., & Öngür, D. (2011). Abnormal Medial Prefrontal Cortex Resting-State Connectivity in Bipolar Disorder and Schizophrenia. *Neuropsychopharmacology*, *36*(10), 2009–2017.
Chamberlain, S. R., Blackwell, A. D., Fineberg, N. A., Robbins, T. W., & Sahakian, B. J. (2005). The neuropsychology of obsessive compulsive disorder: The importance of failures in cognitive and behavioural inhibition as candidate endophenotypic markers. *Neuroscience and Biobehavioral Reviews*, *29*(3), 399–419.
Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.
Chase, H. W., Kumar, P., Eickhoff, S. B., & Dombrovski, A. Y. (2015). Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective & Behavioral Neuroscience*, *15*(2), 435–459.
Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, *29*(39), 12315–12320.
Choi, J., & Kim, K. (2021). The relationship between impulsiveness, self-esteem, irrational gambling belief and problem gambling moderating effects of gender. *International Journal of Environmental Research and Public Health*, *18*(10), 1–13.
Choi, J., Shin, Y., Jung, W. H., Jang, J. H., & Kang, D. (2012). Altered Brain Activity during Reward Anticipation in Pathological Gambling and Obsessive-Compulsive Disorder. *PLoS ONE*, *7*(9), 3–10.
Christiansen, P., Schoenmakers, T. M., & Field, M. (2015). Less than meets the eye: reappraising the clinical relevance of attentional bias in addiction. *Addictive Behaviors*, *44*, 43–50.
Ciccarelli, M., Cosenza, M., Griffiths, M. D., Nigro, G., & D'Olimpio, F. (2019). Facilitated attention for gambling cues in adolescent problem gamblers: An experimental study. *Journal of Affective Disorders*, *252*, 39–46.
Ciccarelli, M., Nigro, G., Griffiths, M. D., Cosenza, M., & D'Olimpio, F. (2016). Attentional biases in problem and non-problem gamblers. *Journal of Affective Disorders*, *198*, 135–141.
Clark, L., Boileau, I., & Zack, M. (2019). Neuroimaging of reward mechanisms in Gambling disorder: an integrative review. *Molecular Psychiatry*, *24*(5), 674–693.
Cohn, A., Engelmann, J. B., Fehr, E., & Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, *105*(2), 860–885.
Coles, M. E., Radomsky, A. S., & Horng, B. (2006). Exploring the boundaries of memory distrust from repeated checking: Increasing external validity and examining thresholds. *Behaviour Research and Therapy*, *44*(7), 995–1006.
Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*(1), 52–64.
Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215.
Cortese, A. (2022). Metacognitive resources for adaptive learning. *Neuroscience Research*, *178*(3), 10–19.
Cortese, A., Lau, H., & Kawato, M. (2020). Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature Communications*, *11*(1), 1–14.
Cosgrove, K. P., Mazure, C. M., & Staley, J. K. (2007). Evolving Knowledge of Sex Differences in Brain Structure, Function, and Chemistry. *Biological Psychiatry*, *62*(8), 847–855.
Cougle, J. R., Salkovskis, P. M., & Wahl, K. (2007). Perception of memory ability and confidence in recollections in obsessive-compulsive checking. *Journal of Anxiety Disorders*, *21*(1), 118–130.
Cowley, E., Briley, D. A., & Farrell, C. (2015). How do gamblers maintain an illusion of control? *Journal of Business Research*, *68*(10), 2181–2188.
Cox, W. M., Fadardi, J. S., Intriligator, J. M., & Klinger, E. (2014). Attentional bias modification for addictive behaviors: Clinical implications. *CNS Spectrums*, *19*(3), 215–224.
Croskerry, P., & Norman, G. (2008). Overconfidence in Clinical Decision Making. *The American Journal of Medicine*, *121*, 24–29.
Cuttler, C., Sirois-Delisle, V., Alcolado, G. M., Radomsky, A. S., & Taylor, S. (2013). Diminished confidence in prospective memory causes doubts and urges to check. *Journal of Behavior Therapy and*

Experimental Psychiatry, *44*(3), 329–334.
da Silva Castanheira, K., Fleming, S. M., & Otto, A. R. (2021). Confidence in risky value-based choice. *Psychonomic Bulletin and Review*, *28*(3), 1021–1028.
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.
Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, *9*(5), 803.
Dar, R. (2004). Elucidating the mechanism of uncertainty and doubt in obsessive-compulsive checkers. *Journal of Behavior Therapy and Experimental Psychiatry*, *35*(2), 153–163.
Dar, R., Rish, S., Hermesh, H., Taub, M., & Fux, M. (2000). Realism of confidence in obsessive-compulsive checkers. *Journal of Abnormal Psychology*, *109*(4), 673–678.
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. In *Psychological Medicine* (Vol. 52, Issue 13, pp. 2404–2412). Cambridge University Press.
Davies, G., Rae, C. L., Garfinkel, S. N., Seth, A. K., Medford, N., Critchley, H. D., & Greenwood, K. (2018). Impairment of perceptual metacognitive accuracy and reduced prefrontal grey matter volume in first-episode psychosis. *Cognitive Neuropsychiatry*, *23*(3), 1–15.
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.
De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *The Journal of Neuroscience*, *37*(25), 6066–6074.
De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2012). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.
de Ruiter, M. B., Veltman, D. J., Goudriaan, A. E., Oosterlaan, J., Sjoerds, Z., & Van Den Brink, W. (2009). Response Perseveration and Ventral Prefrontal Sensitivity to Reward and Punishment in Male Problem Gamblers and Smokers. *Neuropsychopharmacology*, *34*(4), 1027–1038.
De Ruiter, N. M. P., Van Geert, P. L. C., & Kunnen, E. S. (2017). Explaining the “how” of self-esteem development: The self-organizing self-esteem model. *Review of General Psychology*, *21*(1), 49–68.
Den Ouden, L., Suo, C., Albertella, L., Greenwood, L. M., Lee, R. S. C., Fontenelle, L. F., Parkes, L., Tiego, J., Chamberlain, S. R., Richardson, K., Segrave, R., & Yücel, M. (2022). Transdiagnostic phenotypes of compulsive behavior and associations with psychological, cognitive, and neurobiological affective processing. *Translational Psychiatry*, *12*(1), 1–11.
Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778.
Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A postdecisional neural marker of confidence predicts information-seeking in decision-making. *Journal of Neuroscience*, *39*(17), 3309–3319.
Desender, K., Van Opstal, F., Hughes, G., & Van den Bussche, E. (2016). The temporal dynamics of metacognition: Dissociating task-related activity from later metacognitive processes. *Neuropsychologia*, *82*, 54–64.
Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.
Dunning, D., & Story, A. L. (1991). Depression, Realism, and the Overconfidence Effect: Are the Sadder Wiser When Predicting Future Actions and Events? *Journal of Personality and Social Psychology*, *61*(4), 521–532.
Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Mier, D., Esslinger, C., Englisch, S., Meyer-Lindenberg, A., Kirsch, P., & Zink, M. (2015). Metamemory in schizophrenia: Retrospective confidence ratings interact with neurocognitive deficits. *Psychiatry Research*, *225*(3), 596–603.
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.
Eklund, M., Bäckström, M., & Hansson, L. (2003). Personality and self-variables: important determinants of subjective quality of life in schizophrenia out-patients. *Acta Psychiatrica Scandinavica*, *108*(2), 134–

143.
Engelmann, J. B., Berns, G. S., & Dunlop, B. W. (2017). Hyper-responsivity to losses in the anterior insula during economic choice scales with depression severity. *Psychological Medicine*, *47*(16), 2879–2891.
Engelmann, J. B., Meyer, F., Fehr, E., & Ruff, C. C. (2015). Anticipatory anxiety disrupts neural valuation during risky choice. *Journal of Neuroscience*, *35*(7), 3085–3099.
Engelmann, J. B., & Tamir, D. (2009). Individual differences in risk preference predict neural responses during financial decision-making. *Brain Research*, *1290*, 28–51.
Engelmann, J., Hirmas, A., & van der Weele, J. J. (2021). Top Down or Bottom Up? Disentangling the Channels of Attention in Risky Choice. *Tinbergen Institute*, *31*, 1–44.
Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. *Journal of Neuroscience*, *38*(2), 263–277.
Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, *26*(8), 3508–3526.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.
Ferreri, F., Lapp, L. K., & Peretti, C.-S. (2011). Current research on cognitive aspects of anxiety disorders. *Current Opinion in Psychiatry*, *24*(1), 49–54.
Ferris, J., & Wynne, H. (2001). The Canadian Problem Gambling Index : Final report. *Canadian Centre on Substance Abuse*, 38.
Fieker, M., Moritz, S., Köther, U., & Jelinek, L. (2016). Emotion recognition in depression: An investigation of performance and response confidence in adult female patients with depression. *Psychiatry Research*, *242*, 226–232.
Field, M., & Cox, W. M. (2008). Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug and Alcohol Dependence*, *97*(1–2), 1–20.
Figee, M., Pattij, T., Willuhn, I., Luigjes, J., van den Brink, W., Goudriaan, A., Potenza, M. N., Robbins, T. W., & Denys, D. (2016). Compulsivity in obsessive-compulsive disorder and addictions. *European Neuropsychopharmacology*, *26*(5), 856–868.
Fineberg, N. A., Chamberlain, S. R., Goudriaan, A. E., Stein, D. J., Vanderschuren, L. J. M. J., Gillan, C. M., Shekar, S., Gorwood, P. A. P. M., Voon, V., Morein-Zamir, S., Denys, D., Sahakian, B. J., Moeller, F. G., Robbins, T. W., & Potenza, M. N. (2014). New developments in human neurocognition: clinical, genetic, and brain imaging correlates of impulsivity and compulsivity. *CNS Spectrums*, *19*(1), 69–89.
Fineberg, N. A., Menchon, J. M., Zohar, J., & Veltman, D. J. (2016). Compulsivity-A new trans-diagnostic research domain for the Roadmap for Mental Health Research in Europe (ROAMER) and Research Domain Criteria (RDoC) initiatives. *European Neuropsychopharmacology : The Journal of the European College of Neuropsychopharmacology*, *26*(5), 797–799.
Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911.
Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), 1–14.
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.
Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1338–1349.
Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286.
Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117–6125.
Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 1–9.
Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future: quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, 1–12.
Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624.

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.
Foa, E. B., Amir, N., Gershuny, B., Molnar, C., & Kozak, M. J. (1997). Implicit and explicit memory in obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *11*(2), 119–129.
Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496.
Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 1–8.
Fontenelle, L. F., Destrée, L., Brierley, M. E., Thompson, E. M., Yücel, M., Lee, R., Albertella, L., & Chamberlain, S. R. (2022). The place of obsessive–compulsive and related disorders in the compulsive–impulsive spectrum: a cluster-analytic study. *CNS Spectrums*, *27*(4), 486–495.
Fortune, E. E., & Goodie, A. S. (2012). Cognitive distortions as a component and treatment focus of pathological gambling: A review. *Psychology of Addictive Behaviors*, *26*(2), 298–310.
Fowle, H. J., & Boschen, M. J. (2011). The impact of compulsive cleaning on confidence in memory and cleanliness. *Journal of Anxiety Disorders*, *25*(2), 237–243.
Fox, C., Lee, C., Hanlon, A., Seow, T., Lynch, K., Harty, S., Richards, D., Palacios, J., O'Keane, V., Stephan, K., & Gillan, C. (2023). Metacognition in anxious-depression is state-dependent: an observational treatment study. *ELife*, *12*, 1–10.
Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P., & Huppert, J. D. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychological Review*, 1–41.
Fu, T., Koutstaal, W., Fu, C. H. Y., Poon, L., & Cleare, A. J. (2005). Depression, confidence, and decision: Evidence against depressive realism. *Journal of Psychopathology and Behavioral Assessment*, *27*(4), 243–252.
Fullana, M. A., Mataix-Cols, D., Caspi, A., Harrington, H., Grisham, J. R., Moffitt, T. E., & Poulton, R. (2009). Obsessions and compulsions in the community:Prevalence, Interference, Help-Seeking, Developmental Stability, and Co-Occurring psychiatric conditions. *American Journal of Psychiatry*, *166*(3), 329–336.
Garner, D. M., Bohr, Y., & Garfinkel, P. E. (1982). The Eating Attitudes Test: psychometric features and clinical correlates. *Psychological Medicine*, *12*(4), 871–878.
Gaweda, Ł., Moritz, S., & Kokoszka, A. (2012). Impaired discrimination between imagined and performed actions in schizophrenia. *Psychiatry Research*, *195*, 1–8.
Gawęda, 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*, *47*, 67–75.
Gelskov, S. V., Madsen, K. H., Ramsøy, T. Z., & Siebner, H. R. (2016). Aberrant neural signatures of decision-making: Pathological gamblers display cortico-striatal hypersensitivity to extreme gambles. *NeuroImage*, *128*, 342–352.
Genauck, A., Andrejevic, M., Brehm, K., Matthis, C., Heinz, A., Weinreich, A., Kathmann, N., & Romanczuk-Seiferth, N. (2020). Cue-induced effects on decision-making distinguish subjects with gambling disorder from healthy controls. *Addiction Biology*, *25*(6), 1–10.
Gentes, E. L., & Ruscio, A. M. (2011). A meta-analysis of the relation of intolerance of uncertainty to symptoms of generalized anxiety disorder, major depressive disorder, and obsessive–compulsive disorder. *Clinical Psychology Review*, *31*(6), 923–933.
Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, *106*, 134–143.
Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *ELife*, *7*, 1–59.
Giardini, F., Coricelli, G., Joffily, M., & Sirigu, A. (2008). Overconfidence in Predictions as an Effect of Desirability Bias. *Advances in Decision Making Under Risk and Uncertainty*, 163–180.
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.
Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric

symptom dimension related to deficits in goaldirected control. *ELife*, *5*(3), 1–24.
Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & De Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *The American Journal of Psychiatry*, *168*(7), 718–726.
Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 1–11.
Gillan, C. M., & Rutledge, R. B. (2021). Smartphones and the Neuroscience of Mental Health. *Annual Review of Neuroscience*, *44*, 129–151.
Giorgetta, C., Grecucci, A., Rattin, A., Guerreschi, C., Sanfey, A. G., & Bonini, N. (2014). To play or not to play: A personal dilemma in pathological gambling. *Psychiatry Research*, *219*(3), 562–569.
Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making. *Cerebral Cortex*, *19*(2), 483–495.
Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634–645.
Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *ELife*, *7*.
Gobinath, A. R., Choleris, E., & Galea, L. A. M. (2017). Sex, hormones, and genotype interact to influence psychiatric disease, treatment, and behavioral research. *Journal of Neuroscience Research*, *95*(1–2), 50–64.
Goldney, R. D., Fisher, L. J., Dal Grande, E., & Taylor, A. W. (2004). Subsyndromal depression: Prevalence, use of health services and quality of life in an Australian population. *Social Psychiatry and Psychiatric Epidemiology*, *39*(4), 293–298.
Goldstein, R. Z., Craig, A. D. (Bud), Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., & Volkow, N. D. (2009). The Neurocircuitry of Impaired Insight in Drug Addiction. *Trends in Cognitive Sciences*, *13*(9), 372–380.
Goldstein, R. Z., & Volkow, N. D. (2011). Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature Reviews Neuroscience*, *12*(11), 652–669.
Goodie, A. S. (2005). The role of perceived control and overconfidence in pathological gambling. *Journal of Gambling Studies*, *21*(4), 481–502.
Goodie, A. S., & Fortune, E. E. (2013). Measuring cognitive distortions in pathological gambling: Review and meta-analyses. *Psychology of Addictive Behaviors*, *27*(3), 730–743.
Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown Obsessive Compulsive Scale: I. Development, Use, and Reliability. *Archives of General Psychiatry*, *46*(11), 1006–1011.
Goudriaan, A. E., Oosterlaan, J., de Beurs, E., & van den Brink, W. (2006). Psychophysiological determinants and concomitants of deficient decision making in pathological gamblers. *Drug and Alcohol Dependence*, *84*(3), 231–239.
Goudriaan, A. E., Oosterlaan, J., De Beurs, E., & Van Den Brink, W. (2005). Decision making in pathological gambling: a comparison between pathological gamblers, alcohol dependents, persons with Tourette syndrome, and normal controls. *Brain Research. Cognitive Brain Research*, *23*(1), 137–151.
Grabe, H. J., Meyer, C., Hapke, U., Rumpf, H. J., Freyberger, H. J., Dilling, H., & John, U. (2000). Prevalence, quality of life and psychosocial function in obsessive-compulsive disorder and subclinical obsessive-compulsive disorder in northern Germany. *European Archives of Psychiatry and Clinical Neuroscience*, *250*(5), 262–268.
Grant, L. D., & Bowling, A. C. (2015). Gambling Attitudes and Beliefs Predict Attentional Bias in Non-problem Gamblers. *Journal of Gambling Studies*, *31*(4), 1487–1503.
Guggenmos, M. (2021). Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neuroscience of Consciousness*, *2021*(1), 1–14.
Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *ELife*, *5*(3), 1–19.
Haber, S. N., & Behrens, T. E. J. (2014). The Neural Network Underlying Incentive-Based Learning: Implications for Interpreting Circuit Disruptions in Psychiatric Disorders. *Neuron*, *83*(5), 1019–1039.
Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. In *Neuropsychopharmacology* (Vol. 35, Issue 1, pp. 4–26). Nature Publishing Group.
Hales, C. A., Clark, L., & Winstanley, C. A. (2023). Computational approaches to modeling gambling behaviour: Opportunities for understanding disordered gambling. *Neuroscience & Biobehavioral Reviews*, *147*, 1–10.

Hamilton, M. (1959). The Assessment of Anxiety States by Rating. *British Journal of Medical Psychology*, *32*(1), 50–55.
Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*, 56–62.
Hancock, J. A., Moffoot, A. P. R., & O 'carroll, R. E. (1996). Depressive Realism assessed via Confidence in Decision-making. *Cognitive Neuropsychiatry*, *1*(3), 213–220.
Hankin, B. L., Fraley, R. C., Lahey, B. B., & Waldman, I. D. (2005). Is depression best viewed as a continuum or discrete category? A taxometric analysis of childhood and adolescent depression in a population-based sample. *Journal of Abnormal Psychology*, *114*(1), 96–110.
Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing Attention on the Health Aspects of Foods Changes Value Signals in vmPFC and Improves Dietary Choice. *Journal of Neuroscience*, *31*(30), 11077–11087.
Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(22), 5623–5630.
Hauser, T. U., Allen, M., Rees, G., Dolan, R. J., Bullmore, E. T., Goodyer, I., Fonagy, P., Jones, P., Fearon, P., Prabhu, G., Moutoussis, M., St Clair, M., Cleridou, K., Dadabhoy, H., Granville, S., Harding, E., Hopkins, A., Isaacs, D., King, J., … Pantaleone, S. (2017). Metacognitive impairments extend perceptual decision making weaknesses in compulsivity. *Scientific Reports*, *7*(1).
Hauser, T. U., Iannaccone, R., Dolan, R. J., Ball, J., Hättenschwiler, J., Drechsler, R., Rufer, M., Brandeis, D., Walitza, S., & Brem, S. (2017). Increased fronto-striatal reward prediction errors moderate decision making in obsessive-compulsive disorder. *Psychological Medicine*, *47*(7), 1246–1258.
Haushofer, J., & Fehr, E. (2014). On the psychology of poverty. *Science*, *344*(6186), 862–867.
Hawker, C. O., Merkouris, S. S., Youssef, G. J., & Dowling, N. A. (2021). Exploring the associations between gambling cravings, self-efficacy, and gambling episodes: An Ecological Momentary Assessment study. *Addictive Behaviors*, *112*, 1065–1074.
Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2016). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral Cortex*, *26*(1), 118–130.
Heilbron, M., & Meyniel, F. (2019). Confidence resets reveal hierarchical adaptive learning in humans. *PLoS Computational Biology*, *15*(4), 1–24.
Heitmann, J., Bennik, E. C., Van Hemel-Ruiter, M. E., & De Jong, P. J. (2018). The effectiveness of attentional bias modification for substance use disorder symptoms in adults: A systematic review. In *Systematic Reviews* (Vol. 7, Issue 1, pp. 1–21). Syst Rev.
Henriksen, M. G., & Parnas, J. (2014). Self-disorders and Schizophrenia: A Phenomenological Reappraisal of Poor Insight and Noncompliance. *Schizophrenia Bulletin*, *40*(3), 542–547.
Hermans, D., Engelen, U., Grouwels, L., Joos, E., Lemmens, J., & Pieters, G. (2008). Cognitive confidence in obsessive-compulsive disorder: Distrusting perception, attention and memory. *Behaviour Research and Therapy*, *46*(1), 98–113.
Hertz, U., Bell, V., Barnby, J. M., McQuillin, A., & Bahrami, B. (2020). The Communication of Metacognition for Social Strategy in Psychosis: An Exploratory Study. *Schizophrenia Bulletin Open*, *1*(1), 1–10.
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.
Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd Better Think Twice: Post-Decision Perceptual Confidence. *NeuroImage*, *99*, 323–331.
Hiser, J., & Koenigs, M. (2018). The Multifaceted Role of the Ventromedial Prefrontal Cortex in Emotion, Decision Making, Social Cognition, and Psychopathology. *Biological Psychiatry*, *83*(8), 638–647.
Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational Psychiatry Needs Time and Context. *Annual Review of Psychology*, *73*, 243–270.
Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.
Holingue, C., Budavari, A. C., Rodriguez, K. M., Zisman, C. R., Windheim, G., & Fallin, M. D. (2020). Sex Differences in the Gut-Brain Axis: Implications for Mental Health. *Current Psychiatry Reports*, *22*(12), 1–11.
Hollard, G., Massoni, S., & Vergnaud, J. C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, *80*(3), 363–387.
Hønsi, A., Mentzoni, R. A., Molde, H., & Pallesen, S. (2013). Attentional Bias in Problem Gambling: A Systematic Review. *Journal of Gambling Studies*, *29*(3), 359–375.

Hoven, M., Brunner, G., de Boer, N. S., Goudriaan, A. E., Denys, D., van Holst, R. J., Luigjes, J., & Lebreton, M. (2022). Motivational signals disrupt metacognitive signals in the human ventromedial prefrontal cortex. *Communications Biology*, *5*(1), 1–13.

Hoven, M., de Boer, N. S., Goudriaan, A. E., Denys, D., Lebreton, M., van Holst, R. J., & Luigjes, J. (2022). Metacognition and the effect of incentive motivation in two compulsive disorders: Gambling disorder and obsessive–compulsive disorder. *Psychiatry and Clinical Neurosciences*, *76*(9), 437–449.

Hoven, M., Hirmas, A., Engelmann, J. B., & Holst, R. van. (2023). The role of attention in decision-making under risk in gambling disorder: an eye-tracking study. *Addictive Behaviors*, *138*(6), 1–10.

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), 1–18.

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health*, *1*(5), 337–345.

Hoven, M., Mulder, T., Denys, D., van Holst, R. J., & Luigjes, J. (2023). OCD patients show lower confidence and higher error sensitivity while learning under volatility compared to healthy and highly compulsive samples from the general population. *PsyArXiv*.

Hoven, M., Rouault, M., van Holst, R. J., & Luigjes, J. (2023). Differences in metacognitive functioning between obsessive – compulsive disorder patients and highly compulsive individuals from the general population. *Psychological Medicine*, 1–10.

Howe, P. D. L., Vargas-Sáenz, A., Hulbert, C. A., & Boldero, J. M. (2019). Predictors of gambling and problem gambling in Victoria, Australia. *PLoS ONE*, *14*(1).

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

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & 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.

Jacoby, R. J., Fabricant, L. E., Leonard, R. C., Riemann, B. C., & Abramowitz, J. S. (2013). Just to be certain: Confirming the factor structure of the Intolerance of Uncertainty Scale in patients with obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *27*(5), 535–542.

Jaeger, T., Moulding, R., Yang, Y. H., David, J., Knight, T., & Norberg, M. M. (2021). A systematic review of obsessive-compulsive disorder and self: Self-esteem, feared self, self-ambivalence, egodystonicity, early maladaptive schemas, and self concealment. In *Journal of Obsessive-Compulsive and Related Disorders* (Vol. 31, pp. 1–10). Elsevier.

Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320.

Jönsson, F. U., Olsson, H., & Olsson, M. J. (2005). Odor emotionality affects the confidence in odor naming. *Chemical Senses*, *30*(1), 29–35.

Joukhador, J., Maccallum, F., & Blaszczynski, A. (2003). Differences in cognitive distortions between problem and social gamblers. *Psychological Reports*, *92*(3), 1203–1214.

Julien, D., O'Connor, K., & Aardema, F. (2016). The inference-based approach to obsessive-compulsive disorder: A comprehensive review of its etiological model, treatment efficacy, and model of change. *Journal of Affective Disorders*, *202*, 187–196.

Kaare, P. R., Mõttus, R., & Konstabel, K. (2009). Pathological gambling in Estonia: Relationships with personality, self-Esteem, emotional states and cognitive ability. *Journal of Gambling Studies*, *25*(3), 377–390.

Kable, J. W., & Glimcher, P. W. (2009). The Neurobiology of Decision: Consensus and Controversy. *Neuron*, *63*(6), 733.

Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J. D. (2011). Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. *NeuroImage*, *56*(2), 709–715.

Kanen, J. W., Ersche, K. D., Fineberg, N. A., Robbins, T. W., & Cardinal, R. N. (2019). Computational modelling reveals contrasting effects on reinforcement learning and cognitive flexibility in stimulant use disorder and obsessive-compulsive disorder: remediating effects of dopaminergic D2/3 receptor agents. *Psychopharmacology*, *236*(8), 2337–2358.

Karadag, F., Oguzhanoglu, N., Ozdel, O., Atesci, F. C., & Amuk, T. (2005). Memory function in patients with obsessive compulsive disorder and the problem of confidence in their memories: a clinical study. *Croatian Medical Journal*, *46*(2), 282–287.

Kashyap, H., Kumar, J. K., Kandavel, T., & Reddy, Y. C. J. (2014). The dysfunctional inner mirror: Poor insight in obsessive-compulsive disorder, contributions to heterogeneity and outcome. *CNS Spectrums*, *20*(5), 460–462.

Katyal, S., & Fleming, S. M. (2023). Construct validity in metacognition research: balancing the tightrope between rigor of measurement and breadth of construct. *PsyArXiv*.
Katyal, S., Huys, Q. J. M., Dolan, R. J., & Fleming, S. M. (2023). How underconfidence is maintained in anxiety and depression. *PsyArXiv*.
Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory and Cognition*, *28*(1), 92–107.
Kennedy, A. (2016). Eye tracking: A comprehensive guide to methods and measures. *The Quarterly Journal of Experimental Psychology*, *69*(3), 607–609.
Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337.
Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., & Wang, P. S. (2003). National Comorbidity Survey Replication: The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama*, *289*(23), 3095–3105.
Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 593–602.
Kircher, T. T. J., Koch, K., Stottmeister, F., & Durst, V. (2007). Metacognition and reflexivity in patients with schizophrenia. *Psychopathology*, *40*(4), 254–260.
Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, *21*(6), 728–735.
Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173.
Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.
Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *NeuroImage*, *18*(2), 263–272.
Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed Neural Representation of Expected Value. *Journal of Neuroscience*, *25*(19), 4806–4812.
Koellinger, P., & Treffers, T. (2015). Joy leads to overconfidence, and a simple countermeasure. *PLoS ONE*, *10*(12), 1–22.
Kolodny, T., Mevorach, C., Stern, P., Ankaoua, M., Dankner, Y., Tsafir, S., & Shalev, L. (2022). Are attention and cognitive control altered by fMRI scanner environment? Evidence from Go/No-go tasks in ADHD. *Brain Imaging and Behavior*, *16*, 1003–1013.
Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 217–238.
Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *The Lancet Psychiatry*, *3*(8), 760–773.
Koren, D., Scheyer, R., Reznik, N., Adres, M., Apter, A., Parnas, J., & Seidman, L. J. (2017). Basic self-disturbance, neurocognition and metacognition: A pilot study among help-seeking adolescents with and without attenuated psychosis syndrome. *Early Intervention in Psychiatry*, 1–9.
Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592.
Köther, U., Veckenstedt, R., Vitzthum, F., Roesch-Ely, D., Pfueller, U., Scheu, F., & Moritz, S. (2012). "Don't give me that look" - Overconfidence in false mental state perception in schizophrenia. *Psychiatry Research*, *196*(1), 1–8.
Köther, U., Vettorazzi, E., Veckenstedt, R., Hottenrott, B., Bohn, F., Scheu, F., Pfueller, U., Roesch-Ely, D., & Moritz, S. (2017). Bayesian Analyses of the Effect of Metacognitive Training on Social Cognition Deficits and Overconfidence in Errors. *Journal of Experimental Psychopathology JEP*, *8*, 158–174.
Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.
Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology*, *114*(4), 537–550.

Kuhnen, C. M., & Knutson, B. (2011). The Influence of Affect on Beliefs, Preferences, and Financial Decisions. *Journal of Financial and Quantitative Analysis*, *46*(3), 605–626.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.
Ladouceur, R. (2004). Gambling: The hidden addiction. *Canadian Journal of Psychiatry*, *49*(8), 501–503.
Lak, A., Hueske, E., Hirokawa, J., Masset, P., Ott, T., Urai, A. E., Donner, T. H., Carandini, M., Tonegawa, S., Uchida, N., & Kepecs, A. (2020). Reinforcement biases subsequent perceptual decisions when confidence is low: A widespread behavioral phenomenon. *ELife*, *9*, 1–26.
Lakey, C. E., Goodie, A. S., & Campbell, W. K. (2007). Frequent card playing and pathological gambling: The utility of the Georgia Gambling Task and Iowa Gambling Task for predicting pathology. *Journal of Gambling Studies*, *23*(3), 285–297.
Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328.
Laws, K. R., & Bhatt, R. (2005). False memories and delusional ideation in normal healthy subjects. *Personality and Individual Differences*, *39*(4), 775–781.
Lazarov, A., Dar, R., Liberman, N., & Oded, Y. (2012). Obsessive-compulsive tendencies and undermined confidence are related to reliance on proxies for internal states in a false feedback paradigm. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(1), 556–564.
Lazarov, A., Liberman, N., Hermesh, H., & Dar, R. (2014). Seeking proxies for internal states in obsessive–compulsive disorder. *Journal of Abnormal Psychology*, *123*(4), 695–704.
Le Berre, A. P., Pinon, K., Vabret, F., Pitel, A. L., Allain, P., Eustache, F., & Beaunieux, H. (2010). Study of metamemory in patients with chronic alcoholism using a feeling-of-knowing episodic memory task. *Alcoholism: Clinical and Experimental Research*, *34*(11), 1888–1898.
Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167.
Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLoS Computational Biology*, *15*(4), 1–27.
Lebreton, M., Bavard, S., Daunizeau, J., & Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. *Nature Human Behaviour*, *3*(9), 897–905.
Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron*, *64*(3), 431–439.
Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, *4*(5), 1–13.
Lecomte, T., Corbière, M., & Laisné, F. (2006). Investigating self-esteem in individuals with schizophrenia: Relevance of the Self-Esteem Rating Scale-Short Form. *Psychiatry Research*, *143*(1), 99–108.
Ledgerwood, D. M., Dyshniku, F., McCarthy, J. E., Ostojic-Aitkens, D., Forfitt, J., & Rumble, S. C. (2020). Gambling-Related Cognitive Distortions in Residential Treatment for Gambling Disorder. *Journal of Gambling Studies*, *36*(2), 669–683.
Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *ELife*, *10*.
Lee, D. G., & Hare, T. A. (2023). Value certainty and choice confidence are multidimensional constructs that guide decision-making. *Cognitive, Affective and Behavioral Neuroscience*, *1*, 1–19.
Lee, J. K., Rouault, M., & Wyart, V. (2023). Compulsivity is linked to maladaptive choice variability but unaltered reinforcement learning under uncertainty. *BioRxiv Preprint*, *914*, 1–20.
LeGates, T. A., Kvarta, M. D., & Thompson, S. M. (2018). Sex differences in antidepressant efficacy. *Neuropsychopharmacology 2018 44:1*, *44*(1), 140–154.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package*.
Levy, D. J., & Glimcher, P. W. (2011). Comparing Apples and Oranges: Using Reward-Specific and Reward-General Subjective Value Representation in the Brain. *Journal of Neuroscience*, *31*(41), 14693–14707.
Leyton, M., & Vezina, P. (2012). On cue: Striatal ups and downs in addictions. *Biological Psychiatry*, *72*(10), 1–21.
Leyton, M., & Vezina, P. (2013). Striatal ups and downs: Their roles in vulnerability to addictions in humans. *Neuroscience & Biobehavioral Reviews*, *37*(9), 1999–2014.
Liang, S., Ye, D., & Liu, Y. (2021). The Effect of Perceived Scarcity: Experiencing Scarcity Increases Risk Taking. *Journal of Psychology: Interdisciplinary and Applied*, *155*(1), 59–89.
Liebowitz, M. R. (1987). Social Phobia. *Modern Problems of Pharmacopsychiatry*, *22*, 141–173.

Ligneul, R., Sescousse, G., Barbalat, G., Domenech, P., & Dreher, J. C. (2013). Shifted risk preferences in pathological gambling. *Psychological Medicine*, *43*(5), 1059–1068.
Lim, S. L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. *The Journal of Neuroscience*, *31*(37), 13214.
Limbrick-Oldfield, E. H., Cherkasova, M. V., Kennedy, D., Goshko, C. B., Griffin, D., Barton, J. J. S., & Clark, L. (2020). Gambling disorder is associated with reduced sensitivity to expected value during risky choice. *Journal of Behavioral Addictions*, *9*(4), 1044.
Limbrick-Oldfield, E. H., van Holst, R. J., & Clark, L. (2013). Fronto-striatal dysregulation in drug addiction and pathological gambling: Consistent inconsistencies? *NeuroImage: Clinical*, *2*, 385–393.
Lincoln, T. M. (2007). Relevant dimensions of delusions: Continuing the continuum versus category debate. *Schizophrenia Research*, *93*(1–3), 211–220.
Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *35*(5), 1219–1236.
Liu, Y. C., Tang, C. C., Hung, T. T., Tsai, P. C., & Lin, M. F. (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.
Loosen, A., Seow, T. X. F., & Hauser, T. U. (2023). Consistency within change: Evaluating the psychometric properties of a widely-used predictive-inference task. *PsyArXiv*.
Lopez-Persem, A., Bastin, J., Petton, M., Abitbol, R., Lehongre, K., Adam, C., Navarro, V., Rheims, S., Kahane, P., Domenech, P., & Pessiglione, M. (2020). Four core properties of the human brain valuation system demonstrated in intracranial signals. *Nature Neuroscience*, *23*(5), 664–675.
Luigjes, J., Lorenzetti, V., de Haan, S., Youssef, G. J., Murawski, C., Sjoerds, Z., van den Brink, W., Denys, D., Fontenelle, L. F., & Yücel, M. (2019). Defining Compulsive Behavior. *Neuropsychology Review*, *29*(1), 4–13.
Luijten, M., Schellekens, A. F., Kühn, S., MacHielse, M. W. J., & Sescousse, G. (2017). Disruption of reward processing in addiction: An image-based meta-analysis of functional magnetic resonance imaging studies. *JAMA Psychiatry*, *74*(4), 387–398.
MacDonald, P. A., Antony, M. M., MacLeod, C. M., & Richter, M. A. (1997). Memory and confidence in memory judgments among individuals with obsessive compulsive disorder and non-clinical controls. *Behaviour Research and Therapy*, *35*(6), 497–505.
Mallorquí-Bagué, N., Mestre-Bach, G., Lozano-Madrid, M., Granero, R., Vintró-Alcaraz, C., Fernández-Aranda, F., Gómez-Peña, M., Moragas, L., Del Pino-Gutierrez, A., Menchón, J. M., & Jiménez-Murcia, S. (2021). Gender and gambling disorder: Differences in compulsivity-related neurocognitive domains. *Addictive Behaviors*, *113*, 106683.
Mallorquí-Bagué, N., Vintró-Alcaraz, C., Verdejo-García, A., Granero, R., Fernández-Aranda, F., Magaña, P., Mena-Moreno, T., Aymamí, N., Gómez-Peña, M., Del Pino-Gutiérrez, A., Mestre-Bach, G., Menchón, J. M., & Jiménez-Murcia, S. (2019). Impulsivity and cognitive distortions in different clinical phenotypes of gambling disorder: Profiles and longitudinal prediction of treatment outcomes. *European Psychiatry*, *61*, 9–16.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.
Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatry Research*, *38*(2), 143–162.
Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*, 301–306.
Marton, T., Samuels, J., Nestadt, P., Krasnow, J., Wang, Y., Shuler, M., Kamath, V., Chib, V. S., Bakker, A., & Nestadt, G. (2019). Validating a dimension of doubt in decisionmaking: A proposed endophenotype for obsessive-compulsive disorder. *PLoS ONE*, *14*(6), 1–14.
Marzuki, A. A., Tomić, I., Ip, S. H. Y., Gottwald, J., Kanen, J. W., Kaser, M., Sule, A., Conway-Morris, A., Sahakian, B. J., & Robbins, T. W. (2021). Association of Environmental Uncertainty with Altered Decision-making and Learning Mechanisms in Youths with Obsessive-Compulsive Disorder. *JAMA Network Open*, *4*(11).
Marzuki, A. A., Vaghi, M. M., Conway-Morris, A., Kaser, M., Sule, A., Apergis-Schoute, A., Sahakian, B. J., & Robbins, T. W. (2022). Atypical action updating in a dynamic environment associated with adolescent obsessive–compulsive disorder. *Journal of Child Psychology and Psychiatry and Allied*

Disciplines, *63*(12), 1591–1601.
Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*(2–3), 293–296.
Massoni, S. (2014). Emotion as a boost to metacognition: How worry enhances the quality of confidence. *Consciousness and Cognition*, *29*, 189–198.
Mathes, B. M., Morabito, D. M., & Schmidt, N. B. (2019). Epidemiological and Clinical Gender Differences in OCD. *Current Psychiatry Reports*, *21*(5), 1–7.
Mazaika, Whitfield-Gabrieli, & Reiss..., A. (2007). Artifact repair for fMRI data from high motion clinical subjects. *NeuroImage*, 321.
McClintock, S. M., Husain, M. M., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Trivedi, M. H., Cook, I., Morris, D., Warden, D., & Rush, A. J. (2011). Residual symptoms in depressed outpatients who respond by 50% but do not remit to antidepressant medication. *Journal of Clinical Psychopharmacology*, *31*(2), 180–186.
McGorry, P. D., Hartmann, J. A., Spooner, R., & Nelson, B. (2018). Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry*, *17*(2), 133–142.
McGrath, D. S., Meitner, A., & Sears, C. R. (2018). The specificity of attentional biases by type of gambling: An eye-tracking study. *PLoS ONE*, *13*(1).
McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron*, *84*(4), 870–881.
McKay, R., Langdon, R., & Coltheart, M. (2006). Need for closure, jumping to conclusions, and decisiveness in delusion-prone individuals. *Journal of Nervous and Mental Disease*, *194*(6), 422–426.
McLaughlin, K. A., & Nolen-Hoeksema, S. (2011). Rumination as a transdiagnostic factor in depression and anxiety. *Behaviour Research and Therapy*, *49*(3), 186–193.
McNally, R. J., & Kohlbeck, P. A. (1993). Reality monitoring in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *31*(3), 249–253.
McNamee, D., Rangel, A., & O'Doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature Neuroscience*, *16*(4), 479–485.
Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(19), 3859–3868.
Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, *11*(6), 1–25.
Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Perspective Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*, 78–92.
Middlebrooks, P. G., Abzug, Z. M., & Sommer, M. A. (2013). Studying metacognitive processes at the single neuron level. In *The Cognitive Neuroscience of Metacognition* (pp. 225–244). Springer-Verlag Berlin Heidelberg.
Mintzer, M., & Stitzer, M. (2002). Cognitive impairment in methadone maintenance patients. *Drug Alcohol Depend*, *67*(1), 41–51.
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.
Molenberghs, P., Trautwein, F.-M. M., Böckler, A., Singer, T., & Kanske, P. (2016). Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study. *Social Cognitive and Affective Neuroscience*, *11*(12), 1942–1951.
Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, *38*(14), 3534–3546.
Moritz, S., Göritz, A. S., Gallinat, J., Schafschetzy, M., Van Quaquebeke, N., Peters, M. J. V., & Andreou, C. (2015). Subjective competence breeds overconfidence in errors in psychosis. A hubris account of paranoia. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 118–124.
Moritz, S., Göritz, A. S., Van Quaquebeke, N., Andreou, C., Jungclaussen, D., & Peters, M. J. V. (2014). Knowledge corruption for visual perception in individuals high on paranoia. *Psychiatry Research*, *215*(3), 700–705.
Moritz, S., Jacobsen, D., Willenborg, B., Jelinek, L., & Fricke, S. (2006). A check on the memory deficit hypothesis of obsessive-compulsive checking. *European Archives of Psychiatry and Clinical Neuroscience*, *256*(2), 82–86.
Moritz, S., & Jaeger, A. (2018). Decreased memory confidence in obsessive–compulsive disorder for

scenarios high and low on responsibility: is low still too high? *European Archives of Psychiatry and Clinical Neuroscience*, *268*(3), 291–299.
Moritz, S., Kloss, M., von Eckstaedt, F. V., & Jelinek, L. (2009). Comparable performance of patients with obsessive-compulsive disorder (OCD) and healthy controls for verbal and nonverbal memory accuracy and confidence:. Time to forget the forgetfulness hypothesis of OCD? *Psychiatry Research*, *166*(2–3), 247–253.
Moritz, S., Pfuhl, G., Lüdtke, T., Menon, M., Balzan, R., & Andreou, C. (2017). A two-stage cognitive theory of the positive symptoms of psychosis. Highlighting the role of lowered decision thresholds. *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 12–20. https://www.sciencedirect.com/science/article/pii/S0005791616301203#sec2
Moritz, S., Ramdani, N., Klass, H., Andreou, C., Jungclaussen, D., Eifler, S., Englisch, S., Schirmbeck, F., & Zink, M. (2014). Overconfidence in incorrect perceptual judgments in patients with schizophrenia. *Schizophrenia Research: Cognition*, *1*(4), 165–170.
Moritz, S., Rietschel, L., Jelinek, L., & Bäuml, K. H. T. (2011). Are patients with obsessive-compulsive disorder generally more doubtful? Doubt is warranted! *Psychiatry Research*, *189*(2), 265–269.
Moritz, S., Ruhe, C., Jelinek, L., & Naber, D. (2009). No deficits in nonverbal memory, metamemory and internal as well as external source memory in obsessive-compulsive disorder (OCD). *Behaviour Research and Therapy*, *47*(4), 308–315.
Moritz, S., Wahl, K., Zurowski, B., Jelinek, L., Hand, I., & Fricke, S. (2007). Enhanced perceived responsibility decreases metamemory but not memory accuracy in obsessive-compulsive disorder (OCD). *Behaviour Research and Therapy*, *45*(9), 2044–2052.
Moritz, S., & Woodward, T. S. (2002). Memory confidence and false memories in schizophrenia. *Journal of Nervous and Mental Disease*, *190*(9), 641–643.
Moritz, S., & Woodward, T. S. (2006a). Metacognitive control over false memories: A key determinant of delusional thinking. *Current Psychiatry Reports*, *8*(3), 184–190.
Moritz, S., & Woodward, T. S. (2006b). The contribution of metamemory deficits to schizophrenia. *Journal of Abnormal Psychology*, *115*(1), 15–25.
Moritz, S., Woodward, T. S., & Chen, E. (2006). Investigation of metamemory dysfunctions in first-episode schizophrenia. *Schizophrenia Research*, *81*(2–3), 247–252.
Moritz, S., Woodward, T. S., Cuttler, C., Whitman, J. C., & Watson, J. M. (2004). False Memories in Schizophrenia. *Neuropsychology*, *18*(2), 276–283.
Moritz, S., Woodward, T. S., Jelinek, L., & Klinge, R. (2008). Memory and metamemory in schizophrenia: A liberal acceptance account of psychosis. *Psychological Medicine*, *38*(6), 825–832.
Moritz, S., Woodward, T. S., & Rodriguez-Raecke, R. (2006). Patients with schizophrenia do not produce more false memories than controls but are more confident in them. *Psychological Medicine*, *36*(5), 659–667.
Moritz, S., Woodward, T. S., & Ruff, C. C. (2003). Source monitoring and memory confidence in schizophrenia. *Psychological Medicine*, *33*(1), 131–139.
Moritz, S., Woodward, T. S., Whitman, J. C., & Cuttler, C. (2005). Confidence in errors as a possible basis for delusions in schizophrenia. *Journal of Nervous and Mental Disease*, *193*(1), 9–16.
Moritz, S., Woznica, A., Andreou, C., & Köther, U. (2012). Response confidence for emotion perception in schizophrenia using a Continuous Facial Sequence Task. *Psychiatry Research*, *200*(2–3), 202–207.
Moroz, M., & Dunkley, D. M. (2015). Self-critical perfectionism and depressive symptoms: Low self-esteem and experiential avoidance as mediators. *Personality and Individual Differences*, *87*, 174–179.
Moses-Payne, M. E., Habicht, J., Bowler, A., Steinbeis, N., & Hauser, T. U. (2021). I know better! Emerging metacognition allows adolescents to ignore false advice. *Developmental Science*, *24*.
Moses-Payne, M. E., Rollwage, M., Fleming, S. M., & Roiser, J. P. (2019). Postdecision Evidence Integration and Depressive Symptoms. *Frontiers in Psychiatry*, *10*.
Moutoussis, M., Garzón, B., Neufeld, S., Bach, D. R., Rigoli, F., Goodyer, I., Bullmore, E., Fonagy, P., Jones, P., Hauser, T., Romero-Garcia, R., St Clair, M., Vértes, P., Whitaker, K., Inkster, B., Prabhu, G., Ooi, C., Toseeb, U., Widmer, B., … Dolan, R. J. (2021). Decision-making ability, psychopathology, and brain connectivity. *Neuron*, *109*(12), 2025–2040.
Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1), 1–15.
Murch, W. S., Limbrick-Oldfield, E. H., Ferrari, M. A., MacDonald, K. I., Fooken, J., Cherkasova, M. V., Spering, M., & Clark, L. (2020). Zoned in or zoned out? Investigating immersion in slot machine gambling using mobile eye-tracking. *Addiction*, *115*(6), 1127–1138.

Muris, P. (2002). Relationships between self-efficacy and symptoms of anxiety disorders and depression in a normal adolescent sample. *Personality and Individual Differences*, *32*(2), 337–348.
Namkung, H., Kim, S.-H., & Sawa, A. (2017). The Insula: An Underestimated Brain Area in Clinical Neuroscience, Psychiatry, and Neurology. *Trends in Neurosciences*, *40*(4), 200–207.
Nassar, M. R., McGuire, J. T., Ritz, H., & Kable, J. W. (2019). Dissociable Forms of Uncertainty-Driven Representational Change Across the Human Brain. *Journal of Neuroscience*, *39*(9), 1688–1698.
Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.
Navas, J. F., Billieux, J., Perandrés-Gómez, A., López-Torrecillas, F., Cándido, A., & Perales, J. C. (2017). Impulsivity traits and gambling cognitions associated with gambling preferences and clinical status. *International Gambling Studies*, *17*(1), 102–124.
Nelson, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, *26*(C), 125–173.
Nestadt, G., Kamath, V., Maher, B. S., Krasnow, J., Nestadt, P., Wang, Y., Bakker, A., & Samuels, J. (2016). Doubt and the decision-making process in obsessive-compulsive disorder. *Medical Hypotheses*, *96*, 1–4.
Norman, L. J., Taylor, S. F., Liu, Y., Radua, J., Chye, Y., De Wit, S. J., Huyser, C., Karahanoglu, F. I., Luks, T., Manoach, D., Mathews, C., Rubia, K., Suo, C., van den Heuvel, O. A., Yücel, M., & Fitzgerald, K. (2019). Error Processing and Inhibitory Control in Obsessive-Compulsive Disorder: A Meta-analysis Using Statistical Parametric Maps. *Biological Psychiatry*, *85*(9), 713–725.
O'Kearney, E. L., Brown, C. R., Jelinek, G. A., Neate, S. L., Taylor, K. T., Bevens, W., De Livera, A. M., Simpson, S., & Weiland, T. J. (2020). Mastery is associated with greater physical and mental health-related quality of life in two international cohorts of people with multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *38*.
Ochoa, C., Álvarez-Moya, E. M., Penelo, E., Aymami, M. N., Gómez-Peña, M., Fernández-Aranda, F., Granero, R., Vallejo-Ruiloba, J., Menchón, J. M., Lawrence, N. S., & Jiménez-Murcia, S. (2013). Decision-making deficits in pathological gambling: The role of executive functions, explicit knowledge and impulsivity in relation to decisions made under ambiguity and risk. *American Journal on Addictions*, *22*(5), 492–499.
Ong, Q., Theseira, W., & Ng, I. Y. H. (2019). Reducing debt improves psychological functioning and changes decision-making in the poor. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(15), 7244–7249.
Orquin, J. L., Ashby, N. J. S., & Clarke, A. D. F. (2016). Areas of Interest as a Signal Detection Problem in Behavioral Eye-Tracking Research. *Journal of Behavioral Decision Making*, *29*(2–3), 103–115.
Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*(1), 190–206.
Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low Self-Esteem Prospectively Predicts Depression in Adolescence and Young Adulthood. *Journal of Personality and Social Psychology*, *95*(3), 695–708.
Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology. General*, *147*(2), 147–169.
Padoa-Schioppa, C. (2007). Orbitofrontal cortex and the computation of economic value. *Annals of the New York Academy of Sciences*, *1121*, 232–253.
Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226.
Palminteri, S., & Pessiglione, M. (2017). Opponent brain systems for reward and punishment learning: Causal evidence from drug and lesion studies in humans. *Decision Neuroscience: An Integrative Perspective*, 291–303.
Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.
Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review*, *15*(3), 105–130.
Park, K., MinHwa, L., & Seo, M. (2019). The impact of self-stigma on self-esteem among persons with different mental disorders. *International Journal of Social Psychiatry*, *65*(7–8), 558–565.
Parkes, L., Tiego, J., Aquino, K., Braganza, L., Chamberlain, S. R., Fontenelle, L. F., Harrison, B. J., Lorenzetti, V., Paton, B., Razi, A., Fornito, A., & Yücel, M. (2019). Transdiagnostic variations in impulsivity and compulsivity in obsessive-compulsive disorder and gambling disorder correlate with effective connectivity in cortical-striatal-thalamic-cortical circuits. *NeuroImage*, *202*, 116070.

Parkitny, L., & McAuley, J. (2010). The depression anxiety stress scale (DASS). *Journal of Physiotherapy*, *56*(2), 204.
Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface*, *14*.
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.
Pearlin, L. I., & Schooler, C. (1978). The structure of coping. *J Health Soc Behav*, *19*, 2–21.
Perales, J. C., King, D. L., Navas, J. F., Schimmenti, A., Sescousse, G., Starcevic, V., van Holst, R. J., & Billieux, J. (2020). Learning to lose control: A process-based account of behavioral addiction. *Neuroscience and Biobehavioral Reviews*, *108*, 771–780.
Perandrés-Gómez, A., Navas, J. F., van Timmeren, T., & Perales, J. C. (2021). Decision-making (in)flexibility in gambling disorder. *Addictive Behaviors*, *112*, 106534.
Pescetelli, N., Hauperich, A. K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition*, *215*(6).
Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965.
Pessiglione, M., & Lebreton, M. (2015). From the Reward Circuit to the Valuation System: How the Brain Motivates Behavior. In *Handbook of Biobehavioral Approaches to Self-Regulation* (pp. 157–173).
Pessoa, L., & Engelmann, J. B. (2010). Embedding reward signals into perception and cognition. In *Frontiers in Neuroscience* (Vol. 4, Issue 9, p. 17). Frontiers.
Peters, M. J. V., Cima, M. J., Smeets, T., de Vos, M., Jelicic, M., & Merckelbach, H. (2007). Did I say that word or did you? Executive dysfunctions in schizophrenic patients affect memory efficiency, but not source attributions. *Cognitive Neuropsychiatry*, *12*(5), 391–411.
Peters, M. J. V, Hauschildt, M., Moritz, S., & Jelinek, L. (2013). Impact of emotionality on memory and meta-memory in schizophrenia using video sequences. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(1), 77–83.
Petrocelli, J. V., & Crysel, L. C. (2009). Counterfactual thinking and confidence in blackjack: A test of the counterfactual inflation hypothesis. *Journal of Experimental Social Psychology*, *45*(6), 1312–1315.
Petrocelli, J. V., & Sherman, S. J. (2010). Event detail and confidence in gambling: The role of counterfactual thought reactions. *Journal of Experimental Social Psychology*, *46*(1), 61–72.
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 and Cognitive - Behavior Therapy*, *38*(2), 173–183.
Philipp, R., Kriston, L., Lanio, J., Kühne, F., Härter, M., Moritz, S., & Meister, R. (2019). Effectiveness of metacognitive interventions for mental disorders in adults—A systematic review and meta-analysis (METACOG). *Clinical Psychology & Psychotherapy*, *26*(2), 227–240.
Pinciotti, C. M., Riemann, B. C., & Abramowitz, J. S. (2021). Intolerance of uncertainty and obsessive-compulsive disorder dimensions. *Journal of Anxiety Disorders*, *81*, 102417.
Pinheiro, J., Bates, D., & R Core Team. (2022). *nlme: Nonlinear Mixed Effects Models*. https://cran.r-project.org/package=nlme
Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *Journal of Neuroscience*, *27*(37), 9984–9988.
Porras-Segovia, A., Molina-Madueño, R. M., Berrouiguet, S., López-Castroman, J., Barrigón, M. L., Pérez-Rodríguez, M. S., Marco, J. H., Díaz-Oliván, I., de León, S., Courtet, P., Artés-Rodríguez, A., & Baca-García, E. (2020). Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *Journal of Affective Disorders*, *274*, 733–741.
Poser, B. A., Versluis, M. J., Hoogduin, J. M., & Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine*, *55*(6), 1227–1235.
Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.
Prochazkova, L., Parkes, L., Dawson, A., Youssef, G., Ferreira, G. M., Lorenzetti, V., Segrave, R. A., Fontenelle, L. F., & Yucel, M. (2018). Unpacking the role of self-reported compulsivity and impulsivity in obsessive-compulsive disorder. *CNS Spectrums*, *23*(1), 51–58.
Purdon, C., & Clark, D. A. (1999). Metacognition and Obsessions. *Clinical Psychology and Psychotherapy*, *6*(2), 102–110.
Quiles, C., Prouteau, A., & Verdoux, H. (2015). Associations between self-esteem, anxiety and depression

and metacognitive awareness or metacognitive knowledge. *Psychiatry Research*, *230*(2), 738–741.
Raballo, A., Pappagallo, E., Dell Erba, A., Lo Cascio, N., Patane, M., Gebhardt, E., Boldrini, T., Terzariol, L., Angelone, M., Trisolini, A., Girardi, P., & Nastro, P. F. (2016). Self-disorders and clinical high risk for psychosis: An empirical study in help-seeking youth attending community mental health facilities. *Schizophrenia Bulletin*, *42*(4), 926–932.
Radomsky, A. S., & Alcolado, G. M. (2010). Don't even think about checking: Mental checking causes memory distrust. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(4), 345–351.
Radomsky, A. S., Dugas, M. J., Alcolado, G. M., & Lavoie, S. L. (2014). When more is less: Doubt, repetition, memory, metamemory, and compulsive checking in OCD. *Behaviour Research and Therapy*, *59*, 30–39.
Radomsky, A. S., Gilchrist, P. T., & Dussault, D. (2006). Repeated checking really does cause memory distrust. *Behaviour Research and Therapy*, *44*(2), 305–316.
Radomsky, A. S., Rachman, S., & Hammond, D. (2001). Memory bias, confidence and responsibility in compulsive checking. *Behaviour Research and Therapy*, *39*(7), 813–822.
Rahnev, D. (2023). Measuring metacognition: A comprehensive assessment of current methods. *PsyArXiv*.
Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(1), 1–9.
Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*(11), 1664–1680.
Raines, A. M., Oglesby, M. E., Allan, N. P., Mathes, B. M., Sutton, C. A., & Schmidt, N. B. (2018). Examining the role of sex differences in obsessive-compulsive symptom dimensions. *Psychiatry Research*, *259*, 265–269.
Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556.
Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*(2), 262–270.
Richard, A., & Wilhelm, F. (2022). *Package ' careless .'* 1–11.
Rieger, S., Göllner, R., Trautwein, U., & Roberts, B. W. (2016). Low self-esteem prospectively predicts depression in the transition to young adulthood: A replication of Orth, Robins, and Roberts (2008). *Journal of Personality and Social Psychology*, *110*(1), 16–22.
Riesel, A. (2019). The erring brain: Error-related negativity as an endophenotype for OCD—A review and meta-analysis. *Psychophysiology*, *56*, 1–22.
Riesel, A., Endrass, T., Kaufmann, C., & Kathmann, N. (2011). Overactive Error-Related Brain Activity as a Candidate Endophenotype ... *American Journal of Psychiatry*, *168*(3), 317–324.
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.
Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. D. (2014). Cognitive impairment in depression: a systematic review and meta-analysis. *Psychological Medicine*, *44*(10), 2029–2040.
Rogers, R. D., Butler, J., Millard, S., Cristino, F., Davitt, L. I., & Leek, E. C. (2017). A scoping investigation of eye-tracking in Electronic Gambling Machine (EGM) play. *Gambling Aware*, *3*, 1–43.
Rollwage, M., Dolan, J., Fleming, S. M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical political beliefs. *Current Biology*, *344*(24), 419.
Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1).
Romanowska, S., MacQueen, G., Goldstein, B. I., Wang, J., Kennedy, S. H., Bray, S., Lebel, C., & Addington, J. (2018). Neurocognitive deficits in a transdiagnostic clinical staging model. *Psychiatry Research*, *270*, 1137–1142.
Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and Commitment Therapy Measures Package*, *61*(52), 1–18.
Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141.
Rouault, M., & Fleming, S. M. (2020). Formation of global self-beliefs in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27268–27276.
Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, *1*, 1–13.
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.

Rouault, M., Will, G. J., Fleming, S. M., & Dolan, R. J. (2022). Low self-esteem and the formation of global self-performance estimates in emerging adulthood. *Translational Psychiatry*, *12*(1), 1–10.
Rouy, M., Saliou, P., Nalborczyk, L., Pereira, M., Roux, P., & Faivre, N. (2021). Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders. *Neuroscience & Biobehavioral Reviews*, *126*, 329–337.
Ruscio, A. M., Stein, D. J., Chiu, W. T., & Kessler, R. C. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Molecular Psychiatry*, *15*(1), 53–63.
Sadeghi, S., Ekhtiari, H., Bahrami, B., & Ahmadabadi, M. N. (2017). Metacognitive Deficiency in a Perceptual but Not a Memory Task in Methadone Maintenance Patients. *Scientific Reports*, *7*(1).
Salem-Garcia, N., Palminteri, S., & Lebreton, M. (2023). Linking Confidence Biases to Reinforcement-Learning Processes. *Psychological Review*, 1–25.
Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision*, *19*(4), 25–25.
Samuels, J., Bienvenu, O. J., Krasnow, J., Wang, Y., Grados, M. A., Cullen, B., Goes, F. S., Maher, B., Greenberg, B. D., McLaughlin, N. C., Rasmussen, S. A., Fyer, A. J., Knowles, J. A., Nestadt, P., McCracken, J. T., Piacentini, J., Geller, D., Pauls, D. L., Stewart, S. E., … Nestadt, G. (2017). An investigation of doubt in obsessive–compulsive disorder. *Comprehensive Psychiatry*, *75*, 117–124.
San Martín, R., Appelbaum, L. G., Huettel, S. A., & Woldorff, M. G. (2016). Cortical Brain Activity Reflecting Attentional Biasing Toward Reward-Predicting Cues Covaries with Economic Decision-Making Performance. *Cerebral Cortex*, *26*(1), 1–11.
Sancho, M., Bonnaire, C., Costa, S., Casalé-Salayet, G., Vera-Igual, J., Rodríguez, R. C., Duran-Sindreu, S., & Trujols, J. (2021). Impulsivity, Emotion Regulation, Cognitive Distortions and Attentional Bias in a Spanish Sample of Gambling Disorder Patients: Comparison between Online and Land-Based Gambling. *International Journal of Environmental Research and Public Health*, *18*(9).
Sanders, J. I., Hangya, B. B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506.
Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction*, *88*(6), 791–804.
Sax, A.-L., Baddeley, R., & Costa, R. P. (2023). Depression impairs metacognitive biases, but not learning. *PsyArXiv*, 1–27. http://dx.doi.org/10.31234/osf.io/9a7y8
Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*(4), 512–524.
Scheyer, R., Reznik, N., Adres, M., Apter, A., Seidman, L. J., & Koren, D. (2014). Metacognition in Non-psychotic Help-seeking Adolescents: Associations with Prodromal Symptoms, Distress and Psychosocial Deterioration. *Schizophrenia Research*, *51*(1).
Schultz, S. H., North, S. W., & Shields, C. G. (2007). Schizophrenia: A review. *American Family Physician*, *75*(12), 1821–1829.
Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive Computations for Information Search: Confidence in Control. *Psychological Review*, *130*(3), 604–639.
Schwartz, C., Hilbert, S., Schubert, C., Schlegl, S., Freyer, T., Löwe, B., Osen, B., & Voderholzer, U. (2017). Change Factors in the Process of Cognitive-Behavioural Therapy for Obsessive–Compulsive Disorder. *Clinical Psychology & Psychotherapy*, *24*(3), 785–792.
Schwarzer, R., & Jerusalem, M. (1995). Self-efficacy measurement: Generalized Self-Efficacy Scale. In *Measures in Health Psychology: A User's Portfolio* (pp. 35–37).
Seidman, L. J., Giuliano, A. J., Meyer, E. C., Addington, J., Cadenhead, K. S., Cannon, T. D., McGlashan, T. H., Perkins, D. O., Tsuang, M. T., Walker, E. F., Woods, S. W., Bearden, C. E., Christensen, B. K., Hawkins, K., Heaton, R., Keefe, R. S. E., Heinssen, R., & Cornblatt, B. A. (2010). Neuropsychology of the prodrome to psychosis in the NAPLS Consortium: Relationship to family history and conversion to psychosis. *Archives of General Psychiatry*, *67*(6), 578–588.
Seow, T. X. F., & Gillan, C. M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports*, *10*(1), 1–11.
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.
Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *ELife*, *9*, 1–58.

Sescousse, G., Barbalat, G., Domenech, P., & Dreher, J. C. (2013). Imbalance in the sensitivity to different types of rewards in pathological gambling. *Brain*, *136*(8), 2527–2538.
Seymour, B., Maruyama, M., & De Martino, B. (2015). When is a loss a loss? Excitatory and inhibitory processes in loss-related decision-making. *Current Opinion in Behavioral Sciences*, *5*, 122–127.
Shamay-Tsoory, S. G., & Mendelsohn, A. (2019). Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research. *Perspectives on Psychological Science*, *14*(5), 841–859.
Shapiro, A. D., & Grafton, S. T. (2020). Subjective value then confidence in human ventromedial prefrontal cortex. *PLoS ONE*, *15*(2).
Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*(20), 22–33.
Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087.
Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1286–1291.
Shephard, E., Stern, E. R., van den Heuvel, O. A., Costa, D. L. C., Batistuzzo, M. C., Godoy, P. B. G., Lopes, A. C., Brunoni, A. R., Hoexter, M. Q., Shavitt, R. G., Reddy, Y. C. J., Lochner, C., Stein, D. J., Simpson, H. B., & Miguel, E. C. (2021). Toward a neurocircuit-based taxonomy to guide treatment of obsessive–compulsive disorder. In *Molecular Psychiatry* (Vol. 26, Issue 9, pp. 4583–4604). Nature Publishing Group.
Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I Was So Sure! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology*, *7*, 218.
Silverstone, P. H., & Salsali, M. (2003). Low self-esteem and psychiatric patients: Part I - The relationship between low self-esteem an psychiatric diagnosis. *Annals of General Hospital Psychiatry*, *2*(1), 1–9.
Singmann, H., Bolker, B., & Westfall, J. (2015). *Analysis of Factorial Experiments, package "afex."* 1–44. http://cran.r-project.org/package=afex
Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. *New Methods in Cognitive Psychology*, 4–31.
Sip, K. E., Gonzalez, R., Taylor, S. F., & Stern, E. R. (2018). Increased loss aversion in unmedicated patients with obsessive-compulsive disorder. *Frontiers in Psychiatry*, *8*(1), 309.
Snorrason, I., Lee, H. J., de Wit, S., & Woods, D. W. (2016). Are nonclinical obsessive-compulsive symptoms associated with bias toward habits? *Psychiatry Research*, *241*, 221–223.
Soderstrom, N. C., Davalos, D. B., & Vázquez, S. M. (2011). Metacognition and depressive realism: Evidence for the level-of-depression account. *Cognitive Neuropsychiatry*, *16*(5), 461–472.
Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792.
Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin*, *139*(1), 213–240.
Spearman, C. (1923). The nature of "intelligence" and the principles of cognition. In *The nature of "intelligence" and the principles of cognition*. Macmillan London.
Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.
Spurrier, M., & Blaszczynski, A. (2014). Risk Perception in Gambling: A Systematic Review. *Journal of Gambling Studies*, *30*(2), 253–276.
Steenbergh, T. A., Meyers, A. W., May, R. K., & Whelan, J. P. (2002). Development and validation of the Gamblers' Beliefs Questionnaire. *Psychology of Addictive Behaviors*, *16*(2), 143–149.
Stein, D. J. (2002). Obsessive-compulsive disorder. *Lancet*, *360*(9330), 397–405.
Stein, D. J., Costa, D. L. C., Lochner, C., Miguel, E. C., Reddy, Y. C. J., Shavitt, R. G., van den Heuvel, O. A., & Simpson, H. B. (2019). Obsessive–compulsive disorder. In *Nature Reviews Disease Primers* (Vol. 5, Issue 1, pp. 1–21). Nature Publishing Group.
Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92.
Stern, E. R., Welsh, R. C., Fitzgerald, K. D., Gehring, W. J., Lister, J. J., Himle, J. A., Abelson, J. L., & Taylor, S. F. (2011). Hyperactive Error Responses and Altered Connectivity in Ventromedial and Frontoinsular Cortices in Obsessive-Compulsive Disorder. *Biological Psychiatry*, *69*(6), 583–591.
Stewart, N., Scheibehenne, B., & Pachur, T. (2015). Psychological parameters have units : A bug fix for

stochastic prospect theory and other decision models. *PsyArXiv*, 1–11.
Stip, E. E., Letrouneau, G., & Letourneau, G. (2009). Normality and pathology. *La Revue Canadienne de Psychiatrie*, *54*(3), 140–151.
Stone, C., Mattingley, J. B., & Rangelov, D. (2022). On second thoughts: changes of mind in decision-making. *Trends in Cognitive Sciences*, *26*(5), 419–431.
Stone, E. R., Dodrill, C. L., & Johnson, N. (2001). Depressive cognition: A test of depressive realism versus negativity using general knowledge questions. *Journal of Psychology: Interdisciplinary and Applied*, *135*(6), 583–602.
Strauss, G. P., Waltz, J. A., & Gold, J. M. (2014). A review of reward processing and motivational impairment in schizophrenia. In *Schizophrenia Bulletin* (Vol. 40, pp. 107–116). Oxford Academic.
Strayhorn, J. M. (2002). Self-Control: Theory and Research. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(1), 7–16.
Subramaniam, M., Wang, P., Soh, P., Vaingankar, J. A., Chong, S. A., Browning, C. J., & Thomas, S. A. (2015). Prevalence and determinants of gambling disorder among older adults: A systematic review. *Addictive Behaviors*, *41*, 199–209.
Summerfeldt, L. J., Kloosterman, P. H., Antony, M. M., & Swinson, R. P. (2014). Examining an obsessive-compulsive core dimensions model: Structural validity of harm avoidance and incompleteness. *Journal of Obsessive-Compulsive and Related Disorders*, *3*(2), 83–94.
Suzuki, S., Zhang, X., Dezfouli, A., Braganza, L., Fulcher, B. D., Parkes, L., Fontenelle, L. F., Harrison, B. J., Murawski, C., Yücel, M., & Suo, C. (2023). Individuals with problem gambling and obsessive-compulsive disorder learn through distinct reinforcement mechanisms. *PLOS Biology*, *21*(3).
Swedo, S. E., Rapoport, J. L., Leonard, H., Lenane, M., & Cheslow, D. (1989). Obsessive-Compulsive Disorder in Children and Adolescents: Clinical Phenomenology of 70 Consecutive Cases. *Archives of General Psychiatry*, *46*(4), 335–341.
Sylvain, C., Ladouceur, R., & Boisvert, J. M. (1997). Cognitive and behavioral treatment of pathological gambling: A controlled study. *Journal of Consulting and Clinical Psychology*, *65*(5), 727–732.
Szu-Ting Fu, T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 699–704.
Tekcan, A. İ., Topçuoğlu, V., & Kaya, B. (2007). Memory and metamemory for semantic information in obsessive–compulsive disorder. *Behaviour Research and Therapy*, *45*(9), 2164–2172.
Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., & Mohr, P. N. C. (2019). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, *3*(6), 625–635.
Thorsen, A. L., Hagland, P., Radua, J., Mataix-Cols, D., Kvale, G., Hansen, B., & van den Heuvel, O. A. (2018). Emotional Processing in Obsessive-Compulsive Disorder: A Systematic Review and Meta-analysis of 25 Functional Neuroimaging Studies. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(6), 563–571.
Ting, C. C., Palminteri, S., Engelmann, J. B., & Lebreton, M. (2020). Robust valence-induced biases on motor response and confidence in human reinforcement learning. *Cognitive, Affective and Behavioral Neuroscience*, *20*(6), 1184–1199.
Ting, C., Salem-Garcia, N., Palminteri, S., & Engelmann, J. B. (2023). Neural and computational underpinnings of biased confidence in human reinforcement learning. *BioRxiv*.
Tochkov, K. (2008). The role of anticipated regret and risk seeking in gambling behavior. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *69*(1), 702.
Toffolo, M. B. J., van den Hout, M. A., Engelhard, I. M., Hooge, I. T. C., & Cath, D. C. (2016). Patients With Obsessive-Compulsive Disorder Check Excessively in Response to Mild Uncertainty. *Behavior Therapy*, *47*(4), 550–559.
Toledano, S., Guzick, A. G., McCarty, R. J., Browning, M. E., Downing, S. T., Geffken, G. R., & McNamara, J. P. H. (2020). An investigation of self-esteem in the treatment of OCD. *Journal of Obsessive-Compulsive and Related Disorders*, *27*, 100563.
Tolin, D. F., Abramowitz, J. S., Brigidi, B. D., Amir, N., Street, G. P., & Foa, E. B. (2001). Memory and memory confidence in obsessive – compulsive disorder. *Behaviour Research and Therapy*, *39*, 913–927.
Toneatto, T. (1999). Cognitive Psychopathology of Problem Gambling For personal use only. *Substance Use & Misuse*, *34*(11), 595–6399.
Toneatto, T., & Gunaratne, M. (2009). Does the treatment of cognitive distortions improve clinical outcomes for problem gambling? *Journal of Contemporary Psychotherapy*, *39*(4), 221–229.
Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*,

398(6729), 704–708.
Tuna, Ş., Tekcan, A. I., & Topçuoğlu, V. (2005). Memory and metamemory in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *43*(1), 15–27.
Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*.
Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, *96*(2), 348–354.
Van Den Hout, M., & Kindt, M. (2003a). Phenomenological validity of an OCD-memory model and the remember/know distinction. *Behaviour Research and Therapy*, *41*(3), 369–378.
Van Den Hout, M., & Kindt, M. (2003b). Repeated checking causes memory distrust. *Behaviour Research and Therapy*, *41*(3), 301–316.
van Holst, R. J., van den Brink, W., Veltman, D. J., & Goudriaan, A. E. (2010). Why gamblers fail to win: A review of cognitive and neuroimaging findings in pathological gambling. *Neuroscience and Biobehavioral Reviews*, *34*(1), 87–107.
Van Holst, R. J., Veltman, D. J., Bchel, C., Van Den Brink, W., & Goudriaan, A. E. (2012). Distorted expectancy coding in problem gambling: Is the addictive in the anticipation? *Biological Psychiatry*, *71*(8), 741–748.
Van Holst, R. J., Veltman, D. J., Van Den Brink, W., & Goudriaan, A. E. (2012). Right on cue? Striatal reactivity in problem gamblers. *Biological Psychiatry*, *72*(10).
van Leeuwen, W. A., van Wingen, G. A., Luyten, P., Denys, D., & van Marle, H. J. F. (2020). Attachment in OCD: A meta-analysis. *Journal of Anxiety Disorders*, *70*, 102187.
van Maanen, L., Forstmann, B. U., Keuken, M. C., Wagenmakers, E. J., & Heathcote, A. (2016). The impact of MRI scanner environment on perceptual decision-making. *Behavior Research Methods*, *48*(1), 184–200.
Van Marcke, H., Le Denmat, P., Verguts, T., & Desender, K. (2022). Manipulating prior beliefs causally induces under- and overconfidence. *BioRxiv*, 1–22.
van Timmeren, T., Daams, J. G., van Holst, R. J., & Goudriaan, A. E. (2018). Compulsivity-related neurocognitive performance deficits in gambling disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *84*, 204–217.
Van Timmeren, T., Van Holst, R. J., & Goudriaan, A. E. (2023). Striatal ups or downs? Neural correlates of monetary reward anticipation, cue reactivity and their interaction in alcohol use disorder and gambling disorder. *Journal of Behavioral Addictions*, *12*(2), 571–583.
Vanes, L. D., & Dolan, R. J. (2021). Transdiagnostic neuroimaging markers of psychiatric risk: A narrative review. *NeuroImage: Clinical*, *30*(11).
Vinckier, F., Gaillard, R., Palminteri, S., Rigoux, L., Salvador, A., Fornito, A., Adapa, R., Krebs, M. O., Pessiglione, M., & Fletcher, P. C. (2016). Confidence and psychosis: A neuro-computational account of contingency learning disruption by NMDA blockade. *Molecular Psychiatry*, *21*(7), 946–955.
Vizcaino, E. J. V., Fernandez-Navarro, P., Blanco, C., Ponce, G., Navio, M., Moratti, S., & Rubio, G. (2013). Maintenance of attention and pathological gambling. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *27*(3), 861–867.
Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–352.
Vujanovic, A. A., Meyer, T. D., Heads, A. M., Stotts, A. L., Villarreal, Y. R., & Schmitz, J. M. (2017). Cognitive-behavioral therapies for depression and substance use disorders: An overview of traditional, third-wave, and transdiagnostic approaches. In *American Journal of Drug and Alcohol Abuse* (Vol. 43, Issue 4, pp. 402–415). Taylor & Francis.
Wagner, B., Mathar, D., & Peters, J. (2022). Gambling Environment Exposure Increases Temporal Discounting but Improves Model-Based Control in Regular Slot-Machine Gamblers. *Computational Psychiatry*, *6*(1), 142–165.
Warman, D. M. (2008). Reasoning and delusion proneness: Confidence in decisions. *Journal of Nervous and Mental Disease*, *196*(1), 9–15.
Weinstein, A., Dorani, D., Elhadif, R., Bukovza, Y., & Yarmulnik, A. (2015). Internet addiction is associated with social anxiety in young adults. *Annals of Clinical Psychiatry*, *27*(1), 4–9.
Wells, A. (2019). Breaking the Cybernetic Code: Understanding and Treating the Human Metacognitive Control System to Enhance Mental Health. *Frontiers in Psychology*, *10*(12).
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.
Wells, A., & Papageorgiou, C. (1998). Relationships between worry, obsessive-compulsive symptoms and meta-cognitive beliefs. *Behaviour Research and Therapy*, *36*(9), 899–913.
Welte, J. W., Barnes, G. M., Tidwell, M. C. O., & Wieczorek, W. F. (2017). Predictors of Problem Gambling in the U.S. *Journal of Gambling Studies*, *33*(2), 327–342.
Whitton, A. E., Treadway, M. T., & Pizzagalli, D. A. (2015). Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current Opinion in Psychiatry*, *28*(1), 7–12.
Wiehler, A., Chakroun, K., & Peters, J. (2021). Attenuated directed exploration during reinforcement learning in gambling disorder. *Journal of Neuroscience*, *41*(11), 2512–2522.
Wiehler, A., & Peters, J. (2015). Reward-based decision making in pathological gambling: The roles of risk and delay. *Neuroscience Research*, *90*, 3–14.
Williams, L. M., Gatt, J. M., Schofield, P. R., Olivieri, G., Peduto, A., & Gordon, E. (2009). "Negativity bias" in risk for depression and anxiety: Brain-body fear circuitry correlates, 5-HTT-LPR and early life stress. *NeuroImage*, *47*(3), 804–814.
Williams, N. (2014). The GAD-7 questionnaire. *Occupational Medicine*, *64*(3), 224.
Wilson, R. P., Colizzi, M., Bossong, M. G., Allen, P., Kempton, M., Abe, N., Barros-Loscertales, A. R., Bayer, J., Beck, A., Bjork, J., Boecker, R., Bustamante, J. C., Choi, J. S., Delmonte, S., Dillon, D., Figee, M., Garavan, H., Hagele, C., Hermans, E. J., … Bhattacharyya, S. (2018). The Neural Substrate of Reward Anticipation in Health: A Meta-Analysis of fMRI Findings in the Monetary Incentive Delay Task. *Neuropsychology Review*, *28*(4), 496–506.
Wright, A. G. C. (2011). Qualitative and quantitative distinctions in personality disorder. *Journal of Personality Assessment*, *93*(4), 370–379.
Wu, Y., Kennedy, D., Goshko, C. B., & Clark, L. (2021). "Should've known better": Counterfactual processing in disordered gambling. *Addictive Behaviors*, *112*.
Wyckmans, F., Otto, A. R., Sebold, M., Daw, N., Bechara, A., Saeremans, M., Kornreich, C., Chatard, A., Jaafari, N., & Noël, X. (2019). Reduced model-based decision-making in gambling disorder. *Scientific Reports*, *9*(1), 1–10.
Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665.
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1310–1321.
Yücel, M., Carter, A., Harrigan, K., van Holst, R. J., & Livingstone, C. (2018). Hooked on gambling: a problem of human or machine design? *The Lancet Psychiatry*, *5*(1), 20–21.
Yücel, M., Harrison, B. J., Wood, S. J., Fornito, A., Wellard, R. M., Pujol, J., Clarke, K., Phillips, M. L., Kyrios, M., Velakoulis, D., & Pantelis, C. (2007). Functional and biochemical alterations of the medial frontal cortex in obsessive-compulsive disorder. *Archives of General Psychiatry*, *64*(8), 946–955.
Zeelenberg, M., & Pieters, R. (2004). Consequences of regret aversion in real life: The case of the Dutch postcode lottery. *Organizational Behavior and Human Decision Processes*, *93*(2), 155–168.
Zhang, Z., Wang, M., Miao, X., Li, Y., Hitchman, G., & Yuan, Z. (2017). Individuals with high obsessive-compulsive tendencies or undermined confidence rely more on external proxies to access their internal states. *Journal of Behavior Therapy and Experimental Psychiatry*, *54*, 263–269.
Zitterl, W., Urban, C., Linzmayer, L., Aigner, M., Demal, U., Semler, B., & Zitterl-Eglseer, K. (2001). Memory deficits in patients with DSM-IV obsessive-compulsive disorder. *Psychopathology*, *34*(3), 113–117.
Zorowitz, S., Niv, Y., & Bennett, D. (2021). Inattentive responding can induce spurious associations between task behavior and symptom measures. *PsyArXiv*.
Zung, W. W. K. (1965). A Self-Rating Depression Scale. *Archives of General Psychiatry*, *12*(1), 63–70.

PhD portfolio

Name: Monja Hoven

Department: Department of Psychiatry, Amsterdam University Medical Center, Location AMC

Supervisors: Prof. dr. Damiaan Denys, Dr. Ruth van Holst, Dr. Judy Luigjes

PhD period: June 2019 – August 2023

PhD training

Courses	**Year**	**ECT**
UNIX	2020	0.5
Basic Legislation in Science (BROK)	2020	1.5
Practical Biostatistics	2020	1.4
3T MRI scan qualification	2021	0.5
Radiation protection	2021	1.0
Advanced Topics in Biostatistics	2021	1.5
Talents in PhD	2022	0.2
Seminars, workshops and master classes		
Summer School Model-Based Cognitive Neuroscience	2019	2
Computational Psychiatry Course, Zürich, Switzerland (Online Edition)	2020	1.5
BeOnline Research Gorilla Conference	2020	0.2
Metacognition Satellite Workshop, New York City, United States of America	2023	0.2
Analyzing Brain-Body Interaction, New York City, United States of America	2023	0.2
Oral presentations		
The Neural Signature of Confidence and the Influence of Incentives *Spinoza Centre Neuroimaging Meeting, Amsterdam*	2019	0.5
The role of attention and loss aversion in decision-making under risk in gambling disorder: an eye-tracking study *Amsterdam Brain and Cognition Symposium, Amsterdam*	2019	0.2
Neural underpinnings of confidence encoding and the influence of incentives *Research Meeting Psychiatry Department Amsterdam UMC, Amsterdam*	2020	0.5
Confidence in Compulsive Disorders (invited) *Donders Institute BCS Seminar, Nijmegen*	2021	0.5

DOPACON: Gok onderzoek Amsterdam UMC *Jellinek Utrecht*	2021	0.2
De rol van zekerheid in obsessieve compulsieve stoornis *Dag van de Dwang, Nijmegen*	2022	0.2
DOPACON: The relationship between confidence and dopamine synthesis capacity in Gambling Disorder *Onderzoeksdag Academische Werkplaats Verslaving Jellinek, Amsterdam*	2022	0.2
Confidence in psychiatry *Research Day Department of Psychiatry Amsterdam UMC Location AMC, Amsterdam*	2022	0.5
Confidence & Reinforcement Learning in Psychiatry *Lebreton Lab, Paris School of Economics, Paris*	2023	0.5
Confidence & Reinforcement Learning in Psychiatry (invited) *Human Reinforcement Learning Lab, École Normale Supérieure, Paris*	2023	0.5
Poster presentations		
The influence of compulsivity and incentives on confidence in obsessive-compulsive disorder and pathological gambling *Amsterdam Neuroscience Annual Meeting, Amsterdam*	2018	0.5
Confidence and metacognition in psychiatry *Summer School Model-Based Cognitive Neuroscience, University of Amsterdam, Amsterdam*	2019	0.5
Abnormalities of confidence in psychiatry *Amsterdam Neuroscience Annual Meeting, Amsterdam*	2019	0.5
Confidence and the influence of incentive in OCD and GD *European College of Neuropsychopharmacology, Online Edition*	2020	0.5
How do confidence and self-beliefs relate in psychiatry: a trans-diagnostic approach *Society of Biological Psychiatry, New Orleans, United States of America*	2022	0.5
The role of attention in decision-making under risk in gambling disorder: an eye-tracking study *European College of Neuropsychopharmacology, Vienna, Austria*	2022	0.2
Differences in metacognitive functioning between obsessive-compulsive disorder patients and highly compulsive individuals from the general population *Association for the Scientific Study of Consciousness, New York City, United States of America*	2023	0.5

(Inter)national conferences		
Amsterdam Metacognition Symposium, Amsterdam	2019	0.25
Amsterdam Brain and Cognition Research Day, Amsterdam	2019	0.25
European College of Neuropsychopharmacology Virtual Meeting	2020	0.75
Society of Biological Psychiatry Virtual Meeting	2021	0.75
Amsterdam Brain and Cognition Research Day (organizer), Amsterdam	2021	0.25
BCN Symposium Nothing but the Truth, Groningen	2021	0.25
Society of Biological Psychiatry, New Orleans, United States of America	2022	0.75
International Conference on Behavioral Addiction, Nottingham, United Kingdom	2022	0.5
Association for the Scientific Study of Consciousness, New York City, United States of America	2023	0.75
Supervising		
Maura Fraikin (BSc internship) - Unravelling the Role of Confidence and Negative Outcome Anticipation in Obsessive-Compulsive Disorder	2020	2
Channah Osinga (BSc internship) - Unraveling confidence abnormalities and incentivization effects in obsessive-compulsive disorder and gambling disorder: a fMRI study	2020	2
Najoua Marroun (BSc internship) – Confidence and feedback processing in OCD patients	2020	2
Sabine Gnodde (MSc internship) - Exploring the role of chronic stress and loss on confidence abnormalities in obsessive-compulsive disorder	2020	2
Fabiënne Meijboom (BSc internship) - Exploring Confidence Abnormalities during Symptom Provocation in Patients with Obsessive-Compulsive Disorder	2021	2
Katja Cornelissen (MSc internship) - The Formation of Global Self-Performance Estimates from Local Confidence in Obsessive Compulsive Disorder and its Relationship with the Feeling of Autonomy	2021	2
Tosca Mulder (MSc internship) - Are you sure? Dynamics of Action, Confidence and their relationship under uncertainty in Obsessive-Compulsive Disorder	2021-2022	2
Eva van den Assum (MSc internship) - Striatal dopamine synthesis capacity and neuromelanin in the substantia nigra in gambling disorder	2023	2
Savina van Rielova (MSc internship) - The relationship between gambling disorder, dopamine synthesis capacity, and cognitive flexibility	2023	2

Katja Cornelissen (Research Assistant)	2021-2022	2
Other		
Journal club Psychiatry	2019-2023	2
Reviewer for Addictive Behaviors, European Addiction Research, Consciousness and Cognition, Psychological Medicine, Biological Psychiatry, European Archives of Psychiatry and Neuroscience, Scientific Reports, International Gambling Studies	2020-2023	
Co-organizer of Amsterdam Metacognition Symposium	2019	0.2
Co-organizer of Amsterdam Brain and Cognition Research Day	2021	1
Research Visit Paris School of Economics & École Normale Supérieure, Paris, France	2023	1

List of publications

Part of this thesis

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.

Hoven, M., Brunner, G., de Boer, N. S., Goudriaan, A. E., Denys, D., van Holst, R. J., ... & Lebreton, M. (2022). Motivational signals disrupt metacognitive signals in the human ventromedial prefrontal cortex. *Communications Biology*, *5*(1), 244.

Hoven, M., de Boer, N. S., Goudriaan, A. E., Denys, D., Lebreton, M., van Holst, R. J., & Luigjes, J. (2022). Metacognition and the effect of incentive motivation in two compulsive disorders: Gambling disorder and obsessive–compulsive disorder. *Psychiatry and Clinical Neurosciences*, *76*(9), 437-449.

Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health*, 1-9.

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,* 1-10.

Hoven, M., Mulder, T., Denys, D., van Holst, R. J. & Luigjes J. OCD patients show lower confidence and higher error sensitivity while learning under volatility compared to healthy and highly compulsive samples from the general population. *Submitted*.

Hoven, M., Luigjes, J., van Holst R.J. Learning and metacognition under volatility in gambling disorder: lower learning rates and distorted coupling between action and confidence. *Submitted*.

Hoven, M., Hirmas, A., Engelmann, J., & van Holst, R. J. (2023). The role of attention in decision-making under risk in gambling disorder: an eye-tracking study. *Addictive Behaviors*, *138*, 107550.

Hoven, M., Hirmas, A.,Engelmann, J.B., van Holst, R.J. (2023). Confidence and risky decision-making in gambling disorder. *Journal of Behavioral Addictions,* 1-7.

Not part of this thesis

Hoven, M., Schluter, R. S., Schellekens, A. F., van Holst, R. J., & Goudriaan, A. E. (2023). Effects of 10 add-on HF-rTMS treatment sessions on alcohol use and craving among detoxified inpatients with alcohol use disorder: a randomized sham-controlled clinical trial. *Addiction*, *118*(1), 71-85.

Bergamin, J., **Hoven, M.**, van Holst, R.J., Bockting, C.L., Denys, D., Nevicka, B., Luigjes, J. (2023). Development and validation of the Autonomy Scale Amsterdam
Submitted.

van Hooijdonk, C.F.M., van der Pluijm, M., Smith, C., Yaqub, M., van Velden, F.H.P., Horga, G., Wengler, K., **Hoven, M.**, van Holst, R.J., de Haan, L., Selten, J., van Amelsvoort, T.A.M.J., Booij, J., van de Giessen E. (2023). Striatal dopamine synthesis capacity and neuromelanin in the substantia nigra: a multimodal imaging study in schizophrenia and healthy controls. *Under revision at Neuroscience Applied*.

Dankwoord / Acknowledgments

Het boek staat op papier! Nu rest alleen nog het belangrijkste en leukste deel, en tevens het deel waar ik het vaakst over heb gefantaseerd gedurende de lange fietstochten naar het AMC (na al die jaren heb ik zeker zo'n 20.000 km de tijd gehad om er over na te denken). Tijd om te reflecteren op deze mooie jaren, die lang niet zo mooi waren geweest zonder de hulp, toewijding, steun, vriendschap en liefde van zoveel bijzondere mensen, die ik stuk voor stuk enorm dankbaar ben.

Ten eerste wil ik alle deelnemers bedanken die hebben meegedaan aan onze onderzoeken. Ontzettend bedankt voor jullie inzet. Zonder jullie was dit onderzoek en proefschrift er nooit geweest. Ik ben erg dankbaar voor jullie oprechte interesse en bijdrage aan de wetenschap en ik wens jullie allen het allerbeste.

Zonder mijn (co-)promotoren was dit proefschrift er vanzelfsprekend ook nooit geweest. Prof. dr. Denys, beste **Damiaan**, ik wil je van harte bedanken voor de kans om dit promotietraject aan te gaan. Je constructieve feedback en verfrissende kijk op '*confidence*' gaf mij altijd weer stof tot denken. Dr. Luigjes, lieve **Judy**, in het najaar van 2016 kwam ik solliciteren op een stageplek bij jou, en ik ben nooit meer weggegaan. Dat zegt genoeg over de fijne tijd die ik heb gehad, en die heb ik voor een groot deel aan jou te danken. De fijne, veilige, ontspannen sfeer die jij creëert heeft ervoor gezorgd dat ik me meteen op mijn gemak voelde. Het was een warm bad en je deur stond altijd open. Ik heb bewondering voor jouw kijk op het leven (zowel binnen als buiten de wetenschap) en heb enorm veel van je geleerd. Zowel op wetenschappelijk vlak, waarbij je kritische en snelle denkvermogen mij ook altijd weer tot nieuwe ideeën bracht, als op persoonlijk vlak, waarbij je broodnodige relativeringsvermogen alles altijd weer in perspectief plaatste. Dr. Van Holst, lieve **Ruth**, wat heb ik het ook met jou enorm getroffen. Ik ben erg onder de indruk van jouw doorzettingsvermogen, ambitie en drive, welke je feilloos combineert met je open en warme persoonlijkheid. Ook bij jou was het meteen een warm bad en herkende ik de oneindige nieuwsgierigheid naar hoe dingen nou werken en in elkaar zitten. En wat een voorrecht is het geweest om deze jaren samen een stukje van de puzzel te leggen. Ik heb zoveel van je geleerd, binnen de academie op inhoudelijk vlak en interpersoonlijk vlak, als buiten de academie met de altijd wijze levenslessen. Met zijn drieën zijn we een super *confidence* team. Onze meetings met z'n drieën, vaak onder het genot van thee, koffie, chocola of ijsjes, begonnen steevast met de vraag "hoe is het met je?". Ik heb dat enorm gewaardeerd, en ik heb me daardoor altijd gehoord en begrepen gevoeld. Ik wil jullie ontzettend bedanken voor de kans om dit avontuur met jullie aan te gaan, het luisterend oor, het enorme vertrouwen dat jullie al die jaren in mij hebben gehad, de vrijheid om dit pad zelf te bewandelen en alle steun als het even wat

minder ging. Jullie mentorschap heeft me deze jaren op zowel professioneel als persoonlijk vlak gevormd, en ik realiseer me dat dit erg bijzonder is. Ik had me geen betere begeleiders kunnen wensen. Bedankt voor alles.

Beste leden van de promotie commissie, beste prof. dr. Ridderinkhof, prof. dr. de Bruijn, dr. van Gaal, dr. de Wit, dr. Visser, hartelijk bedankt voor het lezen en beoordelen van mijn proefschrift en het opponeren tijdens de verdediging. Ik kijk er naar uit! Prof. dr. Goudriaan, beste **Anneke**, hartelijk dank voor de fijne samenwerking. Ik ben blij dat ik een bijdrage heb kunnen leveren aan het rTMS project, en met onze samenwerking op een aantal *confidence* projecten. Ik kijk met een glimlach terug op ons congres in Nottingham, al duurde de terugweg wel wat lang! Dr. Rouault, dear **Marion**, I am honored that you are part of my committee and very grateful for our collaboration over the past years. Your intelligence and kindness, along with your infectious enthusiasm for our research field have been a big source of inspiration for me and it has been a privilege to learn from you.

Dear **Sofia Bonati**, thank you for creating beautiful works of art and allowing me to use your work on the cover of my thesis. Right before I started this PhD journey I was in Sri Lanka where your *Eudoxia* was hanging in my hotel room. Ever since I returned from that trip she has been hanging on the wall in my small home office.

Wetenschap is een teamsport, en zonder de hulp van ontzettend veel mensen was dit proefschrift dan ook niet tot stand gekomen. **Bas** Brons, ontzettend bedankt voor je tomeloze inzet voor de werving van deelnemers voor onze studies, zonder welke het nooit was gelukt. Ook wil ik de Jellinek, en in het bijzonder **Loes** Marquenie bedanken voor de hulp bij het opzetten van de werving. Hervitas, AGOG en OCDnet: bedankt voor de hulp met werving van onze deelnemers.

Team Spinoza: bedankt voor de hulp en assistentie bij onze MRI onderzoeken. Voor het PET onderzoek, waar helaas geen artikel van in mijn proefschrift staat, maar waar ik wel een groot deel van mijn promotie aan besteed heb, zijn er ook veel mensen te bedanken. Dank aan de radiologie afdeling, **Sandra**, **Paul**, voor jullie hulp bij het leren scannen en het opzetten van de data structuren. Prof. dr. Booij, beste **Jan**, dr. van Giessen, beste **Elsmarieke** en prof. dr. Cools, beste **Roshan**, bedankt voor het delen van jullie schat aan kennis en de hulp bij het opzetten van de PET studie. **Marieke**, bedankt dat je me hebt ingewijd in het reilen en zeilen van PET in de praktijk en inmiddels alle gezelligheid op de psychiatrie. Het PET team: **Meng-Fong**, **Ehsan** en **Martijn**, enorm bedankt voor jullie hulp, expertise, inzet en flexibiliteit tijdens het scannen van onze deelnemers. Martijn, bedankt voor alle gezellige anderhalf uurtjes bij de scanner, ik heb er van genoten: het ga je goed in Delft!

Dear prof. dr. Engelmann, dear **Jan**, thank you for the great collaboration over these past years. As external supervisor, you have graded my first masters thesis and we now co-author several papers! It has been a great pleasure working with you, and I want to thank you for creating a welcoming atmosphere. Dear **Alejandro**, thank you for all the help and collaboration. It has been a pleasure to work with you these past years, and I'm sure you will have a wonderful scientific career ahead of you.

Dear dr. Lebreton, dear **Maël**, I would like to thank you for these past years. Back in 2017, when I started my internship in Amsterdam, I was lucky enough that you had just started collaborating with my supervisors on this very interesting topic of *'confidence'*, on which I have now written an entire thesis! You have been a great inspiration in terms of critical thinking and scientific mindset, but I'd also like to thank you for all the practical help with analyses and writing, without which the first chapters of this thesis would not have been as they are now. My visit to Paris towards the end of my PhD was a wonderful experience, marked by your and your team's warm welcome. Thank you for teaching me all things computational modeling, but also which French beers and Vietnamese restaurants are the best, which of course are all equally important things in life. Dear dr. Palminteri, dear **Stefano**, thank you for your hospitality and sharing your knowledge with me during my stay in Paris. Your team and your work are truly an inspiration to me.

To all the lovely people from room P3.68: **Constance**, **Antonis**, **Clementine**, **Lily**, **Viv**, **Pascale**, **Craig**, **Aurelién**, thank you for welcoming me at PSE and for the croissant Thursdays, lunches and drinks! I wish you all the best during your PhD journeys and beyond.

Ook binnen de psychiatrie afdeling zijn er een hoop collega's te bedanken voor de fijne tijd de afgelopen jaren. De mensen van de AIAR gang tijdens het prille begin van mijn academische carrière op de derde: **Anne Marije**, **Marleen**, **Filipa**, **Anneke**, **Suzan**, **Masha**, **Tim**, bedankt voor de gezellige etentjes, borrels en lunches op het AMC. **Masha**, ik hoop dat we elkaar snel weer tegen komen tijdens de Spaanse les, wie weet onder genot van een sangriaatje! **Tim**, bedankt voor alle gezelligheid, maar ook alle adviezen toen ik aan de wieg stond van mijn PhD. Ik denk met plezier terug op Nottingham, vooral toen we op de terugweg eindelijk ons blikje bier mochten openen toen de wielen van het vliegtuig van de grond kwamen! (Oud) collega's van de andere gang(en): **Gosse**, **Willem**, **Nadine**, **Dilan**, **Paul**, **Gabry**, **Melisse**, **Isidoor**, **Guido**, **Jorien**, **Junus**, **Joost**, **Martine**, **Merel**, **Marieke**, dank voor alle gezellige koffie momentjes, lunches en borrels. Mijn oud-kamergenootjes: **Dominika** en **David**, dank voor de gezellige kletspraatjes op kantoor. Dear **Karoline**, it has been so much fun sharing an office with you at the AMC, on-and-off. I hope we still get to eat an oliebol

this winter together. **Marian**, **Andrea**, **Ingeborg** en **Karin**, ontzettend bedankt voor alle ondersteuning en hulp bij de bureaucratische rompslomp die komt kijken bij promoveren én post ontvangen op het AMC. **Jessy**, bedankt voor de fijne samenwerking bij het ontwikkelen van de autonomie vragenlijst, en vooral voor de gezelligheid en het kunnen delen van onze PhD struggles: you got this!

Zonder de hulp van mijn stagiaires was dit proefschrift er ook niet geweest. **Maura** en **Channah**, jullie waren mijn eerste stagiaires. Bedankt voor jullie hulp met het opzetten van de COCON studie. **Sabine**, bedankt voor al je inzet, hulp en gezelligheid in de begin fase van mijn PhD. Hoe cool dat je nu zelf ook een PhD aan het doen bent, ik hoop dat onze paden elkaar nog zullen kruisen! **Najoua**, ondanks dat je maar een korte periode stage hebt gelopen heb je toch enorm veel geholpen met de opzet van de studies: dank daarvoor. **Fabiënne**, bedankt voor de ontzettend gezellige tijd en jouw aanstekelijke vrolijkheid. Ik heb genoten van jouw stage periode en ons Italiaanse afscheidsdinertje, en ik wil je graag bedanken voor je harde werk bij de werving en het testen van onze deelnemers voor de COCON studie. **Tosca**, jij bent van onschatbare waarde geweest bij zowel de werving, testen als het analyseren van data binnen het COCON project, en we staan dan ook gezellig samen op het artikel. **Eva**, bedankt voor al je hulp tijdens het DOPACON project. Mede dankzij jou waren de testdagen altijd een feestje, en ik heb met veel plezier samen gewerkt. **Savina**, voor jou geldt hetzelfde, zonder jou was het DOPACON project geen succes geworden. Dank voor je tomeloze inzet bij alle aspecten van de studie. Ik wens jullie allemaal enorm veel succes in jullie verdere studie en -wie weet- academische loopbaan!

Lieve **Katja**, waar zou ik toch zijn zonder jou? Je begon als stagiair op het COCON project, waar je me hielp met de opzet en praktische uitvoering. Al snel wisten we dat we je moesten houden, en gelukkig wou je blijven als onderzoeksassistent voor het DOPACON project. Je was de spil in ons team en je was altijd van alles op de hoogte. Ik ben enorm trots dat je nu je eigen onderzoekspad in bent geslagen. Ik heb een super leuke tijd met je gehad op het AMC, waar we over van alles en nog wat konden kletsen en ik hoop dat we elkaar blijven zien. Je bent een topper!

Het is een voorrecht om lieve mensen om je heen te hebben die weten wat het doen van een PhD inhoudt, en zo alle bijkomende struggles begrijpen. Lieve **Carmen**, je was mijn steun en toeverlaat wat betreft de PET studie en een baken op wie ik al mijn vragen kon afvuren. Ik vond het samen werken, maar vooral het samen koffie drinken, kletsen en hapjes eten enorm gezellig. Het was fijn om alle ups en downs die bij een promotietraject horen met jou te kunnen delen, en ik ben ontzettend benieuwd waar onze toekomsten ons zullen brengen. In ieder geval hoop ik dat we elkaar nog geregeld buiten het AMC om zullen zien en wie weet zelfs in Patagonië! Lieve **Hélène**, hoe leuk

dat ik jou heb ontmoet op congres in New York en dat we over van alles kunnen kletsen, en dan óók nog eens over het confidence werkveld. Ik hoop dat we elkaar vaak blijven citeren en blijven bezoeken in Gent en Amsterdam!

Lieve **Nina** en **Renée**, liefste roomies: zonder jullie was ik misschien wel nooit aan m'n PhD begonnen, want wat maakten jullie mijn begintijd op het AMC toch onvergetelijk! Lieve **Nina**, vanaf het begin van onze stages op de AIAR gang waren we meteen matties. Wat een feest was het om samen als kersverse onderzoeksassistenten op hetzelfde project ons eerste grote-mensen kantoor te delen. Van paastakken met kuikens (en konijnenschaaltjes) tot kerstboompjes met ballen en weer terug. Het spijt me dat jij altijd mijn Jesus Christ Superstar vertolking hebt moeten aanhoren, maar volgens mij heeft het je wel overtuigd om samen in 't koor te gaan! Je bent zo'n ongelofelijk lieve schat, en ook nog eens een van de slimste en meest empathische mensen die ik ken. Ik hoop op nog veel koffies, paaseitjes, bitterballen, fiets-kampeer-vakanties en nog veel meer! Lieve **Renée**, vanaf de befaamde borrel waren wij ook grote matties. Zo groot dat we er eigenhandig voor hebben gezorgd dat ik snel bij jou op kantoor kwam te zitten. Wat hadden we het altijd mega gezellig, thee-advent-kalenders, domme filmpjes, Funda zoektochten in Friesland, koffiedates en natuurlijk de weekend nabespreking op maandag ochtend, en weekend voorbespreking op dinsdag ochtend. Jouw positieve energie sleepte me altijd de week door, en je relativeringsvermogen heeft me vaak uit de stress gehaald. Lieve roomies, wat was het een feestje om met z'n drieën een kantoor te delen en heerlijk dat we elkaar nog vaak buiten werk om zien. Wandel dates, privé trompet pizza concerten, festivals: laten we dat zeker zo houden!

Lieve **Laurens** en **Nora**, mijn lieve paranimfen: zonder jullie was mijn promotietijd op het AMC een saaie boel geweest. Wat ben ik enórm blij dat jullie niet veel later dan ik ook van start gingen met jullie promotie. Lieve **Lauri**, wassup lil Goose! Vanaf onze eerste gezamenlijke borrel in de tuinen van het AMC, waar je meteen een ice in je schoot geworpen kreeg, zijn we maten. Wat ben ik blij met zo'n PhD maat die mijn humor begrijpt en waar ik dus altijd mee kan lachen. Onze reis naar Chicago en New Orleans is een van de meest memorabele uit mijn leven. Daar bleek dat je niet zo'n goede verliezer bent (wat was het ook alweer: 20-0?), en niet vies bent van een dansje. Onze fissa in New Orleans was een van de leukste avonden van mijn leven, vooral toen we de dag erna alle booty-shakers weer tegenkwamen op het congres. Bedankt voor alle leuke herinneringen, adviezen en grapjes. Ik weet zeker dat er nog vele zullen volgen. Lieve **Nora**, wij kenden elkaar van horen-zeggen, maar toen wij elkaar ontmoetten was het meteen dikke mik. We hebben zoveel leuke dingen meegemaakt de afgelopen jaren: DIY congressen toen we er niet fysiek heen konden, heerlijke etentjes, biertjes op het strand, wandelingen door de duinen, koffie & cake dates (met bladblazers en honden),

after-work sauna's, met als kers op de taart ons reisje naar Parijs. Je bent zo'n lieverd, en ik ben je enorm dankbaar voor je luisterend oor, al je advies, je nuchtere blik, je support en alle gezellige momenten samen. Zonder had ik het niet gered. Ik ben heel trots op wat en hoe je alles doet, en ik ben benieuwd of onze 5-jaars voorspellingen uit gaan komen! Lieve paranimfen, ik hoop nog op ontzettend veel borrels, vunzige deuntjes, dansjes, ice-jes (voor jullie dan), domme accentjes en gekkigheid samen. You the best!

Ik ben ook gezegend met de liefste vrienden die misschien iets minder snapten van het hele PhD gedoe, maar daardoor juist van onschatbare waarde zijn geweest voor mij gedurende deze periode. Lieve **Thomas**, inmiddels vieren we ons 18+ vriendschapsjubileum vanaf 1Gb. De basis was gelegd en wat een oneindige bron van herinneringen hebben we samen: van drie musketiers studerend aan de keukentafel van m'n moeder tot nu! Vele vakanties, feestjes, karaoke-avondjes, waarvan ik de details maar zal besparen. Je hebt een speciaal plekje in m'n hart, drie musketiers forever. Lieve **Carline**, wát een geluk dat jij in mijn leven bent gekomen. Je bent zo'n vrolijke, fijne, attente, gezellige, roze, lieve schat, en volgens mij komen wij er steeds meer achter dat we enorm op elkaar lijken. Bedankt voor alle fijne avonden, vakanties, etentjes, en alle ontspanning met jullie (en natuurlijk je werk aan het design van het boek)! Ik hoop dat we onze oud-en-nieuw, karaoke en gerechtigheid tradities er voor altijd inhouden.

Allerliefste **Hannah** en **Lynn**. Woorden schieten tekort... Zonder jullie ben ik nergens. Jullie oneindige support, interesse, gezelligheid, en innige vriendschap betekent de wereld voor mij. Zonder alle leuke dingen die we samen de afgelopen jaren hebben gedaan was dit proefschrift er ook nooit geweest: van de bijzondere dingen (Mexico & Peru), tot de meest alledaagse, bij jullie kon ik altijd weer opladen. Lieve **Hannah**, vanaf die eerste geodriehoek tot hier: wie had dat gedacht? Jouw zorgzaamheid, loyaliteit en oprechtheid zijn van enorme betekenis voor mij, en er zijn weinig mensen die mij zo goed kennen als jij. Ik ben super trots op jou als mens, en jou meemaken als de liefste mama van de aller leukste jongen op aarde, kleine Finn, is enorm bijzonder. En hoe mooi dat onze meisjesdroom van het op 10 minuutjes afstand wonen van elkaar in Amsterdam toch is gelukt, ik ben gezegend met jou! Lieve **Lynn**, zonder jou was ik nu niet mij. Een van mijn favoriete plekjes op aarde is niet ergens midden in het Andes gebergte op 4750 meter hoogte, maar op de praatstoel bij jou! Misschien zijn we ook wel een beetje elkaars meubilair, zo vertrouwd voelt het. Ik ben zo trots op het pad dat je hebt gekozen, en ik vind het zó cool dat we nu over wetenschappelijke artikelen kunnen kletsen. Ik zou een boek over onze avonturen kunnen schrijven en ik hoop dat de boerderijdroom ooit werkelijkheid wordt. Lieverds, wat ben ik blij met jullie in mijn

leven, met als bijkomende kadootjes **Thijs**, **Finn**, **Jouke** en **Pip**. Ik wens jullie de wereld. En: het is nog steeds zo gek dat we ineens al zo “volwassen” zijn, terwijl onze puber capriolen voelen als de dag van gister. En zo blijft het waarschijnlijk voelen tot in onze rollators. Ik hou van jullie.

Lieve **Nick**, vanaf de Kamperfoelie tot nu, mijn bonus broer en bamiknul. Wat hebben we een hoop meegemaakt, en wat een voorrecht om dat te mogen delen. Het is fijn om iemand te hebben die mij zo goed kent en begrijpt, en waarvan ik weet dat ik er altijd terecht kan, voor zowel ongein als problemen. En wat een geluk dat we Kana zijn tegengekomen tijdens een van onze kroegentochten. Lieve **Kana**, jij bent de allerliefste persoon die ik ken, en ik ben zo blij dat jij in Nick’s en mijn leven bent gekomen. Ik geniet enorm van onze tijd samen en onze gedeelde liefde voor eten. Het warmt mijn hart om jullie samen als ouders te zien van de allerliefste **Yuna**. あなたたちは最高です!

Lieve **Rowie** en **Richard**, wat bof ik toch met de enorme bonk positiviteit en liefde die jullie uitstralen. Wat hebben we samen prachtige avonturen meegemaakt, van België tot Lloret, Servië en weer terug. Ik kan met niemand anders mijn liefde voor koken en etentjes beter delen dan met jullie. De oprechte interesse die jullie al die jaren hebben getoond in mijn promotie was hartverwarmend en bij jullie voel ik me altijd thuis. Ik heb oneindige bewondering voor hoe jullie in het leven staan.

Lieve **Frances**, bedankt voor alle liefdevolle en gezellige tijden samen, je interesse, de wandelingen, etentjes, feestjes en gesprekken de afgelopen jaren. Ik hoop dat je nog zo lang mogelijk in Amsterdam blijft wonen zodat we elkaar zo vaak mogelijk kunnen zien. Je bent een schat en ik ben heel trots op jouw zelfstandigheid en doorzettingsvermogen.

Lieve **Pelle**, ie bint mien breur! Wat ben je toch een prachtige vent: er is denk ik niemand met wie ik zo hard kan lachen als met jou, maar tegelijkertijd ook zo goed kan praten. Met jou is het altijd een feest en samen dom doen in de kroeg is mijn favoriete activiteit. Waar het leven ons ook brengt: ik weet dat jij er altijd bent. En dan krijg ik ook lieve **Astrid** er nog gratis bij, jij tovert altijd een glimlach op m’n gezicht. Jullie zijn lieverds, big loev!

Lieve **Miron** en **Agnes**, ouwe begaaiers. Wat een mooie herinneringen hebben we, van Deventer tot EXIT en Mexico. Ik vind het tof om te zien hoe we allemaal zo ons eigen pad bewandelen en het is altijd een mega feest om samen te komen. Bedankt voor jullie tomeloze gezelligheid, interesse en geouwehoer, en ik hoop op nog veel meer van dat alles, bij voorkeur aan de Ijssel of op de Brink.

Lieve **Marc** en **Kim**, echte wereldreizigers. Bedankt voor alle fijne tijden samen, de gezelligheid, interesse, en ongekende gastvrijheid tijdens onze talloze bezoekjes aan

Groningen (en zelfs in Peru!). Het is altijd een feestje met jullie. Jullie zijn een grote inspiratiebron en ik hoop dat we ooit een keer samen een verre reis kunnen maken!

Lieve **Amber**, bedankt voor al het lachen, je spontaniteit en gezelligheid samen de afgelopen jaren, al dan niet onder het genot van een lekkere hamburger. Lieve **Karel**, bedankt voor het introduceren van de camembert en brie in mijn leven ten tijden van de B50 en voor alle fijne en leuke tijden die we samen hebben beleefd. Lieve **Aron** en **Claudia**, bedankt voor de fijne feestjes en de gezelligheid. Ik wens jullie het allerbeste toe in Schalkcity!

Mijn lieve vriendengroep uit Deventer tezamen: **Amber**, **Aron**, **Alexandra**, **Carline**, **Chelton**, **Claudia**, **Frances**, **Jeroen**, **Kana**, **Karel**, **Kim**, **Krystyna**, **Marc**, **Richard**, **Rowie**, **Samantha**, **Steven**, **Thanh**, **Thomas**, **Wout**. Ik wil jullie allemaal bedanken voor de enorme steun die ik aan jullie allemaal heb gehad. Het is bijzonder om zo'n fijne groep mensen al zo lang in mijn leven te hebben. Zo lang we naast elkaar blijven staan kunnen we samen alles aan. **Arie** zou stuk voor stuk trots op jullie zijn.

Mijn lieve schoonfamilie, wat is het fijn en bijzonder dat ik al zo veel jaren deel mag zijn van jullie gezin. Mijn lieve schoonouders, **Tuan** en **Phuong**, bedankt voor het warme welkom in jullie gezin al deze jaren. Ik heb me altijd thuis gevoeld bij jullie, en helemaal door al het ontzettend lekkere eten waar jullie ons altijd zo mee verwennen en alle goede zorgen. Bedankt voor alles. Lieve **Hiëp** & **Alyssa**, bedankt voor jullie liefdevolle gastvrijheid en natuurlijk voor de allerliefste en leukste nichtjes en neefjes ter wereld: **Fiënna**, **Alaïna** en **Mason**. Ik ben dol op jullie! Lieve schoonzus, lieve **Huyen**, bedankt voor de ontelbare fijne etentjes, drankjes, warmte en gezelligheid. Wij kunnen altijd bij je terecht en ik hoop dat je weet dat jij dat ook bij ons kan. Je bent een schat.

Mijn allerliefste broer, lieve **Jeroen**. Ik heb warme herinneringen aan onze jeugd samen en al onze geintjes en avonturen. Van pingpongen op de keukentafel en skelteren toen we klein waren tot biertjes en feestjes nu we volwassen zijn. Vlak voor ik aan mijn promotie begon waren we nog samen weg naar Spanje, waar we bijna verdwaalden bovenop een berg (sorry mama), maar uiteindelijk heb je ons toch maar mooi gered. Het is heel fijn om te weten dat je er altijd voor me zult zijn en we altijd samen enorm kunnen lachen. Bedankt voor alles, voor altijd, alle liefde! En natuurlijk ook voor het in mijn leven brengen van **Krystyna**. Lieve schoonzus, bedankt voor alle gezelligheid en warmte. Hoe leuk dat ik iemand in de familie heb met wie ik over de wetenschap kan praten! Op naar nog ontelbaar veel mooie herinneringen samen.

Lieve **papa**, ten eerste, bedankt voor het doorgeven van je doorzettingsvermogen en perfectionisme aan mij: ik heb er enorm veel aan gehad. Jouw vertrouwen in mij en je trots voor mij heeft me enorm gesteund in de afgelopen jaren. Ik kijk met enorm veel

warmte terug aan alle leuke herinneringen die we samen hebben gemaakt en ik hoop op nog veel meer mooie avonturen. Bedankt voor al je liefde!

Lieve **Jacob** en allerliefste **mama**, wat moet ik zeggen? Zonder jullie onvoorwaardelijke liefde en aanmoedigende steun was ik nooit geworden tot de persoon die ik nu ben. Lieve Jacob, je rustige en geduldige karakter, betrokkenheid en behulpzaamheid zijn zo waardevol. Je bent de rots in onze branding. Mijn allerliefste **mama**, ik kan je niet genoeg bedanken voor alles wat jij voor mij hebt gedaan. Je hebt me altijd gesteund en alle ruimte gegeven om mijn dromen na te jagen. Je stond altijd achter me, welke kant ik ook op zou gaan. Bedankt voor de onvoorwaardelijke liefde, zorgen, en geborgenheid. Jouw kracht en positiviteit zijn ongeëvenaard. Wist je nog dat ik vroeger altijd schrijver wilde worden? Hier is mijn eerste boek dan! Alle liefde in de hele wereld is voor jullie.

Liefste **Hoang**, mijn meest favoriete persoon op de wereld. Woorden beperken mijn gedachten. Bedankt dat ik al zo lang mijn lief en leed met je mag delen. Wat is het een bijzondere reis geweest tot nu toe, vanaf de middelbare school naar de grote stad, samen het levenspad bewandelen en volwassen worden. Ik kan nóg een boek vol schrijven met onze zelfbedachte woorden, koosnaampjes, inside jokes, herinneringen en avonturen. Het leven delen met jou is mijn grootste geluk en in jouw armen vind ik rust.

En nu?

Nu gaan we daar waar de wind waait.

About the author

On the 29th of June 1994, Monja Hoven was born in Deventer, the Netherlands. In 2011 she received her high school degree from Etty Hillesum Lyceum in Deventer. To follow her interests in biology, psychology and the brain, she moved to Amsterdam in 2012 to start her bachelor degree in psychobiology at the University of Amsterdam. In 2015 she received her BSc. cum laude, after her first research placement at the psychology department of the University of Amsterdam, studying the effect of non-invasive brain stimulation on attention under supervision of dr. Leon Reteig. After a year of teaching and traveling she started with her master's degree in brain and cognitive sciences in 2016, majoring in cognitive neurobiology. She got acquainted with the psychiatry department of the Academic Medical Center (AMC), now Amsterdam University Medical Center, during her second research placement in 2017 under supervision of dr. Judy Luigjes and dr. Ruth van Holst. Monja obtained her MSc. cum laude in 2019. After working as a research assistant from September 2017 until February 2019, she started her PhD project in June 2019 at the psychiatry department of the Amsterdam UMC, location AMC, under supervision of prof. Damiaan Denys, dr. Ruth van Holst and dr. Judy Luigjes. During her PhD project she studied confidence and metacognition in psychiatry from a neurocognitive perspective, with a special focus on obsessive-compulsive disorder and gambling disorder. After finishing her PhD work, she worked in Paris at École Normale Supérieure and Paris School of Economics in collaboration with dr. Maël Lebreton and dr. Stefano Palminteri. She hopes to receive her PhD degree in the beginning of 2024.