



HAL
open science

Behavioral and electrophysiological characterization of metacognitive deficits in schizophrenia spectrum disorder.

Martin Rouy

► **To cite this version:**

Martin Rouy. Behavioral and electrophysiological characterization of metacognitive deficits in schizophrenia spectrum disorder.. Cognitive Sciences. Université Grenoble Alpes [2020-..], 2023. English. NNT : 2023GRALS002 . tel-04126452

HAL Id: tel-04126452

<https://theses.hal.science/tel-04126452v1>

Submitted on 13 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement

Spécialité : PCN - Sciences cognitives, psychologie et neurocognition

Unité de recherche : Laboratoire de Psychologie et Neuro Cognition

Caractérisation comportementale et électrophysiologique des déficits métacognitifs dans le trouble du spectre schizophrénique

Behavioral and electrophysiological characterization of metacognitive deficits in schizophrenia spectrum disorder

Présentée par :

Martin ROUY

Direction de thèse :

Nathan FAIVRE

Chargé de recherche HDR, CNRS Délégation Alpes

Directeur de thèse

Rapporteurs :

ANNE GIERSCH

Directeur de recherche, INSERM DELEGATION GRAND EST

VALERIAN CHAMBON

Directeur de recherche, CNRS DELEGATION PARIS CENTRE

Thèse soutenue publiquement le **20 janvier 2023**, devant le jury composé de :

NATHAN FAIVRE

Chargé de recherche HDR, CNRS DELEGATION ALPES

Directeur de thèse

ANNE GIERSCH

Directeur de recherche, INSERM DELEGATION GRAND EST

Rapporteuse

VALERIAN CHAMBON

Directeur de recherche, CNRS DELEGATION PARIS CENTRE

Rapporteur

CATHERINE BORTOLON

Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES

Examinatrice

MIRCEA POLOSAN

Professeur des Univ. - Praticien hosp., UNIVERSITE GRENOBLE ALPES

Président

EMMANUEL BARBEAU

Directeur de recherche, CNRS DELEGATION OCCITANIE OUEST

Examineur

Invités :

JEROME SACKUR

Directeur d'études, École des hautes études en sciences sociales

CELINE SOUCHAY

Directeur de recherche, CNRS DELEGATION ALPES



Abstract

During my thesis, we studied the metacognitive abilities of individuals with schizophrenia spectrum disorder using psychophysical and electrophysiological measures. My work had two parts: on the one hand, a fundamental part focused on the characterization of metacognitive abilities in these patients, thanks to a meta-analysis that highlighted the presence of methodological biases in the study of metacognition in patients with schizophrenia. Our work has shown that the protocols and measures used so far did not allow a clear distinction between cognitive and metacognitive components, with the consequence of overestimating the metacognitive deficit in this population. We corroborated this result by analyzing behavioral and electrophysiological data collected in patients performing a perceptual task adjusting cognitive performance to that of healthy volunteers, showing no alteration at the metacognitive level. In the continuity of these results, we have developed an original experimental paradigm that simultaneously quantified metacognition in perception and memory with the ambition of isolating the metacognitive component in our measurements and thus gaining internal validity; on the other hand, a more clinical part focused on remediation, by trying to replicate the effect of a metacognitive training from recent models based on psychophysics, a priori effective in a non-clinical population, and thus potentially exploitable in the case of metacognitive alterations. After controlling for confounding factors of motivational and/or instructional origin, we concluded that this metacognitive training was ineffective. In sum, my thesis has corroborated existing recommendations for the evaluation of metacognition in experimental psychology, in order to strengthen the validity of the measure. The methodological critique that has been made has the effect of downgrading the contribution of the metacognitive component - as captured by signal detection theory - in the symptomatology of schizophrenia; a necessary milestone for a reorientation of research efforts.

Résumé

Au cours de ma thèse, nous avons étudié les capacités métacognitives des individus avec trouble du spectre schizophrénique à l'aide de mesures psychophysiques et électrophysiologiques. Mon travail a comporté deux volets : d'une part un volet fondamental centré sur la caractérisation des capacités métacognitives chez ces patients, grâce à un travail de méta-analyse mettant en évidence la présence de biais méthodologiques dans l'étude de la métacognition chez les patients avec schizophrénie. Notre travail a montré que les protocoles et mesures employés ne permettent pas de distinguer clairement les composantes cognitives et métacognitives, avec pour conséquence la surestimation du déficit métacognitif dans cette population. Nous avons corroboré ce résultat par l'analyse de données comportementales et électrophysiologiques collectées chez des patients effectuant une tâche perceptuelle ajustant la performance cognitive à celle de volontaires sains, ne montrant pas d'altération au niveau métacognitif. Dans la continuité de ces résultats nous nous sommes attachés à développer un paradigme expérimental original qui quantifie simultanément la métacognition en perception et en mémoire avec l'ambition d'isoler la composante métacognitive dans nos mesures et ainsi gagner en validité interne (collecte encore en cours à ce jour); d'autre part un volet plus clinique porté vers la remédiation, en essayant de répliquer l'effet d'un entraînement métacognitif issu de modèles récents basés sur la psychophysique, a priori efficace dans une population non clinique, et ainsi potentiellement exploitable dans le cas d'altérations métacognitives. Après avoir contrôlé pour des facteurs confondus d'origine motivationnelle et / ou d'instructions, nous avons conclu à l'inefficacité de cet entraînement métacognitif. En somme, mon travail de thèse aura permis de corroborer des recommandations existantes pour l'évaluation de la métacognition en psychologie expérimentale, de manière à renforcer la validité de la mesure. La critique méthodologique émise a pour effet de revoir à la baisse la contribution de la composante métacognitive – telle que capturée par la théorie de la détection du signal – dans la symptomatologie propre à la schizophrénie; un jalon nécessaire pour une réorientation des efforts de recherche.

Acknowledgments

First and foremost, I would like to thank Dr. Anne Giersch, Dr. Valérian Chambon, Dr. Catherine Bortolon, Pr. Mircea Polosan, and Dr. Emmanuel Barbeau for having kindly accepted to assess my PhD work.

I would like to take this opportunity to thank Dr. Céline Souchay and Dr. Emmanuel Barbeau for their insightful and encouraging feedback during midterm assessments.

I feel particularly grateful to my thesis director, Nathan Faivre, for his exemplary supervision throughout my thesis. Nathan has been a great support in many ways: by his availability despite his work load, by showing me how to organize and rationalize the steps of each project, by his great technical and conceptual mastery, always full of good advice, and not to forget his friendliness. Add to this his constant relaxed attitude, and it is easy to imagine the friendly yet rigorous atmosphere that reigns within his research team.

Therefore, I would like to express my gratitude to Jérôme Sackur, my former Master thesis director, for having had the good idea, even the prescience, to suggest that I do my thesis with Nathan.

Thanks to all the researchers with whom I collaborated: Elisa Filevich for her great availability, her technical skills and her patience, Paul Roux for the relevance of his remarks and advice, Clément Dondé for facilitating the experimental sessions with the patients as well as for his view of schizophrenia as a psychiatrist, Vincent de Gardelle and Pascal Mamassian for the adaptation of their generative model of confidence to our data.

I would also like to thank all the members of our team for their pleasant company on a daily basis: Ramla, Michael, Marie, Dorian, Audrey, Francois and Lise. More generally, I thank the PhD students of the Laboratory of Psychology and NeuroCognition (LPNC) and in particular Lucile for her warm support and delicious cookies; Méline, Merrick and Elie for their constant effort to socialize me, and Rémi for our exciting philosophical discussions.

I feel grateful to all Master's students whose contributions have supported and facilitated my PhD projects. So, thanks to Pauline, Rémi, Wassila, Hanna, Childéric, and Perrine who carried out the bulk of the experimental runs with the control participants, and Eugénie and Julia who collected data with patients in Versailles, and whose questions helped me deepen my own understanding of metacognition.

I address special thanks to the patients with schizophrenia who agreed to participate in our studies. I sincerely hope that the fruit of this research will be another stone in the edifice of knowledge that will ultimately allow us to understand and cure the handicap that this mental condition represents.

Last but not least, I would like to thank my family, as well as my friends from Dordogne, especially Alexandra, Paltso and Jigmé Rinpoché for their unfailing support.

Table of Contents

Abstract	3
Acknowledgments	5
List of Figures	7
Preamble: Aim of my PhD	8
About the structure of the manuscript	11
PART I - State of the art	13
1. Defining Schizophrenia	13
1.1. A brief history of ideology-driven management of madness	13
1.2. History of schizophrenia diagnosis	16
1.3. Insight and confidence	18
1.4. Toward a diagnosis 2.0 ?	20
1.5. Theories and models	21
1.5.1. A neurodevelopmental model of psychosis: the 22q11.2 deletion syndrome	21
1.5.2. The dopamine hypothesis	22
1.5.3. The aberrant salience hypothesis	23
1.5.4. Models of deficient inferential processing	24
1.5.5. The deficient efference-copy hypothesis	26
1.5.6. Intermediate conclusion	27
2. What is metacognition?	29
2.1. Experimental operationalization	31
2.2. Epistemological considerations	31
2.3. Measures of metacognition	32
2.3.1. Toward a bias-free metacognitive measure	35
2.4. Metacognitive architecture	36
2.5. Dynamic models of confidence	38
2.6. Domain-generality of metacognition	42
2.6.1. Two levels of metacognition	45
2.7. Metacognitive deficits	46
2.8. Remediating metacognitive deficits?	47
PART II - Experimental chapters	49
1. Meta-analytic assessment of metacognitive deficits among individuals with schizophrenia	49
2. Assessment of metaperceptual and metamemory abilities among individuals with schizophrenia	50
3. Exploration of electrophysiological markers of confidence during a metacognitive task	64

4. Improving one's metacognition? Assessment of a metacognitive training efficiency	83
General discussion	84
1. Performance-matching matters	95
2. What about overconfidence?	108
2.1. Overconfidence in errors as a Dunning-Kruger effect	109
2.2. Re-reading of the Dunning-Kruger effect: a type 1 deficit	111
2.3. On the importance of distinguishing sensitivity from bias with psychotic patients	114
2.4. No overconfidence in errors in our samples	115
3. Metacognitive deficits and clinical symptoms	116
4. Reality-monitoring	116
4.1. Hallucination vs hallucinosis	117
4.2. Perception vs imagination	117
4.3. Some peculiar reports of hallucinations	118
4.4. SDT considerations: against the interpretation of "noisy" hallucinations	120
4.5. Sense of reality	122
5. Limitations	126
5.1. Theoretical limitations	126
5.2. Clinical limitations	130
5.2.1. Samples of patients with schizophrenia are heterogeneous	130
5.2.2. Toward a transdiagnostic approach	130
6. Metacognitive training	131
6.1. The explanatory gap	132
6.2. The need for larger collaborations	134
Conclusion	135
References	135
Appendices	155
1. Supplementary information for project 1 (Meta-analysis)	155
2. Supplementary information for project 2 (Assessment of metaperceptual and metamemory abilities)	171
3. Supplementary information for project 3 (Electrophysiological markers of confidence)	183
4. The Dunning-Kruger effect is not a statistical artifact	187

List of Figures

Figure 1. Reverse nosology	21
Figure 2. Example of Bayesian inference	26
Figure 3. Corollary discharge circuit	28
Figure 4. Object-level versus Meta-level	31
Figure 5. Orthogonal dimensions of metacognition	33
Figure 6. Metacognitive bias and sensitivity	34
Figure 7. Meta-d' model	37
Figure 8. Metacognitive architectures	39
Figure 9. Hybrid metacognitive architecture	39
Figure 10. Serial sources of noise for type 1 and type 2 decisions	40
Figure 11. Illustration of the evidence accumulation model	41
Figure 12. Intracranial signature of confidence in monkeys	42
Figure 13. Post-decisional model of evidence accumulation	43
Figure 14. Domain-specific versus domain general architecture	44
Figure 15. Metacognition in discrimination versus detection tasks	45
Figure 16. Levels of metacognitive evaluations	47
Figure 17. Illustration of the confidence gap index	48
Figure 18. Hybrid architecture for metaperception and metamemory	64
Figure 19. Response-locked ERP of performance-monitoring	83
Figure 20. Local versus global metacognition in FCD patients	110
Figure 21. Dunning-Kruger effect	111
Figure 22. Illustration of the regression to the mean	113
Figure 23. Effect of first-order training on metacognition (Dunning and Kruger 1999)	114
Figure 24. Distinguishing noise from dream-like evidence	124
Figure 25. Metacognitive efficiency as a function of sensory noise	129
Figure 26. Correlation between M-ration and decision boundary	130
Figure 27. Consensus goals in the field of visual metacognition	134

Preamble: Aim of my PhD

Psychosis, and more generally madness – as a more encompassing concept regarding history - is associated with irrational thinking. Since everyone is prone to make irrational decisions (cf. system 1 and system 2; Tversky and Kahneman 1986), there should be something more specific about psychotic irrationality. Surprisingly enough in the history of thoughts, this is a man who turned out to suffer from paranoid schizophrenia who made a major theoretical breakthrough about human rationality in complex situations. His name was John Nash, a gifted mathematician from Princeton who received the Nobel prize in economics for his works on game theory (the so-called “Nash equilibrium”, Kreps 1989). So, the question could be narrowed as follows: what was irrational within the notoriously rigorous thinking of such a genius like John Nash? The famous movie adaptation from the life story of John Nash – *A beautiful mind* – gives us a chance to understand from a first-person perspective what a psychotic experience might look like. In this movie, John met with different characters who had some importance to him: there was his eccentric roommate named Charles, and the young girl Marcee who was Charles’ little niece. There was also the mysterious William Parcher, a secret agent from the defense department. John maintained rich relationships with them, and it took a while before he finally realized that anybody but him could see and interact with them. Here is a crucial aspect of psychotic irrationality, which sounds like a double burden: not only do patients have singular thoughts and perceptions which are not shared with other individuals, but they also lack the ability – or at least have some trouble – to identify these thoughts and perceptions as peculiar. In other words, from a first-person perspective, the psychotic irrationality seems perfectly rational, everything seems to be fine. However, from a third-person perspective, and from the clinical perspective in particular, this specific irrationality is referred to as a “lack of insight” (Amador and David 1998), and can be broadly understood as a lack of introspective skill to delineate what pertains to the outside world from contents that are self-generated, or to distinguish between what is real and what is not¹.

In this thesis, we recruited patients with schizophrenia with the aim of understanding one core feature of psychosis. We worked with the working hypothesis that the clinical symptom of “lack of insight” resulted from a metacognitive deficit (David et al. 2012), which made it possible to ground our research in recent theoretical advancements in cognitive science and to develop standardized protocols in laboratory settings. It opens the door for a

¹ On his way to recovery, John Nash pragmatically described his delusional thinking as “essentially a hopeless waste of intellectual effort”. Source: <https://livingwithschizophreniauk.org/john-nash/>

mechanistic and quantitative account of psychosis, which in turn might inform clinical and medical research.

About the structure of the manuscript

This manuscript is divided into three main parts, namely an introduction, empirical articles, and a general discussion. The introductory part consists of 1) a concise literature review about schizophrenia spectrum disorder covering the history of the diagnosis, its etiology and some of its main neural theories; 2) a presentation of metacognition as it is conceived and operationalized in experimental psychology, together with the formulation of the problematic we want to address.

- 1) In order to get closer to what is meant by “schizophrenia” we provide the reader with a brief overview of the historical shaping of the diagnosis of schizophrenia, together with a non-exhaustive list of recent theories trying to explain the psychotic symptomatology. In particular, mentioned theories provide a grasp on the plausible mechanistic underpinnings of positive symptoms such as hallucinations, delusions, and delusions of control. Of note, this part is only aimed at providing contextual elements and theories about schizophrenia, but are not necessary for the understanding of the experimental chapters presented in part II.
- 2) The second section defines what metacognition is and how it is operationalized in experimental psychology. Since it is the cornerstone of all my PhD work, particular attention is devoted to explaining why we should and how we can achieve a bias-free measure of metacognition. This measure will be used to quantify metacognitive deficits in schizophrenia, as well as the efficiency of an online metacognitive training for healthy participants. As we were interested in the potential use of experimental paradigms for clinical purposes, we considered applying this metacognitive training to remedy metacognitive deficits in schizophrenia.

The second part presents four empirical studies that we conducted in an attempt to address the scientific questions expounded in the introduction: 1) a meta-analysis quantifying metacognitive deficits among patients with schizophrenia, 2) a follow-up behavioral study to quantify metacognition in perceptual and memory domains while controlling the relevant parameters, 3) an electrophysiological study to investigate whether markers of confidence would be impaired in schizophrenia, and 4) an attempt to replicate the results of the above-mentioned online metacognitive training study.

This manuscript ends up with a general discussion about our experimental approach and results, where I provide a critical attempt to reframe most of the known metacognitive deficits as stemming from cognitive deficits instead. The reader will note that the

manuscript's narrative sometimes switches from the plural pronoun "we" to the singular pronoun "I", particularly within the discussion part. This is a deliberate intention on my part, in order to distinguish between ideas that came from team thinking and other more personal ideas that I assume to be more speculative.

PART I - State of the art

1. Defining Schizophrenia

1.1. A brief history of ideology-driven management of madness

Since antiquity, the evolution of concepts about madness has been far from linear. The discourse about its origins has oscillated between medical and religious explanations, leading to a plethora of types of treatments, which were sometimes politically instrumentalized (Porter 2003). We can briefly outline three significant historical periods: the Greek medical conception, the Middle Ages demonology, and the post-Renaissance psychiatric approach.

The hippocratic corpus (a collection of medical manuscripts associated with the Greek physician Hippocrates (460 – c. 370 BC)) was fairly opposed to superstition. Madness was conceived as a body imbalance in terms of “humors” – an ancient term designating some bodily fluids – and treated with medicinal plants such as hellebore. Galen (129 - c. 216 AD) even provided psychological counsels, on the theoretical grounds that inappropriate behaviors could derive from untamed or unrecognized passions, thus prefiguring psychotherapy. Efficient or not, we can only imagine that it was far less damaging than forthcoming medieval trepanations.

Throughout the medieval period, the theory of humors was paralleled with the religious interpretation of madness as possessions by the Devil. Depending on the diagnosis, the patient was either insane or possessed, and needed either a doctor or an exorcist, respectively. The historical apogee of the devil's interpretation of madness has been sadly recorded with the publication of the *Malleus Maleficarum* – the hammer of witches – in 1486, which was the starting point for the gloomy “witch-hunting”, the so-called Inquisition throughout European countries. The manuscript detailed what should be understood as witchery, and how to proceed (through torture) to unmask and eliminate witches. A prominent historical figure against Inquisition was Jean Wier (1515-1588). He was a Dutch physician known for his *De Praestigiis daemonum* published in 1563, an influential refutation of the *Malleus Maleficarum* where possession was reinterpreted as mental illness, thus recommending medical care instead of fanatical brutality (Mora 2008). This decisive step away from religious accounts, along with the contemporaneous stream of empiricism and rationalization carried out by illustrious thinkers such as Copernicus, Descartes, and later Newton, reaffirmed mental illness as stemming from body lesions. Descartes'

description of the body as a mechanistic system (Descartes, Rodis-Lewis, and Kambouchner 2010) surely played a significant role in this conceptual reorientation (Brown 2018).

However, this was not the end of political and medical wandering regarding mental illness. There is a long tradition of confining mad people together, starting with the creation of the “Hôpital Général” under Louis XIV, which was officially designed to provide education and medical care to maladapted poor citizens of Paris, but turned out to be a strategy to clean out the streets (Porter 2003). These unfortunate “poor citizens” were provided with an insalubrious place where arbitrary methods of coercion were applied. Although not for the same reasons, this tradition of confinement was prolonged and medically institutionalized in the early 19th century with the creation of asylums under the rationale of alienism, which constituted the first paradigm of psychiatry. Alienism is based on a philosophy of “moral treatment” promoting both isolation and dialog with patients as a cure (Stone 2008). The French promoter of alienism – Philippe Pinel (1745-1826) – is known for having literally removed the chains of hospitalized mental patients and developed a seminal version of psychiatric nosography, a conceptual framework distinguishing mental illness from the rest of the medical world. Despite his initial honorable intention to soften the tough conceptions and treatments related to mental illness, as well as to re-humanize the interactions with patients, asylums were places where hazardous experimental treatments such as bleeding, fastening, cold or hot baths, laxatives, and electroshocks were provided (Porter, 2003). Fortunately, a significant improvement regarding the treatment of mental illness originated from one of these experimental attempts, with the so-called “biological therapies”. Aiming at curing psychopathology by inducing a state of shock, Jean Delay and Pierre Deniker (1955) tried to inoculate chemical substances as a therapy. This is how the first neuroleptic was discovered in 1952, with chlorpromazine being the first efficient treatment to attenuate hallucinations and delusions. In parallel, anti-psychiatry movements arose against asylums, denouncing authority abuse from clinicians, criticizing the consequences of the developing pharmacology, and thus calling for alternative treatments of mental illness. The movement was notably supported by François Tosquelles (1912-1994), who largely contributed to the development of “institutional psychotherapy” designed to restore dignity and autonomy in patients’ lives, encouraging respectful and symmetrical social bonds between therapists and patients. The revendication for getting psychic patients out of the remote and inhospitable asylums as well as the call for the re-establishment of urban care for mental illness also led to the “sectorization” of psychiatry in France (Delion 2014), i.e. the creation of extra-hospital structures aiming at providing psychological support and medical care within the local environment of patients.

Through history, the lack of stability of conceptions about madness as well as the frequent use of inefficient (if not lethal) methods deteriorating both the physical and mental conditions of patients were signs that madness had been poorly understood. The next paragraph tries to outline the efforts made over the past 150 years to improve the psychiatric classifications, resulting in the current conception of what we now call “schizophrenia”.

1.2. History of schizophrenia diagnosis

Here, I will briefly recall the historical shaping of the current schizophrenia diagnosis. One important thing to keep in mind is that the very procedure of delineating diagnostic criteria follows a clinical tradition dating back to the nosography of Philippe Pinel (1809). Indeed, in the absence of a distinctive biological signature, the diagnosis criteria are still based on the identification of observable symptoms, which in turn rely on the clinician’s interpretation of the subjective reports made by the patients.

As exposed by Keshavan, Torous and Tandon (2020), three important early 20th century clinicians – namely Emil Kraepelin, Eugen Bleuler, and Kurt Schneider – proposed different descriptions and criteria, which laid the foundations for the later development of the diagnosis. Their work constituted the most influential concepts adopted in the first version of the American system of classification – the so-called DSM (Diagnostic and Statistical Manual for psychological disorders) – and the 6th version of the international system (ICD: International classification of disorders), and continued to be the main reference point for the successive versions until now.

A widely shared idea by early 20th century clinicians was that mental illness stemmed from a core disturbance, accompanied by various manifestations. This view appeared in reaction to the large proliferation of very specific categories called the “monomanias”, introduced by the alienist Esquirol (1838). Monomania was construed as a disorder of will, rather than a disorder of thinking, since patients were aware of the irresistibility of their intentions (Berrios 2018), hence also called ‘folie lucide’, i.e. insanity accompanied with insight. It was focused on a very local idea (Lefebvre 1988). For instance, erotomania (the conviction that one is loved by someone else), infanticidal monomania (e.g. the macabre illustration from Henriette Cornier who brutally killed a nineteen-month old baby without apparent reason), theomania (the conviction that one is a divine figure), dipsomania (the irresistible drive toward drinking alcohol), pyromania, kleptomania, and so on. The French psychiatrist Jean-Pierre Falret soon criticized monomanias for their uselessness for clinical purpose, in its unequivocal book called “de l’inexistence des monomanies”, published in 1854 (Lepoutre and Dening 2012). In this line, Eugen Bleuler (1857-1939) proposed that

all these various manifestations were best understood as resulting from a common psychic functional split, which he coined “schizophrenia”. He identified four fundamental symptoms (the four A’s²: ambivalence, autism, loosening of associations, and flat affect), and considered hallucinations and delusions as non-specific secondary symptoms. DSM-I (1952) and DSM-II (1968) mainly relied on Bleuler’s description of schizophrenia (Keshavan et al. 2020). Instead, Emil Kraepelin insisted that a useful diagnostic should rely on the description of the course of the illness as well as its outcome in terms of deterioration or improvement. On the basis of hundreds of clinical observations, Kraepelin proposed to distinguish what he called “dementia praecox”, including severe psychotic conditions which were characterized by their chronicity and cognitive deterioration, from “manic-depressive insanity”, which was more episodic with a better prognosis. ICD-7 (1955) and ICD-8 (1968) put the emphasis on Kraepelinian chronicity to define schizophrenia (Keshavan et al. 2020). As a consequence, there was not only a worldwide divergence of definitions between the DSM used in the USA, and ICD used in the rest of the world, but there was also poor inter-rater reliability for the diagnosis since the symptoms described were hard to identify. At this time, Strauss, Carpenter, and Bartko (1974) reviewed a list of the most used criteria by many clinicians, grouped them into six (disorders of content of thought and perception, disorders of affect, disorders of personal relationships, disorder of form of speech and thought, disordered motor behaviors, and lack of insight), and concisely reframed them into three influential categories: positive symptoms, negative symptoms³, and disorders in relating⁴. Positive symptoms are symptoms “that have the appearance of being active processes – for example, delusions, hallucinations, and catatonic motor phenomena, whereas negative symptoms “involve primarily absence of normal functions”, e.g. blunted affect, poverty of speech, or apathy. Importantly, Strauss et al. made the following remark about lack of insight:

“Of the six types of symptoms and signs, one, lack of insight, is obviously important but does not fit neatly into any category; nor are there many data available regarding its antecedents or prognostic implications. Because of the absence of detailed studies regarding its characteristics, this variable will not be discussed further here.”

² These symptoms will be later referred to as “negative symptoms”, i.e. impoverished behaviors and emotions

³ To note, the dichotomy between positive versus negative symptoms already comes from early descriptions of epileptic symptoms by two neurologists: Jackson in 1885 and Reynolds in 1896 (Dollfus, Mach, and Morello 2016).

⁴ The third category was later referred to as “cognitive disorganization” (Liddle 1987)

As a consequence, later classifications heavily relied on positive and negative symptoms, while disregarding lack of insight. Later, Kurt Schneider (1959) enumerated 11 “first-rank symptoms” (mostly positive symptoms) to be diagnostic of schizophrenia such as auditory and somatic hallucinations, thought withdrawal⁵, thought broadcasting⁶ or thought insertion⁷. In order to homogenize the classifications worldwide, as well as to simplify the diagnosis, DSM-III (1980) and ICD-9 (1978) both converged onto the same criteria: they emphasized Kraepelinian chronicity as well as the easy-to-recognize symptoms proposed by Kurt Schneider. With the successful increase in diagnosis reliability, DSM-IV (1994) put even more emphasis on positive symptoms, such that only one type of bizarre delusion or hallucination was sufficient to fall into the new category “Schizophrenia and Other Psychotic Disorders”, now divided into five subtypes: catatonic, disorganized, paranoid, residual, and undifferentiated (Keshavan et al. 2020).

However, the validity of existent classification of schizophrenia has been called into question because the subtypes of schizophrenia had low stability through the course of illness, were not representative of clinical heterogeneity (Mattila et al. 2015), and overall this categorization did not lead to noticeable progress in terms of treatment (Tandon 2012). Indeed, although successful in reducing positive symptoms, antipsychotics were ineffective regarding negative symptoms and cognitive deficits. Therefore, in the DSM-V (2013) subtypes were abandoned and substituted by dimensional descriptions, under the new label “Schizophrenia Spectrum Disorder”. This multidimensional description is a conceptual shift toward a multidimensional illness, where each dimension – reality distortion, negative symptoms, disorganization, cognitive impairment, motor symptoms, and mood symptoms – reflects a distinct pathophysiology with its unique target-treatment. The purpose is to be able to better track individual trajectories (i.e. clinical stages such as prodromal, first-episode, clinical high risk and chronic) and responsiveness to treatment. The C-RDPSS (Clinical-Rated Dimensions of Psychosis Symptom Scale) is a promising clinical tool which has been specially developed to guide treatment according to this tracking (Keshavan et al. 2020).

1.3. Insight and confidence

Although not generally acknowledged among psychiatrists, insight was considered an important - if not determinant - dimension for psychosis diagnosis (David 1990). Several

⁵ Sudden mind-blanking, where thoughts have seemingly vanished

⁶ The belief that one’s thoughts are readable by anyone around

⁷ The belief that some thoughts are imposed or “inserted” in one’s mind by someone else

scales have been elaborated (e.g. SAI⁸, SUMD⁹) to assess three overlapping dimensions related to insight into illness - also called clinical insight - namely the degree to which a patient is aware of having a mental disorder, the ability of the patient to recategorize his or her symptoms such as hallucinations and delusions as pathological, and treatment compliance (David 1990). Mostly associated with positive symptoms, poor insight has been also documented in relation to negative symptoms and construed as a specific executive function deficit, thus leading to the hypothesis that there might be different kinds of insight (Amador and David 2004).

Clinical insight measures robustly correlate with IQ, and with mood (i.e. depressive mood is associated with more insight). In particular, it has been proposed that clinical insight was associated with metacognitive abilities¹⁰, such as the self-monitoring of cognition e.g. monitoring the quality of one's memory or the awareness of perceptual stimuli (David et al. 2012). However, clinical insight has been criticized for its normative character which makes it vulnerable to demand characteristics. Indeed, the more the patient agrees with the doctor's diagnosis, the better his/her insight (David 2020). It has also been criticized for overlooking other aspects such as the adverse treatment effects (e.g. extrapyramidal side effects), the lack of objective criteria for clinical diagnosis (Amador and David 1998), or the fear of stigma (Davis et al. 2020), which are relevant factors that might play against treatment compliance.

A complementary approach to clinical insight is known as "cognitive insight", a more general conceptualization of insight referring to the ability to distance oneself from erroneous beliefs (Beck 2004). Not explicitly referring to clinical assessment, cognitive insight is both relatively immune to the above-mentioned critics and applicable to healthy participants, and thus more amenable for research purposes. Cognitive insight is assessed with the Beck Cognitive Insight Scale (BCIS, Beck 2004), which assesses both self-certainty i.e. the general degree of confidence one has of one's own judgments, and self-reflection i.e. the general knowledge one has of the fallibility of one's judgments. Lower self-reflection and higher self-certainty have been found to be associated with positive symptoms (for a review, see van Camp et al. 2017).

Self-certainty among patients with psychosis has also been found under the form of overconfidence in errors (Moritz et al. 2017; Hoven et al. 2019). This inability to internally adjust the degree of confidence to the correctness of one's decisions has been conceived as a deficit of metacognitive monitoring, which in turn has been associated with a lack of insight (David et al. 2012). Our experimental chapters were devoted to quantify metacognitive

⁸ Schedule for the Assessment of Insight (Sanz et al. 1998)

⁹ Scale to assess Unawareness of Mental Disorder (Amador and Strauss 1993)

¹⁰ Metacognition will be defined in greater detail in part 1.2.

deficits in schizophrenia, construed as an altered ability to calibrate one’s confidence on decision accuracy.

1.4. Toward a diagnosis 2.0 ?

As stated at the beginning of this section, the clinical diagnosis of schizophrenia still relies on subjective interpretations of symptoms, but this clinical syndrome is not properly matched with neurobiological findings in terms of genes, structural and functional brain abnormalities (Keshavan et al. 2020). An emerging avenue to address this issue is the concept of endophenotype. An endophenotype is a set of biomarkers lying between the genetic level and the symptom level, which should at least be measurable, heritable, and state-independent (i.e. one does not need to be in an acute psychotic crisis to be included in a study protocol). The strength of this approach is to enable a reversal of the nosological construction (Keshavan et al. 2013): the common top-down approach from observable symptoms to a speculative biological cause is replaced by a bottom-up classification from a measurable etiology to an emerging symptomatology (see Figure 1).

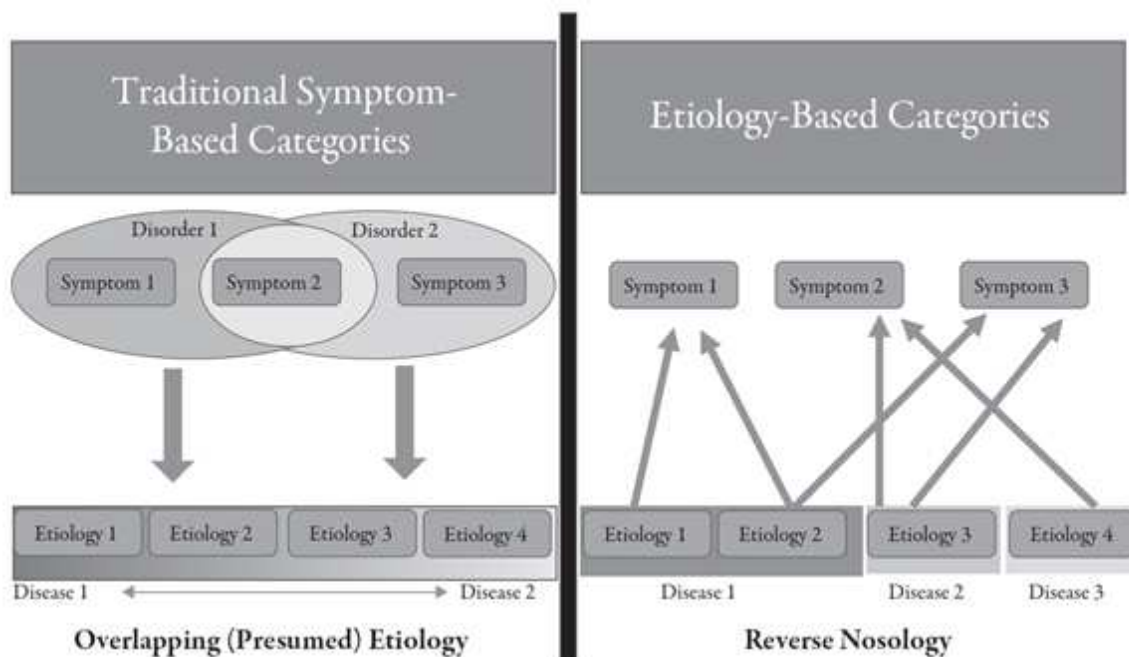


Figure 1. Reverse nosology. Schematic illustration of classical top-down nosology (left), and “reverse” nosology (right). Reprint from Keshavan et al. (2020)

The B-SNIP (Bipolar-Schizophrenia Network for Intermediate Phenotypes) project is an example of application of this promising deconstruction / reconstruction approach of nosology (Clementz et al. 2016). A large dataset of neuropsychological tests, behavioral measures (stop signal task and anti-saccade task) and electrophysiological measures

(N100, P200, P300 among other evoked-potential responses obtained with an auditory oddball paradigm) has been collected among more than 1800 participants, including patients with psychotic disorders (bipolar disorder, schizophrenia, schizoaffective disorder), as well as first-order relatives and healthy controls. Clustering with machine-learning unsupervised algorithms resulted in 3 distinct patterns of biomarkers among patients, called “biotypes”, which were also found in first-order relatives (to a lesser degree), and were transversal to the clinical diagnoses. The biotypes explained more variability in the behavioral and electrophysiological measures than the DSM-IV diagnoses did. Most remarkably, the biotypes have recently demonstrated their robustness. Indeed, the same three biotypes have been obtained in a recent replication of this clustering method applied to a new and equivalently large sample of psychotic patients, with their first-order relatives and healthy control participants (Clementz et al. 2022). Therefore, seeking endophenotypes might be a fruitful approach to identifying the pathological path from risk genes to clinical and neurocognitive observations. However, none of the behavioral and functional markers assessed in this study were related to insight or confidence. Therefore, it remains unclear whether these biotypes are sufficient to explain poor insight as a side-effect of the measured deficits, or if poor insight is part of unexplained variance.

1.5. Theories and models

The classical top-down nosology with its loose categories might have hindered research progress by mixing etiologies together, resulting in unwanted heterogeneity in clinical samples (Figure 1). The field of research on psychosis-related risk genes, based on risk variants¹¹ as well as post-mortem differential gene expression studies between patients with schizophrenia and healthy controls has proven to be cursed with low statistical power and low reproducibility, but these issues are being promisingly addressed by different means (e.g. the use of larger samples of participants, progress in RNA sequencing, and the development of genome-wide association studies, Allen et al. 2020). However, some fortuitous discoveries appeared to be good starting points: the high rate of psychosis in the 22q11.2 deletion syndrome, and the efficiency of antipsychotic medication to regulate positive symptoms (hallucinations and delusions), which in turn influenced a cascade of theories and models, that I briefly review below.

1.5.1. A neurodevelopmental model of psychosis: the 22q11.2 deletion syndrome

The 22q11.2 deletion syndrome (22q11DS) results from a microdeletion (i.e. a loss of a small segment of DNA) on chromosome 22, and is the most common microdeletion

¹¹ comparisons between subjects with risk alleles relative to those without risk alleles

disorder (McDonald-McGinn et al. 2015), with a significant impact on brain development and behavior. Since approximately one third of patients affected by 22q11DS develop a schizophrenia spectrum disorder in adulthood (Schneider et al. 2014), the 22q11DS offers a promising neurodevelopmental model for psychosis. Indeed, disposing of a developmental model of cognitive functions prior to illness onset, based on early cognitive markers instead of late visible clinical symptoms is of capital importance for early tracking of illness and better medical care before the cognitive deterioration is too important. There are two competing models of developmental trajectories in psychosis: the “developmental deterioration” profile, and the “neurodevelopment lag” hypothesis (Hill et al. 2020). The former supposes a normal cognitive development and a rapid deterioration of cognitive abilities from the prodromal stage of illness, while the latter supposes an early deviation from normal cognitive development, leading to a delayed trajectory of cognitive progress. Evidence from 22q11DS studies on verbal IQ and language impairment corroborate both developmental hypotheses (Gur et al. 2021). Other markers such as impaired executive functioning (e.g. attention), or social cognition (e.g. face recognition) are documented in psychosis and 22q11DS (Gur et al. 2021), as well as an impaired sense of agency together with lower confidence calibration (Salomon et al. 2022).

1.5.2. The dopamine hypothesis

There is another body of evidence highlighting a dysfunctional regulation of the dopaminergic system as an etiopathological pathway leading to psychosis. The dopamine hypothesis of schizophrenia resulted from two main observations: 1) antipsychotic drugs such as chlorpromazine or haloperidol owe their effectiveness to their affinity with D2 dopamine receptors (Seeman et al. 1976), 2) the use of dopaminergic drugs such as amphetamine produces psychotic-like states with hallucinations and delusions in healthy individuals (Connell 1957). Thus, already 30 years ago, the abnormal dopamine activity has been linked to positive and negative symptoms of schizophrenia as follows (Davis et al. 1991): the abnormally low prefrontal dopamine activity has been proposed to be the direct cause of negative symptoms, and in turn the lower regulatory (inhibitory) activity of prefrontal dopamine projections causes an increase of subcortical dopamine activity, which is associated to positive symptoms. A specific gene-environment interplay gave rise to the “dual hit” model (Bloomfield and Howes 2020), where the genetic vulnerability responsible for increased capacity of dopamine release (first hit), in conjunction with the presence of environmental and psychosocial stressors (e.g. childhood adversity, social minority group, trauma, drug use, see Howes et al. 2017; van Os, Kenis, and Rutten 2010) causing acute levels of dopamine release (second hit), were conducive to psychosis as a result of

hyperdopaminergy. Interestingly, hyperdopaminergy has been linked to increased confidence in a word detection task (although in conjunction with an increased accuracy, Lou et al. 2011), and overconfidence in errors in another perceptual detection task (Andreou et al. 2015).

In particular, striatal hyperdopaminergy has been shown to induce hallucination-like experiences - i.e. false percepts with abnormally high confidence - in a rodent model of psychosis (Schmack et al. 2021). Moreover, abnormal striatal presynaptic dopamine synthesis has been a robust observation in schizophrenia (see meta-analyses from Fusar-Poli and Meyer-Lindenberg 2013; Howes et al. 2012) and fuelled an influential hypothesis about positive symptoms: the aberrant salience hypothesis.

1.5.3. The aberrant salience hypothesis

The elegance of the aberrant salience hypothesis (Kapur 2003) lies in its attempt to provide a mechanistic explanation filling the gap between the neurobiological (high striatal dopamine activity) and the phenomenological levels (delusional thinking) in schizophrenia. The theory originated with the proposition that striatal dopamine supports the function of attributing motivational salience to reward-related stimuli (Berridge and Robinson 1998). According to this view, dopamine is the mediator which makes someone *wants* (not *likes*, as previously proposed in the hedonic interpretation of dopamine function, Koob and Moal 1997) something by providing a motivational value to a neural representation of either an external object or an internally generated content. Therefore, the rationale behind the aberrant salience hypothesis is that abnormally high synthesis of striatal dopamine – producing stimulus-independent release of dopamine – would result in aberrant attribution of salience to either internal representations (e.g. ideas) or representations of external objects (percepts). Thus, dopamine is metaphorically referred to as “the wind of psychotic fire” (Kapur 2003), emphasizing the dysregulation of dopamine as the core aspect of psychosis. At the phenomenological level, some random stimuli or ideas suddenly grab one’s attention and the irresistible drive felt toward them demands an explanation, and according to this view, the ensuing delusional thinking – i.e. post-rationalization of one’s initial drive – is secondary to this process.

However, there are some limitations to this proposal:

- 1) Elevated levels of striatal dopamine are not always observed in patients with psychosis (Bloomfield and Howes 2020), therefore the “wind of psychotic fire” may not necessarily be dopaminergic. It might explain why 20 to 35% of patients with psychosis are not responsive to D2 antagonist antipsychotics (Demjaha et al. 2012). Other neurotransmitters have been

identified as potential mediators in the pathophysiology of schizophrenia such as serotonin, GABA and glutamate (Berkovich, Dehaene, and Gaillard 2017). Interestingly, the serotonin receptor 5HT_{2a}, which is involved in the formation of visual hallucinations induced by psychedelics such as LSD, is also the receptor targeted by the antagonist activity of Clozapine, the treatment provided in case of refractory-schizophrenia (Nichols 2004).

2) The aberrant salience hypothesis leaves us with two explanatory gaps. Indeed, the aberrant salience hypothesis provides a conceptual framework explaining why the psychotic mind is attracted toward insignificant sensory inputs and ideas, but i) it says nothing about how subsequent delusional world-related statements are formed; and ii) it seems unlikely that other positive symptoms such as hallucinations, thought insertion, or delusion of control might derive from aberrant motivational salience (Bloomfield and Howes 2020). Furthermore, there is no clear link between striatal dopamine and confidence in patients with Parkinson's disease (Bang et al. 2020), which makes it difficult to explain confidence abnormalities that are found in schizophrenia (Hoven et al. 2019).

1.5.4. Models of deficient inferential processing

Progress in neurocomputational modeling has provided interesting theories aiming at explaining both delusional thinking and hallucinations as resulting from a common dysfunctional inferential process. This conceptual step is quite straightforward on the basis of Bayesian theories of cognition, in which perception is conceived as an inferential process resulting from a weighted integration of sensory evidence and prior knowledge, an idea already formulated by von Helmholtz (1911). Thus, under this framework the delineation between abnormal perception (hallucination) and abnormal belief (delusion) is more conceptual than mechanistic, as reminded by the thought-provoking title “perceiving is believing” of the review article from Fletcher and Frith (2009). Regarding the cognitive explanation of delusional thinking, Fletcher and Frith proposed one plausible Bayesian interpretation extending on the aberrant salience hypothesis. But before we delve into the Bayesian explanation of delusional thinking, a concise description of the Bayesian inferential process is needed. In this framework, any unexpected observation produces a “surprise” signal, called a prediction error – i.e. a mismatch between the sensory evidence and the prediction about the cause of the sensory input. The inferential process consists in resolving the prediction error, either by having a second look at the stimulus (bottom-up resampling of sensory evidence), or by choosing an alternative and more adequate interpretation from one's model of the world (top-down shift of expectation). The prediction error can be attenuated at any level of the hierarchical system from the sensory level to the highest level of abstraction (see Figure 2 for an illustration of such levels of processing).

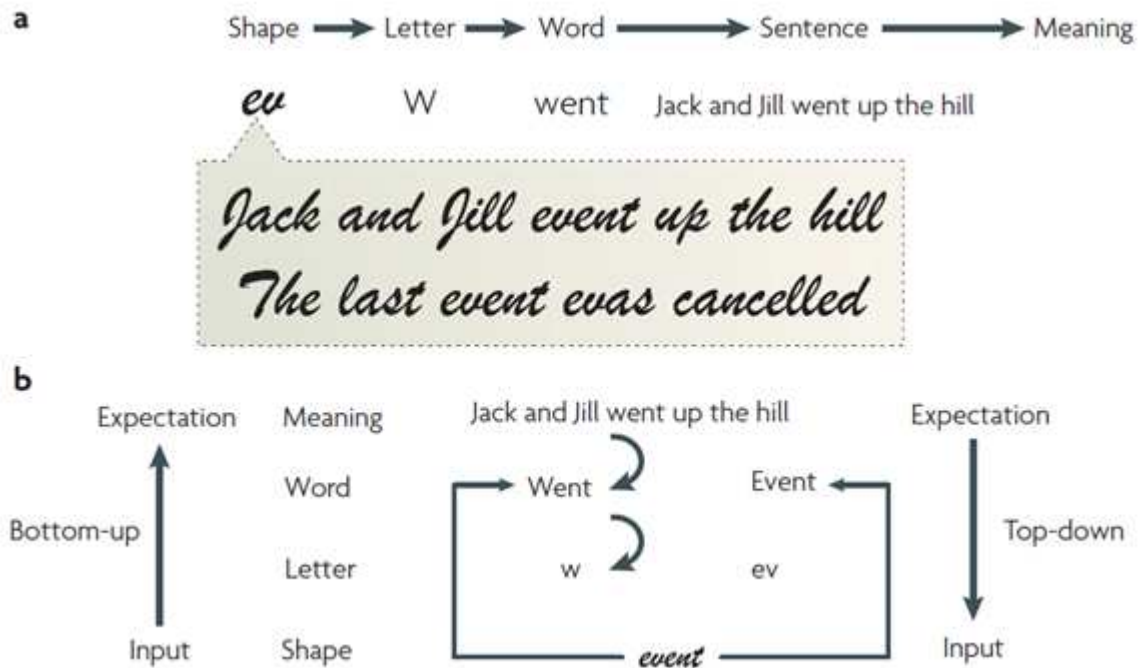


Figure 2. Example of Bayesian inference. In this example, the font gives the letter “w” an ambiguous shape that can be confused with the letters “ev”. Bottom-up processing (shape > letter > word > meaning) might lead to a prediction error since the sentence “Jack and Jill event up the hill” is meaningless. In turn, top-down expectation leads to reinterpreting the shape of the letter as “w”. Reprint from Fletcher and Frith (2009)

Whenever these two strategies are not sufficient to solve the prediction error, a third strategy is to update one’s model of the world (i.e. prior knowledge) by adding a new interpretation to one’s conceptual repertoire. But this strategy is optimal only if the initial surprising signal is informationally relevant. Indeed, under the aberrant salience hypothesis, the excessive level of striatal dopamine provides stimuli with abnormal salience, thus surprisingly grabbing one’s attention. And yet this unpredictable attentional hook originates from a false prediction error – i.e. there is no sensory cause to be inferred to begin with – which in turn can only be solved by enlarging one’s model of the world with new speculative features and properties. Hence, the delusional thinking would be the result of an optimal Bayesian inference which is bypassed by the abnormal salience of stimuli, leading to the false assumption that the surprising stimulus needs further explanations. This explanation sheds light on how “perceiving *involves* believing”, but if the statement “perceiving is believing” is a logical equivalence, then what would be the Bayesian explanation for the reciprocal statement “believing involves perceiving”, referring to hallucinations? And how to explain overconfidence in these false percepts?

The model of “circular inferences” (Jardri and Denève 2013) elegantly explains delusions, hallucinations and overconfidence as resulting from an abnormal Bayesian inferential process. For a hierarchical Bayesian system to work properly, it should be able to clearly disentangle descending prior information from ascending sensory information within the cascade of multiple processing levels and across all the inferential loops. Jardri and Denève proposed that the system could be bypassed in case of excitatory to inhibitory imbalance in mesocortical circuits – reflected by increased levels of dopamine (O’Donnell 2011). As a result, top-down prior knowledge information, once arrived at the sensory processing level, could be mistakenly labeled as sensory information and then reverberated back to higher levels as such, and thus being counted multiple times. In the end, this abnormal circular inferential process might explain hallucinations, where one is projecting his own beliefs as external objects. The opposite is also true for delusions: sensory information could be misunderstood as prior information, and thus be reverberated back down in the inferential loop and counted many times despite the redundancy of the information. In the end, the first interpretation is the one which is retained, discarding additional contextually relevant sensory information and alternative explanations. Even in case of extremely weak sensory evidence, the circularity of inference results in highly confident false percepts, because the evidence is counted multiple times.

1.5.5. The deficient efference-copy hypothesis

There is another aspect which is to be explained in positive symptoms, namely the sense of being passive or under the control of an external influence such as delusion of control, or thought insertion, i.e. the impression that some thoughts do not originate from oneself, and are being imposed by someone else. This resonates with an ancient conception of positive symptoms emphasizing a disorder of will, which was already advanced in mid-XIXth century by Esquirol (1838), Billot (Biéder 2011), and Ribot (2002, cited in Berrios 2018). The tendency among patients with schizophrenia to erroneously attribute one’s thoughts or actions to an external source has been demonstrated using source monitoring¹² tasks (Brébion et al. 2000; Keefe et al. 1999; Vinogradov et al. 2008). Now, how do we normally distinguish between self-produced and externally produced sensations? If we take the visual modality, a retinal motion smear can be caused either by a moving object in the visual field, or by the production of an eye-saccade (or a head movement), then how do we know whether we or the world has moved? This distinction is allowed by a process of self-monitoring which relies on the presence or absence of an internal signal called a

¹² Typically, half of a list of words is read out loud by the experimenter, and the other half by the participant. Then, all the words are presented to the participants and they are asked to remember their source of production.

corollary discharge or an efference copy. As the name suggests, an efference copy is a copy of a motor command, which is meant to anticipate the sensory consequences of the corresponding action. Therefore, when an eye-saccade is made the corresponding perceptual shift is anticipated, such that there is no prediction error, contrary to a perceptual shift produced by an external object (see Figure 3).

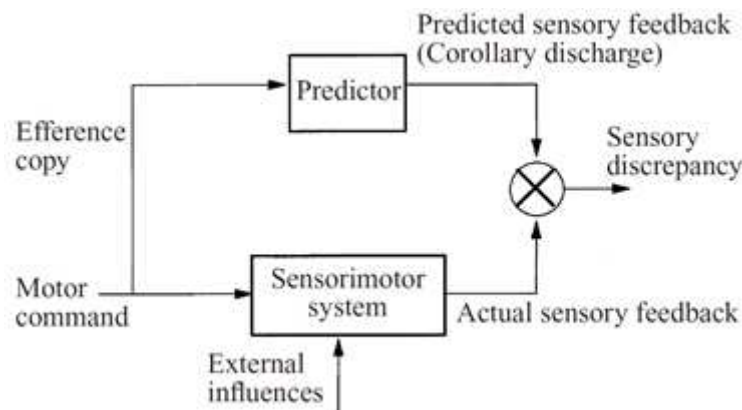


Figure 3. Corollary discharge circuit. The identification of the source of a sensation relies on the comparison between the predicted sensory consequences of one's actions and actual sensory feedback. Reprint from Blakemore et al. (2000)

A testable consequence of this model of self-monitoring in healthy participants is known as “sensory attenuation”. Indeed, since the sensory consequences of one's actions are predicted, the system reduces awareness of the actual sensory feedback. According to this model, this is precisely the reason why self-tickling doesn't work: one's actions can't be surprising. Now, concerning patients with schizophrenia, if source monitoring deficits result from a dysfunctional corollary discharge, then we should expect a reduced sensory attenuation. And this is exactly what has been observed (Shergill et al. 2005), thus providing an explanation for delusions of control: unanticipated sensory consequences of one's actions might result in the delusion that an external agent has caused oneself to move. Furthermore, such deficits of self-monitoring have been linked to an impaired sense of agency (SoA) in psychosis, together with an impaired ability to track SoA judgments with confidence ratings (Krugwasser et al. 2022).

Already in 1974, Feinberg proposed to explain the symptom of thought insertion using the same model. Based on the observation that patients exposed to cortical stimulation can have sudden conscious thoughts that are identified as externally provoked by the surgeon (as reported by Penfield, 1974), Feinberg proposed that thoughts could be considered motor acts, and thus would be accompanied by a thought-like efference copy.

1.5.6. Intermediate conclusion

So far, we have enumerated theories and models which have explanatory power to account for most positive symptoms. In this thesis, we wanted to understand at which level of processing erroneous judgments occur. Is schizophrenia a cognitive disorder, i.e. a disorder related to sensory and/or mnemonic processing as well as the formation of neural representations (Bortolon et al. 2015; Dondé et al. 2019)? Or a higher-order disorder involving self-monitoring, i.e. the ability to assess the precision and relevance of one's own representations? We have already reviewed some evidence in favor of impaired confidence calibration as well as source-monitoring deficits indicating altered abilities to distinguish inner versus external representations. Some authors propose that positive symptoms reflect a deficit at the level of reality-monitoring – referring to a confusion between perceptual and imaginary contents – which would result from an inability to correctly monitor the different properties of both types of cognition (Brébion et al. 2008; Dijkstra, Kok, and Fleming 2022). Understanding at which level deficits occur is crucial for shedding light on the “lack of insight” and targeting better remediation strategies. One important thing to notice about hallucinations is their inherent lack of insight, which is the delineating property from what is called “hallucinoses”. Hallucinoses might arise among individuals without any psychiatric condition (Manford and Andermann 1998) and are conceived as “false” hallucinations, or “hallucinations with insight”, because contrary to hallucinations, the false percepts of hallucinoses are not taken as real. For instance, in the case of Charles Bonnet syndrome (Pang 2016), people with a pathology of the visual pathways or an ocular impairment such as age-related macular degeneration can sometimes experience weird visual percepts without external cause, ranging from simple shapes to complex figures like faces or animals, yet with insight concerning their erroneous character. So, in the case of hallucinoses, despite impaired vision, people are able to monitor the quality and origin of these percepts.

However, in the case of hallucinations, there might be two different reasons for altered source monitoring and confidence abnormalities (eventually leading to poor insight): 1) because of impaired perceptual abilities, making impossible the distinction between perception-like contents (perceived or hallucinated) or 2) because of an impaired ability to correctly monitor the origin of the hallucinatory percept.

In this thesis, we tried to arbitrate between these two propositions by relying on recent developments of metacognitive measures, offering the advantage of disentangling between cognitive and metacognitive performance within experimental settings. The next section describes the details of this operationalization.

2. What is metacognition?

*“Although the term metacognition is a relatively recent invention,
its practice is as old as rational thought.
As long as people have evaluated ideas for their quality and sought
to improve those ideas, they have performed metacognitive operations.”*
Martinez, 2006

Metacognition, as a form of self-knowledge, has been emphasized for millennia as an important means to live a decent life. Among influential thinkers of the past, Buddha Shakyamuni (VIth century BCE) already advised his followers to train in *samprajanya* (Anālayo 2003), a sanskrit term translated as “meta-awareness” and referring to the ability to notice that the mind has wandered away from its initial focus, thus preventing distraction from a virtuous conduct and enabling sustained concentration (Dunne, Thompson, and Schooler 2019). Around the same period, the Greek philosopher Socrates’ unconventional approach to happiness was to self-realize the mortal nature and limits of the human condition. Greek thinkers exhorted citizens to gain knowledge about their own shortcomings and fantasies, which was best summarized in the famous slogan inscribed on the temple of Apollo at Delphi: “*Gnothi seauton*” i.e. “*know thyself*” (Fleming 2021). Indeed, it is one thing to know, yet another to know how well one is knowing.

However, one of the most appealing examples of the ambiguous nature of reflective-knowledge can be dated back to René Descartes’ philosophy. On the one hand, René Descartes deeply acknowledged that beliefs and knowledge were fallible, such that he established skeptical doubt as a method (Descartes 2020). On the other hand he also assumed that introspection was reliable, since he could deduce the indubitability of his existence through his metaphysical meditations (Descartes 2020), an insight captured by his famous “*cogito ergo sum*”. More recently, inspired by Descartes’ methodology, the mathematician Edmund Husserl (1859-1938) undertook to rebuild a rigorous science on the basis of experience by getting “back to things themselves”, i.e. the way things manifest themselves in experience, at a pre-conceptual level (Zahavi 2003). For this purpose, he applied the contemplative gesture known as “*epochè*”, which is described as the suspension of the “natural attitude”, i.e. the suspension of the prior hypothesis of a mind-independent world. For Husserl, the natural attitude refers to the daily non-reflective attitude that reifies phenomena into independent externalities, without the awareness that these phenomena

were constituted as meaningful objects through a mental gesture in the first place. Therefore, Husserl's project was to investigate the foundations of the natural attitude, i.e. the conditions of possibility of the emergence of objectivity and rational thought within experience, and ultimately the very conditions that make logics and mathematics possible. The goal was to understand how experiences were sense-making, in order to be less biased by our interpretative processings (Overgaard 2015).

Parallel to this philosophical enquiry, Wilhem Wundt (1832-1920) used introspection in experimental psychology by relying on the premise that elementary perceptual contents - although transitory events - can be reliably introspected and reported, given an appropriate amount of laboratory training (Overgaard 2008; Sackur 2009). However, the lack of a verifiable ground truth, together with the absence of a metrological science dedicated to assess the reliability of the introspective process itself, led to the failure of introspectionism, notably marked by the imageless thoughts controversy. Since that time, introspection has revealed to be subjected to intrinsic biases which are beyond conscious access, even leading to confabulation about the reasons underlying one's behaviors (Nisbett and Wilson 1977).

The study of reflective knowledge came back into the laboratory due to theoretical and methodological advances that enabled the rigorous study of metacognitive abilities. Reflective knowledge was termed *metacognition* by James H. Flavell (1979), with the broad meaning of "thinking about thinking". This new acceptance assumes a hierarchical organization of cognition with a vertical structure composed of interrelated levels called "object-level" and "meta-level" (Nelson and Narens 1990). The object level forms representations about states of the world, whereas the meta-level forms its own representation of the object-level representations, i.e. a representation of representations. In this metacognitive model, the bottom-up flow of information from the object-level to the meta-level supports the function of cognitive monitoring, while top-down flow of information from the meta-level to the object-level plays a functional role of cognitive control (Figure 4).

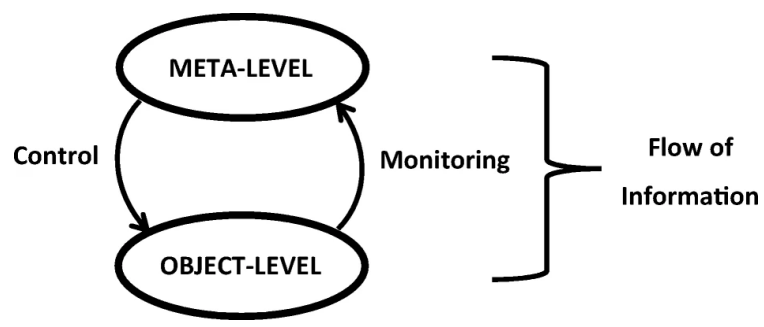


Figure 4. Object-level versus Meta-level. Hierarchical metacognitive model with an object-level subordinated to a meta-level of cognition, with directional flows of information supporting functional roles: monitoring and control. Reprint from Nelson and Narens (1990)

Although metacognition is not necessarily coupled with awareness (Schooler and Smallwood 2009), monitoring was first conceived in the sense of becoming aware of cognitive processes and contents, and cognitive control as an adaptive reaction to such awareness (Nelson 1990). The next section makes it clear that the metacognitive representations we are dealing with in our experimental protocols are explicit.

2.1. Experimental operationalization

Experimental protocols aimed at drawing a line between the object-level and the higher order meta-level by proposing participants to go through two distinct but interrelated tasks: a so-called “type I” (or first-order) task asking participants to detect, recognize or discriminate between two stimuli items and a “type II” (or second-order) task asking participants to rate or predict their performance at the type I task (Galvin et al. 2003). In this manuscript, type II tasks are restricted to confidence judgments tasks. Depending on the sequential order of type I and type II tasks, we talk either about prospective or retrospective metacognition. Prospective metacognition is measured with confidence judgments preceding sensory evidence exposure, such as feelings of knowing (FOK) (Hart 1965) or judgments of learning (JOL) (Leonesio and Nelson 1990). For instance, FOK is measured by asking participants how confident they feel in their ability to recognize in the near future an item which they initially failed to recollect. Retrospective metacognition is typically measured by asking participants to rate how confident they feel about their response on a preceding type I task, on a trial-by-trial basis. Recently, an online confidence database has been created encouraging any researcher to upload their datasets containing retrospective confidence judgments (and also confidence judgments that are simultaneous with the first-order decision) for data preservation, meta-analyses, and to test new hypotheses with sufficient statistical power (145 datasets, ~8700 participants, ~4 million trials, Rahnev et al. 2020).

2.2. Epistemological considerations

Whether or not a type II task such as one involving a confidence rating actually reflects a metacognitive process requires epistemological scrutiny. Confidence judgments would be metacognitive in its epistemological sense, only if we assume that they are meta-representational, i.e. representations “about” a first-order representation. However, a type II task is, strictly speaking, a “behavior about a behavior” (Fleming, Dolan, and Frith

2012), not necessarily cognition about cognition per se. To ensure internal validity, namely that a type II task faithfully reflects a metacognitive process, we need to rule out the possibility that a second-order behavior merely reflects a first-order process directed at the first-order behavior. Actually, whether confidence judgments involve meta-representations is still debated. For this reason, the order of behavior (first-order decision vs. second-order judgment) and the level of representation (object-level vs. meta-level) are still considered as orthogonal dimensions (Figure 5). In what follows, metacognition refers to its empirical definition (i.e. a behavior about a behavior), but we will assume that second order behaviors like confidence judgments actually reflect metacognitive processes. To note, this theoretical position seems justified from existing evidence of dissociations between first-order and second-order performance, such as experimental manipulations increasing confidence while decreasing accuracy (Rahnev et al. 2012) or decreasing metacognitive performance while increasing accuracy (Bang, Shekhar, and Rahnev 2017; Maniscalco et al. 2016). Such dissociations are also found in some neuropsychological cases such as blindsight where confidence poorly tracks accuracy despite a relatively preserved visual discrimination performance (Ko and Lau 2012, but see Phillips 2021 for an alternative account of blindsight), or the opposite pattern found in early Alzheimer disease where patients have poor memory but are nonetheless able to monitor the quality of their memory (Bäckman and Lipinska, 1993; Souchay 2007), suggesting that there is somehow a specific metacognitive process underlying type 2 tasks.

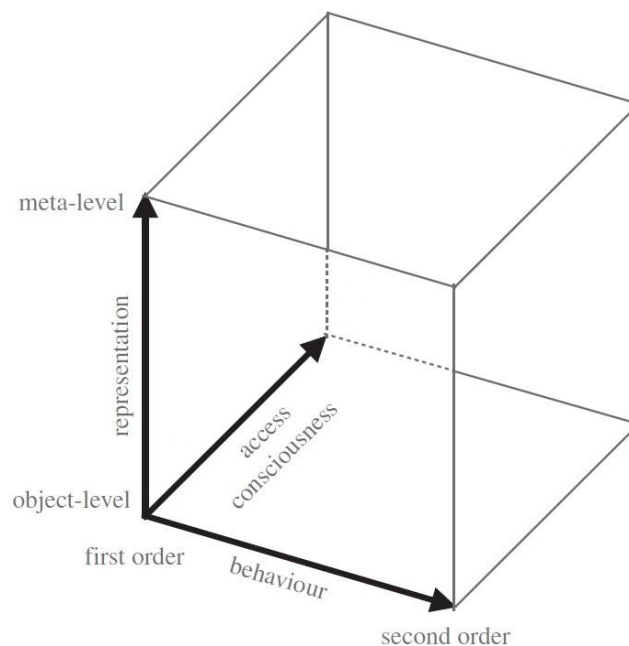


Figure 5. Orthogonal dimensions of metacognition. Three dimensional map of the conceptual landscape in metacognition research: The order of the behavior, the level of the

representation, and the degree to which a stimulus. Reprint from Fleming et al (2012) is explicitly reportable (known as access consciousness) are conceived as orthogonal dimensions.

2.3. Measures of metacognition

All the subtlety of measuring metacognition lies in the ability to disentangle metacognitive sensitivity from metacognitive bias and first-order sensitivity. Here, metacognitive sensitivity refers to the ability to correctly adjust confidence according to task performance, i.e. to correctly discriminate between correct and incorrect responses. Metacognitive bias is the general tendency to be under- or over-confident, regardless of correctness. All combinations of metacognitive sensitivity and metacognitive bias are theoretically possible (Figure 6).

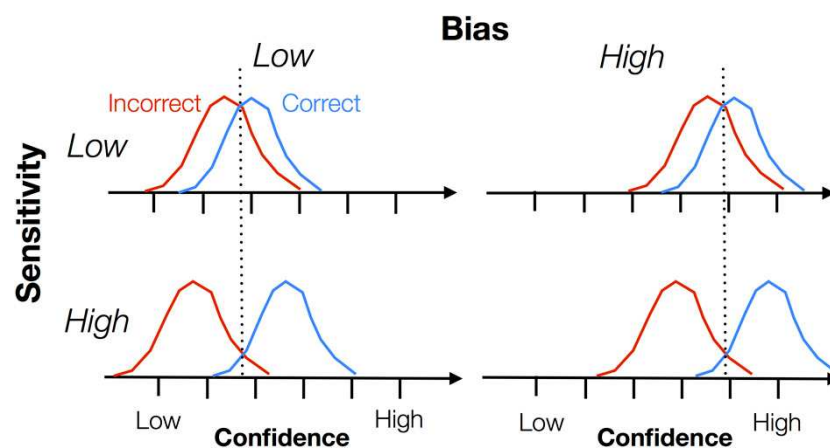


Figure 6. Metacognitive bias and sensitivity. X-axis represents the degree of confidence from low to high. Distributions of confidence ratings for correct (in blue) and incorrect responses (in red) can be characterized by the distance between one another being high or low (i.e. proxy for metacognitive sensitivity), and by an overall shift toward high or low confidence (i.e. proxy for metacognitive bias). Reprint from Fleming and Lau (2014)

Seminal measures of metacognition like the phi-coefficient (Φ) or the Goodman–Kruskall γ -coefficient (Goodman and Kruskal 1979) were correlational measures between raw confidence ratings and trial accuracy, but these measures suffered from a contamination by metacognitive bias. The ability to distinguish between sensitivity (or d') and bias (also called criterion) - at least for type I tasks - is an important feature of the signal detection theory (SDT, Green and Swet 1966, see Box 1). Therefore, attempts were made to

extend SDT to include confidence ratings, in order to benefit from this property at the second-order level (Galvin et al. 2003). However, the assumption of gaussian and equal variance, which was met with sensory distributions, is not met when considering distributions of correct and incorrect responses in type-2 SDT. For this reason, type-2 d' is an invalid measure of metacognition. The ROC (receiver operating characteristic) analysis circumvents this difficulty because it is a non-parametric approach. A type 1 ROC curve is obtained by plotting the proportion of Hits (H) against the proportion of False Alarms (FA, Hits and False Alarms are defined in Box 1) for different values of type 1 criterion. The area under the ROC curve (AUROC) is a measure of type 1 sensitivity. It is possible to plot a type 2 ROC curve from type 2 data, this time by plotting the proportion of high confidence H against the proportion of high confidence FA, for several values of type 2 criterion, from which we can derive a metacognitive measure: the area under the type 2 ROC (AUROC2). Without entering into the details, the limitation of this measure is that it is contaminated by type 1 sensitivity and bias, and since AUROC and AUROC2 are expressed in different spaces, it is not possible to normalize one by the other (Fleming and Lau 2014, Galvin et al. 2003; Maniscalco and Lau 2014). To address this issue, Maniscalco and Lau (2012) developed an influential framework called meta- d' , which offered interesting properties for the study of metacognition.

Box 1: SDT framework

SDT is a theory about how living cognitive systems are able to detect the mere presence of a stimulus, or to discriminate between two different stimuli, by assuming that a threshold mechanism is implemented in the brain. The stimulus is scaled along a decision-axis according to its evidence strength, and the decision is made relative to the position of this evidence strength compared to a threshold. Thus, the first-order response is binary (presence or absence; stimulus 1 or stimulus 2). For instance, in a detection task where a low-intensity signal is flashed on a noisy background, there is detection whenever the signal strength exceeds the threshold (HIT), but the stimulus remains unseen otherwise (MISS). Errors can occur due to additional sensory noise resulting from sensory processing - supposedly sampled from a gaussian distribution - which can accidentally exceed the threshold, causing false alarms. Possible outputs can be summarized with a 2x2 matrix combining the inputs (binary states of the world: stimulus present vs. absent; stimulus 1 vs. stimulus 2) and the responses (also binary, corresponding to input possibilities, see Table 1.A). The ability to detect a stimulus (or to discriminate between two stimuli) is called the sensitivity or d' , and corresponds to the signal-to-noise ratio i.e.

the distance between the distribution of noise (or stimulus 1 in case of a discrimination task) and the distribution of signal + noise (or stimulus 2) divided by the dispersion of noise. Crucially, the sensitivity is independent from bias (also called criterion, i.e. the tendency to provide a response category more often than the other one, and bias (or criterion) - corresponding to the position of the decision threshold along the decision axis, accounting for idiosyncratic strategies such as response-conservatism or response-liberalism). Type-1 SDT was reframed into type-2 SDT to benefit from this property. To do this, the 2x2 matrix of outcomes is adapted by binning the type 2 confidence ratings into two response categories (high versus low), and the inputs consist in type-1 accuracy categories (correct vs. incorrect). For instance, being highly confident in an error is called a type 2 false alarm, and low confidence in a correct answer is interpreted as a type 2 miss (See Table 1.B)

A. TYPE 1 SDT: Detection of an external event		
Stimulus	Detection	No detection
Present	HIT	MISS
Absent	False Alarm	Correct Rejection

B. TYPE 2 SDT: Monitoring of Type 1 accuracy		
Type 1 Accuracy	High confidence	Low confidence
Correct	Type-2 HIT (HIT2)	Type-2 MISS
Incorrect	Type-2 False Alarm (FA2)	Type-2 Correct Rejection

Table 1. A. Type 1 SDT outcomes: correct outcomes are in green, incorrect ones in red. B. Type 2 SDT outcomes: type-2 correct outcomes are in green, and type-2 incorrect ones in red.

2.3.1. Toward a bias-free metacognitive measure

The first-order SDT model has been extended with additional second-order criteria to model the degree of confidence in a given response. As a consequence, the readout of the evidence strength along the decision axis is sufficient to determine both the decision and the level of confidence. This model property is interesting because it enables the computation of the type 2 parameters (sensitivity, called meta-d', and second-order criteria) within the type 1

SDT level (Figure 7.A). These model parameters are determined by fitting this extended SDT model on type 2 data, i.e. the distributions of confidence ratings (Figure 7.B). In other words, meta- d' reflects the expected type 1 sensitivity given the type 2 responses, assuming that the participant is metacognitively ideal, i.e. assuming no information loss between type 1 and type 2 responses (i.e. the ideal case where meta- $d' = d'$). As Maniscalco and Lau (2012) phrased it: “One can think of meta- d' as a measure of the signal that is available for the subject to perform the type 2 task”. The great advantage of expressing type 2 sensitivity at the type 1 level is that it allows comparison between the two estimates. Assuming the presence of a specific metacognitive noise, and in the absence of other sources of metacognitive information, the meta- d' index is therefore bounded by the superior limit of d' : meta- $d' \leq d'$. In order to compute a metacognitive *efficiency*, i.e. to extract the metacognitive component contained in the meta- d' estimate, one can take either the difference between meta- d' and d' ($meta\ d' - d'$) or the ratio $\frac{meta\ d'}{d'}$ (called M-ratio). Thus, having a metacognitive measure relative to first-order accuracy remedies the highlighted issue that type 2 task performance can be contaminated by type 1 task performance (Galvin et al. 2003).

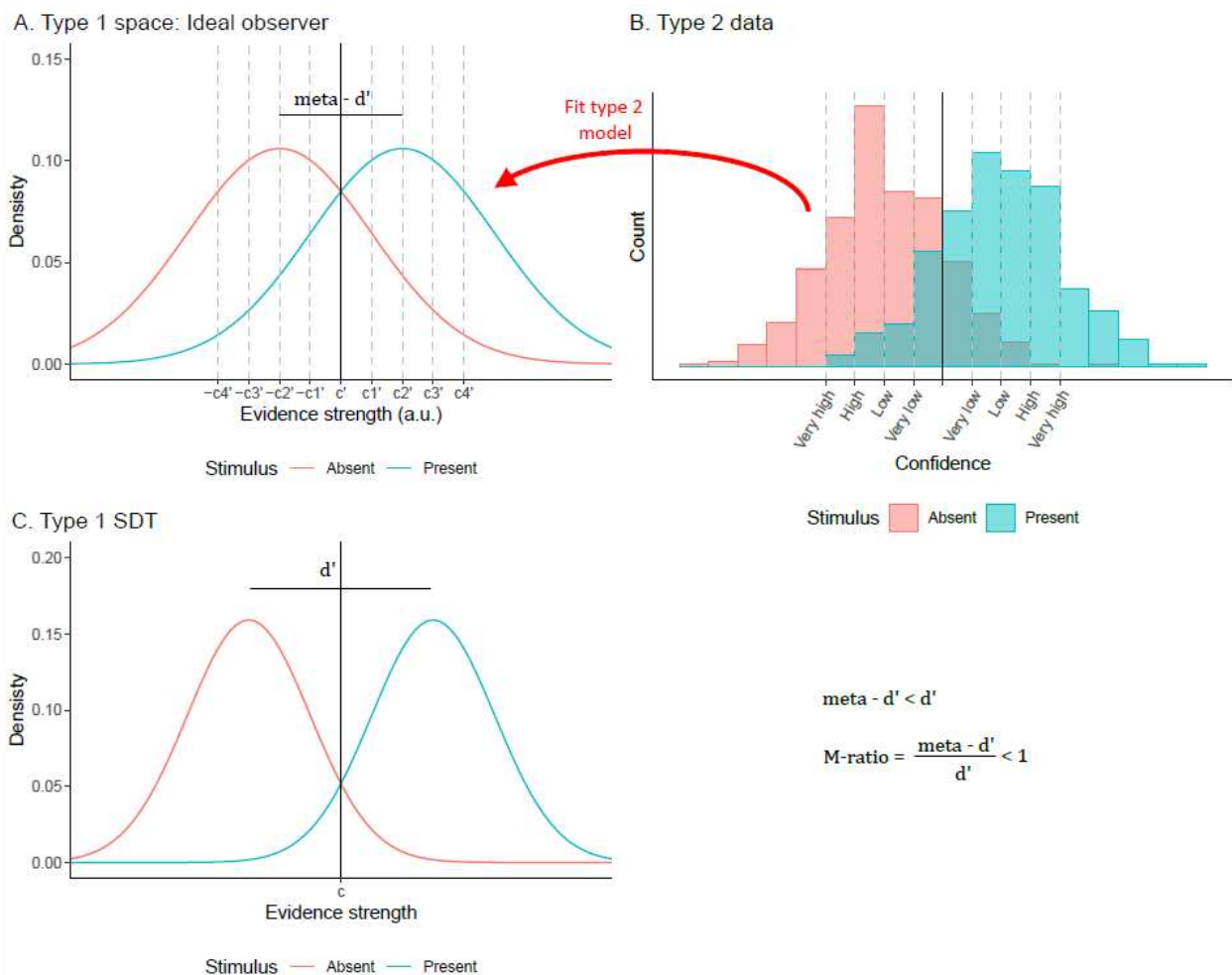


Figure 7. Meta-d' model. Determination of type 2 parameters according to the meta-d' framework. A. Meta-d' model with type 2 parameters: type 2 sensitivity (meta-d') and type 2 criteria ($-c_4', -c_3', -c_2', -c_1', c', c_1', c_2', c_3', c_4'$) expressed within the type 1 space. Meta-d' represents the type 1 sensitivity for an ideal observer (i.e. no metacognitive noise). The red and blue curves are respectively the distributions of evidence for noise samples (i.e. stimulus absent) and noise+signal samples (i.e. stimulus present). B. Type 2 data: distributions of confidence ratings (for absent and present stimulus). Type 2 parameters are determined by fitting the meta-d' model on type 2 data. C. Type 1 parameters: sensitivity (d') and criterion c . To note, d' is a signal-to-noise ratio, rather than the raw distance between the two distributions. In the case of a real observer (metacognitive noise > 0), and assuming no additional source of information for confidence generation, $\text{meta-d}' < d'$.

2.4. Metacognitive architecture

The SDT model posits that type 1 and type 2 decisions rely on the same sensory information through a single process. In a sense, we can say that both decisions are two aspects of the same computation: once the evidence strength is scaled along the decision axis, the algebraic distance between the sensory input and the criterion is sufficient to determine both the type 1 decision - i.e. the sign of the algebraic distance determines the binary decision: Present or Absent, Stimulus 1 or Stimulus 2 - and the type 2 decision, where confidence is proportional to the absolute value of this distance. This is called a "single channel model" (Figure 8). However, under such an account where a common source of noise is shared for type 1 and type 2 decisions, there is a priori no reason why meta-d' should differ from d' . For meta-d' to differ from d' , an additional source of metacognitive noise is needed. In this respect, two other types of models specifying the relations between type 1 and type 2 decisions can be conceived (Maniscalco and Lau, 2016). The second model is called the "dual channel model", in which type 1 and type 2 decisions result from parallel processing streams, each operating on a specific kind of information and characterized by a specific internal noise. The third model is the "hierarchical model", where type 2 decisions result from both the initial sensory processing that gives rise to the type 1 decision and an additional process with its own metacognitive noise. The meta-d' model is an example of a hierarchical model (see Box 2).

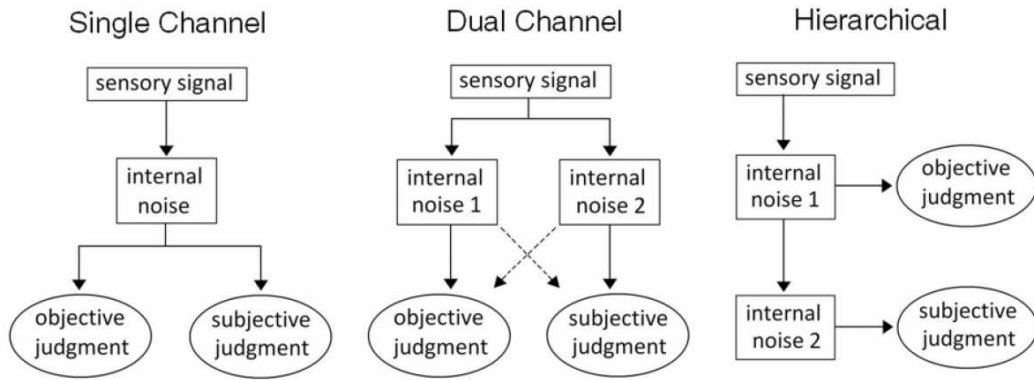


Figure 8. Metacognitive architectures. Three types of models can be conceived to specify the relationship between type 1 decisions (here referred to as “objective judgment”), and type 2 decisions (here referred to as “subjective judgment”). Reprint from Maniscalco and Lau (2016)

Maniscalco and Lau (2016) fitted corresponding computational models on existing data where behavioral dissociations were observed between task-performance and visibility ratings (Lau and Passingham 2006). They found that the hierarchical model best explained the data. However, these results should be taken cautiously, because the models were fitted on one dataset with visibility ratings, and other types of dataset (e.g. using confidence ratings, or probing another cognitive domain) might be best fitted with another model. Moreover, we might want to add a fourth type of model, recently proposed by Mamassian and de Gardelle (2021), which is a hybrid model integrating properties from both the dual channel and the hierarchical models (Figure 9), where confidence is informed both by a confidence noise on top of a sensory noise (i.e. the hierarchical feature), and by an independent stream of information (i.e. the dual channel feature) that contributes to enhancing the reliability of the confidence judgment, called “confidence boost”.

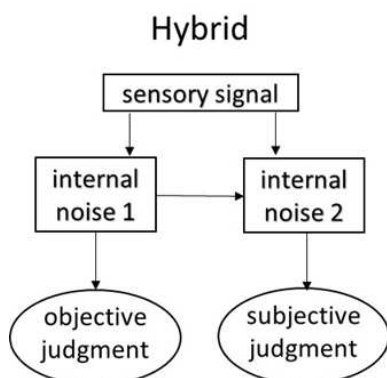


Figure 9. Hybrid metacognitive architecture where subjective judgments benefit from additional and specific sensory evidence, but are degraded by the second order noise that comes on top of the first-order noise.

Box 2: Hierarchical SDT model

Type 1 and type 2 sensitivities (d' and meta- d') depend on hierarchical noisy readouts of the stimulus strength. The first-order decision is corrupted by a gaussian noise: the sensory noise. The second-order decision is a decision about the first-order decision corrupted with an additional gaussian noise: the metacognitive noise.

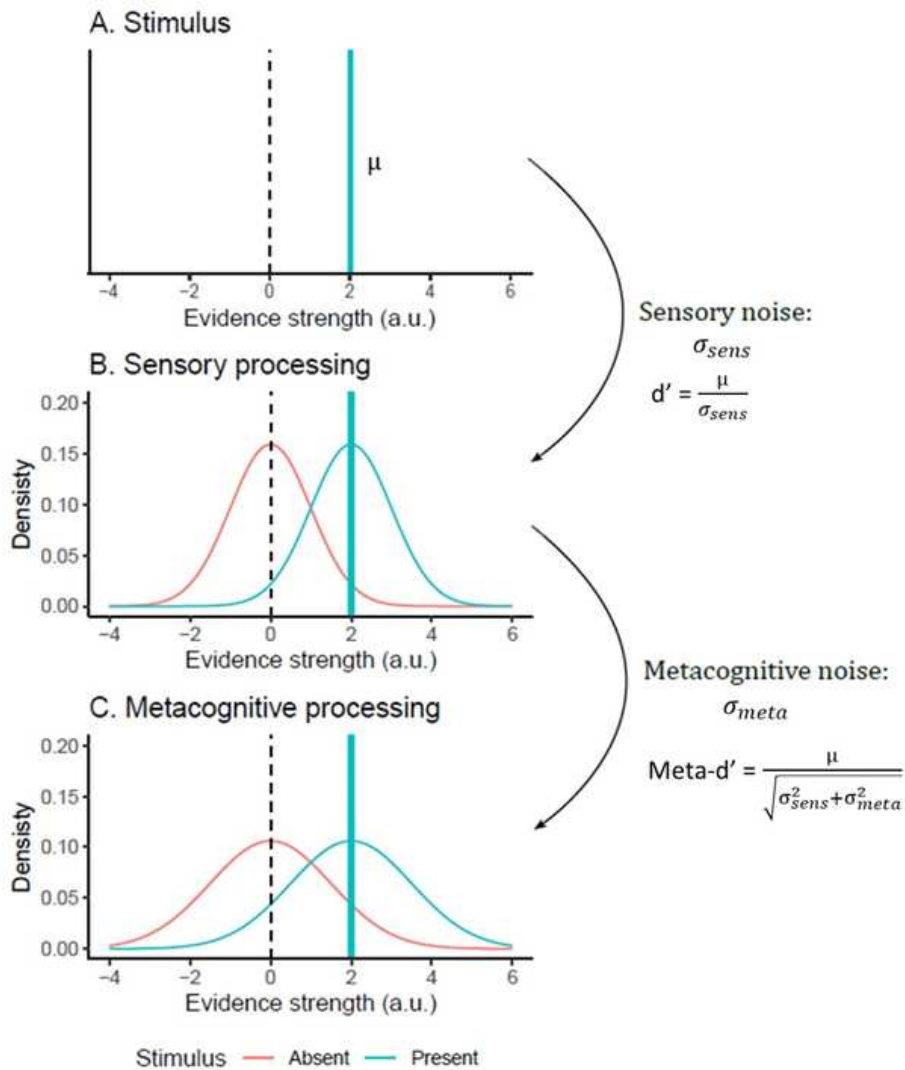


Figure 10. Serial sources of noise for type 1 and type 2 decisions.

2.5. Dynamic models of confidence

How are metacognitive judgments generated? We have seen four types of cognitive architectures that explain the interrelations between type 1 and type 2 processes, but these architectures are silent about the dynamic of the decision processes themselves. A fruitful approach to model the dynamics of decisions, known as “sequential sampling” is to consider decisions as resulting from “accumulator-to-bound” mechanisms, where noisy sensory evidence is sampled progressively until a decision bound is reached. A popular model that implements this principle is known as the “drift-diffusion model” (Ratcliff and McKoon 2008; Ratcliff and Rouder 1998). It was initially developed to predict two-choice decisions and response times, by determining four parameters: the drift rate (i.e. the quality of sensory evidence), the decision bounds (i.e. evidence thresholds from which decisions are made), the bias (i.e. the starting point of the decision variable), and the non-decision time that includes stimulus encoding and response execution (Figure 11).

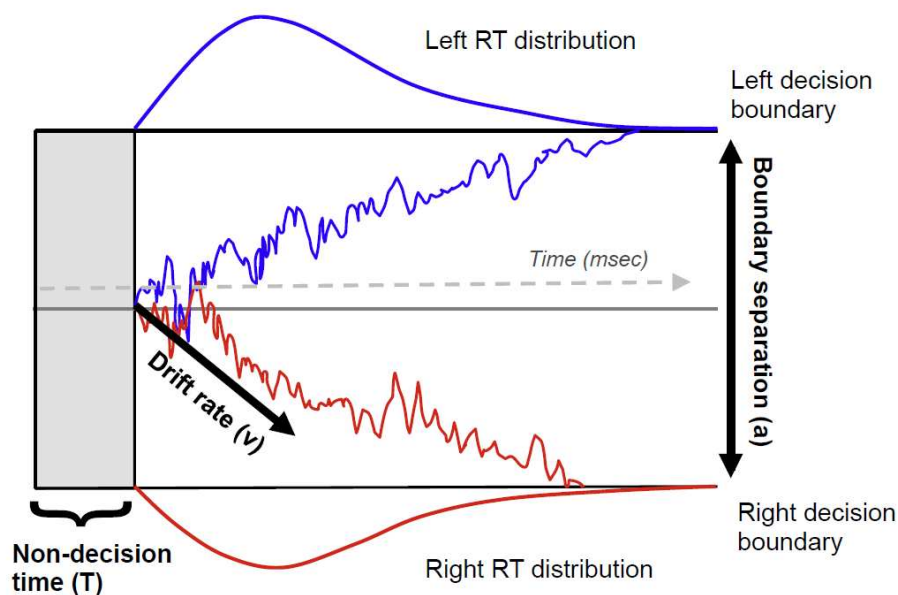


Figure 11. Illustration of the evidence accumulation model. The decision variable accumulates noisy evidence until a decision bound is reached. Reprint from O’Callaghan et al. (2017)

Sequential sampling models were also applied to predict responses and firing rates of populations of neurons in animal studies (Kepecs et al. 2008; Kiani and Shadlen 2009). For instance, Kiani and Shadlen (2009) presented macaques implanted with intracranial electrodes with noisy visual stimuli consisting of random dots moving left or right with a manipulated degree of coherence. The monkeys learned to make an eye-gaze toward the

dominant motion-direction to indicate their response, either left or right, after a delay period. Correct responses were rewarded with juice, and incorrect responses were not rewarded. As expected under an evidence accumulation framework, firing rates of lateral intraparietal (LIP) neurons increased as a function of sensory exposition, with a faster increase when motion coherence was high (i.e. higher evidence quality) compared to low motion coherence, ultimately resulting in an eye-gaze decision. In other words, the firing rates have been interpreted as a decision variable that accumulates until a decision bound is reached. Interestingly, the authors introduced a third response option on a random half of the trials, called the “opt-out” option, that appeared during the delay period. The “opt-out” option is always associated with a juice reward (yet a smaller quantity than in correct answers), irrespective of the correctness of the decision. It has been shown that macaques used the opt-out option adaptively, i.e. in difficult trials, a strategy that maximized their reward. The “opt-out” option can thus be interpreted as a proxy for choice uncertainty or low confidence. Congruently with this interpretation, during opt-out trials the firing rates of LIP neurons reached a middle-point between the two average trajectories that characterized left and right decisions, as if evidence was not accumulated enough to make a decision (Figure 12). Thus, it has been proposed that LIP neurons’ firing rates encode response confidence in macaques.

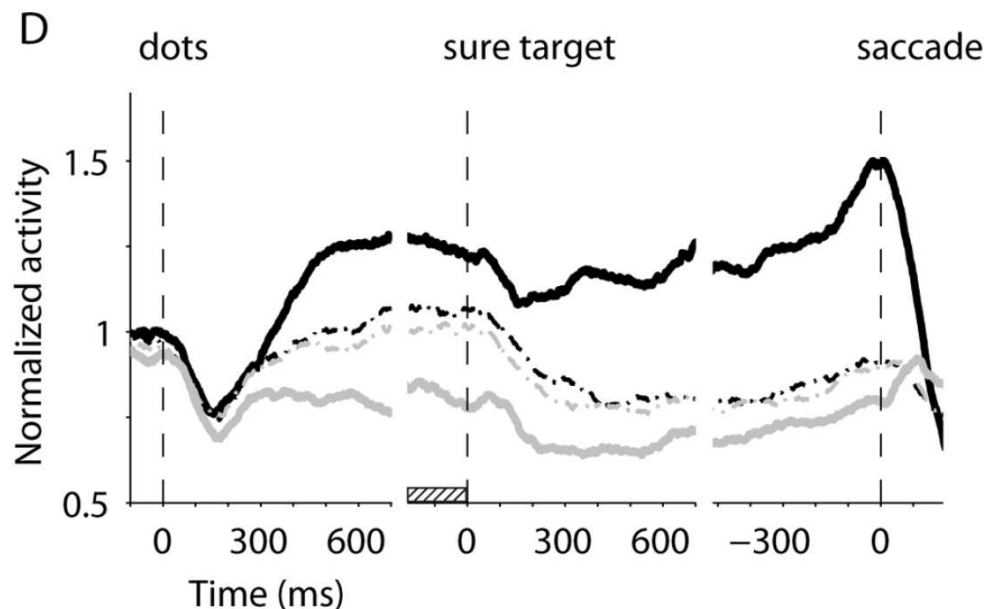


Figure 12. Intracranial signature of confidence in monkeys. Population average responses of 70 LIP neurons from two monkeys. Average firing rates for response option 1 (black) and response option 1 (gray) are shown for all correct choices, during motion viewing and the imposed delay between the stimulus and the response. Averages are aligned to

motion onset (left part of graph) and saccade initiation (right). The dashed lines show neural activity on trials in which the opt-out option was chosen (black and gray, motion toward response 1 and 2, respectively). The middle portion of the graph shows activity in the delay period, aligned to the onset of the opt-out option. Reprint from Kiani and Shadlen (2009)

However, these evidence accumulation models did not account for specific behaviors like “changes-of-mind” (van den Berg et al. 2016; Resulaj et al. 2009). Changes of mind are trials in which participants initiate a motor action toward a response option but suddenly take the opposite trajectory to reach the other response option. These situations suggest that evidence might continue to accumulate after the decision is made, and that post-decisional evidence might also inform confidence (Figure 13, Pereira et al. 2021; Pleskac and Busemeyer 2010; Yeung and Summerfield 2012).

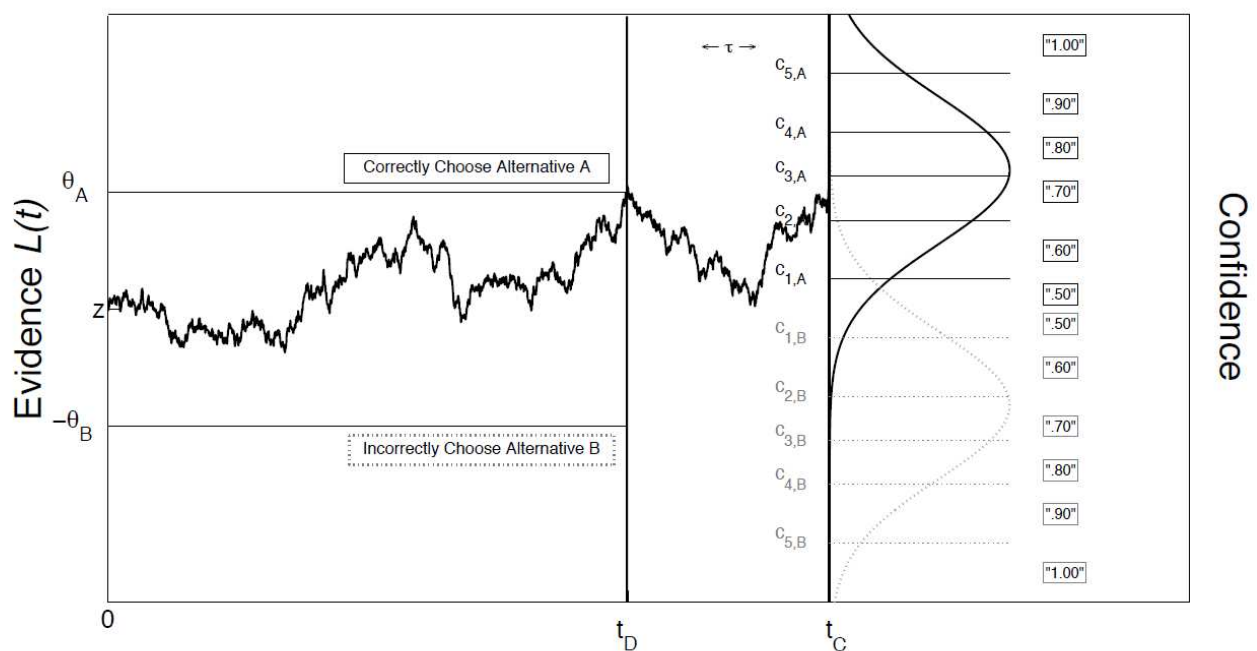


Figure 13. Post-decisional model of evidence accumulation. Evidence is accumulated continuously, the decision is made when the accumulated evidence reaches a decision boundary (time t_D), and the level of confidence is determined by the value of the decision variable at time t_C . Reprint from Pleskac and Busemeyer (2010)

2.6. Domain-generality of metacognition

We have seen that perceptual metacognition was best fitted by a hierarchical model (Maniscalco and Lau 2016), where subjective judgments (confidence ratings or visibility

ratings) are degraded by two sources of noise: a sensory noise and an additional metacognitive noise (see Box 2). Assuming that subjective judgments from other cognitive domains (e.g. memory, agency) are also best explained by a hierarchical model of metacognition, we end up with two possibilities of global architecture: either the metacognitive noise is shared among the cognitive domains, or specific to each domain (Figure 14). In other words: is metacognition a monolithic process, or a set of distributed ones? Indeed, metacognition can be conceived either as one “domain-general” mechanism that processes information coming from different cognitive domains with a shared format, or as an assembly of “domain-specific” processes locally computing metacognitive contents.

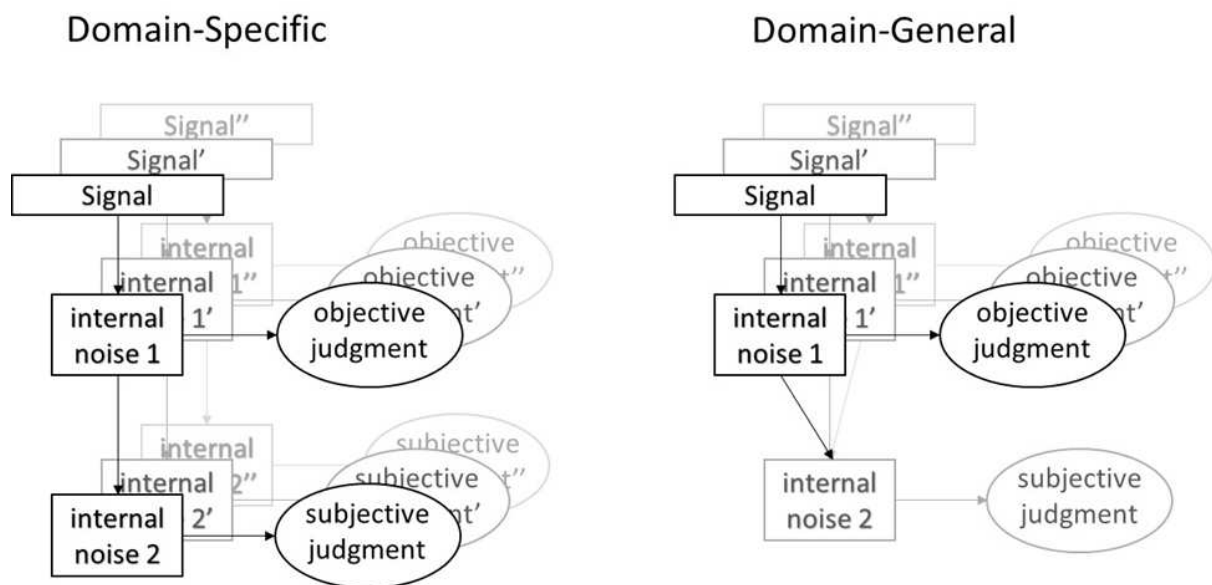


Figure 14. Domain-specific versus domain general architecture. Two types of hierarchical architectures, depending on the specificity (left) or generality (right) of metacognitive processing (internal noise 2). Single and double quotes indicate parallel streams of information processing (e.g. perceptual, memory, agency) with specific types of input signal, internal noise and output judgment.

Thus, if metacognition is domain-general, the shared metacognitive noise between domains should lead to robust correlations of metacognitive performance across tasks. A recent meta-analysis (Rouault et al. 2018) revealed weak evidence of cross-domain correlations of metacognitive performance between memory and perceptual studies. These results are supported by another meta-analysis of 47 fMRI studies including both metaperceptual and metamemory tasks (Vaccaro and Fleming 2018) revealing specific recruitment of brain regions for each task-domain: left dorso lateral prefrontal cortex (dlPFC) in metamemory tasks, and right anterior dlPFC in metaperceptual tasks. However, it has been proposed that such under-estimation of domain-generality could result from task

specific requirements (Hu et al. 2022), or a difference of task types (e.g. discrimination being more used in perceptual studies as opposed to detection tasks more used in memory studies) rather than task domains (Rouault et al. 2018). This interpretation has been supported by a recent fMRI study (Mazor, Friston and Fleming 2020), where participants were asked to perform two metaperceptual tasks: detection and discrimination. Type 2 ROC curves revealed an asymmetry in metacognitive sensitivity between “yes” and “no” responses in the detection task, that was absent in the discrimination task (Figure 15). Furthermore, detection judgments were characterized by a quadratic neural activity in the fronto-polar cortex and right temporo-parietal junction, a pattern which was absent when confidence judgments were made in a discrimination task.

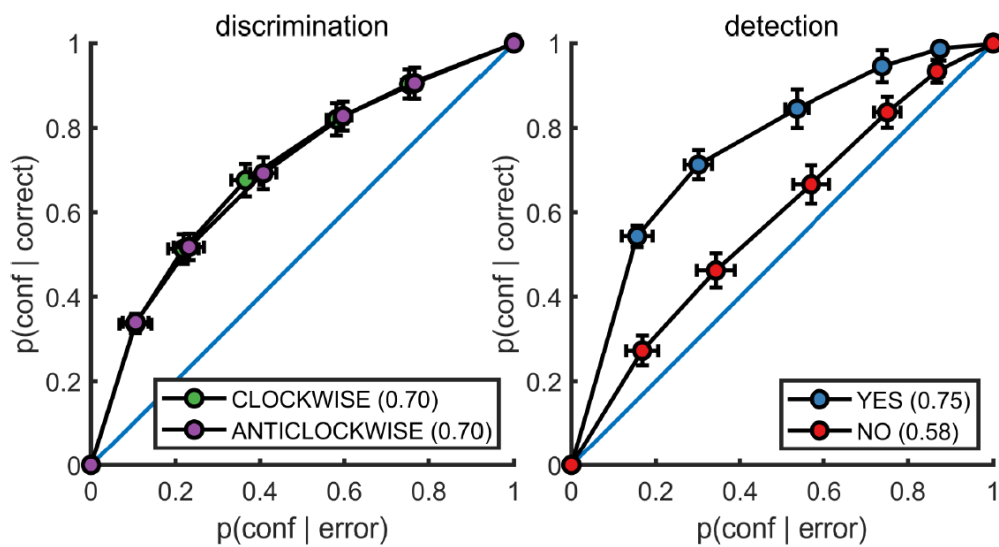


Figure 15. Metacognition in discrimination versus detection tasks. Type 2 ROC curves for a discrimination task (left) and a detection task (right), where colors indicated the response categories. Reprint from Mazor et al. (2020)

Increased homogeneity between first-order tasks led to better cross-domain correlations of metacognitive performance (Mazancieux et al. 2020; Morales, Lau, and Fleming 2018). But more generally, cross-task correlations of meta performance should be interpreted with caution since they could be driven by covarying factors such as gray matter volume (GMV) (McCurdy et al. 2013) or other behavioral factors (IQ, attention, etc). Indeed, McCurdy and colleagues have shown that meta-performance in the memory domain correlated with the volume of the precuneus, whereas visual meta-performance correlated with the volume of prefrontal regions.

The existence of a putative domain-general process for metacognition involves a “common currency” (de Gardelle and Mamassian 2014), i.e. shared cross-domain signals that constitute a common basis for confidence generation. In line with this hypothesis, Faivre and colleagues (2018) designed an experiment involving visual, auditory and tactile sensory modalities, and provided behavioral, modeling and neurophysiological evidence supporting the idea that confidence estimates not only share a supramodal format, but also rely on shared decisional signals across modalities. Other variables have been proposed as plausible shared input signals upon which a general metacognitive process could operate, such as response times or fluency of processing (Boldt, de Gardelle, and Yeung 2017; Rouault et al. 2018), or the precision of the neural representation (Shea and Frith 2019; Yeung and Summerfield 2012).

2.6.1. Two levels of metacognition

Interestingly, a fMRI study (Morales et al. 2018) investigating both metaperception and metamemory provided specific evidence for both domain-specificity and domain-generality. In particular, it was found that generic confidence levels and accuracy were coded within a network centered on the ACC and pre-SMA, and that domain-specific confidence but not accuracy was correlated with the rIPFC activity. The authors proposed the co-existence of hierarchically organized layers for metacognitive processes: one responsible for generic low-level feeling of confidence, and a higher-order level responsible for content-rich metacognitive representations, tagging the generic confidence with contextual information. This type of architecture might also find an echo in the two-layered metacognition proposed by Arango-Muñoz (2011), that distinguished low-level metacognition constituted by non-representational epistemic feelings, and high-level metacognition with meta-representations informed by conceptual contents retrieved from domain-specific beliefs and memory. More recently, a distinction has been made between “local” and “global” metacognition (Rouault and Fleming 2020), where local refers to the monitoring of accuracy on a trial-by-trial basis, which in turn constitute the building blocks for the formation of global judgments, i.e. more general statements of performance about a task or a cognitive domain, ultimately leading to self-beliefs. Local and global levels are hierarchically organized, so that local judgments are the building blocks enabling the summarized global judgments, and global judgments have a reciprocal top-down effect on local metacognition (Figure 16).

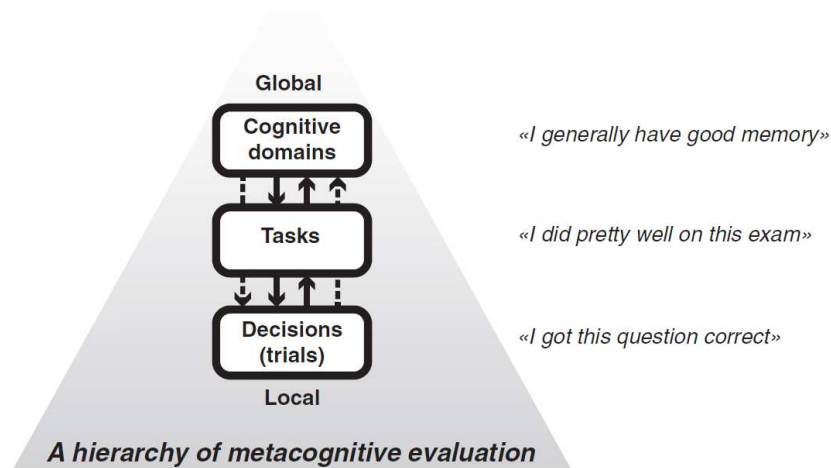


Figure 16. Levels of metacognitive evaluations. Metacognitive evaluations interacting with one another: local metacognition informing higher self-beliefs in a bottom-up stream, and reciprocally higher self-beliefs about performance influences confidence judgments at the local level. Rightward statements illustrate different levels of self-evaluations. Reprint from Seow et al. (2021)

2.7. Metacognitive deficits

Abnormalities of confidence are well documented in schizophrenia, both from questionnaires (Lysaker et al. 2015) and experimental protocols (Hoven et al. 2019). Specific indexes are often used to quantify metacognitive deficits among individuals with schizophrenia, such as overconfidence in errors, the “confidence gap” (Moritz and Woodward 2006; Moritz, Woodward, and Ruff 2003), or the “knowledge corruption index” (Moritz et al. 2004). The “confidence gap” is the difference in confidence levels between correct and incorrect responses. Ideally, confidence should be high for correct responses and low for incorrect responses. Therefore, an inability to correctly calibrate confidence on correctness should decrease the confidence gap (Figure 17). Another index is the “knowledge corruption index”, which is the ratio of high-confident errors over all high-confidence responses.

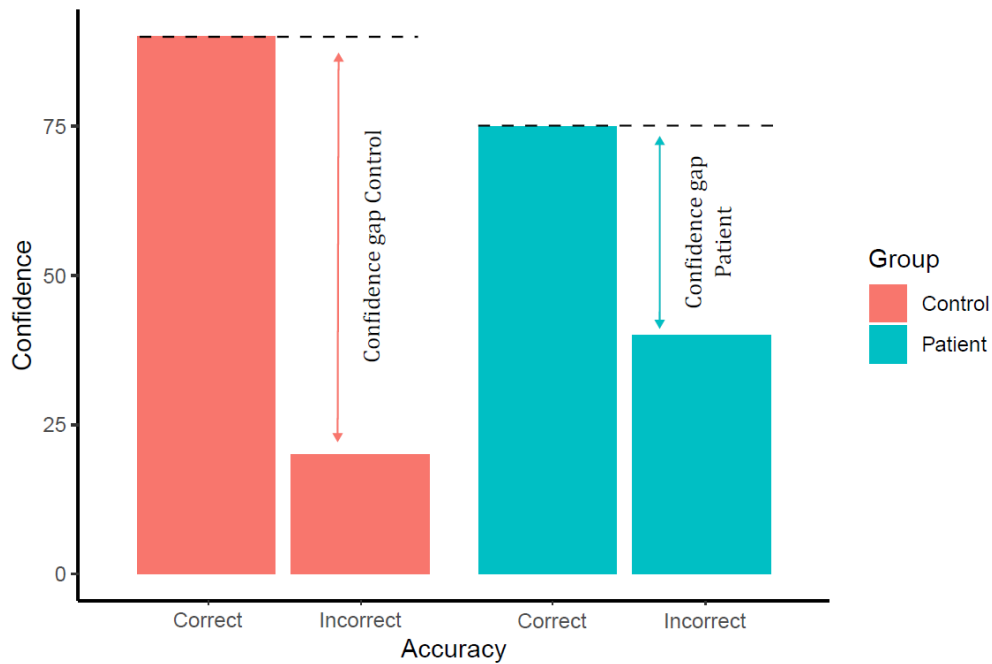


Figure 17. Illustration of the confidence gap index. The y-axis indicates continuous confidence levels from 0 (very low confidence) to 100 (very high confidence). A reduced confidence gap among patients compared to healthy controls is used as a proxy for a metacognitive deficit.

Relying on these measures, numerous studies employing first and second-order tasks have reported lower metacognitive performance in schizophrenia compared to healthy controls across different cognitive domains such as vision (Dietrichkeit et al. 2020; Moritz et al. 2014) audition (Gaweda and Moritz 2019), emotion perception (Kother et al. 2012; Moritz et al. 2012), and memory (Mayer and Park 2012; Moritz and Woodward 2002).

However, this result has been mitigated by a series of recent studies that failed to replicate such a metacognitive deficit in schizophrenia in the perceptual domain (e.g. Pereira et al. 2021; Powers, Mathys, and Corlett 2017; Wright et al. 2020). For instance, in a visual discrimination task, Faivre et al. (2019) showed that metacognitive efficiency (M-ratio) was preserved among patients compared to healthy controls. In addition, parameters of an evidence accumulation model explaining first and second-order responses were similar between groups. How can we make sense of these unexpected results? We noticed that these latter studies all share a common feature: they control for potential group differences in first-order performance, either at the experimental level through adaptive procedures (Levitt 1971), or at the statistical level through the use of the meta-d' framework (Maniscalco and Lau 2012). This is especially important in the case of schizophrenia, where first-order impairments are well documented (for reviews, see Gopal and Variend 2005; Heinrichs and

Zakzanis 1998) and may therefore contaminate metacognitive measures. To determine whether the deficits are specifically related to metacognitive processing, or merely inherited from a first-order deficit, bias-free metacognitive measures should be used (e.g. M-ratio derived from SDT, or confidence efficiency).

In this thesis, we quantified metacognitive deficits using bias-free measures of metacognition (experimental chapters A and B), and investigated neural markers of confidence in a perceptual task that controlled for first-order performance (experimental chapter 3).

2.8. Remediating metacognitive deficits?

Metacognitive training has been developed to help patients gain knowledge about themselves (MCT: Metacognitive Training in schizophrenia, Moritz et al. 2014; MERIT: Metacognitive Reflection and Insight Therapy, de Jong et al. 2019). For instance, MCT is a training designed to help patients to cope with delusional thinking by raising awareness about several cognitive biases such as premature decision-making (“jump to conclusion”), erroneous attributional inferences (e.g. preferring mono-causal rather than pluri-causal explanations in complex situations), overconfidence, or a bias against disconfirmatory evidence. Typically, ambiguous scenarios or incomplete pictures are displayed and patients are asked to provide their interpretation, giving rise to discussions about alternative explanations and the negative effects of giving in to biased reasoning. Thus, MCT and MERIT can be thought of as high-level metacognitive training, tapping into the explicit recognition of one’s shortcomings.

We might also consider the possibility to train metacognition at a lower level. Considering the hierarchical organization of metacognition described above (Rouault and Fleming 2020), where global estimates of confidence are formed through the accumulation of local (i.e. trial-by-trial) evaluations of correctness, it makes sense to try to amend high-level distorted beliefs and global abnormalities of confidence by training patients to better calibrate local confidence judgements on their performance. The very possibility of improving local metacognition and transferring this improvement to other cognitive domains has been suggested among healthy participants (Carpenter et al. 2019). However, this metacognitive training suffered from several confounds, which are detailed and addressed in experimental chapter 4. Considering this training as a potential remediation for metacognitive deficits in individuals with schizophrenia, we aimed at reassessing its efficiency.

In what follows, I present four empirical studies: 1) a meta-analysis which sought to quantify metacognitive deficits in schizophrenia across cognitive domains, while taking first-order deficits into account; 2) a follow-up behavioral study which assessed metaperceptual and metamemory performance in schizophrenia while controlling for first-order performance; 3) an investigation of electrophysiological markers of confidence in schizophrenia; and 4) an attempt to replicate the effect of an online metacognitive training.

PART II - Experimental chapters

1. Meta-analytic assessment of metacognitive deficits among individuals with schizophrenia

We conducted a Bayesian meta-analysis to assess the magnitude of metacognitive deficits across cognitive domains in schizophrenia, while hypothesizing that it would be smaller in studies controlling for first-order performance, compared to studies that did not. This hypothesis was driven by a recent behavioral study from Faivre et al. (2021) where visual performance was titrated with an adaptive staircase, showing that metaperception was preserved among preserved patients with schizophrenia. This result was unexpected since it was at odds with the existing literature on metacognitive deficits among patients. Was it a sampling bias? Asking ourselves why metacognitive abilities would be preserved in this sample of patients, we noticed a pattern of similar results in other recent studies also controlling for first-order performance (Powers et al. 2017; Wright et al. 2020). Therefore, we aimed at quantifying the contribution of this factor (controlling performance versus not controlling performance) on the effect size estimation.

We included 42 studies, and in line with our hypothesis we found that metacognitive deficits among patients with schizophrenia were reduced within studies that controlled for first-order performance compared with studies that did not control for this factor. We also highlighted that the metacognitive deficit was more important in the memory domain, but since all but 1 of the 27 memory studies included in our meta-analysis did not control for memory performance, it was impossible to conclude if the observed deficit was genuinely metacognitive or inherited from impaired memory.

Status of the manuscript: Published in Neuroscience and Biobehavioral Reviews.

Reference: 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.



Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders

Martin Rouy^{a,*}, Pauline Saliou^a, Ladislav Nalborczyk^b, Michael Pereira^a, Paul Roux^{c,1}, Nathan Faivre^{a,1}

^a Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000, Grenoble, France

^b Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000, Grenoble, France

^c Service universitaire de psychiatrie d'adulte et d'addictologie du Centre Hospitalier de Versailles, CESP, Equipe DisAP-DevPsy, INSERM, Université Paris-Saclay et Université de Versailles Saint-Quentin-En-Yvelines, France

ARTICLE INFO

Keywords:

Metacognition
Insight
Psychosis
Schizophrenia
Meta-perception
Meta-memory

ABSTRACT

Metacognitive deficits are well documented in schizophrenia spectrum disorders as a decreased capacity to adjust confidence to performance in a cognitive task. Because metacognitive ability directly depends on task performance, metacognitive deficits might be driven by lower task performance among patients. To test this hypothesis, we conducted a Bayesian meta-analysis of 42 studies comparing metacognitive abilities in 1425 individuals with schizophrenia compared to 1256 matched controls. We found a global metacognitive deficit in schizophrenia ($g = -0.57$, 95 % CrI [-0.72, -0.43]), which was driven by studies which did not control task performance ($g = -0.63$, 95 % CrI [-0.78, -0.49]), and inconclusive among controlled-studies ($g = -0.23$, 95 % CrI [-0.60, 0.16], $BF_{01} = 2.2$). No correlation was found between metacognitive deficit and clinical features. We provide evidence that the metacognitive deficit in schizophrenia is inflated due to non-equated task performance. Thus, efforts should be made to develop experimental protocols accounting for lower task performance in schizophrenia.

1. Introduction

Metacognition is the ability to monitor and control our own mental processes. Metacognitive deficits are thought to play an important role in schizophrenia spectrum disorders (hereafter: schizophrenia) (Hasson-Ohayon et al., 2018). These deficits are inferred both from subjective structured interviews (Semerari et al., 2003) and objective neuropsychological tasks (Koren et al., 2006), and have been linked to core features of schizophrenia including positive and negative symptoms (McLeod et al., 2014), lack of insight into illness (David et al., 2012), disorganisation (Vohs et al., 2014), functioning (Davies and Greenwood, 2020), and quality of life (Arnon-Ribenfeld et al., 2017).

Despite numerous studies, no meta-analysis has yet been conducted to examine metacognition in schizophrenia. Here we sought to conduct a systematic review and meta-analysis of neuropsychological measures of metacognitive performance in schizophrenia compared to matched healthy controls. From an experimental perspective, the gold standard to quantify metacognition is to assess how participants perform an

experimental task (first-order task) and reflect on their own accuracy via confidence ratings (second-order task). Several studies employing this design have reported lower metacognitive performance in schizophrenia compared to healthy controls across different cognitive domains such as vision (Dietrichkeit et al., 2020; Jia et al., 2020; Moritz et al., 2014), audition (Gaweda and Moritz, 2019), emotion perception (Kother et al., 2012; Moritz et al., 2012; Pinkham et al., 2018), and memory (Berna et al., 2019; Mayer and Park, 2012; Moritz and Woodward, 2006a). However, these results are mitigated by recent studies that failed to reveal such metacognitive deficits (Faivre et al., 2019; Powers et al., 2017; Wright et al., 2020). Noticeably these studies controlled for potential group differences in first-order performance, either at the design level through adaptive staircase procedures (Levitt, 1971), or at the metric level through indices of metacognitive performance which are independent of first-order performance (Maniscalco and Lau, 2012). This is especially important in schizophrenia where cognitive impairments are well documented (Gopal and Variend, 2005; Heinrichs and Zakzanis, 1998) and associated with metacognitive deficits (Davies and

* Corresponding author at: Laboratoire de Psychologie et Neurocognition, CNRS UMR 5105, UGA BSHM, 1251 Avenue Centrale, 38058 Grenoble Cedex 9, France.
E-mail address: martin.rouy@univ-grenoble-alpes.fr (M. Rouy).

¹ Equal contribution.

Greenwood, 2020). This known issue in the field of metacognition (Galvin et al., 2003; Maniscalco and Lau, 2012) can be stated as follows: because it is easier to finely adjust confidence ratings following an easy task than a difficult one, any comparison of metacognitive performance between two conditions that differ in terms of task difficulty is non-specific: should a difference in metacognitive performance be observed, it is impossible to tell if it stems from first-order (i.e., task difficulty), or second-order origins (i.e., metacognitive processes per se). As first-order performance is typically lower in patients vs. controls, a putative metacognitive deficit may be merely inherited from a deficit at the first-order level, and thus not specific to second-order processing. To determine whether schizophrenia involves specific deficits in metacognitive abilities, we conducted a systematic review followed by a Bayesian meta-analysis on a sample of 42 studies. Our main hypothesis was that metacognitive deficits would be smaller in studies controlling for first-order performance. Following a pre-registered plan, we conducted additional subgroup analyses and meta-regressions to explore if metacognitive deficits vary across cognitive domains, the severity of schizophrenia symptoms, and antipsychotic dosage. We had preregistered two additional directional hypotheses regarding the influence of diagnosis (first-episode vs. chronic schizophrenia) and symptomatology (depression, insight) on metacognitive abilities, but these analyses were not conducted due to the scarcity of the available data.

2. Methods

This meta-analysis followed the PRISMA recommendations (Moher et al., 2009). The protocol was registered on PROSPERO (CRD42020188614) on May 26th 2020, before data extraction.

2.1. Inclusion criteria

Inclusion criteria followed the PICO framework.

- **Population:** individuals with schizophrenia or related disorders (schizoaffective, schizophreniform), as defined by standard diagnostic criteria (DSM-III, DSM-III-R, DSM-IV, DSM-IV-R, DSM-IV-TR, DSM 5, ICD-10).
- **Intervention:** a computerized or manual experimental task with self-reported retrospective confidence judgments as behavioral measures on a confidence scale with more than one trial.
- **Comparison:** healthy controls.
- **Outcome:** meta-performance defined as the strength of the relationship between first-order performance (accuracy on a neuropsychological task in perception, memory, executive functions, social cognition, and agency) and retrospective confidence judgments in the first-order performance, repeated for each trial. Meta-performance indices included: meta- d' , M-Ratio, AUROC2, logistic regression, confidence gap, knowledge corruption index, gamma correlation (for details on these measures, see Fleming and Lau, 2014).

2.2. Search strategy

We retrieved English written preprints and peer-reviewed articles in three databases – Pubmed, Web of Science, Scopus – with the following query applied to the title, abstract and keywords:

(schizophrenia OR schizophrenic OR schizo-affective OR schizo-affective) AND (confident OR confidence OR metacognition OR meta-cognitive OR "error awareness" OR "error monitoring").

2.3. Screening and data extraction

The search was performed on April 24th 2020, and no new search before analysis was performed. This query could not identify one article previously known by a co-author (Powers et al., 2017) as it contained

non-matching key-words and reported metacognitive performance in supplementary materials. It was manually included in the list of publications. Two authors (MR and PS) screened studies for inclusion in parallel, using Cadima (<https://www.cadima.info>; see supplementary information (SI) for details). For each study group, MR and PS extracted the following primary outcomes:

- whether the study controlled for first-order performance between groups (TRUE or FALSE)
- metacognitive performance indices (see above)
- first-order accuracy (% correct, d')

Depending on the data available, either the mean and standard deviation, or raw statistics (t and F values) were extracted (SI). The following secondary outcomes were extracted:

- cognitive domain
- clinical characteristics including Positive and Negative Syndrome Scale scores (PANSS total, positive, and negative) and antipsychotic dosage (chlorpromazine equivalent).
- age (mean and standard deviation)
- sample size

2.4. Statistical analyses

All analyses were conducted in R. We used the brms package (Bürkner, 2017) based on the Stan framework (Carpenter et al., 2017) to fit Bayesian meta-analytic multilevel models.

Before testing our main hypothesis regarding the influence of equating first-order performance on metacognitive abilities, we fitted a global model M1 with fixed and random effects as follows:

$$M1: G_i | \sigma_i \sim \text{Intercept} + (\text{Intercept} | \text{study})$$

Where G_i denotes the Hedge's g effect size of study i , σ_i denotes the standard error of the effect size from study i , thereby accounting for different sample sizes across studies (SI). M1 estimated the overall effect-size of a difference in metacognitive performance between groups (the grand intercept of the model) while accounting for the between-study variability (random intercept per study; see SI for prior definition). To test the existence of a metacognitive deficit in schizophrenia (H1), we compared the estimations of M1 to the estimations of an alternative model M0 assuming that metacognitive deficit was inexistent (i.e., fixing the intercept at 0; H0).

$$M0: G_i | \sigma_i \sim 0 + (\text{Intercept} | \text{study})$$

Hypothesis testing:

Results were interpreted based on the relative evidence toward H0 (absence of a metacognitive deficit in schizophrenia) or H1 (presence of a metacognitive deficit in schizophrenia) given by the Bayes factor (BF), and the summary statistics of the posterior distribution (mean and 95 % credible interval, CrI). The BF is the ratio of the marginal likelihoods of each hypothesis. We note BF_{10} the ratio of evidence in favour of H1 and BF_{01} the ratio of evidence in favour of H0. We used the interpretation of BFs given by Wagenmakers et al. (2018), which translates continuous BF values into a categorical scheme. Thus, we considered the relative strength of evidence in favor of hypothesis H1 over H0 (resp. H0 over H1), to be anecdotal if $BF_{10} \in [1, 3]$ (resp. $[\frac{1}{3}, 1]$), moderate if $\in [3, 10]$ (resp. $[\frac{1}{10}, \frac{1}{3}]$), strong if $\in [10, 30]$ (resp. $[\frac{1}{30}, \frac{1}{10}]$), very strong if $\in [30, 100]$ (resp. $[\frac{1}{100}, \frac{1}{30}]$) and extremely strong if > 100 (resp. < 0.01).

For subgroup analyses, we retrieved the summary statistics (mean and 95 % CrI) of the difference between the two posterior distributions obtained in each group. Then we assessed in each case under which hypothesis (H0: absence of deficit or H1: existence of a deficit) the data was the most plausible.

To test our main hypothesis, we assessed the influence of equating

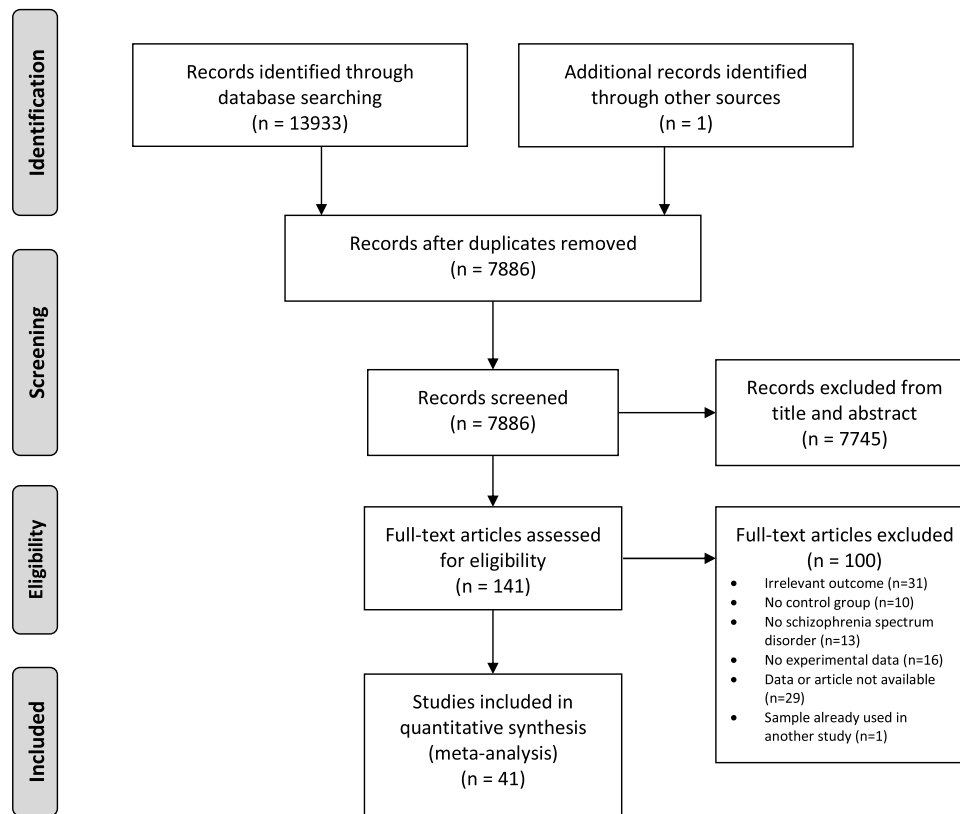


Fig. 1. Flow diagram of the selection process.

first-order performance with a model M2, identical to M1 including performance-matching as an additional binary predictor as follows:

$$M2: G_i | \sigma_i \sim \text{Intercept} + \text{control_type1} + (\text{Intercept} | \text{study})$$

Where *control_type1* is a binary predictor, TRUE for controlled-studies, FALSE otherwise.

All moderator analyses were first motivated by heterogeneity assessments. Three measures of heterogeneity were computed: the Q-statistic (Card and Little, 2016), the Q-between statistic (Borenstein et al., 2010), and the I^2 index for the percentage of the total variation due to between-studies variability (Higgins and Thompson, 2002). I^2 values between 0 and 0.25 suggest small magnitudes of heterogeneity, 0.25 to 0.50 medium magnitudes, and > 0.50 large magnitudes. Exploratory subgroup analyses and meta-regressions were performed in case of significant Q-between and I^2 above 25 % (Huedo-Medina et al., 2006). Namely, we assessed the metacognitive deficit amplitude across cognitive domains by fitting a model identical to M1 with the between-study variable “cognitive domains” (perception, memory, others) as an additional categorical covariate. We also explored the correlation between metacognitive performance among patients and continuous variables by adding standardized (z-scores) PANSS scores and chlorpromazine equivalent as meta-regressors to M1.

2.5. Quality assessment

To quantify the risk of bias in individual studies, we assessed whether our selection contained extreme effect size values via a leave-one-out sensitivity analysis (SI). We also assessed the risk of bias according to the Newcastle-Ottawa Scale (NOS) adapted for case-control studies (SI). Publication bias was assessed using a funnel plot of observed outcomes against corresponding standard errors (Sterne and Harbord, 2004). The distribution of p-values was analyzed using the R package *dmeter* (Harrer et al., 2019) to examine whether some of the studies were

subject to p-hacking (p-curve: Simonsohn et al., 2014).

3. Results

3.1. Study characteristics

Our search retrieved 13933 records, 7886 after duplicates removal. 7745 records were excluded after title and abstract screening (Fig. 1). Another 100 articles were excluded based on full-text screening, resulting in a selection of 41 articles.

One article was excluded because of a strongly deviant outcome identified via a leave-one-out analysis performed on the metacognitive deficit effect-size (SI). Among the 40 remaining articles, two were split into two independent studies as they involved different populations (young versus old: Gaweda (2015); hallucination-prone versus non-hallucination-prone: Powers et al. (2017)). The final selection consisted of 42 studies, with a total population of 2681 participants (1425 patients) (Table 1).

Our selection included 10 perception (auditory and visual), 27 memory, 4 social cognition, and 1 agency studies. Because of their low number, social cognition and agency studies were regrouped into a generic category termed “others”.

3.2. Global metacognitive deficit in schizophrenia

The meta-analytic model M1 revealed lower metacognitive performance in the schizophrenia vs. control groups with an effect size $g = -0.57$, 95 % CrI $[-0.72, -0.43]$ (Fig. 2). Comparison against the null hypothesis (i.e., absence of a metacognitive deficit in schizophrenia modelled by M0) resulted in a Bayes factor favoring the alternative hypothesis $BF_{10} = 36.56 \times 10^6$, indicating extremely strong evidence in favor of a metacognitive deficit in schizophrenia. Of note, this pattern of results was robust to prior variations (SI).

Table 1

Study characteristics. KCI: knowledge corruption index; AUROC2: area under the type 2 receiver operating characteristic curve.

Study	Sample size		Age		Matched performance	Cognitive domain	Metacognitive index	NOS
	SCZ	HC	SCZ	HC				
Dietrichkeit et al. (2020)	39	20	34.72 ± 8.68	30.55 ± 8.54	no	perception	KCI	4.0
Jia et al. (2020)	38	38	22.6 ± 8.3	23 ± 4.6	yes	perception	AUROC2	5.0
Jones (2020)	215	151	41.72 ± 11.64	41.95 ± 12.42	no	others	confidence gap	5.0
Wright et al. (2020)	50	68	27.17 ± 1.3	25.7 ± 6.6	yes	perception	M-ratio	6.0
Berna et al. (2019)	10	10	36.3 ± 7.5	36.2 ± 8.4	yes	memory	meta-d' - d'	3.5
Faivre et al. (2019)	21	20	38.8 ± 8.77	42.6 ± 3.35	yes	perception	M-ratio	7.0
Gaweda et al. (2019)	33	33	35.82 ± 11.22	41.33 ± 14.8	no	perception	false perception	5.0
Davies et al. (2018)	31	18	26.16 ± 5.69	24.06 ± 4.87	yes	perception	M-ratio	7.0
Gawęda et al. (2018)	25	33	20.36 ± 2.16	20.27 ± 2.11	no	memory	KCI	5.7
Mayer et al. (2018)	24	24	40.67 ± 11.65	38.88 ± 9.66	no	memory	false memories	5.5
Pinkham et al. (2018)	31	32	35.65 ± 7.52	35.41 ± 7.07	no	others	AUROC2	3.5
Charles et al. (2017)	13	13	28.8 ± 5.9	28.8 ± 4.7	no	perception	meta-d'	5.0
Powers et al. (2017)	15	15	39.4 ± 13.47	46.07 ± 12.96	yes	perception	M-ratio	5.5
Powers et al. (2017)	14	15	38.29 ± 14.4	40.53 ± 13.04	yes	perception	M-ratio	5.5
Balzan et al. (2016)	25	50	39.96 ± 10.04	42.8 ± 15.46	no	memory	confidence in errors	4.0
Eifler et al. (2015)	29	25	37.22 ± 9.68	38.12 ± 10.72	no	memory	confidence gap	4.0
Eisenacher et al. (2015)	21	38	26.52 ± 5.57	25.08 ± 6.55	no	memory	confidence gap	3.5
Gaweda (2015)	13	17	22.08 ± 1.93	23.59 ± 1.87	no	memory	KCI	4.0
Gaweda (2015)	10	10	53.9 ± 3.21	57.4 ± 3.72	no	memory	KCI	4.0
Akdogan et al. (2014)	23	23	38 ± 8	37.5 ± 7.2	no	memory	gamma correlation	3.3
Mayer et al. (2014)	31	28	40.23 ± 9.1	37.89 ± 8.35	no	memory	false memories	4.0
Moritz et al. (2014)	55	45	38.22 ± 8.61	37.24 ± 13.93	no	perception	KCI	4.5
Gaweda et al. (2013)	54	34	35.17 ± 10.43	33.21 ± 11.33	no	memory	KCI	4.5
Peters et al. (2013)	27	24	37.96 ± 12.86	34.21 ± 11.33	no	memory	KCI	4.5
Gaweda et al. (2012)	32	32	32.81 ± 8.36	31.78 ± 11.67	no	memory	KCI	4.7
Kother et al. (2012)	76	30	34.26 ± 11.41	32.97 ± 10.88	no	others	KCI	4.0
Mayer et al. (2012)	28	29	38.32 ± 9.29	37.28 ± 8.41	no	memory	false memories	5.0
Metcalfe et al. (2012)	22	20	42.3 ± 11.1	38.1 ± 11.3	no	others	correlation perf-confidence	5.0
Moritz et al. (2012)	23	29	35.17 ± 11.12	34.24 ± 16.14	no	others	KCI	4.5
Peters et al. (2012)	47	47	35.72 ± 11.63	36.87 ± 11.89	no	memory	KCI	5.0
Bhatt et al. (2010)	25	20	47 ± 8.65	44.5 ± 8.81	no	memory	KCI	2.0
Kim et al. (2010)	12	13	40.2 ± 10.23	40.4 ± 9.34	no	memory	false memories	4.5
Moritz et al. (2008)	68	25	33.94 ± 10.45	32.04 ± 10.23	no	memory	confidence gap	4.0
Kircher et al. (2007)	27	19	32.8 ± 11.4	33.4 ± 13.4	no	memory	correlation perf-confidence	5.0
Peters et al. (2007)	23	20	36.3 ± 13.13	35.2 ± 9.71	no	memory	confidence gap	6.0
Moritz et al. (2006a)	31	61	33.77 ± 9.9	31.05 ± 8.75	no	memory	confidence gap	4.5
Moritz et al. (2006b)	30	15	24.73 ± 8.73	24.8 ± 8.99	no	memory	confidence gap	3.5
Moritz et al. (2006c)	35	34	36.29 ± 11.34	34.29 ± 11.38	no	memory	KCI	4.0
Moritz et al. (2005)	30	17	37.3 ± 10.16	37.67 ± 12.47	no	memory	KCI	4.5
Moritz et al. (2004)	20	20	33.2 ± 9.28	29.2 ± 12.51	no	memory	KCI	4.5
Moritz et al. (2003)	30	22	31.08 ± 8.3	27 ± 10.7	no	memory	confidence gap	4.0
Bacon et al. (2001)	19	19	31.7 ± 8.4	30.7 ± 8.2	no	memory	confidence gap	3.5

3.3. Metacognitive deficit in studies controlling for first-order performance

Our main hypothesis stipulated that metacognitive deficits would be decreased in studies controlling for first-order performance. The following analysis was further justified by a heterogeneity analysis which produced a significant Q-statistic (124.1, $df = 41$, $p < .001$) and a high amount of heterogeneity (I^2 statistic 0.66, 95 % CI [0.54, 0.76]), suggesting this moderator analysis was appropriate. Because metacognitive performance is known to depend on first-order performance (Maniscalco and Lau, 2012), and because the latter differed between groups ($g = -0.64$, 95 % CrI [-0.77, -0.52], $BF_{10} = 2.06 \times 10^{10}$), we sought to assess whether metacognitive deficits could stem from cognitive impairments that are well documented in schizophrenia (Gopal and Variend, 2005; Heinrichs and Zakzanis, 1998). Distinguishing studies controlling for first-order performance ($N = 7$) from those which did not ($N = 35$) revealed a significant moderation effect ($Q_{\text{between}} = 6.82$, $df = 1$, $p = 0.009$). Thus, we assessed the influence of performance-matching with a model M2, identical to M1 including performance-matching as an additional binary predictor. The sub-group of non-controlled studies had an overall metacognitive deficit of magnitude $g = -0.63$, 95 % CrI [-0.78, -0.49], which was reduced to $g = -0.23$, 95 % CrI [-0.60, 0.16] in the sub-group of controlled studies (Fig. 3A). Accordingly, the evidence ratio supporting our directional

hypothesis that controlling for first-order performance decreases the magnitude of the metacognitive deficit was very strong ($BF_{10} = 51$) (Fig. 3B). Comparison against the null hypothesis among controlled studies revealed inconclusive evidence in favor of a metacognitive deficit in schizophrenia ($BF_{01} = 2.2$). Finally, a positive correlation between cognitive and metacognitive deficits was found among non-controlled studies (SI). Sub-group analyses reduced heterogeneity which however remained significant (SI).

3.4. Metacognitive deficits across cognitive domains

Next, in line with our pre-registered analysis plan and a significant moderation effect of cognitive domains ($Q_{\text{between}} = 38.5$, $df = 2$, $p < .001$), we assessed how metacognitive deficits varied across cognitive domains (i.e., perception, memory, others). A subgroup analysis revealed the largest metacognitive deficit among memory studies, compared to perception and others. Mean value of the metacognitive deficit in the memory domain ($g = -0.74$, 95 % CrI [-0.89, -0.58], $BF_{10} = 7.74 \times 10^{156}$) was twice higher than in the perception domain ($g = -0.33$, 95 % CrI [-0.63, -0.04], $BF_{10} = 2.16$), and three times higher than in other domains ($g = -0.26$, 95 % CrI [-0.62, 0.10], $BF_{10} = 0.40$; see Figs. 4 and SI). Sub-group analyses reduced heterogeneity which however remained significant (SI).

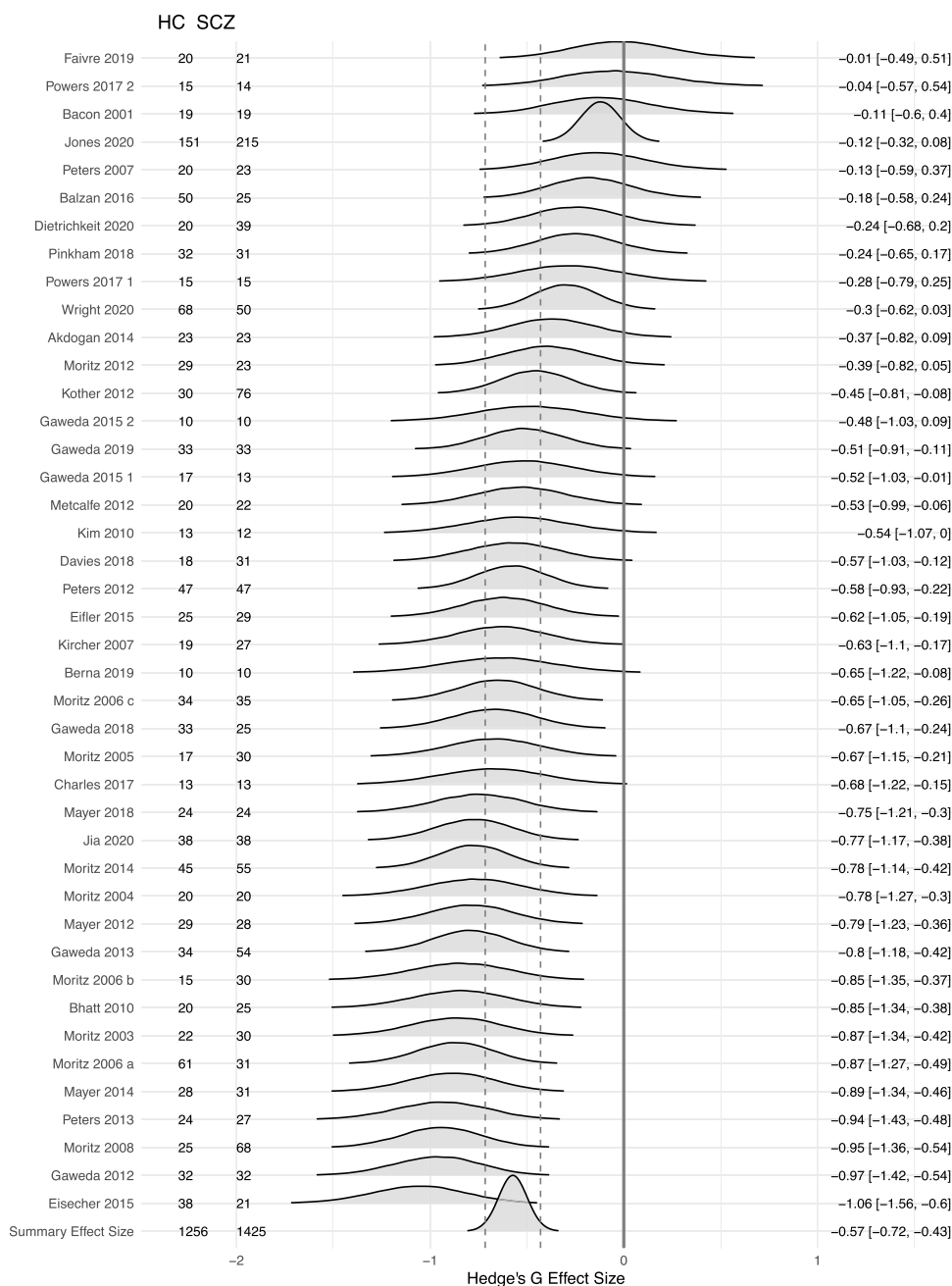


Fig. 2. Forest plot of the metacognitive deficit in schizophrenia. Left: Authors with publication year and sample sizes; Middle: posterior distribution of the effect size; Right: mean and 95 % CrI of the posterior distribution. The summary effect size is displayed on the last row: the solid vertical grey line is centred on zero (i.e., equivalent metacognitive performance between groups), and the dashed vertical lines depict the boundaries of the 95 % CrI.

3.5. Meta-regression analyses

Finally, we performed further meta-regressions to explore how metacognitive deficits co-varied with the severity of positive and negative symptoms (PANSS equivalent scores) and antipsychotic dosage (chlorpromazine equivalent), with a prior of mean 0 and SD = 1. We had pre-registered the hypothesis of a negative correlation between meta-performance and PANSS positive scores. However, meta-regression analyses provided inconclusive evidence regarding the influence of symptom severity on the metacognitive deficit: $BF_{10} = 0.88$ for PANSS total scores (N = 35), $BF_{10} = 0.91$ for PANSS positive scores (N = 32) and $BF_{10} = 0.75$ for PANSS negative scores (N = 33) (see SI, Fig. S6).

Similarly, we found no evidence for an association between

metacognitive performance and pharmacological treatment (N = 20), with an evidence ratio ($BF_{10} = 0.99$) suggesting inconclusive data (see SI, Fig. S6).

3.6. Risk of bias in selected studies

A quality evaluation using the Newcastle-Ottawa Scale suggested that about half the studies had a relatively high risk of bias with scores < 5/9 (SI and (Luchini et al., 2017)). The shape of the funnel plot revealed no asymmetry (Egger's test: $z = -0.07$, $p = 0.94$; Figs. 5A and SI), suggesting no clear publication bias. Plus, testing the right-skewness of the P-curve (Fig. 5B) with Stouffer's method revealed that both the half (p 's < 0.025) and full p-curves (p 's < 0.05) were right-skewed with $p < .001$,

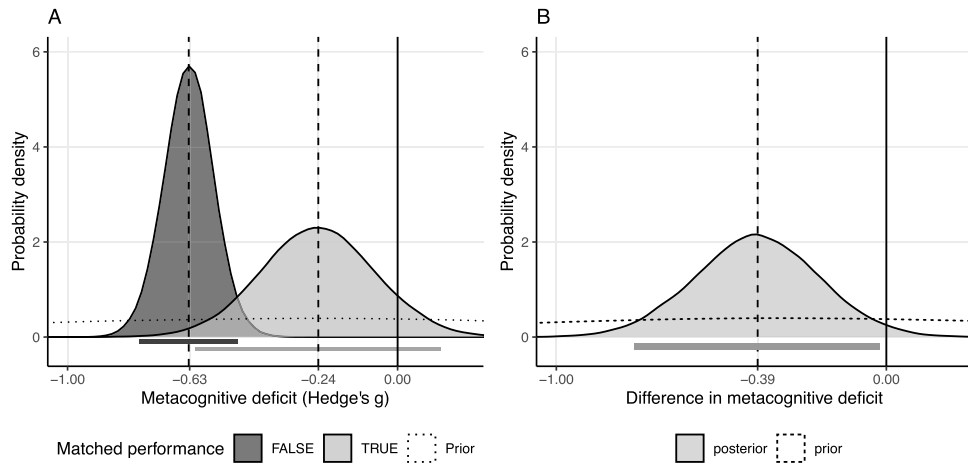


Fig. 3. A: Posterior distributions of the metacognitive deficit. Dark gray: non-controlled first-order performance (n = 35), Light gray: controlled first-order performance (n = 7). B: Posterior distribution of the difference in effect size between studies which did or did not control for first-order performance. In both panels, dotted lines represent the prior distributions, vertical dashed lines the mean posterior values, and the horizontal bars the 95 % CrI.

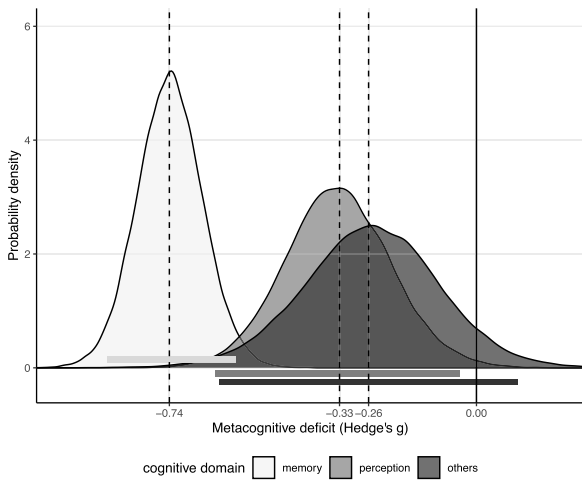


Fig. 4. Posterior distributions of the metacognitive deficit (Hedge's g effect size) according to each cognitive domain. The vertical dashed lines represent mean values and the horizontal bars the 95 % CrI.

suggesting that our study sample was not contaminated by p-hacking.

4. Discussion

The present meta-analysis based on 42 studies and 2681 individuals aimed at synthesizing the literature on the metacognitive abilities among individuals with schizophrenia. At first sight, our findings confirmed a deficit in metacognitive abilities in schizophrenia, but with high heterogeneity. The effect was of medium magnitude, which is smaller than the large effects reported in prior meta-analyses regarding cognitive impairments (Schaefer et al., 2013). The leave-one-out sensitivity analysis confirmed this effect was robust to outliers. We found several sources for heterogeneity that we describe hereafter.

4.1. Main result

Because patients' first-order cognitive deficits risked to artificially inflate metacognitive deficits (Galvin et al., 2003), our main hypothesis was that metacognitive deficits would be reduced in studies equating first-order performance between groups. Results indicated strong evidence in favor of our hypothesis, as metacognitive deficits were twice smaller in studies controlling for first-order performance, most of them

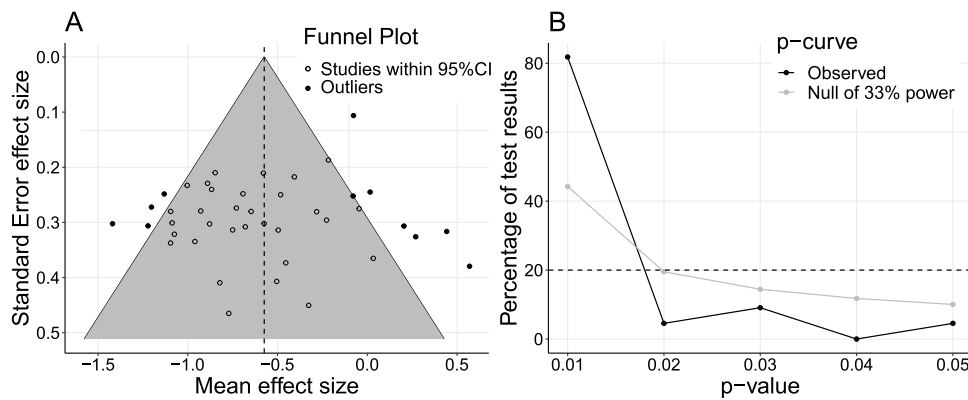


Fig. 5. A: Funnel plot centered on the overall effect size. The vertical dashed line represents the global metacognitive deficit. The gray area represents the 95 % CI of the overall effect size. Each dot represents a study, full dots represent outliers. B: Observed p-curve (black) and theoretical p-curve expected for low-powered (33 %) studies (gray). Horizontal dashed line: Expected uniform distribution for null effects.

concerning the perceptual domain. In this subset of studies, assessing the presence of a metacognitive deficit revealed inconclusive evidence. By contrast, a correlation between cognitive and metacognitive deficits was found among non-controlled studies, indicating that first-order performance is a critical moderator of metacognition which should be controlled for when assessing metacognitive deficits in schizophrenia.

4.2. Metacognitive deficits across cognitive domains

We also explored possible differences in metacognitive deficits across cognitive domains (perception, memory, others), and found the most prominent deficits among memory studies. As such, this result is not sufficient to confirm the presence of a specific meta-memory deficit in schizophrenia, as all the memory studies but one did not control for differences in first-order performance between groups. Given that the magnitude of the meta-memory deficit we found is lower than the one of episodic verbal memory (range between -1.53 and -1.11 SD) (Gopal and Variend, 2005; Heinrichs and Zakzanis, 1998; Schaefer et al., 2013), arbitrating between the existence of a specific meta-memory deficit or the side effect of a non-controlled first-order factor will require the development of more robust experimental protocols. Of note, this meta-analysis did not examine the literature based on judgments of learning or feeling of knowing, which may reveal different patterns of results (Souhay et al., 2006).

4.3. Unexplained heterogeneity

Despite moderation analyses, heterogeneity remained high even after clustering studies according to performance matching and cognitive domains. This heterogeneity may be explained by the different diagnoses included in our selection of studies. The category of first episode of psychosis may be particularly problematic, as it included variable diagnoses (mania with psychosis, bipolar disorder with psychosis, depression with psychosis, delusional disorder, substance-induced psychotic disorder, psychosis not otherwise specified, acute and transient psychotic disorder, brief psychotic disorder). Heterogeneity may also come from the use of idiosyncratic first-order tasks (e.g., memory performance was quantified using recognition, source memory and spatial delayed response tasks) and confidence scales (e.g., ordinal vs. continuous scales, full vs. half scales, etc.). Finally, one should consider that the same research group co-contributed a large number of selected studies, with metacognitive deficits of larger magnitudes than the one estimated by other authors (SI). With this in mind, it will be important to use more systematic paradigms among more diverse study samples in the future.

4.4. Perspectives

Additional analyses evaluating how metacognitive deficits varied as a function of clinical scores (PANSS total, positive, negative) and antipsychotic dosage (chlorpromazine equivalent) revealed inconclusive evidence for correlation in each case. As we had no access to individual data, correlations were based on summary statistics extracted from each experimental group, which is suboptimal. As with all meta-analyses, our findings are shaped and limited by selection and analytical methods, and the information made available to researchers in the studies selected for review. Thus, they may be contradicted by other relevant studies referenced in non-searched databases. The scarcity of data prevented us from running planned analyses regarding the link between metacognitive performance and clinical/cognitive insight. Establishing this link is of crucial importance to validate confidence calibration as a valid empirical construct for clinical practice, and to refine current strategies to improve insight in schizophrenia. We encourage authors to share anonymized individual data similar to what is done for healthy controls (Rahnev et al., 2020) on a dedicated repository (<https://osf.io/cfm5d/>). Our findings point to several areas for future research. First, few studies

included in this meta-analysis measured mood, despite it being an important determinant of metacognition (Lin et al., 2019), with a bias toward underconfidence in depression (Hoven et al., 2019). No study included in this meta-analysis focused on the metacognition of executive function. Further studies are needed because meta-executive functions have been linked with attenuated psychosis syndrome (Koren et al., 2019). Further studies should also investigate whether metacognitive abilities are associated with insight, relapse and psychosocial functioning before using it in clinical settings.

5. Conclusion

This is the first meta-analysis to examine metacognitive deficits based on confidence judgments in schizophrenia. Our results show that this deficit is inflated due to non-equated first-order performance, and varies across cognitive domains. Importantly, metacognitive deficits may also be overestimated in other psychiatric and neurological conditions involving cognitive impairments. Efforts should be made to develop experimental protocols accounting for lower first-order performance in schizophrenia before including the accuracy of confidence judgments as a cognitive dimension in neuropsychological batteries for clinical applications.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Author contributions

MR, PS, MP, PR and NF developed the study concept and contributed to the study design. Data selection and extraction were performed by MR and PS. MR, LN and NF analyzed data. MR and NF drafted the paper; all authors provided critical revisions and approved the final version of the paper for submission.

Data availability statement

Bibliographic data and analyses scripts are publicly available: https://gitlab.com/nfaivre/meta_analysis_scz_public.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

NF has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 803122).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.neubiorev.2021.03.017>.

References

- David, A.S., Bedford, N., Wiffen, B., Gilleen, J., 2012. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philos. Trans. Biol. Sci.* 367 (1594), 1379–1390. <https://doi.org/10.1098/rstb.2012.0002>.
- Akdogan, E., Izaute, M., Bacon, E., 2014. Preserved strategic grain-size regulation in memory reporting in patients with schizophrenia. *Biol. Psychiatry* 76 (2), 154–159. <https://doi.org/10.1016/j.biopsych.2013.09.004>.

- Arnon-Ribenfeld, N., Hasson-Ohayon, I., Lavidor, M., Atzil-Slonim, D., Lysaker, P.H., 2017. The association between metacognitive abilities and outcome measures among people with schizophrenia: a meta-analysis. *Eur. Psychiatry* 46, 33–41. <https://doi.org/10.1016/j.eurpsy.2017.08.002>.
- Bacon, E., Danion, J.-M., Kauffmann-Muller, F., Bruant, A., 2001. Consciousness in schizophrenia: a metacognitive approach to semantic memory. *Conscious. Cogn.* 10 (4), 473–484. <https://doi.org/10.1006/cog.2001.0519>.
- Balzan, R.P., Woodward, T.S., Delfabbro, P., Moritz, S., 2016. Overconfidence across the psychosis continuum: a calibration approach. *Cogn. Neuropsychiatry* 21 (6), 510–524. <https://doi.org/10.1080/13546805.2016.1240072>.
- Berna, F., Zou, F., Danion, J.-M., Kwok, S.C., 2019. Overconfidence in false autobiographical memories in patients with schizophrenia. *Psychiatry Res.* 279, 374–375. <https://doi.org/10.1016/j.psychres.2018.12.063>.
- Bhatt, R., Laws, K.R., McKenna, P.J., 2010. False memory in schizophrenia patients with and without delusions. *Psychiatry Res.* 178 (2), 260–265. <https://doi.org/10.1016/j.psychres.2009.02.006>.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (Eds.), 2010. *Introduction to Meta-Analysis*. Wiley (Reprinted).
- Bürkner, P.-C., 2017. Brms: an R package for Bayesian multilevel models using stan. *J. Stat. Softw.* 80 (1) <https://doi.org/10.18637/jss.v080.i01>.
- Card, N.A., Little, T.D., 2016. *Applied Meta-analysis for Social Science Research*. Paperback edition. The Guilford Press.
- Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. *Stan*: a probabilistic programming language. *J. Stat. Softw.* 76 (1) <https://doi.org/10.18637/jss.v076.i01>.
- Charles, L., Gaillard, R., Amado, I., Krebs, M.-O., Bendjema, N., Dehaene, S., 2017. Conscious and unconscious performance monitoring: evidence from patients with schizophrenia. *NeuroImage* 144, 153–163. <https://doi.org/10.1016/j.neuroimage.2016.09.056>.
- Davies, G., Greenwood, K., 2020. A meta-analytic review of the relationship between neurocognition, metacognition and functional outcome in schizophrenia. *J. Ment. Health* 29 (5), 496–505. <https://doi.org/10.1080/09638237.2018.1521930>.
- 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. *Cogn. Neuropsychiatry* 23 (3), 165–179. <https://doi.org/10.1080/13546805.2018.1444597>.
- Dietrichkeit, M., Grzella, K., Nagel, M., Moritz, S., 2020. Using virtual reality to explore differences in memory biases and cognitive insight in people with psychosis and healthy controls. *Psychiatry Res.* 285 <https://doi.org/10.1016/j.psychres.2020.112787>, 112787–112787.
- 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 Res.* 225 (3), 596–603. <https://doi.org/10.1016/j.psychres.2014.11.040>.
- Eisenacher, S., Rausch, F., Ainsler, 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. *Psychol. Med.* 45 (15), 3329–3340. <https://doi.org/10.1017/S0033291715001373>.
- Faivre, N., Roger, M., Pereira, M., de Gardelle, V., Vergnaud, J.-C., Passerieux, C., Roux, P., 2019. Confidence in perceptual decision-making is preserved in schizophrenia [Preprint]. *Psychiatry Clin. Psychol.* <https://doi.org/10.1101/2019.12.15.19014969>.
- Fleming, S.M., Lau, H.C., 2014. How to measure metacognition. *Front. Hum. Neurosci.* 8. <https://doi.org/10.3389/fnhum.2014.00443>.
- Galvin, S.J., Podd, J.V., Drga, V., Whitmore, J., 2003. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.* 10 (4), 843–876. <https://doi.org/10.3758/BF03196546>.
- Gawęda, L., 2015. Does ageing affect source monitoring and cognitive confidence in schizophrenia? Preliminary results. *Psychiatry Res.* 228 (3), 936–940. <https://doi.org/10.1016/j.psychres.2015.06.024>.
- Gawęda, L., Moritz, S., 2019. The role of expectancies and emotional load in false auditory perceptions among patients with schizophrenia spectrum disorders. *Eur. Arch. Psychiatry Clin. Neurosci.* <https://doi.org/10.1007/s00406-019-01065-2>.
- Gawęda, L., Moritz, S., Kokoszka, A., 2012. Impaired discrimination between imagined and performed actions in schizophrenia. *Psychiatry Res.* 195 (1–2), 1–8. <https://doi.org/10.1016/j.psychres.2011.07.035>.
- Gawęda, L., Woodward, T.S., Moritz, S., Kokoszka, A., 2013. Impaired action self-monitoring in schizophrenia patients with auditory hallucinations. *Schizophr. Res.* 144 (1–3), 72–79. <https://doi.org/10.1016/j.schres.2012.12.003>.
- Gawęda, L., 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. *Eur. Psychiatry* 47, 67–75. <https://doi.org/10.1016/j.eurpsy.2017.09.003>.
- Gopal, Y.V., Variend, H., 2005. First-episode schizophrenia: review of cognitive deficits and cognitive remediation. *Adv. Psychiatr. Treat.* 11 (1), 38–44. <https://doi.org/10.1192/apt.11.1.38>.
- Harrer, M., Cuijpers, P., Ebert, D., 2019. Doing Meta-Analysis in R. <https://doi.org/10.5281/ZENODO.2551803>.
- Hasson-Ohayon, I., Goldzweig, G., Lavi-Rotenberg, A., Luther, L., Lysaker, P.H., 2018. The centrality of cognitive symptoms and metacognition within the interacting network of symptoms, neurocognition, social cognition and metacognition in schizophrenia. *Schizophr. Res.* 202, 260–266. <https://doi.org/10.1016/j.schres.2018.07.007>.
- Heinrichs, R.W., Zakzanis, K.K., 1998. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology* 12 (3), 426–445. <https://doi.org/10.1037/0894-4105.12.3.426>.
- Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21 (11), 1539–1558. <https://doi.org/10.1002/sim.1186>.
- 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. *Transl. Psychiatry* 9 (1), 268. <https://doi.org/10.1038/s41398-019-0602-7>.
- Huedo-Medina, T.B., Sánchez-Meca, J., Marín-Martínez, F., Botella, J., 2006. Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychol. Methods* 11 (2), 193–206. <https://doi.org/10.1037/1082-989X.11.2.193>.
- Jia, W., Zhu, H., Ni, Y., Su, J., Xu, R., Jia, H., Wan, X., 2020. Disruptions of frontoparietal control network and default mode network linking the metacognitive deficits with clinical symptoms in schizophrenia. *Hum. Brain Mapp.* 41 (6), 1445–1458. <https://doi.org/10.1002/hbm.24887>.
- Jones, M.T., Deckler, E., Laurrari, C., Jarskog, L.F., Penn, D.L., Pinkham, A.E., Harvey, P.D., 2020. Confidence, performance, and accuracy of self-assessment of social cognition: a comparison of schizophrenia patients and healthy controls. *Schizophr. Res. Cogn.* 19 <https://doi.org/10.1016/j.scog.2019.01.002>, 2–2.
- Kim, J., Matthews, N.L., Park, S., 2010. An event-related fMRI study of phonological verbal working memory in schizophrenia. *PLoS One* 5 (8). <https://doi.org/10.1371/journal.pone.0012068>.
- Kircher, T.T.J., Koch, K., Stottmeister, F., Durst, V., 2007. Metacognition and reflexivity in patients with schizophrenia. *Psychopathology* 40 (4), 254–260. <https://doi.org/10.1159/000101730>.
- Koren, D., Seidman, L.J., Goldsmith, M., Harvey, P.D., 2006. Real-world cognitive–and metacognitive–dysfunction in schizophrenia: a new approach for measuring (and remediate) more ‘Right stuff’. *Schizophr. Bull.* 32 (2), 310–326. <https://doi.org/10.1093/schbul/sbj035>.
- Koren, D., Scheyer, R., Stern, Y., Adres, M., Reznik, N., Apter, A., Seidman, L.J., 2019. Metacognition strengthens the association between neurocognition and attenuated psychosis syndrome: preliminary evidence from a pilot study among treatment-seeking versus healthy adolescents. *Schizophr. Res.* 210, 207–214. <https://doi.org/10.1016/j.schres.2018.12.036>.
- Kother, 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 Res.* 196 (1), 1–8. <https://doi.org/10.1016/j.psychres.2012.03.004>.
- Levitt, H., 1971. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* 49 (2B), 467–477. <https://doi.org/10.1121/1.1912375>.
- Lin, X., Lu, D., Huang, Z., Chen, W., Luo, X., Zhu, Y., 2019. The associations between subjective and objective cognitive functioning across manic or hypomanic, depressed, and euthymic states in Chinese bipolar patients. *J. Affect. Disord.* 249, 73–81. <https://doi.org/10.1016/j.jad.2019.02.025>.
- Luchini, C., Stubbs, B., Solmi, M., Veronesi, N., 2017. Assessing the quality of studies in meta-analyses: advantages and limitations of the Newcastle Ottawa Scale. *World J. Metaanal.* 5 (4), 80. <https://doi.org/10.13105/wjma.v5.i4.80>.
- Maniscalco, B., Lau, H., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* 21 (1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>.
- Mayer, J.S., Park, S., 2012. Working memory encoding and false memory in schizophrenia and bipolar disorder in a spatial delayed response task. *J. Abnorm. Psychol.* 121 (3), 784–794. <https://doi.org/10.1037/a0028836>.
- Mayer, J.S., Kim, J., Park, S., 2014. Failure to benefit from target novelty during encoding contributes to working memory deficits in schizophrenia. *Cogn. Neuropsychiatry* 19 (3), 268–279. <https://doi.org/10.1080/13546805.2013.854199>.
- Mayer, Jutta S., Stablein, M., Oertel-Knochel, V., Fiebach, C.J., 2018. Functional dissociation of confident and not-confident errors in the spatial delayed response task demonstrates impairments in working memory encoding and maintenance in schizophrenia. *Front. Psychiatry* 9. <https://doi.org/10.3389/fpsy.2018.00202>, 202–202.
- McLeod, H.J., Gumley, A.I., MacBeth, A., Schwannauer, M., Lysaker, P.H., 2014. Metacognitive functioning predicts positive and negative symptoms over 12 months in first episode psychosis. *J. Psychiatr. Res.* 54, 109–115. <https://doi.org/10.1016/j.jpsychres.2014.03.018>.
- Metcalf, J., van Snellenberg, J.X., DeRosse, P., Balsam, P., Malhotra, A.K., 2012. Judgements of agency in schizophrenia: an impairment in autoecognitive metacognition. *Philos. Trans. Biol. Sci.* 367 (1594), 1391–1400. <https://doi.org/10.1098/rstb.2012.0006>.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6 (7) <https://doi.org/10.1371/journal.pmed.1000097> e1000097.
- Moritz, S., Woodward, T.S., 2006a. The contribution of metamemory deficits to schizophrenia. *J. Abnorm. Psychol.* 115 (1), 15–25. <https://doi.org/10.1037/0021-843X.115.1.15>.
- Moritz, S., Woodward, T.S., Ruff, C.C., 2003. Source monitoring and memory confidence in schizophrenia. *Psychol. Med.* 33 (1), 131–139. <https://doi.org/10.1017/S0033291702006852>.
- Moritz, S., Woodward, T.S., Cuttler, C., Whitman, J.C., Watson, J.M., 2004. False memories in schizophrenia. *Neuropsychology* 18 (2), 276–283. <https://doi.org/10.1037/0894-4105.18.2.276>.
- Moritz, S., Woodward, T.S., Whitman, J.C., Cuttler, C., 2005. Confidence in errors as a possible basis for delusions in schizophrenia. *J. Nerv. Ment. Dis.* 193 (1), 9–16. <https://doi.org/10.1097/01.nmd.0000149213.10692.00>.

- Moritz, S., Woodward, T.S., Chen, E., 2006b. Investigation of metamemory dysfunctions in first-episode schizophrenia. *Schizophr. Res.* 81 (2–3), 247–252. <https://doi.org/10.1016/j.schres.2005.09.004>.
- Moritz, S., Woodward, T.S., Rodriguez-Raecke, R., 2006c. Patients with schizophrenia do not produce more false memories than controls but are more confident in them. *Psychol. Med.* 36 (5), 659–667. <https://doi.org/10.1017/S0033291706007252>.
- Moritz, S., Woodward, T.S., Jelinek, L., Klinge, R., 2008. Memory and metamemory in schizophrenia: a liberal acceptance account of psychosis. *Psychol. Med.* 38 (6), 825–832. <https://doi.org/10.1017/S0033291707002553>.
- Moritz, S., Woznica, A., Andreou, C., Köther, U., 2012. Response confidence for emotion perception in schizophrenia using a Continuous Facial Sequence Task. *Psychiatry Res.* 200 (2–3), 202–207. <https://doi.org/10.1016/j.psychres.2012.07.007>.
- 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. *Schizophr. Res. Cogn.* 1 (4), 165–170. <https://doi.org/10.1016/j.scog.2014.09.003>.
- 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. *Cogn. Neuropsychiatry* 12 (5), 391–411. <https://doi.org/10.1080/13546800701470145>.
- Peters, Maarten J.V., Engel, M., Hauschildt, M., Moritz, S., Jelinek, L., Otgaar, H., 2012. Investigating the corrective effect of forewarning on memory and meta-memory deficits in schizophrenia patients. *J. Exp. Psychopathol.* 3 (4), 673–687. <https://doi.org/10.5127/jep.022011>.
- Peters, M.J.V., Hauschildt, M., Moritz, S., Jelinek, L., 2013. Impact of emotionality on memory and meta-memory in schizophrenia using video sequences. *J. Behav. Ther. Exp. Psychiatry* 44 (1), 77–83. <https://doi.org/10.1016/j.jbtep.2012.07.003>.
- Pinkham, A.E., Klein, H.S., Hardaway, G.B., Kemp, K.C., Harvey, P.D., 2018. Neural correlates of social cognitive introspective accuracy in schizophrenia. *Schizophr. Res.* 202, 166–172. <https://doi.org/10.1016/j.schres.2018.07.001>.
- Powers, A.R., Mathys, C., Corlett, P.R., 2017. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science* 357 (6351), 596–600. <https://doi.org/10.1126/science.aan3458>.
- Rahnev, D., Desender, K., Lee, A.L.F., Adler, W.T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L.Y., Balci, F., Bang, J.W., Bègue, I., Birney, D.P., Brady, T.F., Calder-Travis, J., Chetverikov, A., Clark, T.K., Davranche, K., Denison, R.N., Dildine, T.C., et al., 2020. The confidence database. *Nat. Hum. Behav.* 4 (3), 317–325. <https://doi.org/10.1038/s41562-019-0813-1>.
- Schaefer, J., Giangrande, E., Weinberger, D.R., Dickinson, D., 2013. The global cognitive impairment in schizophrenia: consistent over decades and around the world. *Schizophr. Res.* 150 (1), 42–50. <https://doi.org/10.1016/j.schres.2013.07.009>.
- Semerari, A., Carcione, A., Dimaggio, G., Falcone, M., Nicolò, G., Procacci, M., Alleva, G., 2003. How to evaluate metacognitive functioning in psychotherapy? The metacognition assessment scale and its applications: assessing Metacognitive Functions in Psychotherapy. *Clin. Psychol. Psychother.* 10 (4), 238–261. <https://doi.org/10.1002/cpp.362>.
- Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* 143 (2), 534–547. <https://doi.org/10.1037/a0033242>.
- Souchay, C., Bacon, E., Danion, J.M., 2006. Metamemory in schizophrenia: an exploration of the feeling-of-knowing state. *J. Clin. Exp. Neuropsychol.* 28 (5), 828–840. <https://doi.org/10.1080/13803390591000846>.
- Sterne, J.A.C., Harbord, R.M., 2004. Funnel plots in meta-analysis. *Stata J.* 4 (2), 127–141. <https://doi.org/10.1177/1536867X0400400204>.
- Vohs, J.L., Lysaker, P.H., Francis, M.M., Hamm, J., Buck, K.D., Olesek, K., Outcalt, J., Dimaggio, G., Leonhardt, B., Liffick, E., Mehdiyoun, N., Breier, A., 2014. Metacognition, social cognition, and symptoms in patients with first episode and prolonged psychoses. *Schizophr. Res.* 153 (1–3), 54–59. <https://doi.org/10.1016/j.schres.2014.01.012>.
- Wright, A., Nelson, B., Fowler, D., Greenwood, K., 2020. Perceptual biases and metacognition and their association with anomalous self experiences in first episode psychosis. *Conscious. Cogn.* 77 <https://doi.org/10.1016/j.concog.2019.102847>, 102847–102847.

2. Assessment of metaperceptual and metamemory abilities among individuals with schizophrenia

This study was a follow-up to the meta-analysis included above. Since our meta-analysis revealed that the most important alteration of metacognition was in the memory domain, but that memory performance was not controlled for in almost all included memory studies, it was still unclear if metamemory was specifically impaired among patients with schizophrenia. For this reason, we aimed at developing a memory task that controlled experimentally for memory performance. This protocol had many versions and we ran several pilot studies. We wanted to apply a staircase procedure on memory trials, so we first designed an experiment where a fixed number of four pairs of human face stimuli were encoded at the beginning of each trial, and where participants had to recognize the seen pair among two pairs of human faces. The difficulty of each trial depended on the sequential position of the target pair during the encoding phase: the first pairs memorized during encoding were considered harder to recognize, compared to the last memorized pairs. Unfortunately, this staircase procedure was not sensitive enough: the first level of difficulty (i.e. target pair = last pair seen during encoding) was too easy, and the second level was already too difficult. This staircase oscillated only between two difficulty levels, so we abandoned the staircase idea and opted instead for a statistical control of first-order performance. Then, we conceived an easier task based on a detection protocol (instead of 2AFC), showing only one face stimulus at a time. Since we were also interested in testing the domain-generalty of metacognition, our protocol includes three tasks: a visual detection task and two memory tasks: familiarity and recollection.

Distinguishing between memory processes

In the memory literature, there is a long-standing debate about whether familiarity and recollection rely upon a unique common process or upon two distinct processes (Rotello 2017; Wixted 2007; Yonelinas 2002). Although reviewed neuropsychological evidence suggests that familiarity and recollection processes are distinguishable in space and time by the recruitment of distinct neural substrates at different time stamps (Moulin et al. 2013), whether they are functionally connected or not is still unclear. Relying on the SDT framework, this debate can be informed by comparing how well different models fit data obtained from protocols combining type 1 and type 2 tasks (Rotello 2017). Indeed, a single process account can be modeled either by an equal or an unequal variance SDT model, whereas a dual-process account can be modeled either by adding a high-threshold for recollection (Yonelinas, 1994), or by two separate decision axes with their own SDT

parameters, known as the “continuous dual-process model” (Wixted and Mickes 2010). Considerations of goodness of fit suggest that the unequal variance SDT model is the most adequate to account for data obtained from a wide range of recognition memory tasks, advocating for the single process account (Rotello 2017). However, considerations based on the latencies of intracranial signals recorded in 18 epileptic patients (Barbeau et al. 2008) - more than 2000 sites including memory and visual regions - during a *famous face recognition task* directly challenged the previous conclusion. Indeed, a common evoked response potential (ERP) at 240 ms (N240) after stimulus onset was found in multiple regions pertaining to the visual ventral pathway, plus the perirhinal cortex which is commonly known to support recognition processes, hence pertaining to memory processes. This synchronization of the perirhinal cortex with visual areas suggested a functional coupling supporting a perceptivo-mnesic process. Interestingly, the hippocampus did not synchronize with these regions at this time point, suggesting that contextualized memory pertains to another functional network. These results are advocating for the dual process account of familiarity and recollection, and are further supported by ERP findings revealing distinct neural markers for familiarity and recollection: a mid-frontal signature of familiarity peaking 300-500 ms after stimulus onset, and a parietal signature of recollection manifesting at 400-800 ms after stimulus onset (for a review see Rugg and Curran 2007). From these considerations, we inferred a metacognitive architecture (Figure 18) that would predict better correlations of metacognitive performance between perception and familiarity domains compared to the cross-task correlations of metacognitive performance with the recollection task.

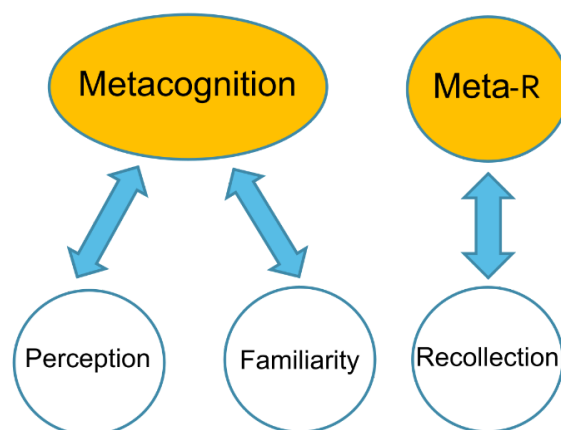


Figure 18. Hybrid architecture for metaperception and metamemory. Inferred metacognitive architecture relying on empirical results showing a continuity between perceptual and familiarity processes (domain-general confidence), while recollection was functionally separated (domain-specific confidence).

Status of the manuscript: Under revision in *Schizophrenia Bulletin*

Preprint available on medRxiv: <https://doi.org/10.1101/2023.03.28.23287851>

Reference: Martin Rouy, Michael Pereira, Pauline Saliou, Rémi Sanchez, Wassila el Mardi, Hanna Sebban, Eugénie Baqué, Childéric Dezier, Perrine Porte, Julia Micaux, Vincent de Gardelle, Pascal Mamassian, Chris J.A. Moulin, Clément Dondé, Paul Roux, and Nathan Faivre, (2023). Confidence in visual detection, familiarity and recollection judgements is preserved in schizophrenia spectrum disorder

Confidence in visual detection, familiarity and recollection judgements is preserved in schizophrenia spectrum disorder

Martin Rouy¹⁺, Michael Pereira¹, Pauline Saliou¹, Rémi Sanchez¹, Wassila el Mardi¹, Hanna Sebban¹, Eugénie Baqué², Childéric Dezier¹, Perrine Porte¹, Julia Micaux², Vincent de Gardelle³, Pascal Mamassian⁴, Chris J.A. Moulin¹, Clément Dondé^{5,6*}, Paul Roux^{2*}, Nathan Faivre^{1*}

1 Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

2 Centre Hospitalier de Versailles, Service Hospitalo-Universitaire de Psychiatrie d'Adultes et d'Addictologie, Le Chesnay; Université Paris-Saclay; Université de Versailles Saint-Quentin-En-Yvelines; DisAP-DevPsy-CESP, INSERM UMR1018, Villejuif, France

3 Centre d'Économie de la Sorbonne, CNRS and Paris School of Economics

4 Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, PSL University, CNRS

5 Univ. Grenoble Alpes, Inserm, U1216, Adult Psychiatry Department CHU Grenoble Alpes, Grenoble Institut Neurosciences, 38000 Grenoble, France

6 Adult Psychiatry Department, CH Alpes-Isère, F-38000 Saint-Egrève, France

* shared authorship

+ Corresponding author:

Martin Rouy

Laboratoire de Psychologie et Neurocognition

CNRS UMR 5105 UGA BSHM

1251 Avenue Centrale

38058 Grenoble Cedex 9

martinrouy03@gmail.com

Keywords: metacognition, meta-perception, meta-memory, confidence, schizophrenia spectrum disorder, psychosis

Author Contributions: MR, CM and NF developed the study concept. MR implemented experiments with the collaboration of MP and RS. Pilot data was collected and analyzed by MR, PS, RS, WM, HS. Patients and healthy controls were recruited by EB, JM, CDo, PR, CDe, PP. Data collection was performed by MR, CDe, PP and EB & JM. VdG and PM provided analytical tools. MR and NF analyzed data and drafted the paper; all authors provided critical revisions and approved the final version of the paper for submission. The authors declare no competing interests.

Word count: Abstract: 183; Text body: 4385

Preregistration (<https://osf.io/k4p79>), data and analysis scripts are publicly available

(https://gitlab.com/nfaivre/metaface_scz_public).

Abstract

An effective way to quantify metacognitive abilities is to ask participants to estimate their confidence in the accuracy of their response during a cognitive task. A recent meta-analysis¹ raised the issue that most assessments of metacognitive abilities in schizophrenia spectrum disorders may be confounded with cognitive deficits, which are known to be present in this population. Therefore, it remains unclear whether the reported metacognitive deficits are metacognitive in nature, or rather inherited from cognitive deficits. Arbitrating between these two possibilities requires equating task performance between experimental groups. Here, we aimed to characterize metacognitive performance among individuals with schizophrenia across three tasks (visual detection, familiarity, recollection) using a within-subject design, while controlling experimentally for intra-individual task performance and statistically for between-subject task performance. In line with our hypotheses, we found no metacognitive deficit for visual detection and familiarity judgements. While we expected metacognition for recollection to be specifically impaired among individuals with schizophrenia, we found evidence in favor of an absence of a deficit in that domain also. The clinical relevance of our findings is discussed in light of a hierarchical framework of metacognition.

Introduction

Confidence abnormalities in the form of overconfidence in errors in schizophrenia spectrum disorder have been documented in multiple cognitive domains, including memory, perception, and emotion recognition². Yet, the hierarchical level at which these abnormalities occur is still unclear. In line with the terminology proposed by Galvin and colleagues³, cognitive performance is referred to as first-order performance (i.e. how well one is able to detect or discriminate between probed stimuli), and metacognitive performance is referred to as second-order performance (i.e. how well one is able to discriminate between correct and incorrect responses). Properly quantifying metacognitive abilities requires controlling for variations of cognitive performance that are not metacognitive in nature³⁻⁵. This concern is of particular relevance in schizophrenia, where cognitive deficits are well documented^{6,7}. In a meta-analysis we recently conducted¹, metaperception was mostly preserved when first-order performance was controlled for. Yet, conclusions about the metamemory deficit could not be drawn in this meta-analysis since the medium to large effect size resulted from studies where memory performance was not equated between patients and healthy controls (except for one study⁸). In these conditions, metamemory deficits were likely to be confounded with memory deficits.

To compare metaperceptual and metamemory deficits in individuals with schizophrenia while controlling for perceptual and memory deficits, we developed a novel experimental paradigm including three randomly interleaved perceptual and memory tasks attempting to experimentally match first-order performance at the intra-individual level across tasks, and to statistically control for performance at the inter-individual level.

We preregistered our main predictions based on current knowledge regarding the cognitive architecture of perception and memory and their impairments in schizophrenia (see⁹ for a meta-analysis). Individuals with schizophrenia typically have preserved performance in familiarity judgements (i.e. decontextualized memory¹⁰) but impaired performance in recollection judgments (i.e. episodic/recollection memory necessitating multimodal integration via hippocampal activity¹¹), which may be explained by impaired hippocampus recruitment¹² and hippocampal atrophy¹³. Our main preregistered hypothesis was that metamemory was globally more impaired than metaperception, assuming that previous reports of deficits in metamemory were not only driven by deficits taking place at the first-order level. Furthermore, since familiarity can be considered a perceptual-mnemonic process storing decontextualized perceptual elements¹⁴, we hypothesized domain-generalty between perception and familiarity processes, and expected that meta-recollection would be specifically impaired. Besides this preregistered hypothesis, we explored the links between metacognitive performance and clinical traits such as positive, negative and disorganization syndromes.

Methods

The present design, hypotheses, and analyses were preregistered prior to data collection and analysis (<https://osf.io/k4p79>). Data and analysis scripts are available online (https://gitlab.com/nfaiivre/metaface_scz_public).

Participants

Following a preregistered open-ended sequential Bayes Factor design (see SI for details), we recruited 38 individuals with schizophrenia and 39 healthy control participants matched for age, sex, education level and premorbid IQ (see Table 1 for demographic and clinical information). After exclusions according to preregistered criteria (essentially due to ceiling performance, see SI for details), the analyses were conducted on a sample of 34 individuals with schizophrenia and 36 healthy controls. Two licensed psychiatrists (CD and PR) confirmed the diagnoses in the schizophrenia group according to the DSM-V criteria for schizophrenia (details about the recruitment procedure are provided in SI).

Experimental design

A video description of each task is available online (https://gitlab.com/nfaivre/metaface_scz_public/-/tree/main/videos). All participants were naive to the purpose of the study, gave written informed consent in accordance with institutional guidelines and the Declaration of Helsinki, and received monetary compensation (10€ / h) except those participants under legal protection. The study was approved by the ethical committee *Sud Méditerranée II* on April the 3d 2020 (217 R01 MS1).

Stimuli

4000 copyright-free artificially generated faces were downloaded from the open platform <https://generated.photos>. Two independent observers screened the stimuli to exclude children's faces, unrealistic faces, and faces with salient features (e.g. sunglasses, hats). The remaining 1700 male and 1700 female adult faces were converted to grayscale and equalized in contrast and luminosity (SHINE Matlab toolbox¹⁵). Each face was presented against a visual background noise consisting of its phase-scrambled version. The background was colorized in blue or red (balanced for luminosity) to provide a contextual cue. Size and gaze position were kept identical across all stimuli.

Procedure

Memory tasks

The familiarity and recollection tasks shared the same timeline (Figure 1). Each trial started with an encoding phase consisting of four successive face stimuli presented during 400ms each (random combination of 2 male and 2 female faces) on a blue or red background (context), with a 500 ms inter-stimulus interval. To avoid learning effects and familiarity confounds, each face was presented only once throughout the whole experiment. Following the encoding phase, the test phase consisted in presenting a fifth face on a gray background, and asking a task-specific question. In familiarity trials, the participant was asked to indicate whether the face had already been seen (80% of the trials, to obtain a uniform distribution across "stimulus strength" levels, see next paragraph) or not (20% of the trials); in recollection trials the fifth face was always a seen face (i.e. a face presented during the encoding phase), and the participant was asked whether the context of this stimulus was blue (80% of the trials) or not (20% of the trials) during the encoding phase. Participants provided their answers with

a mouse click on “no” or “yes” buttons respectively displayed at the top left and top right of the screen.

The difficulty of the familiarity and recollection tasks was manipulated by changing the serial position of the target stimulus during the encoding phase. Accordingly, there were four levels of stimulus strength - ranging from 1 to 4 -, corresponding to each of the four faces displayed sequentially within the encoding phase (Figure 1). Because this variable corresponds to the temporal distance between the target stimulus and the test stimulus, we refer to it as a “lag”. For instance, if the target face was the first face displayed during the encoding sequence, then the temporal distance between the target and the test was maximal, and the trial was categorized as “lag 4”. On the contrary, if the target was the last face of the encoding phase, the temporal distance between the target and the test was minimal, and the trial was categorized as “lag 1”. A fifth lag-level “lag 0” was used to indicate catch trials (20% of the trials): i.e. new faces in familiarity trials, faces presented in the red context in recollection trials.

Visual detection task

Participants had to indicate whether a face was present (80% of the trials) or not (catch trials: 20%). The face could be presented at four contrast levels, chosen to match performances obtained in the memory tasks for each of the four lags (See SI Figure S2 D). A fifth level - stimulus strength 0 - was used to tag catch trials: trials where no face was presented (20% of the trials). As for memory trials, participants provided their answer with a mouse click on “no” or “yes” buttons displayed at the top left and top right of the screen.

Trial exclusions

A time limit of 6 seconds was set on all trials to avoid differences in response rates between patients and controls. When the time limit was reached, an error-like sound was produced along with a visual warning in red characters asking participants to respond quicker. Proportions of non-responses were comparable between individuals with schizophrenia (mean \pm SD: 2.91% \pm 3.60%) and controls (mean \pm SD: 2.32% \pm 7.66%, BF = 0.26). These non-response trials were excluded from our analyses.

Confidence rating

For all three tasks, participants were asked to provide confidence judgments. After each first-order response (i.e. responses given to the familiarity, recollection and visual detection tasks) participants were asked to report their subjective confidence regarding the correctness of their decision by moving a slider with the mouse on a visual analog scale (see Figure 1) ranging from 0% (“Sure incorrect”) to 100% (“Sure correct”). The initial position of the cursor for each trial corresponded to 50% confidence (“Not sure”).

Structure of the experiment

This protocol aimed to match intra-individual performance across familiarity, recollection, and visual detection tasks. Participants were asked to perform two sessions of one hour each. Session 1 allowed us to measure memory performance at four difficulty levels (according to

the variable “lag”, see *Memory Tasks*). We then matched perceptual performance to memory performance by determining four adequate contrast levels for the visual detection task for each participant (see SI for details). Thus, session 1 provided four levels of stimulus strength, i.e. 4 memory lags and 4 visual contrast levels, corresponding to matched performance for each participant. Based on these individual parameters, session 2 contained 10 blocks of 30 randomly interleaved trials (familiarity, recollection and visual detection task), totalizing 300 trials (100 trials per task), each followed by a confidence rating task. Task order and stimulus strength were randomized, so participants could not predict which task they were going to perform on each trial.

Importantly, this paradigm was designed to match first-order performance between tasks, which is convenient to compare metacognitive deficits across tasks. Although we also attempted to match first-order performance between groups, pilot experiments revealed this was not possible using adaptive staircases. Therefore, differences in task performance between groups were accounted for at the statistical level using the confidence efficiency metric¹⁶, taking advantage of our design with different levels of difficulty.

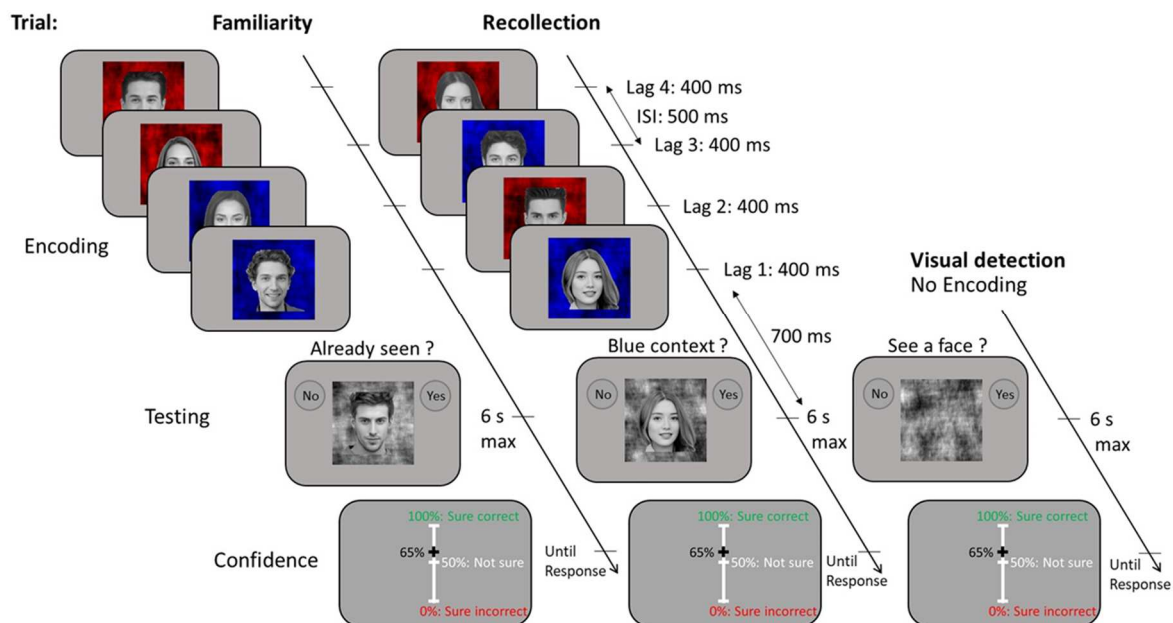


Figure 1. Experimental Design. Timeline of the familiarity, recollection and visual detection tasks. The timeline was identical in the familiarity and recollection task, except for the testing phase where the question was task-specific: “Already seen?” for familiarity, and “Blue context?” for recollection. No encoding took place in the visual detection task. In the present illustration, the correct answers to the familiarity, recollection and perceptual questions are respectively: “No”, “Yes”, and “No”. Lag is an ordinal variable corresponding to the temporal distance between the target stimulus and the test stimulus.

Statistical analyses

Analyses were performed with R¹⁷, using notably the brms¹⁸, BayesFactor¹⁹, ggplot2²⁰ and lme4²¹ packages. Confidence efficiency scores were computed with Matlab (Mathworks, 2017a).

Socio-demographic and neuropsychological characterization

The groups' socio-demographic (age, sex, education), neuropsychological (National Adult Reading Test measuring patients' premorbid IQ²², matrix reasoning subtest from the Wechsler Adult Intelligence Scale version IV²³ and mood (Calgary Depression Scale²⁴) characteristics were compared using the Student t test or χ^2 test when appropriate. Patients were characterized in terms of cognitive insight (using the self-reported Beck Cognitive Insight Scale²⁵), schizophrenia symptomatology (using the clinician-evaluated Positive And Negative Syndrome Scale²⁶, with factorial scores²⁷) and subjective evaluation of cognitive functioning (using the self-reported Subjective Scale To Investigate Cognition in Schizophrenia²⁸). As additional analyses, we explored whether metacognitive performance was correlated with demographic characteristics and clinical scores.

Metacognitive performance

We quantified metacognitive sensitivity with population-level estimates of confidence efficiency¹⁶. This model accounts for potential differences in first-order performance and relies on an explicit generative model of confidence. We also quantified metacognitive sensitivity with a measure of the strength of the relationship between first-order accuracy and confidence (via individual regression slopes), obtained from Bayesian mixed-effects logistic regressions. Importantly, this second model did not take first-order performance into account. Thus, comparing the two measures of metacognitive sensitivity, we assessed the importance of controlling for first-order performance.

Bayesian mixed-effects logistic regressions

We conducted two Bayesian mixed-effects logistic regressions on first-order accuracy (binary categorical variable) as a function of standardized confidence (continuous variable): one model (1a, see below) for *hit* vs *miss* responses (i.e. stimulus strength [1:4]), and one model (1b) for *false alarms* vs. *correct rejections* (i.e. stimulus strength = 0). We analyzed trials with 0 stimulus strength separately (i.e., 0 versus [1:4]) assuming that stimulus strength 0 involved different processes (e.g. detecting a new stimulus may not be based on the same information as detecting an old stimulus. This was corroborated by pilot experiments showing that task-performance at stimulus strength 0 was hardly extrapolated from stimulus strength > 0). Model 1a included group (binary categorical variable: controls vs patients), stimulus strength (ordinal variable with 4 levels: 1 to 4), task (categorical variable: visual detection, familiarity, recollection) as fixed effects, and a full random effect structure (see SI for priors' specifications). Model 1b included the same variables except that stimulus strength was fixed to 0. Results were interpreted on the basis of the Bayes factor (BF) according to Wagenmakers and colleagues²⁹. The BF is the ratio of the marginal likelihoods of each hypothesis, therefore BF > 3 indicates evidence toward H1 (existence of a difference between conditions) and BF < 1/3 indicates evidence toward H0 (absence of difference between conditions). Effects were further characterized by the summary statistics of the posterior distribution (mean and 95 % credible interval, CrI).

Formulae:

accuracy ~ confidence * group * task * evidence + (confidence*task*evidence | participant)

(1a)

accuracy ~ confidence * group * task + (confidence*task | participant)

(1b)

Confidence efficiency

As preregistered, we assessed metacognitive performance while accounting for first-order performance and task difficulty with a recently developed metacognitive index called “confidence efficiency”¹⁶, here adapted to confidence ratings. This index is based on a generative model of confidence judgments, based on Signal Detection Theory, where observers’ confidence judgments are not only subject to metacognitive noise but may also incorporate additional information from the stimulus. Interestingly for us, this method enables the simultaneous modeling of confidence responses across different levels of task difficulty, unlike other methods such as M-ratio^{4,5}.

We estimated confidence efficiency by collapsing all participants into one global population, after normalizing for variations in task performance across individuals, and we quantified its dispersion using a bootstrapping procedure. Namely, we computed 1000 confidence efficiency estimates based on a random resampling of our pool of participants (with replacement), resulting in one estimation distribution per task and group.

Our predictions regarding metacognitive performance (i.e., confidence efficiency and slopes of mixed-effects logistic regressions) were as follows: 1) A metamemory deficit for individuals with schizophrenia compared to healthy controls. 2) A significant interaction effect between group and task reflecting a larger deficit in recollection metamemory among individuals with schizophrenia compared to other tasks, whereas healthy controls show no differences in metacognitive performances across tasks.

We also expected intra-individual first-order performances to be matched (assessed with model 1a), as reflected by equivalent accuracy across the three tasks among patients and healthy controls. Since we did not experimentally adapt task performance between groups, we expected lower task performances among patients compared to controls.

Results

Clinical and neuropsychological variables

Groups were balanced for sex ($\chi^2 = 0.25$, $p = 0.62$) and comparable for age, education level, premorbid IQ, and scores on the WAIS matrix subtest (Table 1). However, individuals with schizophrenia had higher depression scores (mean \pm SD: 4.7 ± 3.9) than healthy controls (mean \pm SD: 1.5 ± 1.7 , $t = 4.20$, $p < 0.001$, $BF = 209$). Descriptive statistics regarding false alarms, hits and confidence are described in Table 2 and show that in both groups, participants were performing all tasks correctly (i.e., better than chance).

	Control N = 36 (mean ± SD)	Schizophrenia N = 34 (mean ± SD)	t-statistic	p-value	Bayes Factor
Age, yr	34.5 ± 14.3	38.3 ± 11.1	1.26	0.21	0.48
Education Level, yr	12.9 ± 1.3	12.8 ± 2.7	-0.16	0.87	0.26
Premordid IQ	108.2 ± 6.8	104.9 ± 12.5	-1.30	0.20	0.55
WAIS Matrix subtest	9.6 ± 2.9	8.4 ± 2.5	-1.87	0.07	1.07
Calgary Depression Scale, score	1.5 ± 1.7	4.7 ± 3.9	4.20	<0.001	209.00
Illness duration, yr		14.7 ± 9.3			
BCIS, composite score		8.2 ± 6.2			
SSTICS, total		30.1 ± 16.1			
SSTICS, working memory		4.6 ± 2.7			
PANSS, positive		13.2 ± 6.3			
PANSS, negative		11.7 ± 9.8			
PANSS, disorganization		21.7 ± 7.8			
PANSS, total		48.8 ± 33.7			

Table 1: Sociodemographic and clinical characteristics of individuals with schizophrenia and control participants. WAIS: Wechsler Adult Intelligence Scale (standardized scores); BCIS: Beck Cognitive Insight Scale; SSTICS: Subjective Scale To Investigate Cognition in Schizophrenia; PANSS: Positive And Negative Symptoms in Schizophrenia. p-values are not corrected for multiple comparisons. Bayes factors are based on Bayesian t-tests with a scaling factor of 0.7.

	Task	Control N = 36 (mean ± SD)	Schizophrenia N = 34 (mean ± SD)	t-statistic	p-value	Bayes Factor
% False alarms	Visual detection	20.0 ± 20.7	16.9 ± 24.5	-0.57	0.57	0.28
	Familiarity	17.1 ± 14.1	24.6 ± 21.1	1.74	0.09	0.92
	Recollection	27.9 ± 17	50.6 ± 23.3	4.63	0.00	1,211.87
% Hits	Visual detection	80.2 ± 13	70.6 ± 18.4	-2.50	0.02	3.53
	Familiarity	81.3 ± 10.4	72.8 ± 20.1	-2.20	0.03	2.02
	Recollection	78.5 ± 15.9	72.3 ± 14.9	-1.68	0.10	0.82
% Confidence in errors	Visual detection	82.8 ± 11.5	85.6 ± 11.4	1.02	0.31	0.39
	Familiarity	73.1 ± 12	72.7 ± 11.2	-0.13	0.89	0.25
	Recollection	73.2 ± 12.1	70.5 ± 11	-0.97	0.33	0.37
% Confidence in correct responses	Visual detection	95.0 ± 5.3	91.6 ± 9.7	-1.82	0.07	1.06
	Familiarity	89.2 ± 6.8	84.6 ± 11	-2.07	0.04	1.57
	Recollection	87.4 ± 10	80.9 ± 10.9	-2.58	0.01	4.04

Table 2: Experimental characteristics of individuals with schizophrenia and healthy control participants. p-values are not corrected for multiple comparisons. Bayes factors are based on Bayesian t-tests with a scaling factor of 0.7.

First-order performance

Model 1a revealed that patients had lower performance than healthy controls in the visual detection, familiarity and recollection tasks, and these first-order deficits were similar across tasks (i.e. no first-order interactions, see Table 3, Figure 2A).

		Estimate [95% CrI]	Bayes Factor
First-order deficits	Schizophrenia-Control (Visual detection)	-0.82 [-1.43, -0.21]	8.40
	Schizophrenia-Control (Familiarity)	-0.81 [-1.49, -0.13]	3.07
	Schizophrenia-Control (Recollection)	-0.62 [-1.27, 0.02]	1.23
First-order interactions	group x task (Familiarity - Visual detection)	0.01 [-0.57, 0.62]	0.31
	group x task (Recollection - Visual detection)	0.2 [-0.38, 0.78]	0.38
	group x task (Recollection - Familiarity)	0.19 [-0.41, 0.81]	0.27
Intra-individual performance-matching	Familiarity - Visual detection (Control)	0.13 [-0.31, 0.56]	0.26
	Recollection - Visual detection (Control)	-0.2 [-0.62, 0.21]	0.32
	Recollection - Familiarity (Control)	-0.33 [-0.78, 0.12]	0.50
	Familiarity - Visual detection (Schizophrenia)	0.14 [-0.31, 0.58]	0.20
	Recollection - Visual detection (Schizophrenia)	0 [-0.44, 0.43]	0.16
	Recollection - Familiarity (Schizophrenia)	-0.14 [-0.59, 0.31]	0.14

Table 3: First-order deficits across tasks. We report posterior distributions' summary statistics (mean and 95% Credible interval) along with Bayes factors.

Differences in performance were expected as task performance was not experimentally controlled between groups. However, our procedure was designed to match intra-individual performance across tasks. Accordingly, pairwise first-order task performances were similar among patients and among control participants (Table 3). This confirms that our procedure globally matched intra-individual performance across tasks, although it did not match intra-individual performance for each stimulus strength (see Table S1).

Patients and controls were sensitive to task manipulation of stimulus strength as indicated by a strong effect of stimulus strength in all tasks (See Table S1 and Figure 2A).

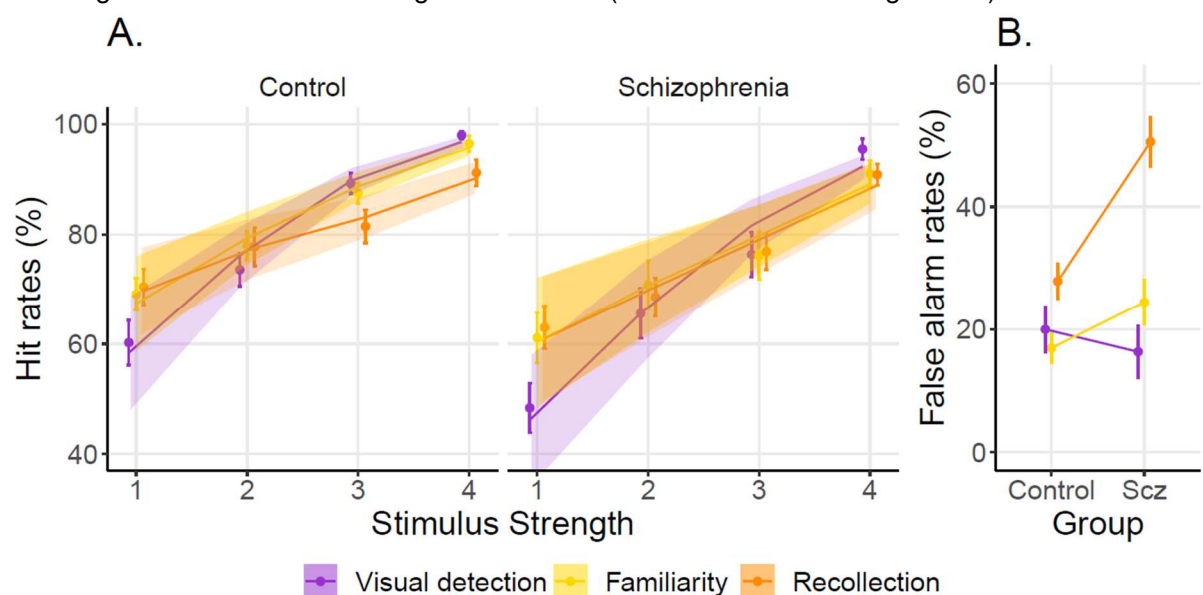


Figure 2: A. Hit rates (i.e., rates of “yes” responses following stimuli with stimulus strength > 0) across stimulus strengths in the visual detection (purple), familiarity (yellow), and recollection tasks (orange). Points and error bars indicate average accuracy and standard error of the mean, respectively; solid lines and shaded areas represent model fit mean and

95% confidence interval, respectively. B. False-alarm rates (i.e., rates of “yes” responses following stimuli with 0 stimulus strength) across groups. Points and error bars indicate average accuracy and standard error of the mean, respectively. Same color description as panel A.

False alarms:

Compared to healthy controls, patients had a similar false alarm rate in both the visual detection task (i.e., reporting seeing a face when none was presented: -0.22 [$-1.02, 0.57$], $BF = 0.45$), and in the familiarity task (i.e., reporting having seen the test face during the encoding phase when presented with a new face: 0.78 [$-0.07, 1.64$], $BF = 2.21$) but they committed significantly more false alarms in the recollection task (i.e., reporting having seen the test face in a given context during the encoding phase when presented in another context: 1.54 [$0.61, 2.46$], $BF = 96.7$) (Figure 2B).

Second-order performance

Confidence

Confidence levels were similar between patients and controls, except for the recollection task where patients were underconfident in correct responses (Table 1, confidence mean \pm SD: 80.9 ± 10.9) compared to controls (confidence mean \pm SD: 87.4 ± 10.0 , $t = -2.58$, $p < 0.05$, $BF = 4.04$).

Metacognitive sensitivity

When quantifying metacognitive sensitivity as the slope between accuracy and confidence in mixed-effects logistic regressions (model 1a), individuals with schizophrenia were not found to underperform compared to healthy controls (Figure 3A). Although qualitatively, the results could suggest a metacognitive deficit in the visual detection task, the evidence was statistically inconclusive (-0.41 [$-0.84, 0.01$], $BF = 1.33$). By contrast, we obtained moderate evidence in favor of an absence of a deficit both in meta-familiarity (-0.24 [$-0.59, 0.12$], $BF = 0.32$), and meta-recollection (-0.13 [$-0.51, 0.27$], $BF = 0.17$). Moreover, there was no difference of deficit between tasks (Familiarity - Recollection: 0.11 [$-0.33, 0.56$], $BF = 0.18$); Familiarity - Perception: 0.17 [$-0.26, 0.62$], $BF = 0.28$; Recollection - Perception: 0.28 [$-0.18, 0.75$], $BF = 0.47$). As discussed above, metacognitive sensitivity can be contaminated by differences in terms of first-order performance, which was only partially controlled in our paradigm. To estimate metacognitive performance independently of first-order performance, we turned to another metric called the confidence efficiency.

When quantifying metacognitive performance using the confidence efficiency measure of metacognition - which controls for first-order deficits - individuals with schizophrenia had similar confidence efficiency in the detection (-0.17 [$-0.45, 0.06$]), familiarity (-0.00 [$-0.44, 0.31$]) and recollection tasks (-0.10 [$-0.58, 0.30$]) (Figure 3B). Within each group, metacognitive performance was comparable across tasks (Controls: Visual detection - Familiarity: -0.11 [$-0.40, 0.20$], Visual detection - Recollection: -0.18 [$-0.51, 0.17$], Recollection - Familiarity: 0.07 [$-0.26, 0.41$]; Patients: Visual detection - Familiarity: -0.28 [$-0.55, 0.14$], Visual detection - Recollection: -0.24 [$-0.57, 0.19$], Recollection - Familiarity: -0.03 [$-0.50, 0.42$]).

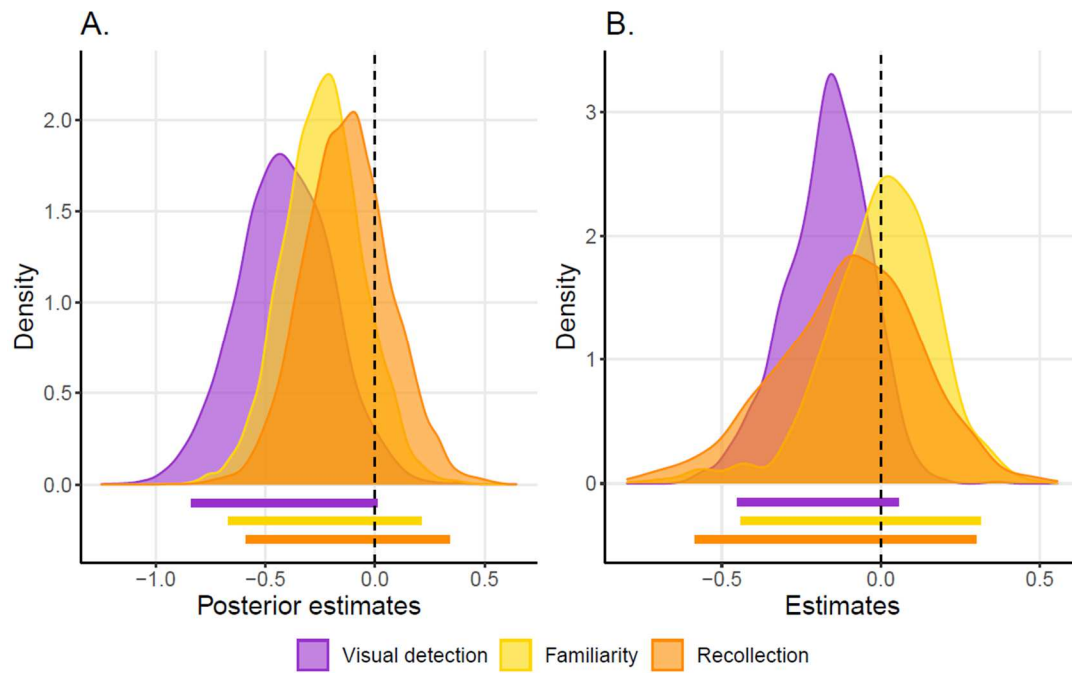


Figure 3: A. Bayesian posterior distributions of differences of regression slope estimates between patients and controls (i.e. distributions of metacognitive deficits estimations): meta-perceptual difference (purple), meta-familiarity difference (yellow), meta-recollection difference (orange). Vertical dashed line (estimate = 0) represents no difference between patients and controls. Horizontal colored bars indicate 95% credible intervals. B. Distributions of differences in confidence efficiency estimates between patients and controls. Horizontal colored bars indicate 95% confidence intervals. Same color description as panel A.

Contrary to our predictions about domain-generality, metacognitive performance as measured with mixed-effects logistic regressions did correlate across tasks, neither did we find correlations with clinical traits such as positive, negative and disorganization syndromes and cognitive insight (see SI, Figure S7, 8).

Discussion

The present study aimed at characterizing metamemory and metaperception in people with schizophrenia while controlling for first-order deficits. In particular, we assessed metacognition in visual detection, familiarity, and recollection tasks. We hypothesized that people with schizophrenia would be specifically impaired in the meta-memory domain. At the first-order level, we found that people with schizophrenia had lower first-order performance in the three tasks compared to healthy controls, which confirms the importance of accounting for first-order deficits to quantify second-order processes specifically. When doing so, contrary to our hypothesis we found that metacognitive sensitivity was preserved among individuals with schizophrenia in the three tasks. In what follows, we discuss technical and conceptual aspects of our paradigm that should be considered to interpret this result, and then examine its clinical and theoretical significance.

A key contribution of this study is our attempt to match first-order performance between tasks for each participant using adaptive procedures, and between groups of participants using a generative model of confidence. We note that our adaptive procedure to match performance between tasks was successful when considering average performance, but not when considering task performance across levels of stimulus strength. In other words, we equated the overall performance but not the slopes between tasks in Figure 2A (see SI for details). A plausible explanation for this is a contextual effect. In session 1, blocks of visual detection trials were separated from blocks of memory trials, whereas in session 2 the three tasks were interleaved within each block of trials. Thus, the visual detection psychometric curve (SI, Figure S2c) from which we determined four visual contrast levels was obtained from a sequence of low-contrast perceptual stimuli (3 x 80 stimuli in a row), whereas during session 2 these low-contrast visual stimuli were interleaved with high-contrasted memory stimuli. This contextual effect might have resulted in a rightward shift (See Figure S3) of the visual detection psychometric curve, leading to underperformance in both groups in the visual detection task compared to the familiarity and recollection tasks.

Regarding between-groups task performance matching, early pilot versions of the present protocol aimed at equating memory performance between participants using adaptive staircases that manipulated either the number of encoding items, or the lag variable, but these attempts were not successful (no convergence). Instead, we accounted for differences in task-performance between groups by relying on measures of confidence efficiency from a recent generative model of confidence¹⁶, which enables the estimation of metacognitive abilities in factorial designs. Although this framework is recent and has not been fully benchmarked yet, we note that we found qualitatively similar results using a Bayesian logistic mixed-effects regression, which does not consider possible cognitive deficits but has the advantage of providing hierarchical estimates of metacognitive sensitivity, dealing with unbalanced data, and considering prior knowledge to compute Bayes factors. In contradiction to existing literature, both frameworks revealed no evidence for a metacognitive deficit in any of the three tasks. In fact, we found evidence for an absence of metacognitive deficit in memory tasks, and only inconclusive evidence in the perceptual domain. The absence of metacognitive deficit in schizophrenia was corroborated by an absence of difference regarding confidence biases. Indeed, contrary to several studies which did not control for first-order performance³⁰⁻³², we found no overconfidence in errors nor underconfidence in correct responses. One possibility is that the confidence biases previously reported in schizophrenia also stem from first-order deficits differences. Furthermore, contrary to previous behavioral results showing a positive link between false alarms and positive symptoms or proneness to hallucinations³³⁻³⁵, our sample of patients had comparable rates of false alarms compared to healthy controls in the visual detection task. They committed more false-alarms in the memory tasks, interpreted as false recognitions, but no relationships were found between rates of false alarms and PANSS positive score (see SI).

At a conceptual level, the framing of our memory tasks in terms of familiarity and recollection processes may be questionable. Indeed, although our recognition memory tasks shared some features with usual familiarity and recollection tasks (in particular the testing questions which are respectively context-independent and context-dependent), there was no delay between encoding and testing phases, as we manipulated task difficulty with a variable lag. Therefore, one may consider our tasks to reflect working memory, which is also known to involve familiarity and recollection processes³⁶. To our knowledge, no study assessed

metacognition related to short-term memory in schizophrenia. At first sight, our results seem to be in contradiction with the study by Berna et al and colleagues⁸, which reported impaired metamemory in schizophrenia in a long-term (autobiographical) memory task, while controlling statistically for first-order performance. Yet, if our results are construed as evidence for preserved “short-term” metamemory in schizophrenia, the contradiction might be only apparent. A full taxonomy of metamemory processes is beyond the scope of the present study, and developing new paradigms to assess metacognitive performance in distinct subdomains of memory while controlling for first-order performance is one of the numerous challenges the metacognitive field is facing³⁷.

With these technical and conceptual considerations in mind, we can contextualize our findings and assess their clinical relevance. Our protocols focus on “in-the-moment” metacognition³⁸, i.e. confidence in trial-by-trial decisions, also known as “local” metacognition as opposed to more “global” evaluations^{39–41}. Metacognitive evaluations have been construed as hierarchically organized, where aggregated local judgments give rise to global self-beliefs about one’s performance within a cognitive task or domain⁴². Interestingly, it has been shown that global metacognitive evaluations can be altered independently from the local monitoring processes⁴³. Yet, as recently discussed⁴⁴, both local and global measures of metacognition may give an incomplete picture of metacognitive abilities from a clinical perspective. This concern is corroborated by the fact that our metacognitive measures are not correlated with several clinical dimensions of interest for schizophrenia (symptoms, cognitive insight, self-reported cognitive functioning see SI). Only perceptual reasoning assessed with WAIS matrix subtest scores were positively correlated with metacognitive performance, as reported previously⁴⁵. The need for paradigms that do justice to the breadth of the metacognition construct, i.e. including more cognitive domains, larger timescales, and theory of mind is now becoming acknowledged by the field.

Acknowledgments: NF has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 803122).

References

1. Rouy M, Saliou P, Nalborczyk L, Pereira M, Roux P, Faivre N. Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders. *Neurosci Biobehav Rev.* 2021;126:329-337. doi:10.1016/j.neubiorev.2021.03.017
2. Hoven M, Lebreton M, Engelmann JB, Denys D, Luigjes J, van Holst RJ. Abnormalities of confidence in psychiatry: an overview and future perspectives. *Transl Psychiatry.* 2019;9(1):268. doi:10.1038/s41398-019-0602-7
3. Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychon Bull Rev.* 2003;10(4):843-876. doi:10.3758/BF03196546
4. Fleming SM, Lau HC. How to measure metacognition. *Front Hum Neurosci.* 2014;8. doi:10.3389/fnhum.2014.00443
5. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn.* 2012;21(1):422-430. doi:10.1016/j.concog.2011.09.021

6. Gopal YV, Variend H. First-episode schizophrenia: review of cognitive deficits and cognitive remediation. *Adv Psychiatr Treat*. 2005;11(1):38-44. doi:10.1192/apt.11.1.38
7. Schaefer J, Giangrande E, Weinberger DR, Dickinson D. The global cognitive impairment in schizophrenia: Consistent over decades and around the world. *Schizophr Res*. 2013;150(1):42-50. doi:10.1016/j.schres.2013.07.009
8. Berna F, Zou F, Danion JM, Kwok SC. Overconfidence in false autobiographical memories in patients with schizophrenia. *Psychiatry Res*. 2019;279:374-375. doi:10.1016/j.psychres.2018.12.063
9. Libby LA, Yonelinas AP, Ranganath C, Ragland JD. Recollection and Familiarity in Schizophrenia: A Quantitative Review. *Biol Psychiatry*. 2013;73(10):944-950. doi:10.1016/j.biopsych.2012.10.027
10. Mishkin M, Suzuki WA, Gadian DG, Vargha-Khadem F. Hierarchical organization of cognitive memory. Burgess N, Oapos;Keefe J, eds. *Philos Trans R Soc Lond B Biol Sci*. 1997;352(1360):1461-1467. doi:10.1098/rstb.1997.0132
11. Moulin CJA, Souchay C, Morris RG. The cognitive neuropsychology of recollection. *Cortex*. 2013;49(6):1445-1451. doi:10.1016/j.cortex.2013.04.006
12. Heckers S, Rauch S, Goff D, et al. Impaired recruitment of the hippocampus during conscious recollection in schizophrenia. *Nat Neurosci*. 1998;1(4):318-323. doi:10.1038/1137
13. van Erp TGM, Hibar DP, Rasmussen JM, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol Psychiatry*. 2016;21(4):547-553. doi:10.1038/mp.2015.63
14. Barbeau EJ, Taylor MJ, Regis J, Marquis P, Chauvel P, Liegeois-Chauvel C. Spatio-temporal Dynamics of Face Recognition. *Cereb Cortex*. 2008;18(5):997-1009. doi:10.1093/cercor/bhm140
15. Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW. Controlling low-level image properties: The SHINE toolbox. *Behav Res Methods*. 2010;42(3):671-684. doi:10.3758/BRM.42.3.671
16. Mamassian P, de Gardelle V. Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol Rev*. Published online July 29, 2021. doi:10.1037/rev0000312
17. R Core Team. R: A Language and Environment for Statistical Computing. <https://www.R-project.org/>
18. Bürkner PC. **brms** : An R Package for Bayesian Multilevel Models Using Stan. *J Stat Softw*. 2017;80(1). doi:10.18637/jss.v080.i01
19. Morey RD, Rouder JN. BayesFactor: Computation of Bayes Factors for Common Designs. Published online 2018. <https://CRAN.R-project.org/package=BayesFactor>
20. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Published online 2016. <https://ggplot2.tidyverse.org>
21. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models using lme4. Published online 2014. doi:10.48550/ARXIV.1406.5823
22. Mackinnon A, Mulligan R. Estimation de l'intelligence prémorbide chez les francophones. *L'Encéphale*. 2005;31(1):31-43. doi:10.1016/S0013-7006(05)82370-X
23. Wechsler D. Wechsler Adult Intelligence Scale--Fourth Edition. Published online November 12, 2012. doi:10.1037/t15169-000
24. Addington D, Addington J, Maticka-tyndale E. Assessing Depression in Schizophrenia: The Calgary Depression Scale. *Br J Psychiatry*. 1993;163(S22):39-44. doi:10.1192/S0007125000292581
25. Beck A. A new instrument for measuring insight: the Beck Cognitive Insight Scale. *Schizophr Res*. 2004;68(2-3):319-329. doi:10.1016/S0920-9964(03)00189-0
26. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr Bull*. 1987;13(2):261-276. doi:10.1093/schbul/13.2.261
27. Vandergaag M, Hoffman T, Remijsen M, et al. The five-factor model of the Positive and Negative Syndrome Scale II: A ten-fold cross-validation of a revised model. *Schizophr Res*. 2006;85(1-3):280-287. doi:10.1016/j.schres.2006.03.021

28. Stip E, Caron J, Renaud S, Pampoulova T, Lecomte Y. Exploring cognitive complaints in schizophrenia: the subjective scale to investigate cognition in schizophrenia. *Compr Psychiatry*. 2003;44(4):331-340. doi:10.1016/S0010-440X(03)00086-5
29. Wagenmakers EJ, Marsman M, Jamil T, et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev*. 2018;25(1):35-57. doi:10.3758/s13423-017-1343-3
30. Eifler S, Rausch F, Schirmbeck F, et al. Metamemory in schizophrenia: retrospective confidence ratings interact with neurocognitive deficits. *Psychiatry Res*. 2015;225(3):596-603. doi:10.1016/j.psychres.2014.11.040
31. Garcia CP, Sacks SA, Weisman de Mamani AG. Neurocognition and Cognitive Biases in Schizophrenia. *J Nerv Ment Dis*. 2012;200(8):724-727. doi:10.1097/NMD.0b013e3182614264
32. Eisenacher S, Zink M. The Importance of Metamemory Functioning to the Pathogenesis of Psychosis. *Front Psychol*. 2017;8:304. doi:10.3389/fpsyg.2017.00304
33. Bentall RP, Slade PD. Reality testing and auditory hallucinations: A signal detection analysis. *Br J Clin Psychol*. 1985;24(3):159-169. doi:10.1111/j.2044-8260.1985.tb01331.x
34. Moseley P, Fernyhough C, Ellison A. The role of the superior temporal lobe in auditory false perceptions: A transcranial direct current stimulation study. *Neuropsychologia*. 2014;62:202-208. doi:10.1016/j.neuropsychologia.2014.07.032
35. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*. 2017;357(6351):596-600. doi:10.1126/science.aan3458
36. Oberauer K. Binding and Inhibition in Working Memory: Individual and Age Differences in Short-Term Recognition. *J Exp Psychol Gen*. 2005;134(3):368-387. doi:10.1037/0096-3445.134.3.368
37. Rahnev D, Balsdon T, Charles L, et al. *Consensus Goals in the Field of Visual Metacognition*. PsyArXiv; 2021. doi:10.31234/osf.io/z8v5x
38. Palmer-Cooper EC, Wright AC, McGuire N, et al. Metacognition and psychosis-spectrum experiences: A study of objective and subjective measures. *Schizophr Res*. Published online January 2023:S0920996422004613. doi:10.1016/j.schres.2022.12.014
39. Lee ALF, de Gardelle V, Mamassian P. Global visual confidence. *Psychon Bull Rev*. 2021;28(4):1233-1242. doi:10.3758/s13423-020-01869-7
40. Rouault M, Fleming SM. Formation of global self-beliefs in the human brain. *Proc Natl Acad Sci*. 2020;117(44):27268-27276. doi:10.1073/pnas.2003094117
41. Cavalan Q, Vergnaud JC, de Gardelle V. From local to global estimations of confidence in perceptual decisions. *Forthcoming in JEP General*.
42. Seow TXF, Rouault M, Gillan CM, Fleming SM. Reply to: Metacognition, Adaptation, and Mental Health. *Biol Psychiatry*. 2022;91(8):e33-e34. doi:10.1016/j.biopsych.2021.11.005
43. Bhome R, McWilliams A, Price G, et al. Metacognition in functional cognitive disorder. *Brain Commun*. 2022;4(2):fcac041. doi:10.1093/braincomms/fcac041
44. Schnakenberg Martin AM, Lysaker PH. Metacognition, Adaptation, and Mental Health. *Biol Psychiatry*. 2022;91(8):e31-e32. doi:10.1016/j.biopsych.2021.09.028
45. Faivre N, Roger M, Pereira M, et al. Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *J Psychiatry Neurosci*. 2021;46(1):E65-E73. doi:10.1503/jpn.200022

3. Exploration of electrophysiological markers of confidence during a metacognitive task

From the results obtained in our meta-analysis showing preserved metacognitive abilities among patients with schizophrenia, we had two hypotheses in mind. The most parsimonious interpretation was to take these results at face value, i.e. confidence calibration processes are preserved among patients with schizophrenia, at least in visual discrimination tasks. But we could also consider the possibility that behavioral measures were not fine-grained enough to capture subtle differences in the underlying neural processes. For instance, it could be that confidence calibration was actually impaired, but that patients had recruited additional resources compensating for their difficulties. To arbitrate between these two hypotheses - normal processing vs a compensating mechanism - we recorded electrophysiological (EEG) data and isolated neural correlates of performance monitoring. These data were already available, as EEG had been recorded on participants who were included in the published behavioral study from Faivre and colleagues (2021), and were still waiting to be analyzed.

A recent review (Kirschner et al. 2021) has highlighted that performance monitoring impairments in schizophrenia were essentially related to two electrophysiological markers of error monitoring processing: an early blunting of error-related negativity (ERN) and a late blunting of error-positivity (Pe). The ERN and Pe are event-related potentials that are time-locked to the first-order responses. The ERN occurs in error trials rapidly after the response (50-100 ms), and has been interpreted either as an error-monitoring or a conflict-monitoring signal (Ullsperger et al. 2014). The Pe occurs is known to reflect awareness of errors and is modulated by confidence (Figure 19, Boldt and Yeung 2015).

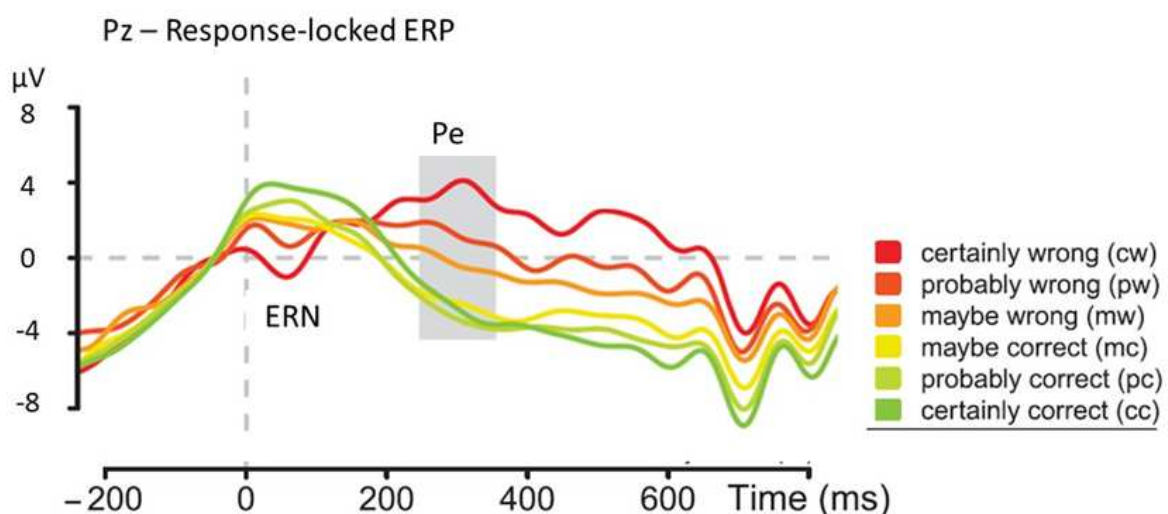


Figure 19. Response-locked ERP of performance-monitoring. ERP from electrode Pz, in a visual discrimination task where healthy participants had to indicate which of the two briefly flashed squares contained more dots. Then participants had to indicate their confidence. Colors indicate the degree of confidence from “certainly wrong” (red) to “certainly correct” (green). Reprint from Boldt and Yeung (2015)

The meta-analysis from Martin et al. (2018) had shown that the blunting of the ERN was a robust marker of internal monitoring impairment in schizophrenia, but not the Pe blunting. In this study, we tested whether markers of error-monitoring were impaired in schizophrenia, even when a procedure controlling for task-performance is used.

Status of the manuscript: Published in *Schizophrenia*

Reference: Rouy, M., Roger, M., Goueytes, D., Pereira, M., Roux, P., & Faivre, N. (2023). Preserved electrophysiological markers of confidence in schizophrenia spectrum disorder. *Schizophrenia*, 9(1), 12. <https://doi.org/10.1038/s41537-023-00333-4>

Preserved electrophysiological markers of confidence in schizophrenia spectrum disorder

Martin Rouy¹✉, Matthieu Roger², Dorian Goueytes¹, Michael Pereira¹ , Paul Roux³  and Nathan Faivre¹ 

A large number of behavioral studies suggest that confidence judgments are impaired in schizophrenia, motivating the search for neural correlates of an underlying metacognitive impairment. Electrophysiological studies suggest that a specific evoked response potential reflecting performance monitoring, namely the error-related negativity (ERN), is blunted in schizophrenia compared to healthy controls. However, attention has recently been drawn to a potential confound in the study of metacognition, namely that lower task-performance in schizophrenia compared to healthy controls involves a decreased index of metacognitive performance (where metacognitive performance is construed as the ability to calibrate one's confidence relative to response correctness), independently of metacognitive abilities among patients. Here, we assessed how this confound might also apply to ERN-blunting in schizophrenia. We used an adaptive staircase procedure to titrate task-performance on a motion discrimination task in which participants ($N = 14$ patients and 19 controls) had to report their confidence after each trial while we recorded high density EEG. Interestingly, not only metaperceptual abilities were preserved among patients at the behavioral level, but contrary to our hypothesis, we also found no electrophysiological evidence for altered EEG markers of performance monitoring. These results bring additional evidence suggesting an unaltered ability to monitor perceptual performance on a trial by trial basis in schizophrenia.

Schizophrenia (2023)9:12; <https://doi.org/10.1038/s41537-023-00333-4>

INTRODUCTION

Schizophrenia spectrum disorder (SSD) is a mental condition with severe consequences in terms of cognitive abilities^{1,2}, social abilities^{3,4}, and more broadly on quality of life^{5,6}. For two decades, an increasing attention has been drawn to metacognitive abilities in individuals with SSD, with numerous behavioral studies suggesting an impaired ability to calibrate confidence judgments according to performance compared to healthy controls^{7,8}, paralleled with a substantial number of electrophysiological studies showing performance monitoring impairments in this population⁹. In particular, electrophysiological studies highlighted specific evoked response potentials (ERPs) which are blunted in individuals with SSD, such as the error-related negativity (ERN), the error positivity (Pe), or the feedback-related negativity (FRN)⁹. The ERN is a response-locked ERP peaking around 100 ms on frontal midline electrodes following errors^{10,11} and mostly elicited in choice reaction time tasks (e.g. Flanker task, Simon task) where participants are pressured to respond quickly (typically under 1 s), although an ERN is also found in non speeded tasks¹². The function reflected by the ERN is not clear, whether it reflects a response conflict or an error-monitoring signaling is still debated^{13–15}. It has been shown that ERN amplitude increased with confidence that one made an error¹⁶. Boldt and Yeung¹⁷ have also demonstrated a similar gradation of ERN amplitude as a function of confidence, but their multivariate analysis indicated more robust confidence modulations of Pe amplitude - a subsequent neural marker of error awareness occurring 200–300 ms after errors are committed^{9,18}. The FRN peaks 200–300 ms after negative performance feedback. Among these ERPs, the blunted ERN is considered the most robust candidate as a trait marker predictive of symptomatology^{9,19}. Here, we were interested in the link between electrophysiological markers of

performance-monitoring - such as ERN and Pe - and explicit metacognitive judgments such as the adequation between confidence ratings and task-performance. Since ERN and Pe are related to confidence but blunted among patients with SSD, we might expect a decreased ability to form relevant confident judgments (i.e. confidence judgments accurately reflecting task-performance) among patients, hence a decreased metacognitive performance.

Considering that individuals with SSD typically underperform in cognitive tasks compared to matched controls, it is important to assess if ERN-blunting simply reflects poorer behavioral performance, or if performance monitoring mechanisms are specifically impaired in SSD. In this respect, Kirschner and Klein⁹ mentioned that among the 21 reviewed studies showing a blunted ERN in individuals with SSD, 12 studies reported comparable performance between groups, while the other 9 studies reported underperformance among patients. The meta-analysis from Martin et al.¹⁹ revealed a similar pattern of results suggesting that group performance was not predictive of ERN blunting. Yet, it has been shown in healthy participants that task difficulty decreases the amplitude of the ERN²⁰, so we reasoned that *comparable* group performance, which may include differences in performance between groups that did not reach statistical significance, might still underlie subtle discrepancies between individuals that are not captured behaviorally, but that might nevertheless contaminate measures such as ERPs. In turn, little can be said about differences in ERN amplitudes between groups that behave similarly on average, but that include individuals with varying degrees of performance. Instead, we argue that performance *matching*, which uses a procedure designed to equate performance between each participant is necessary to assess the specificity of electrophysiological markers of performance monitoring in SSD. In the present

¹Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France. ²Centre d'Economie de la Sorbonne, Paris, France. ³Centre Hospitalier de Versailles, Service Hospitalo-Universitaire de Psychiatrie d'Adultes et d'Addictologie, Le Chesnay; Université Paris-Saclay; Université de Versailles Saint-Quentin-En-Yvelines; DisAP-DevPsy-CESP, INSERM UMR1018, Villejuif, France. ✉email: martin.rouy@univ-grenoble-alpes.fr

preregistered study (https://gitlab.com/nfaivre/meta_scz_public), task performance was controlled via a staircase procedure adapting the amount of sensory evidence (motion coherence) according to individual perceptual abilities. This procedure enabled us to match performance between groups and individuals, and therefore to discuss further whether the typical blunted ERN in individuals with SSD is dependent on task-performance or not.

Here we present the results from EEG data collected in patients and matched healthy controls who performed a visual motion discrimination task. Behavioral analyses of the same cohort of participants revealed preserved metacognitive abilities among individuals with SSD²¹. Building upon previous findings showing a blunted ERN in individuals with SSD^{9,19}, we conducted EEG analyses: 1) to investigate the occurrence of a blunted ERN in individuals with SSD with matched task-performance, and 2) to explore whether compensatory neural activity related to confidence (ERN-like or Pe) are found among individual with SSD, which could possibly explain why their metacognitive abilities are preserved²¹. Besides matching for performance, we also employed a paradigm which did not enforce speeded responses. By giving participants sufficient time to provide a response, we sought to quantify the electrophysiological correlates of performance monitoring without confounding our results with the detection of “slips” - a category of errors corresponding to incorrect executions of appropriate motor programs^{22,23}. Slips typically occur when participants provide a speeded response and immediately realize they pressed the wrong button, a process which differs from evaluating the probability that a decision about noisy sensory information is correct^{24,25}.

METHODS

Methods and analyses were pre-registered (NCT03140475; https://gitlab.com/nfaivre/meta_scz_public). The study was approved by the ethical committee Sud Méditerranée II (217 R01).

Participants

Twenty individuals with schizophrenia spectrum disorder (schizophrenia or schizoaffective disorder, 16 males, 4 females) and 22 healthy participants (15 males, 7 females) from the general population took part in this study. Schizophrenia and schizoaffective disorders were diagnosed by M.R. based on the Structured Clinical Interview for assessing the DSM-5 criteria. Another licensed psychiatrist (patients' treating psychiatrist) confirmed the diagnosis for each patient according to the DSM-5 criteria. The control group was screened for current or past psychiatric illness, and individuals were excluded if they met the criteria for a severe and persistent mental disorder. Five patients were excluded for having excessive artifactual EEG activity (see below), and one for having 208/300 trials with a movement onset <100 ms. Three control participants were excluded based on the following criteria: one because of a non-converging staircase, one with an estimated IQ lower than our inclusion criterion of 70, and one because no EEG data was available. In the end, the EEG analyses presented in this article were conducted on 14 individuals with SSD and 19 healthy controls.

Neuropsychological and clinical evaluation

Both individuals with schizophrenia spectrum disorders and healthy controls were evaluated on several neuropsychological domains. In particular, we assessed perceptual reasoning with the standardized score on the matrices subtest of the Wechsler Adult Intelligence Scale 4th version (WAIS-IV²⁶), verbal reasoning with the standardized score on the vocabulary subtest of WAIS-IV, working memory with the standardized score on the letter-number sequencing subtest of WAIS-IV, depressive symptoms

with the Calgary Depression Scale (CDS²⁷), cognitive insight with the composite index on the Beck Cognitive Insight Scale (BCIS²⁸). The composite index of the BCIS reflects the cognitive insight and is calculated by subtracting the score for the self-certainty scale from that of the self-reflectiveness scale. The French National Adult Reading Test (fNART²⁹) provided an estimate of pre morbid IQ.

The intensity of schizophrenia symptoms was evaluated on patients with the Positive And Negative Syndrome Scale (Kay et al.³⁰).

Experimental design

We used a visual discrimination task. Stimuli consisted of 100 moving dots within a circle (3° radius) at the center of the screen. On each trial, participants indicated whether the motion direction of the dots was to the left or to the right by reaching and clicking on one of two choice targets (3° radius circle) at the top corners of the screen with a mouse (Fig. 1A). After 6 s without response, a buzz sound rang and a message was displayed inviting the participant to respond quicker. Motion coherence was adapted at the individual level via a 1up/2down staircase procedure in order to match task-performance between groups. Following each perceptual decision, participants were asked to report their confidence about their response using a vertical visual analog scale from 0% (Sure incorrect) to 100% (Sure correct), with 50% confidence meaning “Not sure at all”. (For more details, see ref. ²¹).

In the original study²¹, we used mouse trajectories instead of button presses to investigate how the kinematics of mouse movements (velocity and acceleration) related to confidence. Here, we focused on movement initiation rather than response click as a proxy for decision time to avoid the temporal jitter due to kinematic noise in mouse trajectories (i.e. overshoots and undershoots plus small adjustments to reach the response box). We reasoned that time-locking on the initiation of the movement rather than on the response click was of particular relevance when dealing with patients with SSD, who are prone to various motor impairments, either due to medication³¹ or illness³². Movement onset was defined as the time needed from stimulus onset to reach 20% of maximum mouse velocity on each trial, from which we subtracted an arbitrary offset of five frames (~83 ms) to better capture the moment of movement initiation. Mouse movements with velocity peaks lower than 20% of the maximum velocity were discarded as non-decisional, noisy mouse movements. Visual inspection of the corresponding mouse velocity profiles showed that this procedure was effective in finding the movement onset (see Supplementary Fig. S1).

Trials with early (<100 ms) or late (>6 s) mouse movements were excluded (6.7 ± 7.1% and 15.3 ± 9.8% of trials in controls and patients, $t = -3.06$, $p < 0.01$, BF = 13.2). Changes of mind occurring before the response (i.e., indicated by a change in mouse trajectory, 7.1 ± 5.4% and 8.3 ± 8.1% of trials in controls and patients, resp., $t = 0.52$, $p = 0.61$, BF = 0.36) or after the response (i.e., indicated by a confidence lower than 50%, 10.0 ± 12.0% and 10.7 ± 11% of trials in controls and patients, resp. $t = -0.17$, $p = 0.87$, BF = 0.33) were excluded, avoiding contamination from additional noise and cognitive processes. We counted as changes of mind trials in which the mouse trajectory - after having reached 20% of the total distance between the initial position of the mouse and the y-projection of the target position - crossed the midline between the two response targets.

Data analysis

Behavioral analysis. Behavioral analyses were performed using R (2020), to ensure that our behavioral conclusions in the original study²¹ were still valid on this subset of participants. In particular, we assessed whether our groups of participants were comparable both in terms of demographic and neuropsychological

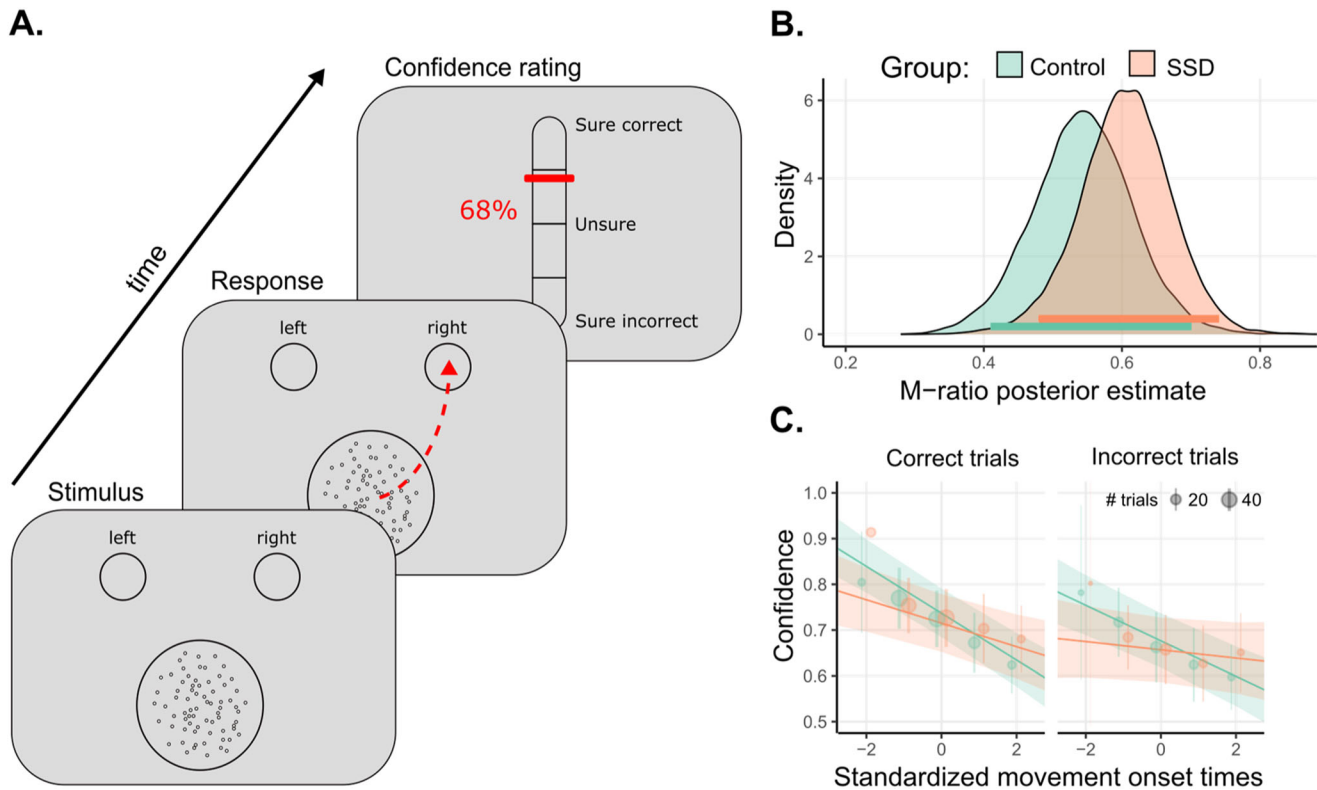


Fig. 1 Experimental task and behavioral results. **A** Experimental task: participants indicated the direction of the random-dot kinematogram by clicking in the corresponding response circle, and then estimated their confidence in their response. **B** Distribution of posterior estimates of metacognitive efficiency (M-Ratio), horizontal lines depict 95% Highest density intervals. **C** Confidence ratings regressed on standardized movement onsets, as a function of groups. Error bars indicate 95% confidence intervals. Orange: Participants with SSD; Green: Control participants. Adapted from ref. ²¹.

characteristics, and metacognitive performance (i.e. how well participants were able to calibrate their confidence judgments on performance, by computing an index of metacognitive efficiency or M-ratio³³, in a Bayesian framework³⁴ (See pre-registered plan for more details).

EEG recording and preprocessing. During the task, EEG activity was continuously recorded using a 64-channels Hlamp system (g.tec, Schiedlberg, Austria), sampled at 1200 Hz. Electrodes were positioned according to the international 10-10 system with AFz as the reference site. Impedance of electrodes was kept below 5k Ω .

Pre-processing was conducted with Matlab³⁵ scripts including functions from the EEGLAB toolbox (v2021.0, EEGLAB³⁶). EEG signal was downsampled at 128 Hz, then high-pass filtered at 0.5 Hz and low-pass filtered at 45 Hz. Bad channels were removed manually through visual inspection. Continuous EEG data were locked on mouse movement onset and epoched between -1s pre-movement to 2s post-movement. All channels were re-referenced to the common average, i.e. average over all scalp channels. Horizontal and vertical electro-oculograms were estimated by subtracting AF7 from AF8, and AFz from FPz for later identification and correction of eye-induced artifacts. At this point, noisy epochs with non stereotypical patterns of activity (as opposed to identified non-neural artifactual activity) were excluded through visual inspection. Within each individual data set, the number of components to keep for subsequent independent component analysis (ICA³⁷) was obtained through a principal component analysis, keeping only the first components contributing up to 99% of signal variance. ICA was conducted to identify artifactual components before automatic rejection using the EEG artifact detector ADJUST³⁸. Remaining noisy

epochs were excluded by visual inspection. Previously excluded channels were re-interpolated from surrounding channels using spherical splines³⁹.

To avoid spurious effects of baseline-correction, we only subtracted the average signal per subject and channel over the window from 700 to 200 ms before the movement onset. Finally, to get rid of the remaining noisy trials, we excluded the first and last percentiles of trials by individual, in terms of maximum amplitude.

Among the aforementioned exclusion of five patients for excessive artifactual EEG activity, one patient was excluded because automatic ICA rejection failed to get rid of artifactual components (170/277 trials with regular bursts of voltage amplitude remaining after ICA-based rejection). One patient was excluded because 180 trials and 9 electrodes were rejected after visual inspection. Three patients were excluded for having more than 10 channels with a variance exceeding what was found in the pool of participants by two standard deviations.

EEG data analysis. Voltage amplitude was analyzed with linear mixed-effects regressions using R⁴⁰ together with the 'lme4'⁴¹ and 'lmerTest' packages⁴². This method allows analyzing single trial data, with no averaging across conditions or participants, and no discretization of confidence ratings⁴³. Models were applied to each time sample and electrode for individual trials, to explain broadband EEG amplitude with group and correctness (resp. confidence) as fixed effects, and a random intercept per participant. False discovery rate (FDR) -correction⁴⁴ for multiple testing was applied to adjust *p*-values using the built-in R package 'stats'. Bayesian mixed-effects regressions with full random-effects structures were fitted using 'brms' R package⁴⁵.

Analyses were conducted in three steps: 1) we searched for responsive electrodes, 2) we determined the duration and amplitude of the effect at the level of the cluster of responsive electrodes, and 3) we characterized the robustness of these effects by computing evidence ratios at the cluster level (i.e. the ratio of the evidence supporting the hypothesized direction of the effect, over evidence supporting the non-hypothesized direction of the effect).

1. Search for responsive electrodes: For each time sample, electrode and independent variable of interest (i.e., correctness and confidence), we identified significant effect on voltage amplitude (FDR-corrected) within a time window from 0 to 500 ms post-movement onset with the following mixed-effects linear regressions:

$$\text{amplitude} \sim \text{correctness} * \text{group} + (1|\text{participant})$$

$$\text{amplitude} \sim \text{confidence} * \text{group} + (1|\text{participant})$$

Of note, random slopes were not added at this step as they resulted in convergence failures.

The behavioral result of a link between movement onset times and confidence ratings invited us to explore how the ERN-like ERP was modulated by movement onsets with the additional model:

$$\text{amplitude} \sim \text{movementRT} * \text{group} + (1|\text{participant})$$

where movementRT refers to movement onset times.

2. Cluster analyses: Building on the literature on ERN, we focused only on central and fronto-central electrodes as regions of interest. To avoid redundant analyses performed on each electrode separately (which are spatially close to each other), and to gain statistical power, we conducted mixed-effects linear regression restrained to the electrodes selected at step 1 which fell within our scalp regions of interest. Regressions were performed at each time sample, taking participants as random intercepts, with electrode nested within participants:

$$\text{amplitude} \sim \text{correctness} * \text{group} + (1|\text{participant}/\text{electrode}) \quad (1)$$

$$\text{amplitude} \sim \text{z_confidence} * \text{group} + \text{z_movementRT} + (1|\text{participant}/\text{electrode}) \quad (2)$$

$$\text{amplitude} \sim \text{z_movementRT} * \text{group} + \text{z_confidence} + (1|\text{participant}/\text{electrode}) \quad (3)$$

where z_confidence is standardized confidence ratings, and z_movementRT is standardized movement onset times for each participant. Since response times are known to correlate with confidence, z_movementRT was added in model (2), and z_confidence is added in model (3) as covariables of non-interest.

Of note, random slopes were not added at this step as they resulted in convergence failures. FDR-correction was applied on the resulting p-values. Only periods with significant adjacent samples extending over more than 50 ms were considered genuine effects. The voltage amplitude of the effect of correctness was computed as the average difference between correct and incorrect trials over a 50 ms window centered on the peak of the main effect of correctness. The voltage amplitude of the effect of confidence was computed as the average difference between 'Very sure' and 'Unsure' tertiles, over a 50 ms window centered on the peak of the main effect of confidence.

3. Bayesian analyses: An evidence ratio was computed on the averaged voltage amplitude over each significant spatio-

temporal cluster found in step 2:

$$\text{amplitude} \sim \text{correctness} * \text{group} + (\text{correctness}|\text{participant}/\text{electrode}) \quad (4)$$

$$\text{amplitude} \sim \text{z_confidence} * \text{group} * \text{z_movementRT} + (\text{z_confidence} * \text{z_movementRT}|\text{participant}/\text{electrode}) \quad (5)$$

Bayesian models were created in Stan computational framework (<http://mc-stan.org/>) accessed with the brms package, based on four chains of 2000 iterations including 1000 warmup samples.

For model (4) we made assumptions leading to the following prior specifications: 1) we assumed that voltage amplitude would be higher for correct versus incorrect responses with a mildly informative Gaussian prior (Mean = 1, SD = 1); 2) we assumed no difference in voltage amplitude between groups with a mildly informative Gaussian prior (Mean = 0, SD = 1); 3) we assumed a blunted ERN among patients with SSD, leading to an interaction effect of group x correctness on voltage amplitude (Gaussian prior Mean = -1, SD = 1). Other priors were by default according to the brms package in R.

For model (5) we specified the following priors: 1) we assumed that voltage amplitude would be higher (resp. lower) for higher (resp. lower) confidence ratings with a mildly informative Gaussian prior (Mean = 0.5, SD = 1); 2) we assumed no difference of voltage amplitude between groups with a mildly informative prior (Mean = 0, SD = 1); 3) Because we expected a compensatory electrophysiological signal related to confidence among patients with SSD, we assumed an increased confidence ERP among patients with SSD, leading to an interaction effect of group x confidence on voltage amplitude (Gaussian prior Mean = 0.5, SD = 1); 4) We assumed a decreasing amplitude as a function of movement onset times (Mean = -1, SD = 1); and 5) we assumed an interaction effect on amplitude between movement onset times and groups, such that movement onset times from patients with SSD would correlate less with voltage amplitude compared to healthy controls (Mean = 0.5, SD = 1). Other priors were by default according to the brms package in R.

Operationalised hypotheses were as follows:

1. The presence of an ERN is indicated by a significant main effect of correctness over frontocentral sites following movement onset,
2. The presence of ERN-blunting is indicated by a correctness x group interaction within the above mentioned cluster of electrodes, characterized by a dampened difference of voltage amplitude between correct and incorrect trials among individuals with SSD compared to healthy controls,
3. The presence of confidence-related compensatory mechanisms in patients is indicated by a confidence x group interaction, characterized by an increased difference of voltage amplitude between confidence levels in patients compared to healthy controls. We did not expect any particular localization or time-window for this effect.

All analysis scripts and data (behavioral and EEG) are publicly available (https://gitlab.com/nfaivre/meta_scz_public).

RESULTS

Demographic and neuropsychological variables

Participants with SSD and healthy controls had similar age, education level and premorbid IQ (see Table 1). However,

Table 1. Clinical and neuropsychological characteristics of our sample of participants.

	Control, M ± 95% CI (N = 19)	Schizophrenia, M ± 95% CI (N = 14)	t-statistic	p-value	Bayes factor
Age, yr	43.8 ± 5.0	38.3 ± 6.5	1.33	0.196	0.68
Education Level, yr	12.4 ± 0.4	13.0 ± 1.5	-0.69	0.502	0.43
BCIS, self-reflectiveness score	8.6 ± 1.3	16.0 ± 2.4	-5.33	<0.001	3531.51
BCIS, self-certainty score	9.5 ± 1.2	10.3 ± 2.5	-0.56	0.585	0.39
BCIS, composite score	-0.9 ± 1.7	5.7 ± 4.3	-2.81	0.012	9.10
Calgary Depression Scale, score	0.5 ± 0.5	4.4 ± 2.5	-3.01	0.009	17.84
Premordid IQ	104.2 ± 4.0	101.8 ± 4.6	0.80	0.431	0.43
WAIS Matrix subtest, score	10.1 ± 1.1	8.8 ± 1.3	1.43	0.165	0.74
WAIS letter-number sequencing subtest, score	10.0 ± 1.4	10.7 ± 1.7	-0.65	0.524	0.40
WAIS vocabulary subtest, score	8.9 ± 1.3	7.5 ± 1.2	1.62	0.115	0.88
Illness duration, yr		13.7 ± 4.2			
PANSS positive, score		16.6 ± 2.2			
PANSS negative, score		20.1 ± 2.8			
PANSS total, score		76.3 ± 8.9			
Chlorpromazine equivalent, mg		419.1 ± 159.4			

Bayes Factor >3 (resp. <0.33) indicates moderate evidence for H1 (resp. for H0).

CI Confidence Interval, BCIS Beck Cognitive Insight Scale, IQ Intelligence Quotient, WAIS Wechsler Adult Intelligence Scale, PANSS Positive And Negative Syndrome Scale.

dividuals with SSD were more depressed, with higher levels of cognitive insight (composite score) than healthy control participants, explained by a higher self-reflectiveness score.

Behavioral results

Most task-related cognitive variables were comparable between groups. In particular, both groups had similar accuracy levels (SSD: Mean = 0.73, SD = 0.03; controls: Mean = 0.74, SD = 0.02; difference between groups: $t = 1.22$, $p = 0.23$, BF = 0.61), which indicated that the staircase procedure was successful in adapting perceptual difficulty (motion variance among SSD: Mean = 1.53, SD = 0.37; controls: Mean = 2.04, SD = 0.40; difference between groups: $t = 3.75$, $p < 0.01$, BF = 35.8), with very low dispersion in task performance between participants. There was no difference in average confidence between groups (SSD: Mean = 0.71, SD = 0.12; controls: Mean = 0.71, SD = 0.14; difference between groups: $t = 0.0$, $p = 1$, BF = 0.34), indicating no confidence bias, and confidence ratings' variability was comparable between groups (SSD: Mean = 0.14, SD = 0.05; controls: Mean = 0.16, SD = 0.05; difference between groups: $t = 0.72$, $p = 0.48$, BF = 0.41). Furthermore, there was no difference in movement onsets between groups, neither for correct trials (SSD: Mean = 1.23 s, SD = 0.72; controls: Mean = 1.32 s, SD = 0.66 s, difference between groups: $t = 0.38$, $p = 0.71$, BF = 0.36) nor for incorrect trials (SSD: Mean = 1.38 s, SD = 0.84; controls: Mean = 1.53 s, SD = 0.73 s, difference between groups: $t = 0.53$, $p = 0.60$, BF = 0.38). However, there was a response side bias towards the left in patients with SSD, yet with inconclusive evidence given the BF < 3 (SSD: Mean = -0.32, SD = 0.32; controls: Mean = 0.02, SD = 0.43; difference between groups: $t = 2.56$, $p < 0.05$, BF = 2.98).

Concerning metacognitive performance, the Bayesian hierarchical model provided moderate evidence for an absence of difference between the two groups in terms of M-ratio (SSD: Mean = 0.60, 95% highest posterior density interval [95% HDI] = [0.48, 0.74]; controls: Mean = 0.55, 95% HDI = [0.41, 0.70], BF = 0.20), indicating no metacognitive deficits in our sample of participants with SSD (Fig. 1B).

Next, we investigated the relationship between confidence and movement onset. We found a negative relationship between

confidence and standardized movement onset (estimate = -0.04 [-0.05 to -0.02]; evidence ratio > 4000), indicating that confidence was higher following earlier movements. This relationship was modulated by the correctness of the responses (interaction correctness × movement onset: estimate = -0.01 [-0.02 to 0.00], evidence ratio = 124) indicating steeper slopes for correct responses, and by the group (interaction group × movement onset: estimate = 0.03 [0.01 to 0.05]; evidence ratio = 67) indicating that confidence ratings were less correlated with movement onset in participants with SSD, compared to healthy controls (Fig. 1C). Yet, the relationship between confidence and movement onset did not interact with correctness × group (interaction correctness × group × movement onset: estimate = 0.00 [-0.02, 0.01], evidence ratio = 1.83). Together, these results suggest that in this subsample of patients also, movement onsets are less predictive of confidence than among healthy controls, irrespective of correctness.

EEG analysis

Effect of correctness. Electrodes Cz, C1 and C2 responded significantly to response correctness within the 0–500 ms post-movement window, and were thus selected for further analyses. At the cluster level (Cz, C1, C2), there was a main effect of response-correctness (effect of correctness = -0.34 ± 0.55 μV, estimate = 0.05 ± 95% CI [0.02, 0.08], evidence ratio = 165) starting 10 ms and until 330 ms after movement onset (Fig. 2). The peak of the effect occurred 266 ms after movement onset. However, there was no effect of group (estimate = 0.04 ± 95% CI [-0.13, 0.21], BF₀₁ = 10.81), nor a correctness × group interaction effect (estimate = 0.01 ± 95% CI [-0.05, 0.06], BF₀₁ = 56.47), providing evidence for an absence of blunted ERN in patients. The qualitative topographical differences seen in Fig. 2 did not reach significance.

Effect of confidence, for correct trials only. To test for a specific effect of confidence irrespective of task performance, we analyzed EEG signals as a function of confidence among correct trials only (thus including 73.7% ± 2.25% of trials in controls and 72.7% ± 2.49% of trials among patients in the analysis, $t = -1.10$,

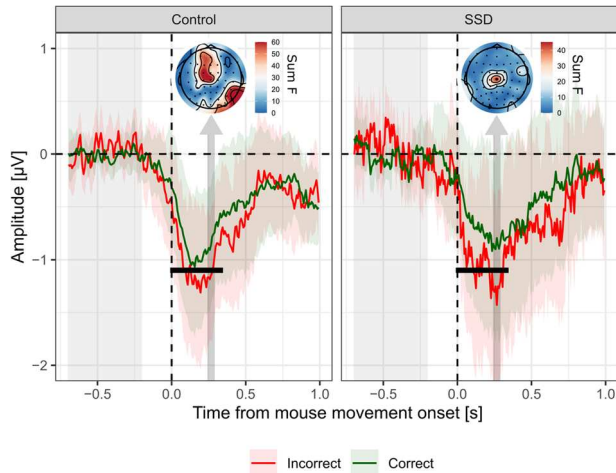


Fig. 2 ERP locked on mouse movement onset. Average signal amplitude from central electrodes (Cz, C1, C2) for incorrect responses (red), and correct responses (green). Shaded-areas represent 95% confidence intervals. Light gray shaded area on the left indicates the baseline correction window. Dark gray vertical arrows indicate the 50 ms window centered on the peak of the main effect, pointing at the corresponding topographies, scaled according to the magnitude of the main effect of correctness (summed F-values over the 50 ms window). Black horizontal lines indicate significant adjacent samples with a significant main effect of correctness following FDR correction.

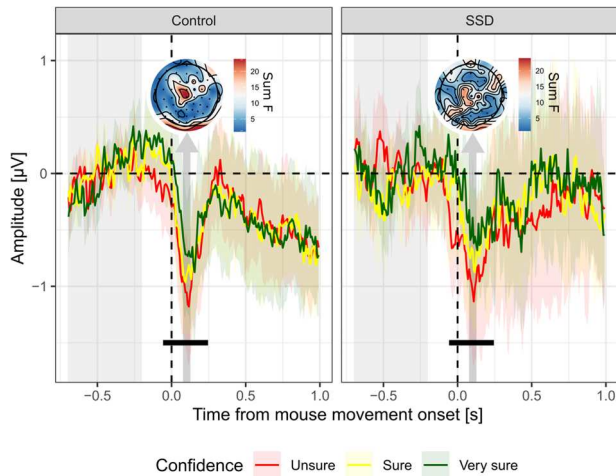


Fig. 3 ERP locked on mouse movement onset. Average signal amplitude from fronto-central electrodes (F3, F1, Fz, F2, FC3, FC1, FCz, FC2, C1, Cz, C2) on tertiles of confidence (Unsure, Sure and Very sure trials are plotted in red, yellow and green, respectively). Note that although the graphical representation is based on confidence tertiles, statistical models considered raw continuous confidence ratings. Shaded-areas represent 95% confidence intervals. Light gray shaded area on the left indicates the baseline correction window. Vertical arrows indicate the 50 ms windows centered on the peaks of the main and interaction effects (dark gray and blue, respectively), pointing at the corresponding topography, scaled according to the magnitude of the main and interaction effects of confidence (summed F-values over the 50 ms window). Black lines indicate significant adjacent samples (main and interaction effect of confidence, respectively), following FDR-correction.

$p = 0.28$, $BF_{10} = 0.54$).

Electrodes Cz, C1, C2, FCz, FC1, FC2, FC3, Fz, F1, F2, F3 responded significantly to confidence ratings within the 0–500 ms post-movement window, and were thus selected for further

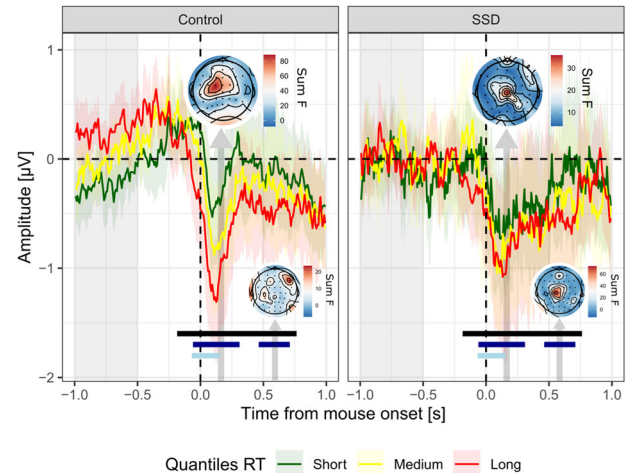


Fig. 4 ERP locked on mouse movement onset. Average signal amplitude from frontocentral electrodes (F1, Fz, FC3, FC1, FCz, FC2, C3, C1, Cz, C2) on tertiles of response times (Short, Medium and Long trials are plotted in green, yellow and red, respectively). Note that although the graphical representation is based on tertiles of movement onset times, statistical models considered continuous response times. Shaded-areas represent 95% confidence intervals. Light gray shaded area on the left indicates the baseline correction window. Vertical arrows indicate the 50 ms windows centered on the peaks of the main and interaction effects (dark gray and blue, respectively), pointing at the corresponding topography (N.B.: Occipital electrodes have been deliberately removed from the interaction topography, to focus on the activity of the ROI), scaled according to the magnitude of the main and interaction effects of response times (summed F-values over the 50 ms window). Black, dark blue, and light blue lines indicate significant adjacent samples (main and interaction effect of movement onsets, and main effect of confidence, respectively), following FDR-correction.

analyses. At the cluster level, there was a main effect of confidence peaking 102 ms after movement onset (effect of confidence = $0.41 \pm 0.61 \mu\text{V}$, estimate = $0.07 \pm 95\% \text{ CI} [-0.01, 0.15]$, evidence ratio = 12.5) and ranging from -40 to 227 ms after movement onset (see Fig. 3). There was no group difference in voltage amplitude (estimate = $0.02 \pm 95\% \text{ CI} [-0.56, 0.61]$, $BF_{01} = 3.68$), neither there was an interaction effect between z_confidence and group within this cluster (estimate = $-0.02 \pm 95\% \text{ CI} [-0.16, 0.12]$, $BF_{01} = 14.8$). The qualitative topographical differences seen in Fig. 3 did not reach significance.

Overall, these EEG results are consistent with our previous behavioral results. The comparable ERN-like ERPs observed in both groups (Fig. 2) reflect the comparable correctness rates measured behaviorally, and the comparable confidence-related ERPs (Fig. 3) reflect the similarity in confidence mean and metacognitive efficiency between groups.

Effect of movement onset. As a final analysis step, we sought to investigate the relationship between response times and voltage amplitude, since response times were found to be less correlated with confidence in individuals with SSD at the behavioral level (Fig. 1.C). Electrodes Cz, C1, C2, C3, FCz, FC1, FC2, FC3, Fz, F1 responded significantly to response times and were thus selected for further analyses. At the cluster-level, there was a main effect of response times peaking 133 ms after movement onset (effect of movement onset = $0.61 \pm 0.95 \mu\text{V}$, evidence ratio = 12000) and ranging from -164 to 742 ms after movement onset (Fig. 4). Interestingly, an interaction effect between movement onset times and groups was found in two time clusters (depicted with dark blue lines in Fig. 4): the first interaction cluster ranged from -39 to 289 ms after movement onset, with moderate evidence (evidence ratio = 5.3).

This interaction cluster was characterized by a lesser relationship between EEG amplitude and movement onset times among individuals with SSD (Effect of movement onset = $0.33 \pm 0.76 \mu\text{V}$, see topography in Fig. 4), compared to control participants (Effect of movement onset = $0.67 \pm 0.71 \mu\text{V}$). The second interaction cluster ranged from 484 to 688 ms after movement onset, with moderate evidence (evidence ratio in favor of the alternative hypothesis = 7.1). This second interaction cluster described the opposite pattern compared to the first one: it was characterized by a steeper increase in voltage as a function of movement onset times among individuals with SSD (Effect of movement onset = $0.38 \pm 0.64 \mu\text{V}$, see topography in Fig. 4), compared to control participants (Effect of movement onset = $0.34 \pm 0.74 \mu\text{V}$). Furthermore, and in line with the previous analysis on confidence, there was a significant effect of the confidence covariate (depicted as light blue line in Fig. 4) from -47 to 133 ms (evidence ratio = 12.5).

DISCUSSION

In the present study, we sought to investigate EEG data recorded on patients with SSD while they performed a visual discrimination task followed by a confidence rating task²¹. Building on the literature on ERN among individuals with SSD⁹ we expected the ERN-like ERP to be blunted in the group of patients, indicating a performance monitoring deficit under matched levels of task performance between individuals of each group. Then, to make sense of the preserved metacognitive abilities at the behavioral level despite an anticipated performance monitoring deficit among patients with SSD, we hypothesized a distinctive confidence-related ERP among patients, which would constitute evidence for the existence of a compensatory mechanism helping them to provide confidence ratings that are as accurate as those provided by control participants.

We found a negative ERP over frontocentral electrodes that was larger for errors in both groups. Although the peak of this effect occurred later (266 ms) than typical ERN obtained with standard response-conflict tasks, it started 10 ms after the movement onset, consistent with the literature. This difference might be attributed to the fact that we time-locked our analysis onto the initiation of the mouse movement, which might have led to a slightly larger temporal spreading of the ERP. This has the advantage of capturing the very beginning of the decisional process instead of its end-point indicated by a button press^{46,47}. However, this may be less precise as the definition of a movement onset is temporally more ambiguous than a button press.

In line with our behavioral results, but contrary to our initial hypothesis, EEG analyses revealed evidence for unaltered neural correlates of confidence among individuals with SSD, which is consistent with the absence of a confidence bias as well as comparable variability in confidence ratings we found behaviorally between the two groups. Future research efforts with more sensitive measures and bigger sample sizes are necessary to consolidate our conclusion. We argue that such research efforts should consider matching performance experimentally between individuals of each group as comparable performance between groups may in itself not be sufficient to disambiguate ERN-blunting from poorer task performance among individuals with SSD. This argument is supported by the finding from Van der Borgh et al.²⁰ showing that ERN decreases with task-difficulty among healthy participants.

Another reason why no evidence in favor of altered correlates of confidence was found in the present study might be that the ERPs we measured are sensitive to the type of task, i.e. motion discrimination versus response-conflict tasks as commonly employed in the literature¹³. Indeed, previous studies on ERN used speeded response-conflict tasks during which participants had to suppress a prepotent response. Here, we were interested in studying

errors that arise from the slow accumulation of incorrect noisy sensory information, which are not detected as errors, by giving participants enough time to respond (6 s). It might be that the ERN-blunting is specific to fast errors committed in response-conflict tasks, which are known to involve specific mechanisms both in the memory⁴⁸ and perceptual domains^{24,25}. The difference between fast versus slow errors may be considered as involving "slips" versus "mistakes", a terminology proposed by Reason²³. A speeded context increases the proportion of so-called "slips" - a category of errors corresponding to "incorrect executions of appropriate motor programs"²² - as opposed to "mistakes", reflecting inaccurate intentions due to erroneous knowledge. Slips are therefore obvious errors due to participants executing the wrong motor command (pressing A and soon realizing they meant to press B). However, in a non-speeded context, errors are more likely to result from "mistakes" rather than "slips". In sum, the ERN-blunting in speeded experiments might reflect a specific impairment regarding the monitoring of fast errors or slips, whereas the absence of ERN-blunting in the present non-speeded task might be evidence for a preserved monitoring of genuine mistakes reflected by "slow errors". Finally, the only notable behavioral difference between the two groups was that confidence was less predicted by response times among patients with SSD (see Fig. 1D). Now extending this aspect to EEG, we found that voltage amplitudes were distinctively modulated by movement onset times among patients, within two spatio-temporal clusters (Fig. 4). At first sight, it appears striking that patients with SSD have comparable average response times, confidence levels, metacognitive abilities, and yet lower correlations between confidence ratings and response times compared to healthy controls. Zheng and colleagues⁴⁹ have also observed a lower correlation between confidence and response times among patients with SSD in a metamemory task, and their results suggest that this pattern is partly explained by a stronger reliance on previous confidence ratings (called confidence history) for the estimation of confidence in the current trial compared to healthy controls. However, when conducting the same analysis on our data, we could not find a stronger correlation between confidence and confidence history among patients compared to healthy controls, indicating that the result obtained by Zheng and colleagues did not extend to our perceptual task (Supplementary Fig. S1). Although speculative at this stage, one possibility would be that patients rely more on internal evidence than on additional cues such as response times (e.g. ref. ⁵⁰) or sensorimotor cues^{51,52} to rate their confidence.

Of note, our sample of patients with SSD was more depressed than healthy control participants. Interestingly, depression is known to enhance the amplitude of the ERN (for a review, see ref. ⁵³), and one might think that it could have compensated for the ERN blunting. Since depression is a very frequent comorbidity in schizophrenia (~50%, for a review see ref. ⁵⁴), this confound is most likely present in every ERN study including individuals with SSD, even though it is usually not discussed. Thus, depression is not sufficient to explain the absence of ERN blunting in our sample. However, depression may be sufficient to explain higher levels of self-reflectiveness (cognitive insight) among patients. Indeed, the link between depression and increased insight is now well established⁵⁵ and constitutes the "insight paradox"⁵, namely that improved insight degrades patients' quality of life. Yet, the relationship between cognitive insight and behavioral measures of metacognition is still unclear⁵⁶. Future research experiments should assess the degree of correlation between insight and behavioral metacognition.

To sum up, we propose two alternative interpretations to explain why we found no evidence for altered neural correlates of performance monitoring among individuals with SSD: 1) Such alterations are confounded with altered task-performance in patients with SSD and are not observed anymore when task-performance is experimentally matched between individuals of each group; 2) Such alterations are specific to "fast errors" committed in response-conflict tasks, which would suggest a

specific impairment to detect fast errors or suppress prepotent responses among individuals with SSD (Morey and Rouder⁵⁷ and Addington et al.⁵⁸).

CONCLUSION

In our sample of individuals with SSD showing no metacognitive deficit at the behavioral level, we found evidence for an absence of deficits in performance-monitoring at the electrophysiological level. Larger scale studies assessing distinct types of errors while finely controlling for task performance are needed to better understand performance monitoring in SSD.

DATA AVAILABILITY

All analysis scripts and data (behavioral and EEG) are publicly available (https://gitlab.com/nfaivre/meta_scz_public).

Received: 20 October 2022; Accepted: 19 January 2023;

Published online: 23 February 2023

REFERENCES

- Gopal, Y. V. & Variend, H. First-episode schizophrenia: review of cognitive deficits and cognitive remediation. *Advan. Psychiatric Treatment* **11**, 38–44 (2005).
- Schaefer, J., Giangrande, E., Weinberger, D. R. & Dickinson, D. The global cognitive impairment in schizophrenia: consistent over decades and around the world. *Schizophr. Res.* **150**, 42–50 (2013).
- Lysaker, P. H. et al. Social cognition and metacognition in schizophrenia: evidence of their independence and linkage with outcomes. *Acta Psychiatrica Scand.* **127**, 239–247 (2013).
- Lysaker, P. H. et al. Social dysfunction in psychosis is more than a matter of misperception: advances from the study of metacognition. *Front. Psychol.* **12**, 723952 (2021).
- Davis, B. J., Lysaker, P. H., Salyers, M. P. & Minor, K. S. The insight paradox in schizophrenia: a meta-analysis of the relationship between clinical insight and quality of life. *Schizophr. Res.* **223**, 9–17 (2020).
- Hasson-Ohayon, I. et al. Metacognitive and social cognition approaches to understanding the impact of schizophrenia on social quality of life. *Schizophr. Res.* **161**, 386–391 (2015).
- Hoven, M. et al. Abnormalities of confidence in psychiatry: an overview and future perspectives. *Transl. Psychiatry* **9**, 268 (2019).
- Rouy, M., et al. Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders. *Neurosci. Biobehav. Rev.* <https://doi.org/10.1016/j.neubiorev.2021.03.017> (2021).
- Kirschner, H. & Klein, T. A. Beyond a blunted ERN - biobehavioral correlates of performance monitoring in schizophrenia. *Neurosci. Biobehav. Rev.* **133**, 104504 (2022).
- Falkenstein, M. Effects of errors in choice reaction time tasks on the ERP under focussed and divided attention in Brunia C H M, Gallard A W K, Kok A. (Eds), *Psychophysiol. Brain Res.* (pp 192–195) Tilburg, The Netherlands Tillburg University Press (1990).
- Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E. & Donchin, E. A neural system for error detection and compensation. *Psychol. Sci.* **4**, 385–390 (1993).
- Rausch, M., Zehetleitner, M., Steinhauser, M. & Maier, M. E. Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage* **218**, 116963 (2020).
- Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In *The Oxford handbook of event-related potential components* (eds. Luck, S. J. & Kappenman, E. S.) 231–291 (Oxford University Press, 2012).
- Ullsperger, M., Fischer, A. G., Nigbur, R. & Endrass, T. Neural mechanisms and temporal dynamics of performance monitoring. *Trends Cognit. Sci.* **18**, 259–267 (2014).
- Vidal, F., Burle, B. & Hasbroucq, T. Errors and action monitoring: errare humanum est sed corrigere possibile. *Front. Hum. Neurosci.* **13**, 453 (2020).
- Scheffers, M. K. & Coles, M. G. H. Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 141–151 (2000).
- Boldt, A. & Yeung, N. Shared neural markers of decision confidence and error detection. *J. Neurosci.* **35**, 3478–3484 (2015).
- Murphy, P. R., Robertson, I. H., Allen, D., Hester, R., & O'Connell, R. G. An electrophysiological signal that precisely tracks the emergence of error awareness. *Front. Hum. Neurosci.* **6**, <https://doi.org/10.3389/fnhum.2012.00065> (2012).
- Martin, E. A. et al. ERP indices of performance monitoring and feedback processing in psychosis: a meta-analysis. *Int. J. Psychophysiol.* **132**, 365–378 (2018).
- Van der Borgh, L., Houtman, F., Burle, B. & Notebaert, W. Distinguishing the influence of task difficulty on error-related ERPs using surface Laplacian transformation'. *Biological Psychol.* **115**, 78–85 (2016).
- Favre, N. et al. Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *J. Psychiatry Neurosci.* **46**, E65–E73 (2021).
- Dehaene, S., Posner, M. I. & Tucker, D. M. Localization of a neural system for error detection and compensation. *Psychol. Sci.* **5**, 303–305 (1994).
- Reason, J. T. *Human error* (Cambridge University Press, New York, 1990).
- Desender, K., Ridderinkhof, K. R. & Murphy, P. R. Understanding neural signals of post-decisional performance monitoring: an integrative review. *ELife* **10**, e67556 (2021).
- Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychol. Sci.* **9**, 347–356 (1998).
- Wechsler, D., Coalson, D. L., Raiford, S. E. WAIS-IV: Wechsler adult intelligence scale (Pearson San Antonio, TX, 2008).
- Addington, D., Addington, J., Maticka-Tyndale, E. & Joyce, J. Reliability and validity of a depression rating scale for schizophrenics. *Schizophr. Res.* **6**, 201–208 (1992).
- Beck, A. A new instrument for measuring insight: the Beck cognitive insight scale. *Schizophr. Res.* **68**, 319–329 (2004).
- Nelson, H. E. & O'Connell, A. Dementia: the estimation of premorbid intelligence levels using the new adult reading test. *Cortex* **14**, 234–244 (1978).
- Kay, S. R., Fiszbein, A. & Opler, L. A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* **13**, 261–276 (1987).
- Weiden, P. J. EPS profiles: the atypical antipsychotics: are not all the same. *J. Psychiatric Pract.* **13**, 13–24 (2007).
- Osborne, K. J., Walther, S., Shankman, S. A. & Mittal, V. A. Psychomotor slowing in schizophrenia: implications for endophenotype and biomarker development. *Biomarkers Neuropsychiatry* **2**, 100016, <https://doi.org/10.1016/j.bionps.2020.100016> (2020).
- Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Human Neurosci.* **8**, <https://doi.org/10.3389/fnhum.2014.00443> (2014).
- Fleming, S. M. HMeta-d: hierarchical bayesian estimation of metacognitive efficiency from confidence ratings. *Neurosci. Conscious.* <https://doi.org/10.1093/nc/nix007> (2017).
- MATLAB. *9.7.0.1471314 (R2019b)* (The MathWorks Inc., Natick, Massachusetts, 2019).
- Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
- Delorme, A., Sejnowski, T. & Makeig, S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage* **34**, 1443–1449 (2007).
- Mognon, A., Jovicich, J., Bruzzone, L. & Buiatti, M. ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features: automatic spatio-temporal EEG artifact detection. *Psychophysiology* **48**, 229–240 (2011).
- Perrin, F., Pernier, J., Bertrand, O. & Echallier, J. F. Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol.* **72**, 184–187 (1989).
- R Core Team. *R: a language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2020), <https://www.R-project.org/>.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. Fitting linear mixed-effects models using lme4. <https://doi.org/10.48550/ARXIV.1406.5823> (2014).
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, <https://doi.org/10.18637/jss.v082.i13> (2017).
- Bagiella, E., Sloan, R. P. & Heitjan, D. F. Mixed-effects models in psychophysiology. *Psychophysiology* **37**, 13–20 (2000).
- Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
- Bürkner, P.-C. Brms: an R package for bayesian multilevel models using stan. *J. Stat. Softw.* **80**, <https://doi.org/10.18637/jss.v080.i01> (2017).
- Pereira, M., Sobolewski, A. & Millán, J. D. R. Action monitoring cortical activity coupled to submovements. *ENEURO*.0241-17.2017 (2017).
- Tafuro, A., Vallesi, A. & Ambrosini, E. Cognitive brakes in interference resolution: a mouse-tracking and EEG co-registration study. *Cortex* **133**, 188–200 (2020).
- Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59 (1978).
- Zheng, Y. et al. Atypical meta-memory evaluation strategy in schizophrenia patients. *Schizophr. Res. Cognit.* **27**, 100220 (2022).
- Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
- Favre, N. et al. Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex* **124**, 224–234 (2020).

52. Pereira, M. et al. Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proc. Natl Acad. Sci.* **117**, 8382–8390 (2020).
53. Bruder, G. E., Kayser, J., & Tenke, C. E. Event-related brain potentials in depression: clinical, cognitive, and neurophysiological implications. In *The Oxford handbook of event-related potential components* (eds Luck, S. J. & Kappenman, E. S.) pp. 563–592 (Oxford University Press, 2012).
54. Buckley, P. F., Miller, B. J., Lehrer, D. S. & Castle, D. J. Psychiatric comorbidities and schizophrenia. *Schizophr. Bull.* **35**, 383–402 (2009).
55. Murri, M. B. et al. Is good insight associated with depression among patients with schizophrenia? Systematic review and meta-analysis. *Schizophr. Res.* **162**, 234–247 (2015).
56. David, A. S. Insight and psychosis: the next 30 years. *Brit. J. Psychiatry* **217**, 521–523 (2020).
57. Morey, R. D. & Rouder, J. N. BayesFactor: computation of bayes factors for common designs. R package version 0.9.12-4.2. <https://CRAN.R-project.org/package=BayesFactor> (2018).
58. Addington, D., Addington, J. & Maticka-tyndale, E. Assessing depression in schizophrenia: the Calgary depression scale. *Brit. J. Psychiatry* **163**, 39–44 (1993).

ACKNOWLEDGEMENTS

M.P. is supported by a Postdoc.Mobility fellowship from the Swiss National Science Foundation (P400PM_199251). N.F. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 803122). We thank Vincent de Gardelle and Jean-Christophe Vergnaud for their support during data acquisition.

AUTHOR CONTRIBUTIONS

M.Roger, P.R., and N.F. designed the study and acquired the data. M.Rouy, M.P., D.G., and N.F. analyzed the data. M.Rouy and N.F. wrote the article, which all authors reviewed. All authors approved the final version to be published and can certify that no other individuals not listed as authors have made substantial contributions to the

paper. This work has been presented at the Association for the Scientific Study of Consciousness in Amsterdam.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41537-023-00333-4>.

Correspondence and requests for materials should be addressed to Martin Rouy.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

4. Improving one's metacognition? Assessment of a metacognitive training efficiency

A recent study suggested that training metacognitive skills among healthy individuals is possible, so that metacognitive improvements can be transferred to other cognitive domains (Carpenter et al. 2019). We aimed at assessing the efficiency of this metacognitive training as a potential remediation strategy for metacognitive deficits among individuals with schizophrenia. In the study by Carpenter and colleagues, participants took part in an online longitudinal study consisting of ten sessions, one per day. On the first day, they performed a memory task as well as a perceptual task. From the second to the ninth session, they performed 108 trials of a visual discrimination task followed by a confidence judgment task. The experimental group received feedback on the calibration of their confidence judgments to their performance, whereas the control group received feedback on their accuracy, in the form of monetary bonuses. In the last session, participants performed both the memory and the perceptual tasks as they did in the first session. Results showed higher metacognitive efficiency in the last session compared to the first session in the metacognitive group, both in memory and perceptual tasks, whereas there was no improvement in the control group, providing evidence for a training effect on metacognition.

As we explain in the manuscript below, we identified several potential confounds which could explain these results, including motivation and ambiguous instructions. We sought to replicate this online study while better controlling for these factors. We hypothesized no improvement of the metacognitive efficiency as a result of this metacognitive training.

Status of the manuscript: Published in the Journal of Experimental Psychology: General

Reference: Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J. C., Filevich, E., & Faivre, N. (2022). Metacognitive improvement: Disentangling adaptive training from experimental confounds. *Journal of Experimental Psychology: General*.

Journal of Experimental Psychology: General

Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds

Martin Rouy, Vincent de Gardelle, Gabriel Reyes, Jérôme Sackur, Jean Christophe Vergnaud, Elisa Filevich, and Nathan Faivre

Online First Publication, February 14, 2022. <http://dx.doi.org/10.1037/xge0001185>

CITATION

Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J. C., Filevich, E., & Faivre, N. (2022, February 14). Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds. *Journal of Experimental Psychology: General*. Advance online publication. <http://dx.doi.org/10.1037/xge0001185>

Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds

Martin Rouy¹, Vincent de Gardelle², Gabriel Reyes³, Jérôme Sackur⁴, Jean Christophe Vergnaud⁵,
Elisa Filevich^{6, 7}, and Nathan Faivre¹

¹ Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC

² Paris School of Economics and CNRS

³ Faculty of Psychology, Universidad Del Desarrollo

⁴ Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure, PSL University

⁵ Centre d'Economie de la Sorbonne, Paris, France

⁶ Department of Psychology, Humboldt Universität zu Berlin

⁷ Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany


Metacognition is defined as the capacity to monitor and control one's own cognitive processes. Recently, Carpenter and colleagues (2019) reported that metacognitive performance can be improved through adaptive training: healthy participants performed a perceptual discrimination task, and subsequently indicated confidence in their response. Metacognitive performance, defined as how much information these confidence judgments contain about the accuracy of perceptual decisions, was found to increase in a group of participants receiving monetary reward based on their confidence judgments over hundreds of trials and multiple sessions. By contrast, in a control group where only perceptual performance was incentivized, metacognitive performance remained constant across experimental sessions. We identified two possible confounds that may have led to an artificial increase in metacognitive performance, namely the absence of reward in the initial session and an inconsistency between the reward scheme and the instructions about the confidence scale. We thus conducted a preregistered conceptual replication where all sessions were rewarded and where instructions were consistent with the reward scheme. Critically, once these two confounds were corrected we found moderate evidence for an absence of metacognitive training. Our data thus suggest that previous claims about metacognitive training are premature, and calls for more research on how to train individuals to monitor their own performance.


Keywords: cognitive training, confidence, introspection, metacognition

Metacognition is defined as the capacity to monitor and control one's own cognitive processes (Flavell, 1979; Nelson & Narens, 1994). Metacognitive monitoring is imperfect: Under- or overestimations regarding the accuracy of one's own judgments are frequent, both in healthy individuals (Shekhar & Rahnev, 2021a, 2021b) and in individuals with neurological or

psychiatric disorders (Hoven et al., 2019; Rouy, Saliou, et al., 2021). Thus, one outstanding issue is whether one can design training protocols to help individuals improve their abilities to evaluate their own performances.

Recently, Carpenter and colleagues (2019) proposed that metacognitive abilities can be improved through adaptive training. In

Martin Rouy  <https://orcid.org/0000-0003-4280-4683>

Nathan Faivre  <https://orcid.org/0000-0001-6011-4921>

Elisa Filevich and Nathan Faivre contributed equally.

All authors developed the study concept and contributed to the study design. Modifications in the original code were implemented by Elisa Filevich. Data collection was performed by Martin Rouy. Martin Rouy and Nathan Faivre analyzed data. Martin Rouy and Nathan Faivre drafted the article; all authors provided critical revisions and approved the final version of the article for submission. The authors declare no competing interests.

Nathan Faivre has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 803122). Elisa Filevich was supported by a Freigeist

Fellowship from the Volkswagen Foundation (Grant 91620). Jérôme Sackur received support from the Agence Nationale de la Recherche, ANR-17-EURE-0017. We thank Steve Fleming for sharing the materials of the original study and commenting on a first version of this article.

This work has been presented at the occasion of the 24th annual meeting of the Association for the Scientific Study of Consciousness.

Preregistration is publicly available: <https://osf.io/gak2t>.

Data and analysis scripts are publicly available: <https://doi.org/10.17605/OSF.IO/RQ967> (Rouy, de Gardelle, et al., 2021).

Correspondence concerning this article should be addressed to Martin Rouy, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC UMR 5105 UGA BSHM, 1251 Avenue Centrale, 38058 Grenoble Cedex 9, France. Email: martinrouy03@gmail.com

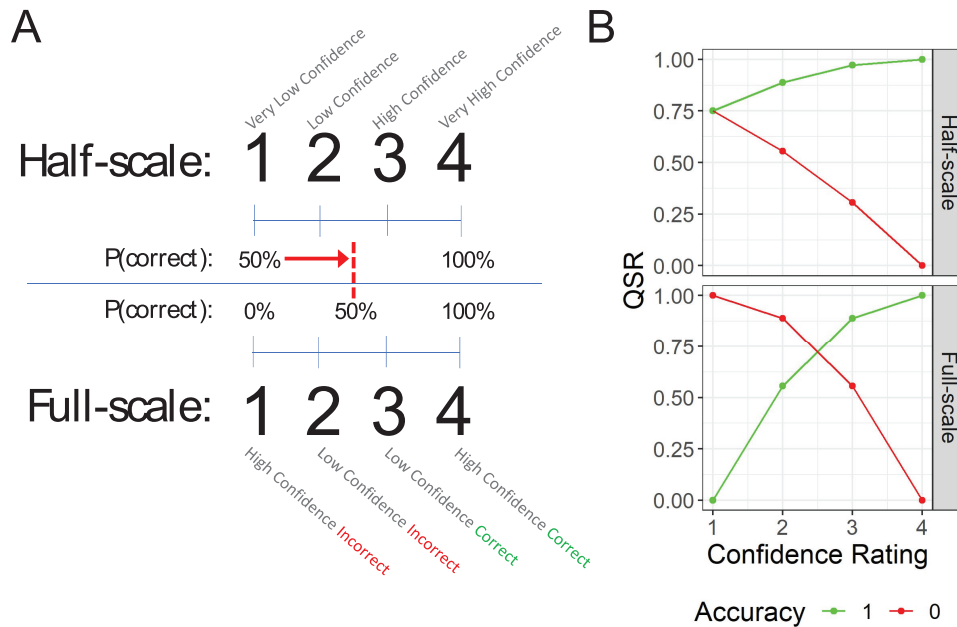
their study, healthy participants were asked to perform both a memory and a perceptual discrimination task, either with shapes or words stimuli, and subsequently report their confidence in their response. They used a longitudinal protocol in 10 sessions (see Figure 2A), where the first session (S1, or pretraining session) served as a baseline, followed by eight sessions of training (S2–S9) on the perceptual task, and finally a posttraining session (S10). In the training sessions, participants received feedback and monetary reward on the basis of their confidence evaluations, after each block of 27 trials: the better the confidence ratings reflected perceptual accuracy in that block, the higher the reward. The pretraining and posttraining sessions had no feedback.

Importantly, Carpenter and colleagues reported that metacognitive efficiency, defined as the adequacy between task performance and confidence, increased between pre- and posttraining sessions in the experimental group where participants received monetary reward on their metacognitive performance, but remained constant in a control group rewarded on their perceptual performance.

In their article, Carpenter et al. argued that the increase in metacognitive efficiency that they observed in the posttraining session (S10) was mediated by an increase in overall confidence between the pretraining session (S1) and the following session (S2) occurring only in the experimental group. A close inspection of these results reveal that confidence indeed sharply increased from the very beginning of S2, and remained constant afterward. Likewise, metacognitive performance increased between the pretraining session and S2 but remained constant from S2 onward. This sudden increase in confidence and metacognitive performance suggests that they might have occurred due to factors other than training.

We identified two potential confounding factors which we thought could lead to apparent increases in metacognitive efficiency, without involving a real improvement as a result of training. First, because no reward was offered during the pretraining session, it is possible that the sharp increase in average confidence in S2 reflects a response bias due to the introduction of incentives. Indeed, recent research shows that positive (resp. negative) reward increased (resp. decreased) confidence irrespective of task performance or metacognitive abilities (Lebreton et al., 2018). Second, the increase in confidence may be driven by differences in the definition of the possible confidence ratings across groups. Indeed, in the pretraining session participants in both the experimental and control groups were instructed to report confidence on a four-level scale, defined as 1 = *very low confidence*, 2 = *low confidence*, 3 = *high confidence* and 4 = *very high confidence*. Importantly no explicit mapping from confidence levels to subjective probabilities was given to participants. In this context, the correct interpretation of the lowest confidence rating is that of a 50% chance of being correct, that is, being unsure of the accuracy of their response, and therefore that participants are provided with a half-scale of confidence (Figure 1A). Yet, from S2 to S9, the experimental group (but not the control group) was presented with a full confidence scale, that is, confidence was mapped onto a probability of a response being correct from 0 to 1. As a result, confidence ratings 1 and 2 were to be used in case subjects thought they made an error (level 1 would be used when they were certain that they made an error, see Figure 1A), which rarely occurs in such experimental settings. This full-scale was explained to participants at the beginning of S2 and implemented in the reward scheme. For instance, according to a full-scale, rating confidence 1 (i.e., “sure

Figure 1
Confidence Rating Scales



Note. (A) Meaning of each confidence rating depending on the type of confidence scale (Half vs. Full), along with the corresponding probability of being correct (P(correct)). (B) Reward schemes depending on the type of confidence scale (Half vs. Full). QSR = Quadratic Scoring rule. See the online article for the color version of this figure.

incorrect”) when incorrect is maximally rewarded ($QSR = 1$, see Method) while rating confidence 1 on a half-scale (i.e., “not sure”) is equally rewarded regardless of accuracy (Figure 1B). Using a full-scale, participants should mostly use the highest ratings, as one can assume that the confident detection of errors is rare in nonspeeded perceptual tasks. Thus, ratings should increase from the first to the second session.

Thus, the introduction of incentives and the switch from a half-scale to a full-scale may have led to an artificial increase in confidence bias. Importantly, this upward shift in confidence ratings may also be expected to produce an artificial increase in metacognitive efficiency. Indeed, precise confidence criteria might be easier to maintain across two levels on a full scale than four levels on a half-scale. In addition, as suggested in recent works (Shekhar & Rahnev, 2021a, 2021b; Xue et al., 2021) criteria for high confidence are noisier than criteria for low confidence and thus a merge of high confidence categories can artificially increase metacognitive efficiency.

To assess the contribution of these potential confounds to the observed effects, we reanalyzed the original data, and collected a new dataset attempting to replicate the original findings while controlling for both keeping the incentives and reward scheme constant across sessions (Figure 2B). Assuming that the original procedure involves genuine metacognitive training, we reasoned that metacognitive efficiency should increase between the first and last session in the experimental group even when issues related to incentives and reward scheme are corrected. Instead, based on a preregistered sample size of 18 participants, we provide moderate evidence in favor of the null hypothesis according to which adaptive training in the present form does not improve metacognitive ability.

Method

Metacognitive Performance Measurement

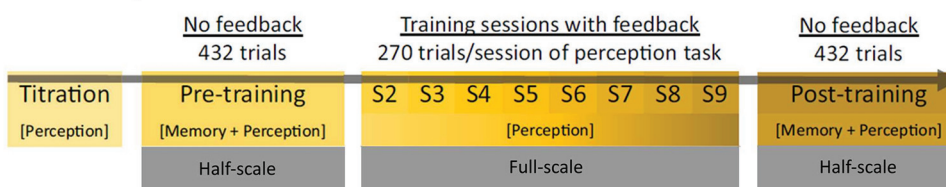
To evaluate metacognitive performance, we relied on the M-Ratio measure, derived from the *meta-d'* framework by Maniscalco and Lau (2012). In signal detection theory, the sensitivity d' quantifies the ability to detect or discriminate a stimulus from the distributions of correct and incorrect responses. Likewise, the metacognitive sensitivity *meta-d'*, quantifies the expected discriminability between two stimuli, if sensory evidence were not degraded between the discrimination decision and confidence rating. Thus, *meta-d'* refers to the sensory evidence available for metacognition, just as d' is the sensory evidence available for decision-making. It is then possible to quantify how much information was available for the metacognitive task, relative to the information available for the type I task, using the ratio *meta-d'*/ d' . This measure, called M-Ratio, is considered as the efficiency of metacognition for each observer.

Reanalysis of Original Data

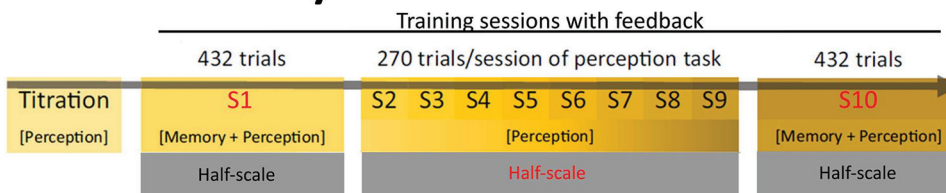
We retrieved the original data from the authors and further characterized the evolution of metacognitive performance across sessions with additional mixed-model ANOVAs with Training (Pretraining session vs Posttraining session) and group (Control vs Experimental) as factors. In line with the original mediation analysis, we expected to find a significant increase in metacognitive performance between pretraining session and S2. Furthermore, we compared S2 and S9 to assess the effect of training itself irrespective of the difference in incentives between pretraining session and

Figure 2
Comparison of the Original Study by Carpenter et al and the Present Study

A. Carpenter et al.



B. Present study



Note. (A) Original version of the protocol, with pre- and posttraining sessions providing no feedback, and rewards from S2 to S9 mapped onto a full-confidence scale. (B) Present version of the protocol, with S1 and S10 providing feedback, and rewards from S2 to S9 mapped onto a half-confidence scale. From “Domain-general enhancements of metacognitive ability through adaptive training”, by Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M., 2019, *Journal of Experimental Psychology: General*, 148(1), 51. Copyright 2019 by the authors. Adapted with permission. See the online article for the color version of this figure.

S2. Statistical analyses were conducted on $\log(\text{meta-}d'/d')$, like in the original study.

Conceptual Replication

Methods and hypotheses were preregistered (<https://osf.io/gak2t>) prior to data collection.

Modifications From the Original study

First, to test the possibility that a difference in terms of incentives between the pre and posttraining sessions might have artificially inflated metacognitive performance, we kept the incentives constant throughout the 10 sessions of the experiment. Accordingly, we refer to the first and last sessions as S1 and S10, instead of the original “pretraining” and “posttraining” sessions, respectively (see Figure 2). In the pre- and posttraining sessions, participants in the original study could either start with the memory tasks or the perception tasks. As a consequence of rewarding S1 and S10, participants always started with the perception task. This is to allow for continuity in the explanation of how points were calculated and assigned to participants.

Detailed instructions on how to map confidence to correct and incorrect trials were provided after the titration tasks in S1 but before any task where participants rated confidence. As in the original study, these instructions included a predefined set of demonstration trials and a series of practice trials with trial-wise feedback about whether confidence ratings were correctly assigned to correct or incorrect trials. However, here we made sure that the instructions were consistent with the reward scheme, and that both corresponded to a half-scale.

Second, to assess whether the increase in metaperformance observed in the original study stemmed from an incongruence between instructions regarding the confidence scale and reward, we provided reward that was consistent with instructions in all sessions: Participants were instructed to report confidence on a four-point scale with 1 = *very low confidence*, 2 = *low confidence*, 3 = *high confidence* and 4 = *very high confidence*, in all sessions including S1 and S10 (see Figure 2B). As opposed to the original study, we mapped confidence onto a probability of being correct between .5 and 1, as follows: $P(\text{correct}) = \frac{\text{conf} + 2}{6}$. Subsequently the quadratic scoring rule (QSR) was defined as $1 - (\text{accuracy} - P(\text{correct}))^2$, for each trial (see Figure 1B).

We also performed minor modifications to the experiment with no consequence on the experimental design: for example, Carpenter and colleagues ran the initial titration staircase until a fixed number of reversals was reached, or a maximum of 60 trials. We ran the titration staircase for a fixed number of 60 trials. We also fixed a small error in the code shared by Carpenter and colleagues in the memory task resulting in images being presented more than once in each block, and other images to never be displayed. All corresponding details are provided in our preregistration document (<https://osf.io/gak2t>).

Participants

The sample size was determined according to a preregistered stopping rule, using an open-ended sequential Bayes Factor (BF) design. Thus, we tested our effect of interest, namely the interaction between groups (Control vs. Experimental) and sessions (S1

vs. S10) on metacognitive efficiency until moderate evidence toward H1 or H0 was reached, that is, $\text{BF} > 5$ or $\text{BF} < .2$, respectively. As in the original study, we recruited participants through Amazon’s MTurk participant marketplace. Sixty-nine participants completed at least the first session. Of these, 11 participants dropped out from the study before the end of the tenth session. Nine participants were excluded for responding incorrectly to screening questions related to the understanding of the tasks, before the beginning of the training (for details, see Carpenter et al., 2019). Nineteen participants were excluded for technical issues during the first session, leading them to drop at least one experimental condition. Further, 11 participants were excluded for either floor ($< 55\%$) or ceiling ($> 95\%$) performance in at least one condition/session. Finally, one participant was excluded for reporting the same confidence level on at least 95% of the trials over three sessions or more. Trials where participants did not respond in time ($> 2,000$ ms) or responded too quickly (< 200 ms) were excluded from further analyses (1.61% of the trials).

The analyses were conducted on a sample of 18 participants (10 women, mean age = 40.4 years, range age = 19–59). All participants received monetary compensation in U.S. dollars (range = \$37.6–\$41.8). An upper bound for sample size was determined using a design analysis with Bayes factors as index of evidence (Schönbrodt & Wagenmakers, 2018). Data simulations with an expected increase in metacognitive efficiency between S1 and S10 of small effect size (Cohen’s $d = .3$) revealed that a maximal sample of 100 participants would lead to conclusive evidence under H1 in 74% of cases ($\text{BF} > 5$), and under H0 in 89% of cases ($\text{BF} < .2$). However, the stopping rule criterion was already met when performing the first Bayes Factor sequential analysis after a first group of 18 participants had completed all ten sessions (see Figure 5). We recruited participants in the experimental group only (i.e., participants receiving reward according to metacognitive performance), and compared their data with those of participants in the original control group, who received reward according to their perceptual performance. As in the original study, bonuses were distributed pseudorandomly to ensure equivalent financial motivation irrespective of performance. The study was approved by the ethics committee from the Paris School of Economics (#2019 021).

Procedure

Save from the modifications to the code, we used the same HTML/JS/CSS scripts, and therefore the very same stimuli, as in the original study by Carpenter et al. The study ran on a JATOS server (www.jatos.org; Lange et al., 2015).

Statistical Analysis

We ran the same analyses as Carpenter and colleagues. We tested for potential changes in metacognitive efficiency ($\log(\text{meta-}d'/d')$) and metacognitive bias (average confidence) using mixed-design ANOVAs in Rstudio Version 1.3.1093 (RStudio Team, 2020) using notably the packages tidyverse (Wickham et al., 2019), afex (Singmann et al., 2015), and metaSDT (Craddock, 2018). Bayesian ANOVAs were computed with default prior (Cauchy distribution centered on the effect size, with a scaling parameter set to $\frac{\sqrt{2}}{2}$) using the BayesFactor package (Morey et al., 2018).

Results

Reanalysis of Carpenter et al. (2019)

After confirming the results reported by Carpenter et al. we extended the analyses reported in the original paper in two ways. First, to account for a potential effect of a change in instructions in S2 versus Pretraining, we compared metacognitive efficiency between S2 and the posttraining session S10 (instead of between pre- and posttraining sessions, as originally reported). Here, we found no significant interaction effect between group and Training, $F(1, 58) = .71, p = .40, BF = .27$. When comparing S2 with S9 (i.e., the first and the last of the training sessions), the Group \times Training interaction remained nonsignificant, $F(1, 59) = .49, p = .49, BF = .39$ (Figure 3A and 3B). These results suggest that the improvement of metacognitive efficiency occurred not during the extended training part of the protocol, but quite abruptly at the beginning of the training phase.

Second, we studied the abrupt changes in metacognitive efficiency between the pretraining session and S2. We first found a significant interaction between group and Training, $F(1, 59) = 4.64, p = .035$. Perhaps more strikingly, we found in the original data an abrupt increase in average confidence between the last five trials of the pretraining session and the first five trials of S2 (Figure 4E), in the experimental group only, $F(1, 28) = 22.14, p < .001$. Together, these results suggest that this increase in metacognitive efficiency could be driven by the changes introduced from S2 to S9, also influencing participants' strategy on the posttraining session (S10).

A Preregistered Replication Study

Sequential Bayes Factor Analysis

Informed by the reanalysis of the original data, we then turned to our conceptual replication study. To assess the efficiency of metacognitive training while accounting for incentives and confidence scale confounding factors, we conducted the same analysis

as in the original study comparing metacognitive efficiency ($\log(\text{meta-}d'/d')$) between sessions (S1 and S10) and groups (experimental vs. control).

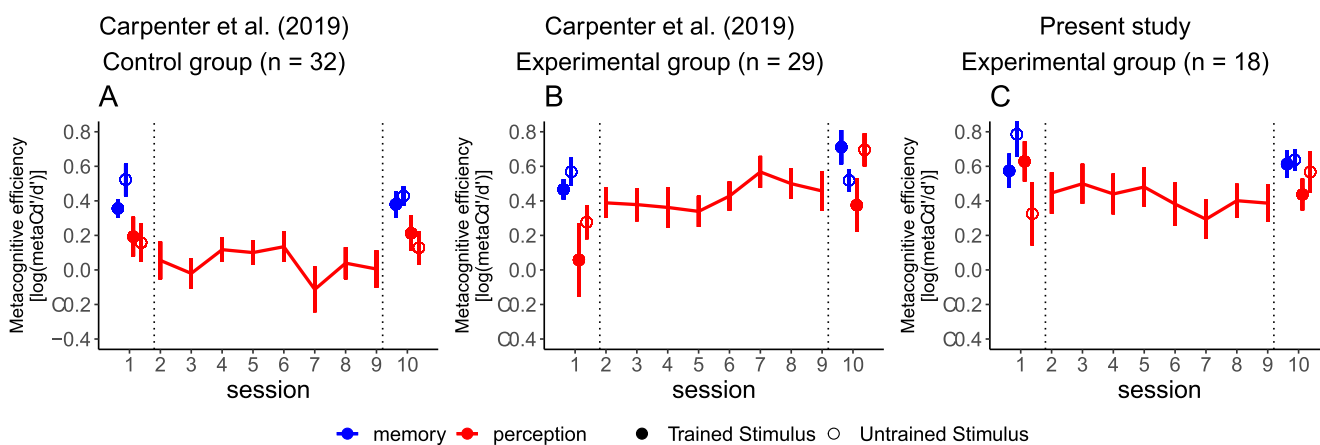
We had preregistered recruiting participants until moderate evidence toward H1 or H0 was reached.

Metacognitive Efficiency

We compared metacognitive efficiency in S1 and S10 in our new experimental group (Figure 3C) with those in the control group from Carpenter et al. (2019). (Figure 3A). Contrary to the original results, the group \times Training interaction was not significant in this analysis, $F(1, 45) = .083, p = .93, BF = .17$. Moreover, assessing the linear trend of metacognitive efficiency between S2 and S9 in the three groups, we found no main effect of the training sessions, $F(7, 490) = .25, p = .97, BF = .13$, and no interaction effect between the training sessions and groups (control vs. experimental group in the original study: $F[7, 399] = 1.61, p = .13, BF = 2.50$; control vs. our experimental group: $F[7, 294] = .90, p = .51, BF = .24$). In other words, once we kept the reward scheme constant across all sessions, we found no evidence for metacognitive training in our study. This suggests that previous results might have been confounded by effects of incentives and/or confidence scale, as we detailed in the Introduction.

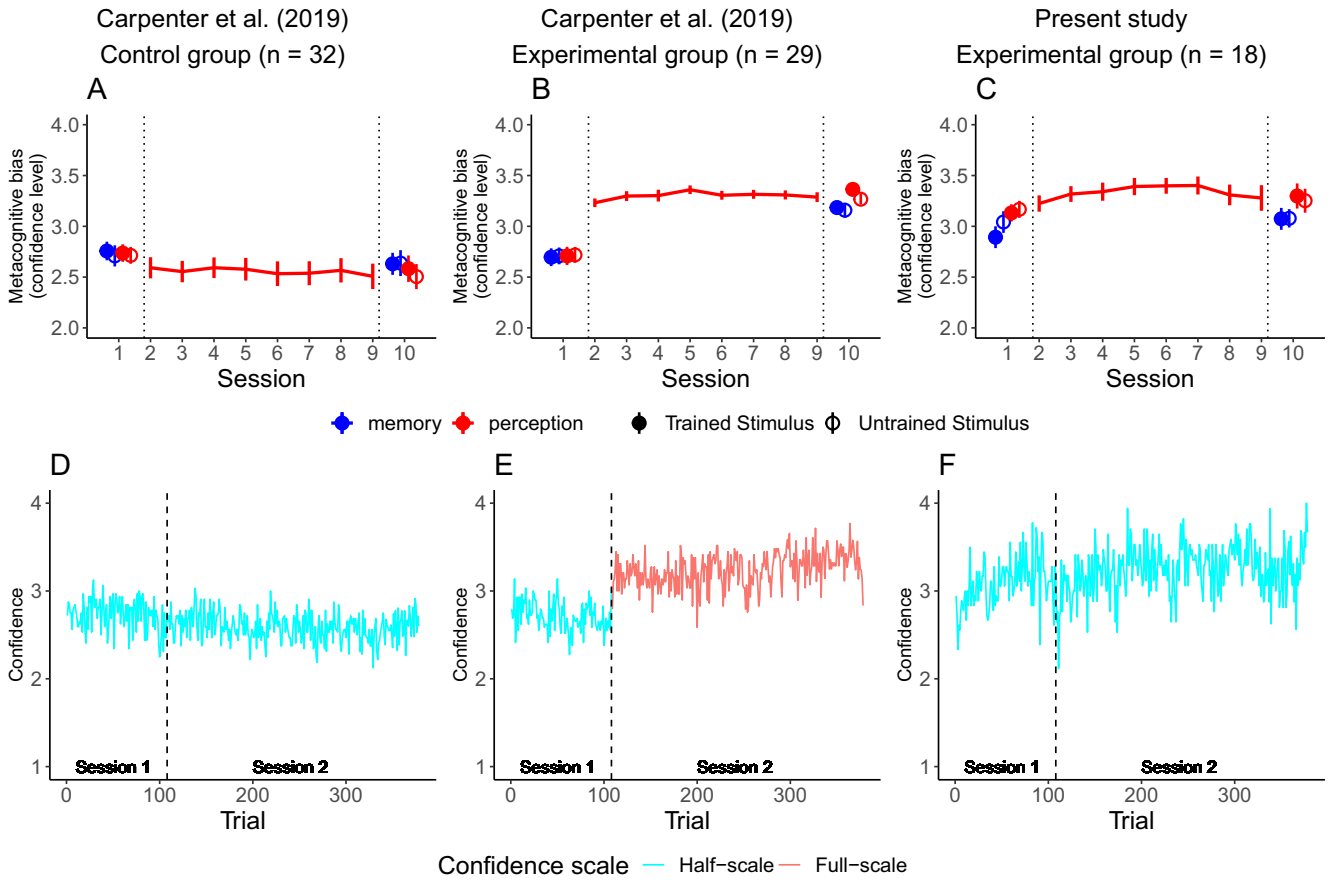
In their study, Carpenter and colleagues also reported that the peak change in metacognitive efficiency occurred systematically later than the peak change in confidence bias. To assess if a similar pattern was present in our replication group, we conducted an ANOVA with peak session as dependent variable, and outcome (metacognitive efficiency vs. confidence bias) and group (experimental: original vs. replication) as fixed effects. This analysis revealed a main effect of outcome, $F(1, 45) = 11.37, p = .02$, but no interaction with group, $F(1, 45) = .01, p = .98$, indicating that in both groups the peak change in metacognitive efficiency occurred systematically later than the peak change in confidence bias. Because this temporal pattern was also found in our replication group in the absence of global increase in metacognitive efficiency, the extent to which those dynamics are important for

Figure 3
Metacognitive Efficiency ($\log(\text{meta-}d'/d')$) Over the Ten Experimental Sessions



Note. (A and B) Results reproduced from the original data by Carpenter et al, control group, and experimental group, respectively. (C) Results from the present study. Error bars represent standard error of the mean. See the online article for the color version of this figure.

Figure 4
Confidence Level Across Sessions and Trials



Note. A–C: Metacognitive bias across sessions in the control (A) and experimental groups (B) from Carpenter et al., and in the experimental group from our sample (C). Note. Error bars represent standard error of the mean. D–F: Evolution of average confidence across participants and trials in S1 and S2 in the control (D) and experimental groups (E) from Carpenter et al., and in the experimental group from our sample (F). Colors indicate the type of confidence scale in use. Blue: Half-scale, Red: Full-scale. See the online article for the color version of this figure.

metacognitive training remains unclear. Of note, these results are based on a rather small sample size, in compliance with the stopping rule we preregistered prior to data collection.

Exploring the Origin of the Metacognitive Bias

Next, we assessed in an exploratory analysis which of the two confounds, incentives or confidence scale, was the main contributor of the confidence increase. This also relates to the question of metacognitive training, as Carpenter and colleagues reported that the increase in metacognitive efficiency was in fact mediated by the increase in metacognitive bias, and as an increase in confidence bias might result in an increase in metacognitive efficiency (Shekhar & Rahnev, 2021).

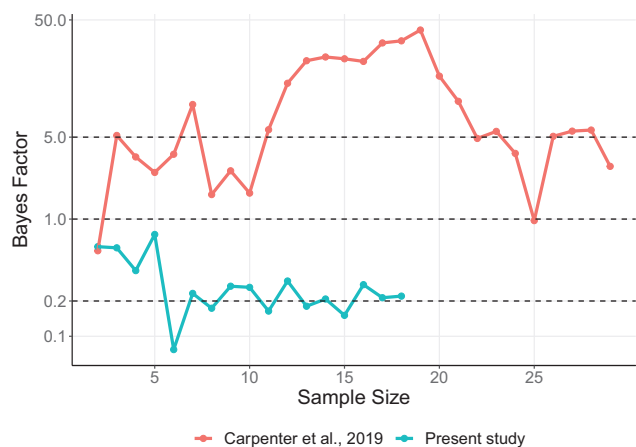
If this abrupt increase in confidence ratings was due to the introduction of incentives at S2, we would expect the same average confidence in our experimental group (Figure 4C) and from S2 to S9 in the original experimental group, as these conditions are similar in terms of reward. We would also expect these two conditions to show higher levels of confidence than the control group. This is what we found in the data. When comparing average confidence in

S2–S9 between the three groups (control vs. original experimental vs. replication) with an ANOVA, we found a main effect of group, $F(2, 72) = 24.61, p < .001$, driven by significantly higher levels of confidence both in our replication group, $t(72) = -5.05, p < .001$, and in the original experimental group, $t(72) = -6.43, p < .001$, compared with the control group, with no difference between the experimental group and the replication group, $t(72) = .10, p = .995, BF = .46$. However, we are cautious in interpreting confidence biases that might not be comparable between groups and studies.

One other possibility is that this abrupt increase in average confidence was due to a shift in the type of confidence scale (i.e., half-scale in the pretraining session, and full-scale from S2 to S9, see Figure 2A). If this were true, then we would expect the average confidence in our replication group (which used a half-confidence scale) to be lower than the level of confidence obtained from S2 to S9 in the original experimental group. As just mentioned, however, these two conditions were not different in terms of average confidence.

Furthermore, because the increased levels of confidence described above are not accompanied by an increase in first-order

Figure 5
Bayes Factor (BF) Sequential Analysis of the Interaction Effect Between Sessions (S10 and S1) and Groups (Control Versus Experimental) on $\log(\text{meta-}d'/d)$



Note. The BF assesses whether the effect of interest (interaction Group \times Training for metacognitive efficiency) is more plausible under H0 or under H1. $BF > 1$ is evidence supporting H1. $0 < BF < 1$ is evidence supporting H0. The dashed lines mark the ratios where the evidence is five-fold more likely under each hypothesis, which we took as boundaries for moderate evidence. Red curve: Carpenter et al., 2019. Blue curve: Present study. See the online article for the color version of this figure.

performance (as assessed through difficulty levels across the three groups, $F(2, 72) = .17$, $p = .84$, $BF = .18$) it is unlikely that the metacognitive bias can simply be explained by a generic motivation effect.

Altogether, these analyses thus suggest that the presence of incentives might be the main reason for the increase in confidence ratings, which in turn would have led to an increase of metacognitive efficiency, as recently proposed (Shekhar & Rahnev, 2021a). Nonetheless, because our analyses relied on comparing confidence biases between studies in relatively small samples, these conclusions on the specific mechanism at stake should be taken with caution.

Discussion

In the present work, we aimed at reassessing the effectiveness of a protocol designed by Carpenter and colleagues (2019) to improve metacognitive abilities. We noticed that the increase in metacognitive efficiency found by Carpenter and colleagues might be unspecific, owing to an artificial increase in confidence bias, triggered by two confounding factors: In the original study, reward was not held constant throughout all sessions, so that participants might have been more incentivized to perform the task not only during rewarded sessions (S2–S9), but also in the posttraining session (S10), as a spillover effect. Also, the instructions provided to the participants in the experimental group were not congruent with the reward scheme, encouraging them to use high confidence ratings (i.e., ratings 3 and 4) from S2 onward but not in the pretraining session. To evaluate our claim that the original results may be due to confounding factors, we performed additional analyses on the original data set. First, when restricting the analysis to training

sessions only (i.e., S2 to S9, instead of pretraining and posttraining sessions), thus controlling for incentives, we found no evidence for an improvement in metacognitive performance in the experimental group. By contrast, this increase was already significant between S1 and S2. This sharp increase in metacognitive performance was accompanied by an abrupt increase in average confidence between the last trials of the pretraining session and the first trials of S2. In our view, the fact these behavioral changes occurred rapidly in time at the very beginning of the experimental procedure casts doubts on the possibility that they arose as a result of a genuine improvement in metacognitive performance. Instead, we suspect that they may have been attributable to either, or both, of the two possible experimental confounds mentioned above.

To further assess the validity of this training procedure, we conducted a conceptual replication controlling for both incentives and confidence-related factors by, first, providing reward in all sessions (i.e., including S1 and S10) and, second, rewarding the experimental group on the basis of a half-confidence scale, in line with the instructions received by participants (and instead of a full-scale as in the original study). We reasoned that, if the training method was effective in improving metacognition, estimates of metacognitive efficiency should increase between S1 and S10 in the experimental group, even when issues related to incentives and confidence scale were corrected. Instead, we obtained moderate evidence in favor of H0 (following a preregistered open-ended sequential Bayes factor analysis), indicating that no increase in metacognitive efficiency occurred. Thus, we suggest that the increase in metacognitive efficiency reported by Carpenter et al. (2019) resulted from a global change in the use of the confidence scale, possibly owing to incentives or instructions regarding the confidence scale, rather than from an improved sensitivity to trial-wise fluctuations in the quality of the decision. While such a global adjustment of confidence ratings might be adaptive and useful (e.g., when communicating confidence to reach joint decisions), it is important to distinguish this effect from a genuine improvement of metacognitive monitoring, conceptually and empirically. Of note, post hoc analyses revealed that metacognitive efficiency in S1 was higher in the replication compared with the original experimental group with marginal significance ($p = .11$), probably attributable to the fact that S1 in our replication group was rewarded, pushing participants to perform better. Yet, it might be that metacognitive efficiency in the replication group reached a ceiling early in the procedure, leaving little room for improvement even if training were in fact possible under this new protocol.

In recent years, the field of metacognition has seen a dramatic increase in popularity, in part due to the development of new statistical tools that allow quantifying metacognitive performance independently from typical confounds such as first-order performance (Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012). Moreover, metacognitive deficits are prevalent in several psychiatric and neurological disorders, with severe consequences in terms of medical observance and quality of life (Hasson-Ohayon et al., 2015; Lysaker et al., 2015). This is why developing robust, efficient, and cost-effective remediation procedures to improve metacognitive performance is important. Several studies already provided evidence suggesting that monitoring abilities can be trained: A two-week meditation training was found to enhance metacognitive accuracy in the memory domain (Baird et al., 2014), and knowledge about cognitive biases is held to reduce

delusions and positive symptoms in schizophrenia (for a review, see Eichner & Berna, 2016). More recently, preliminary results from a virtual-reality assisted training consisting in frequently questioning the reality of wakeful experiences augmented the rate of lucid dreaming experiences (Gott et al., 2021). Despite pioneering experiments showing promising results (Adams & Adams, 1958; Sharp et al., 1988), to our knowledge, no recent remediation procedure based on feedback has been successful in improving the quality of confidence ratings (for a recent attempt based on single-trial feedback, see Haddara and Rahnev, 2019, 2020).

Future attempts to improve the quality of confidence ratings may be informed by recent findings regarding the definition of metacognitive noise (Shekhar & Rahnev, 2021a, 2021b; Xue et al., 2021), as a way to provide more information to participants regarding the qualitative nature of their metacognitive deficits. They could also rely on elicitation methods that encourage participants to report optimal confidence estimates, such as measuring participants' willingness to trade a gamble based on the accuracy of their response against a lottery with known probabilities (Dienes & Seth, 2010; Massoni et al., 2014). Another way of refining confidence ratings may be to provide participants with feedback regarding the temporal dynamics with which first-order decisions are made. Indeed, becoming aware of how the decision-making process unfolds in time may help to better judge the accuracy of a given decision. Practically, this could simply consist in presenting participants with feedback about their own response times for correct and incorrect responses, or more ambitiously with parameter estimates from mouse-tracking (Dotan et al., 2019; Faivre et al., 2021) or postdecisional evidence accumulation models (Pleskac & Busemeyer, 2010; Pereira et al., 2020, 2021). Other strategies may consist in training participants to better detect their attentional lapses (Baird et al., 2014; Recht et al., 2021), or to regulate brain networks associated with over or underconfidence (Cortese et al., 2016). Given the complexity of this endeavor, and the societal and clinical issues at stake, effective metacognitive training will probably require collective efforts rather than individual initiatives (Rahnev et al., 2021). In that regard, we highlight the openness from the authors of the original study, who publicly shared their valuable code and data and discussed these results openly with us, as those are the first necessary steps toward collective research on metacognition.

Context of the Research

We were interested in the possibility to train metacognitive abilities in the broader context of our research on schizophrenia. A rich clinical literature suggests the existence of metacognitive deficits in individuals with schizophrenia, and efforts had already been made to alleviate symptoms and improve quality of life through metacognitive training. Existing metacognitive training procedures rely on explicit and high-level strategies, notably by encouraging patients to bring unnoticed beliefs and cognitive biases to awareness. As a complementary intervention, we were enthusiastic about the metacognitive training proposed by Carpenter and colleagues, which targeted lower-level mechanisms involved in learning how to properly estimate confidence on a trial-to-trial basis. If successful in healthy participants, we were hoping to adapt this procedure to clinical settings.

References

- Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*(4), 747–751. <https://doi.org/10.2307/1420334>
- Baird, B., Mrazek, M. D., Phillips, D. T., & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology: General*, *143*(5), 1972–1979. <https://doi.org/10.1037/a0036882>
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51–64. <https://doi.org/10.1037/xge0000505>
- Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, *7*(1), 13669. <https://doi.org/10.1038/ncomms13669>
- Craddock, M. (2018). metaSDT: Calculate Type 1 and Type 2 Signal Detection Measures. (R package version 0.5.0). <https://github.com/craddm/metaSDT>
- Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674–681. <https://doi.org/10.1016/j.concog.2009.09.009>
- Dotan, D., Pinheiro-Chagas, P., Al Roumi, F., & Dehaene, S. (2019). Track it to crack it: Dissecting processing stages with finger tracking. *Trends in Cognitive Sciences*, *23*(12), 1058–1070. <https://doi.org/10.1016/j.tics.2019.10.002>
- Eichner, C., & Berna, F. (2016). Acceptance and efficacy of metacognitive training (MCT) on positive symptoms and delusions in patients with schizophrenia: A meta-analysis taking into account important moderators. *Schizophrenia Bulletin*, *42*(4), 952–962. <https://doi.org/10.1093/schbul/sbv225>
- Faivre, N., Roger, M., Pereira, M., de Gardelle, V., Vergnaud, J. C., Passerieux, C., & Roux, P. (2021). Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *Journal of Psychiatry & Neuroscience: JPN*, *46*(1), E65–E73. <https://doi.org/10.1503/jpn.200022>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. <https://doi.org/10.3758/BF03196546>
- Gott, J., Bovy, L., Peters, E., Tziouridou, S., Meo, S., Demirel, Ç., Esfahani, M. J., Oliveira, P. R., Houweling, T., Orticoni, A., Rademaker, A., Boutilik, D., Varatheeswaran, R., van Hooijdonk, C., Chaabou, M., Mangiaruga, A., van den Berge, E., Weber, F. D., Ritter, S., & Dresler, M. (2021). Virtual reality training of lucid dreaming. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1817), 20190697. <https://doi.org/10.1098/rstb.2019.0697>
- Haddara, N., & Rahnev, D. (2019). Trial-by-trial feedback does not improve performance or metacognition in a large-sample perceptual task. *Journal of Vision*, *19*(10), 27. <https://doi.org/10.1167/19.10.27>
- Haddara, N., & Rahnev, D. (2020, March 26). The impact of feedback on perceptual decision making and metacognition: Reduction in bias but no change in sensitivity. *PsyArXiv*. <https://doi.org/10.31234/osf.io/p8zyw>
- Hasson-Ohayon, I., Avidan-Msika, M., Mashiach-Eizenberg, M., Kravetz, S., Rozencwaig, S., Shalev, H., & Lysaker, P. H. (2015). Metacognitive and social cognition approaches to understanding the impact of schizophrenia on social quality of life. *Schizophrenia Research*, *161*(2–3), 386–391. <https://doi.org/10.1016/j.schres.2014.11.008>

- Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry*, 9(1), 268. <https://doi.org/10.1038/s41398-019-0602-7>
- Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, 10(6), e0130834. <https://doi.org/10.1371/journal.pone.0130834>
- 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), eaaq0668. <https://doi.org/10.1126/sciadv.aaq0668>
- Lysaker, P. H., Vohs, J., Minor, K. S., Irrazaval, L., Leonhardt, B., Hamm, J. A., & Dimaggio, G. (2015). Metacognitive deficits in schizophrenia: Presence and associations with psychosocial outcomes. *Journal of Nervous and Mental Disease*, 203(7), 530–536. <https://doi.org/10.1097/NMD.0000000000000323>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Massoni, S., Gajdos, T., & Vergnaud, J. C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, 5, 1455. <https://doi.org/10.3389/fpsyg.2014.01455>
- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12-4.1) [Computer software]. Comprehensive R Archive Network. <https://CRAN.R-project.org/package=BayesFactor>
- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing about Knowing*, 13, 1–25.
- Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., & Millán, J. D. R. (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 8382–8390. <https://doi.org/10.1073/pnas.1918335117>
- Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., & Faivre, N. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nature Communications*, 12(1), 3261. <https://doi.org/10.1038/s41467-021-23540-y>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3), 864–901. <https://doi.org/10.1037/a0019737>
- Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R. N., Desender, K., Faivre, N., Filevich, E., Fleming, S., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odgaard, B., Peters, M. A. K., Reyes, G., Rouault, M., Sackur, J., . . . Zylberberg, A. (2021). Consensus goals for the field of visual metacognition. *PsyArXiv*. <https://doi.org/10.31234/osf.io/z8v5x>
- Recht, S., de Gardelle, V., & Mamassian, P. (2021). Metacognitive blindness in temporal selection during the deployment of spatial attention. *Cognition*, 216, 104864. <https://doi.org/10.1016/j.cognition.2021.104864>
- Rouy, M., de Gardelle, V., Vergnaud, J.-C., Reyes, G., Filevich, E., & Faivre, N. (2021). *Replication: Domain-general enhancements of metacognitive ability through adaptive training*. <https://doi.org/10.17605/OSF.IO/RQ967>
- 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 and Biobehavioral Reviews*, 126, 329–337. <https://doi.org/10.1016/j.neubiorev.2021.03.017>
- RStudio Team. (2020). RStudio: Integrated Development Environment for. <http://www.rstudio.com/>
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, 42(3), 271–283. [https://doi.org/10.1016/0749-5978\(88\)90001-5](https://doi.org/10.1016/0749-5978(88)90001-5)
- Shekhar, M., & Rahnev, D. (2021a). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, 128(1), 45–70. <https://doi.org/10.1037/rev0000249>
- Shekhar, M., & Rahnev, D. (2021b). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, 25(1), 12–23.
- Singmann, H., Bolker, B., & Westfall, J. (2015). Afex: Analysis of Factorial Experiments (R Package Version 0.15-2). <http://CRAN.R-project.org/package=afex>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, 95, 103196.

Received August 5, 2021

Revision received November 21, 2021

Accepted December 14, 2021 ■

General discussion

We have seen in the introduction that schizophrenia is a severe mental condition and that the last 150 years have witnessed much effort from clinicians and researchers to improve the reliability of the diagnosis as well as treatment efficacy. A lot of research has been dedicated to understanding the underpinnings of the symptomatology of schizophrenia which comprises phenomena like delusions and hallucinations, accompanied by the absence of awareness of such symptoms. In this thesis, we used advanced methods of psychophysics as well as recent models and measures from cognitive psychology to try to better understand at which level of the cognitive hierarchy (i.e. first or second-order) these deficits occurred in schizophrenia. In particular, we relied on the hypothesis that the lack of insight stemmed from an impairment situated at the second-order - i.e. metacognitive - level of processing (David et al. 2012), as opposed to the first-order level. In this sense, patients with schizophrenia would have difficulties monitoring and reflecting on their own mental states, eventually leading to impaired reality monitoring (Dijkstra et al. 2022) and in turn several positive symptoms including hallucinations and delusions (Aynsworth et al. 2017; Mondino et al. 2019).

1. Performance-matching matters

We have seen that a major issue regarding the measure of metacognition is to get rid of undesired sources of variance that are non-metacognitive in nature. In particular, patients with schizophrenia suffer from first-order deficits that have to be controlled for, experimentally or statistically, in order to capture what is genuinely metacognitive. If those confounds are not controlled for, it is impossible to determine if abnormal confidence calibration reflects metacognitive impairment or simply results from greater difficulties felt by patients when performing the experimental tasks.

In this thesis, we aimed at quantifying metacognitive abilities among individuals with schizophrenia while controlling for first-order deficits. In experimental chapter 1, our meta-analysis revealed an overall metacognitive deficit of medium effect size, the strongest deficits being found in metamemory studies. However, the effect size turned out to be twice smaller and statistically inconclusive among studies that controlled for first-order performance. Because studies controlling for first-order performance were almost uniquely found in the perceptual domain, it remained unclear whether the measured metamemory deficit was inherited from memory deficits or reflected a genuine and domain-specific metacognitive alteration. We addressed this issue in our experimental chapter 2, where we

developed an experimental procedure involving metaperceptual and metamemory tasks, and used a recent metacognitive measure called confidence efficiency (Mamassian and de Gardelle 2021) to statistically control for first-order performance. Quite unexpectedly, we found that metamemory abilities in familiarity and recollection tasks were relatively preserved compared to metaperceptual abilities in a visual detection task. It was even more unexpected when considering results from our experimental chapter 3, where we found intact electrophysiological markers of confidence on fronto-central electrodes in the context of a visual discrimination task where behavioral results revealed no metacognitive deficits among patients. The seemingly contradictory results opposing meta-perceptual deficits in chapter 2 and preserved meta-perceptual abilities in chapter 1 and 3 drove our attention to the type of stimuli we used, namely complex face stimuli in chapter 2 as opposed to low-level random dot kinetogram in chapter 3. However, if the metaperceptual deficit derived from an impaired face processing, then it is not clear why metamemory abilities relying on the same face stimuli would be preserved. Alternatively, we wondered whether the incongruence could result from the type of task i.e. altered metaperception in detection tasks versus preserved metaperception in discrimination tasks, but this hypothesis is not supported by the data from our meta-analysis. As discussed in chapter 2, in spite of all precautions and efforts to control first-order deficits, the metaperceptual deficit we found might also derive from the under optimization of our intra-individual performance-matching procedure, and should be taken with caution. Furthermore, given the heterogeneity of profiles in schizophrenia (further developed in section 5.2.1), analyses conducted on our limited sample size might be subject to false positives (e.g. the metaperceptual deficit) or false negatives (e.g. preserved metamemory abilities). More evidence is required to further discuss these effects.

All in all, our work has provided a quite different picture from the existing literature regarding metacognitive deficits among patients with schizophrenia. Indeed, combining a meta-analysis and fine-grained methodological and statistical tools, we have provided some evidence that metacognitive abilities were relatively preserved across cognitive domains, when first-order deficits are accounted for. What does it mean to have preserved metacognitive abilities in spite of first-order deficits? It points to a modular architecture of the mind, where first-order deficits are not reverberated to higher order functional levels, and thus can be accurately monitored. In this sense, the efficiency of clinical metacognitive training like MCT (Moritz et al. 2014) might not result from a genuine “training” and reinforcement of an initially weak metacognitive ability, but rather from the suggestion by the clinician to use this neglected but preserved ability.

The following sections discuss the relevance of our results regarding existing literature in clinical research, in an effort to articulate conceptual frameworks with one another (sometimes with a fair amount of speculation).

2. What about overconfidence?

We hope that our considerations about the delineation between first-order and second-order levels of processing made it clear that overconfidence in errors does not necessarily involve a metacognitive deficit. At least, not a “local” metacognitive deficit. Indeed, it is important to distinguish between local metacognition and global metacognition, the former referring to the ability to calibrate confidence on performance on a trial-by-trial basis, whereas the latter consists of more global estimations of self-performance within a given task or domain (Rouault and Fleming 2020). In this sense, Seow et al (2021) proposed that the high-level construct of self-confidence ultimately results from a progressive learning to estimate one’s performance, starting with the building block of local metacognition, then estimating one’s confidence in a group of trials, up to the most global estimations summarizing accumulated experience over a whole cognitive task or domain (Figure 16). Reciprocal interactions are assumed to be at play between each intermediate level.

One can understand from this hierarchical presentation of metacognitive evaluations that the level of metacognition involved in the present thesis (local) is far-distant from the level captured by clinical questionnaires (global metacognition) such as the Beck Cognitive Insight Scale (Beck 2004). Most relevant for our purpose here, results from a recent study (Bhome et al. 2022) with individuals with a functional cognitive disorder (FCD) have shown preserved local metacognition using both a perceptual task and a memory task, but an impaired global metacognition. Contrary to the hierarchical and reciprocal model of metacognitive evaluations (Seow et al. 2021), this result suggests the existence of independent sources of information in the formation of confidence estimates at the different hierarchical levels. Bhome and colleagues speculatively proposed that the incongruence came from abnormal priors influencing specifically the higher metacognitive evaluations (Figure 20).

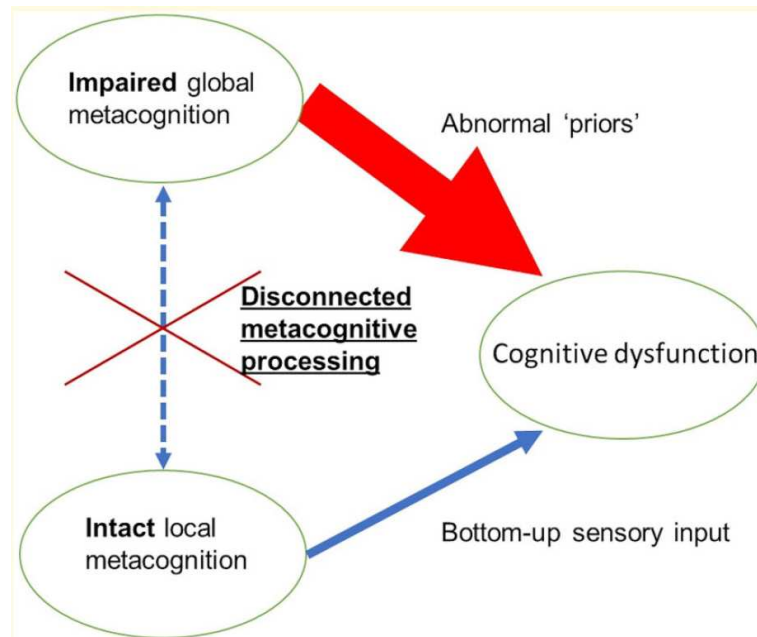


Figure 20. Local versus global metacognition. Incongruity between intact local metacognition and impaired global metacognition, reflecting a disconnected metacognitive processing. Reprint from Bhome et al. (2022)

Alternatively, global metacognition can be impaired as a result of poor first-order performance. In the next section, I will argue that first-order deficits are sufficient to produce a seeming impairment of global metacognition. The idea is to conceive the global metacognition impairment as a Dunning-Kruger effect (DKE, 1999) and to rely on recent conceptual and statistical considerations that question the metacognitive nature of the DKE.

2.1. Overconfidence in errors as a Dunning-Kruger effect

Dunning and Kruger became famously known for their eponymous effect (1999). Through four experiments involving various domains of investigation (grammar, logical reasoning, and humor), the authors analyzed how well people's subjective evaluations of performance (global estimations) were calibrated on objective performance, depending on their competence with the task. In all experiments, participants were grouped into four quartiles according to their competence (Figure 21). The authors came to the general conclusion that "unskilled" people (the bottom quartile) suffered a double-burden: "Not only do these people reach erroneous conclusions and make unfortunate choices, but their incompetence robs them of the metacognitive ability to realize it" (Dunning and Kruger 1999).

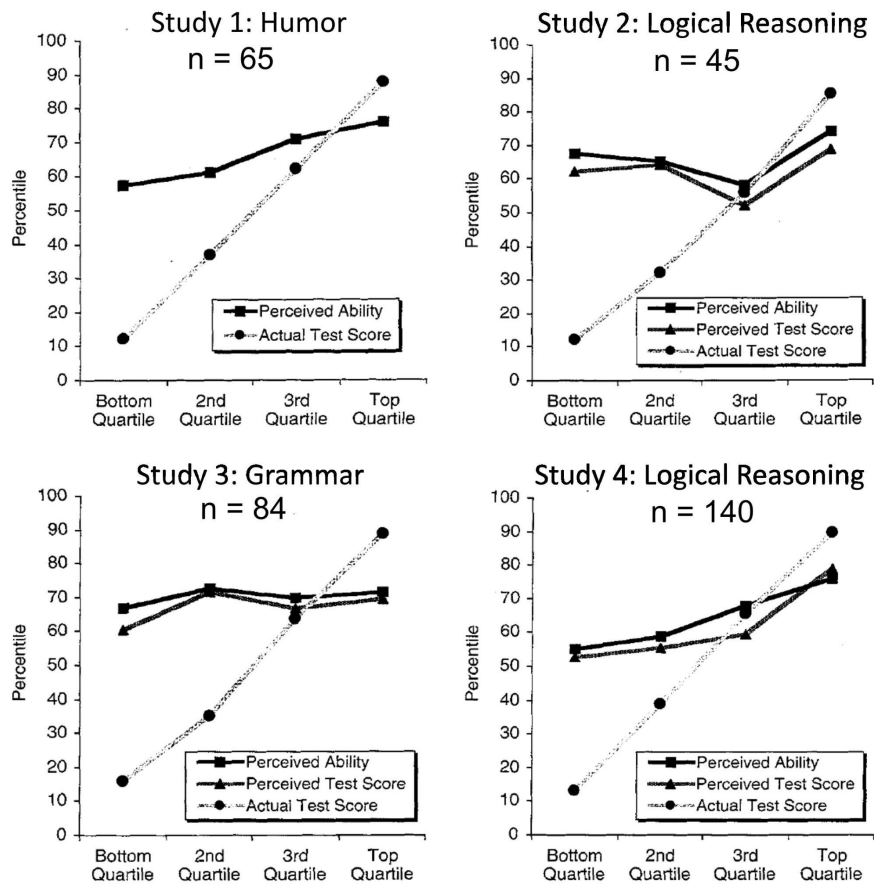


Figure 21. Dunning-Kruger effect. The vertical axis is the percentile ranking, a comparative measure where participants indicate their subjective ranking of competence compared to others (proxy for perceived ability). The horizontal axis indicates quartiles of objective performance to the tasks. Solid black lines are averaged percentile rankings for each quartile. The gray diagonals are averaged actual scores projected onto the percentile axis. It became apparent that low performers (bottom quartile) were overestimating their own abilities. Reprint from Dunning and Kruger (1999).

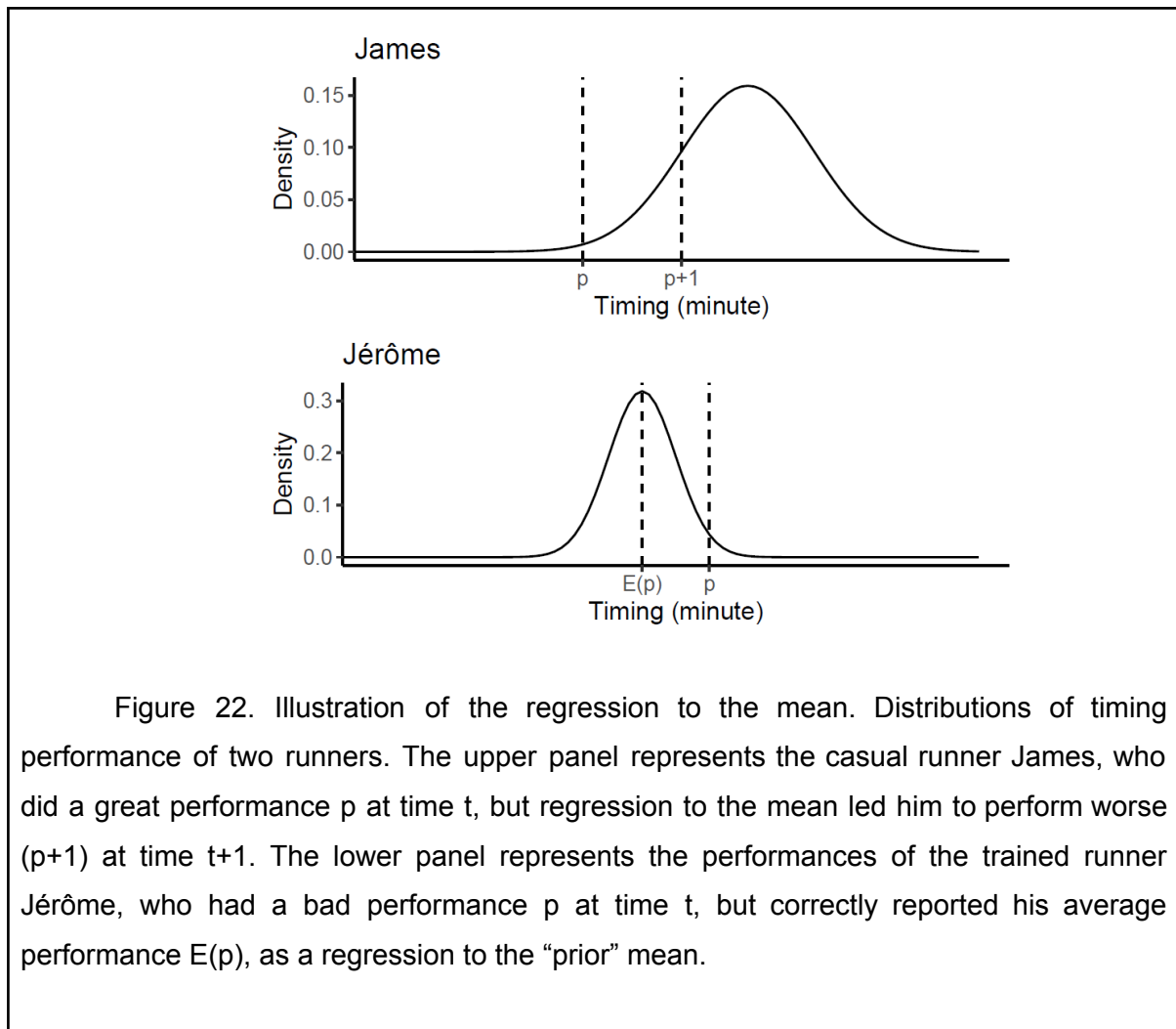
Of note, the DKE has long been debated, between defenders of the effect as genuine, and interpretations in terms of a regression to the mean (i.e. a statistical artifact, see Box 3). However, the recent study by Jansen, Rafferty, and Griffiths (2021) applied model comparisons on a large sample ($N \sim 3500$) and ruled out an explanation solely phrased in terms of a regression to a “prior” mean. On the basis of the large dataset made available online by the authors, we have conducted other analyses corroborating this conclusion (see Appendix 1).

Importantly for us, Dunning and Kruger illustrated clearly how being unskilled was related to overconfidence, even among healthy psychology students.

Box 3. Regression to the mean

The regression to the mean effect is illustrated in Figure 22. Let's consider the example of James, a casual runner who routinely does the same course in the neighborhood. For some reason, James outperformed last week: he finished the course 5 minutes quicker than usual. Happy with his seeming progress, James expected to do as well - if not better - the next week. Contrary to his expectations, James turned out to be upset the following week, since he did not manage to reproduce the performance. This is what we call a "regression to the mean" effect. Given that no progress had really been achieved (i.e. the mean and dispersion of the distribution of performances are kept constant), the probability that the next performance will be better than an already good performance is low.

In the case of the Dunning-Kruger effect, the story is slightly different since the authors did not compare performance p with performance $p+1$, rather they compared performance p with the expected performance $E(p)$. For the sake of illustration, let's now consider the case of Jérôme, a trained runner who knows how he should perform on average. Although he did a terrible performance at the last edition of the "10 km Paris Centre", he accurately knows that this performance was not representative. So, if queried about how well he usually performs, Jérôme would be correct to report a higher performance (closer to $E(p)$) than what it did that day. In this sense, it is more accurate to talk about regression to the "prior" mean, as suggested by Jansen et al. (2021). To note, in case of a regression to the "prior" mean, then the category of "unskilled" participants (the bottom quartile in Dunning and Kruger experiments) would be better described as "skilled participants who were on a bad day".



2.2. Re-reading of the Dunning-Kruger effect: a type 1 deficit

The critical question in relation to the goal of this PhD concerns the level of the deficit. The authors explicitly claimed that the deficit was metacognitive in nature. Yet, some results presented in the original article already favored an interpretation in terms of impaired task-competence, rather than impaired metacognition. For instance, in study 4, the authors manipulated competence and measured how it influenced metacognitive skills. There were 2 phases. In phase 1, participants ($N = 140$) were asked to solve ten problems based on the Wason selection tasks, and then had to estimate how many problems they were able to solve. In phase 2, half of the participants (the test group, $N = 70$) received a short training in logical reasoning, while the other half did an unrelated task (control group, $N = 70$). Then, all the participants were given their own tests and had the opportunity to indicate whether their initial responses were correct or not. I reproduced and summarized the results from the original article within a single figure for more clarity (Figure 23).

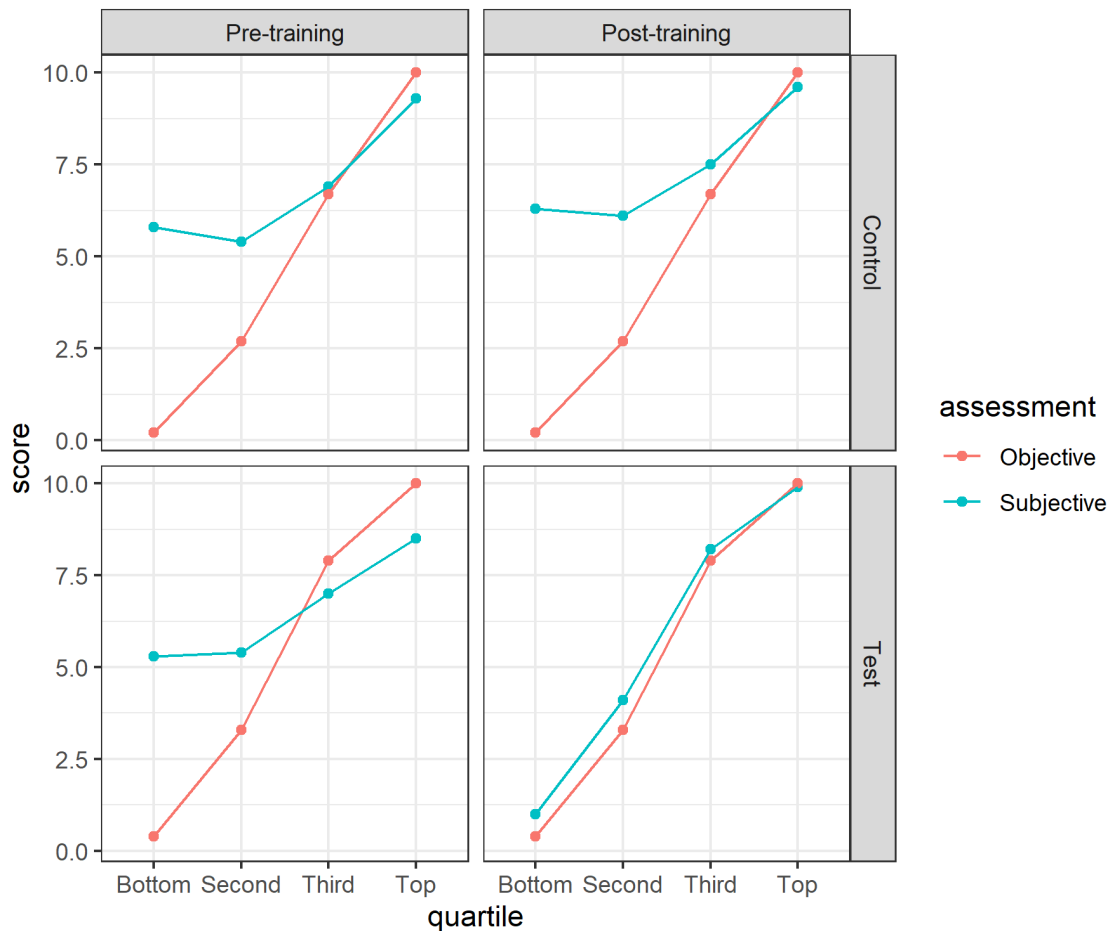


Figure 23. Effect of first-order training on metacognition. Graphic summarizing the results reported in Table 2 from Dunning and Kruger (1999). Objective scores are in red (number of solved Wason problems out of 10), subjective scores in blue (global estimations of the number of solved problems out of 10). Pre-training refers to scores obtained in phase 1, common to both groups. In post-training, the pattern of results shows a better calibration of subjective assessments to objective assessments in the test group (trained in the Wason task), compared to the control group (not trained).

The pre-training results displayed a typical Dunning-Kruger effect (DKE) in both groups, with low performers overestimating their own performance. In post-training, the DKE remained in the control group, but it disappeared in the test group, where trained participants were able to better identify their own errors. The authors originally formulated their conclusion in the following terms: “mediational analyses revealed that it was by means of their improved metacognitive skills that incompetent individuals arrived at their more accurate self-appraisals [in the test group]”.

Given these patterns of results, contrary to Dunning and Kruger I would conclude that metacognitive skills per se were not impaired in unskilled participants. Indeed, what the authors found is that by gaining the first-order skill to do the task the participants automatically regained the ability to properly discriminate their errors from their correct responses, thus suggesting that the metacognitive ability was not impaired to begin with.

Furthermore, the dependence on task-performance of the DKE has been recently demonstrated with a model of rational self-assessment fitted on a large data set collected online (N ~ 3500 participants) on two tasks: grammar and logical reasoning (Jansen et al. 2021). They compared two models: one model of Bayesian inference based on prior beliefs about ability, and a second model implementing a direct dependence between task-performance and the monitoring of correctness. Having shown that the second model explained best the DKE, the authors ruled out a DKE interpretation in terms of metacognitive deficits. It would be interesting to test in future work if a similar mechanism could explain overconfidence in errors among patients with schizophrenia, which would go in hand with our findings of preserved metacognitive performance under matched-levels of first-order performance.

In the next section, I further develop the conceptual reasons why overconfidence in errors should not in principle be interpreted as a metacognitive deficit.

2.3. On the importance of distinguishing sensitivity from bias with psychotic patients

“Instead of saying that an hallucination is a false exterior percept, one should say that the external percept is a true hallucination” Taine, 1870 (cited in Corlett et al. 2019)

When assessing metacognition in the specific case of positive symptomatology such as hallucinations and delusions, it is crucial to define confidence as a subjective probability of being correct, as opposed to an objective probability. Indeed, the ground truth for confidence estimation is what the patient perceives within its first-person perspective, and not what the experimenter thinks the patient should perceive from its third-person perspective. For this very reason, it is of primary importance to use a measure able to distinguish sensitivity from bias (e.g. psychometric curve, SDT model, confidence efficiency model), where the bias is a proxy for the tendency to hallucinate (Bentall and Slade 1985). In this context, as Mamassian and de Gardelle highlighted (2021), metacognitive sensitivity is

best described as self-consistency (i.e. the tendency to provide similar confidence judgments under similar internal signal) according to idiosyncratic criteria (i.e. regardless of the first-order bias). And since a liberal criterion can be interpreted as a tendency to hallucinate (i.e. an inclination to report having perceived stimuli which were objectively absent, Bentall and Slade 1985; Moritz et al. 2017), these measures are adequate to disentangle metaperceptual sensitivity from positive symptoms occurring at the first-order level. To illustrate this point, a patient hallucinating voices might be very confident that a voice has been heard. Should we conclude that confidence is abnormally processed? As far as a voice is actually heard (whether the origin of the voice is objective or hallucinated is irrelevant here), the answer is no. Being confident in hearing an objectively absent noise does not necessarily imply a metacognitive distortion since the subjective phenomenological space is the ground truth. Quoting Mamassian and de Gardelle (2021): “the bias arises here at the perceptual level and not at the metacognitive level per se”. Therefore, assessing confidence while taking into account first-order representations (either true perceptions or hallucinations) is a promising way to single out the metacognitive component of confidence generation.

Of note, in the context of social interactions where communicating one’s confidence is an adaptive strategy for collective decision-making and social learning (Bahrami et al. 2010; Frith and Frith 2012), the ground truth should be the intersubjective rather than the subjective experience. Therefore, persistently communicating an abnormal level of confidence in a social context might be linked to the inability to infer the perspective of others (as suggested by recent studies showing impairments of perspective-taking in schizophrenia Eack et al. 2017; Kronbichler et al. 2019), or the inability to properly update one’s model of the world (Corlett et al. 2019; Nassar et al. 2021), rather than internal monitoring impairments.

2.4. No overconfidence in errors in our samples

Last but not least, I was surprised to notice that contrary to the existing literature our samples of patients (chapters 2 and 3) were not overconfident. I have already discussed the possibility that depression was a mediator in lowering confidence levels, but this hypothesis was difficult to test since most studies did not report depression.

Another possible reason why patients were not overconfident is the fact that high-level remediation strategies like metacognitive trainings (Moritz et al. 2014), or psychosocial rehabilitation (Franck 2021) become more frequent in clinical practices. This could have generated a form of “auto-stigmatization” (Dubreucq 2020) where patients lower

their confidence ratings as a consequence of internalizing shared stereotypes about their mental conditions.

3. Metacognitive deficits and clinical symptoms

Further evidence that local metacognition did not capture clinically relevant dimensions of schizophrenia came from the absence of correlation between individual metacognitive abilities and clinical scores (PANSS positive, negative and disorganized scores; and insight score with BCIS scale) as reported in our experimental chapters 1 and 2, as well as reported by Faivre et al. (2021).

We reasoned that the null correlations obtained in our meta-analytic work (chapter 1) might have been due to the fact that we only had access to summary statistics, namely the average and standard deviation for each variable of interest at the group level. To remedy this issue, we created a dedicated online repository (<https://osf.io/cfm5d/>) similar to the confidence database in healthy participants (Rahnev et al. 2020), and we invited all contributors of our meta-analysis (42 studies included) to share their anonymized original datasets with single-trial granularity. The expected gain in statistical power would have enabled us to further explore the correlations between metacognitive performance and clinical scores. Despite a handful of enthusiastic answers to this proposition, this project was unfortunately aborted due to a lack of positive responses among the contributors.

4. Reality-monitoring

During my thesis, I wanted to dedicate an experiment to delve deeper into the construct of *insight*. I reasoned that insight was more about reality-monitoring than performance-monitoring. Going back to the phenomenology of poor insight, the experiential framework of patients is characterized by a “loosening of ‘common sense’ constraints”, in the sense that their experience might “no longer be ruled by the ‘natural’ certitudes concerning space, causality, and noncontradiction” (Henriksen and Parnas 2014). In experimental settings, it is close to the notion of “liberal acceptance”, namely that patients are less prone than healthy controls to dismiss delusional interpretations concerning administered pictures despite their implausibility (Moritz et al. 2017). Furthermore, in a reality evaluation task where pictures depicting either real or unreal scenes are presented to the participants, patients were less accurate than healthy controls (Lee et al. 2015). In their framework, Lee and colleagues propose that reality evaluation relies on three sequential phases, namely “context appraisal” based on memory processes, “relational reasoning” where the stimulus is compared to background knowledge about reality or “norms of reality”, and “declarative

memory”. As episodic memory impairments are consequent in patients (Gopal and Variend 2005; Heinrichs and Zakzanis 1998; Schaefer et al. 2013), and because it is challenging to control for memory performance between groups in experimental protocols (Rouy et al. 2021), it might be difficult to isolate the metacognitive component of reality evaluation. One way to get rid of the memory processes involved in reality evaluation would be to investigate the “sense of reality”, as opposed to “judgments of reality”, the former being non-reflective, automatic, and non-propositional (Fortier 2018).

Therefore, I planned to go on a 5-month in-doc under the co-supervision of Dr. Roy Salomon who carries out a specific project called “UnReal” at Bar-Ilan University in Israel (<https://salomonlab.org/unreal/>), where the sense of reality is investigated using virtual reality and fine-grained psychophysics. In particular, I aimed at elaborating a protocol enabling to disentangle local metacognition from the sense of reality, thus providing evidence for different and partially independent hierarchical levels of metacognitive evaluations. Unfortunately, this project was aborted due to the sanitary crisis.

This part is a speculative section resulting from dialogs I had with patients, as well as thought experiments. Regarding the phenomenology of hallucinations, I will argue against the view that hallucinations reflect a deficit in perceptual reality monitoring (Dijkstra et al. 2022).

4.1. Hallucination vs hallucinosis

As mentioned in the introduction, contrary to hallucinosis which are conceived as “false hallucinations” or hallucinations with insight (Carota and Bogousslavsky 2019), true hallucinations are - by definition - endowed with a sense of reality, such that patients have no insight that what they perceive do not reflect a physical state of the world. Under the framework of SDT, the tendency to hallucinate has been linked to a perceptual bias leading to higher rates of false alarms (Bentall and Slade 1985), and in turn false alarms triggered by conditioned hallucinations have been associated with overconfidence and positive symptoms (Powers et al. 2017). While this framing of hallucinations in terms of SDT parameters suggests a first-order deficit, many authors endorse the view that hallucinations result from a metacognitive deficit in terms of reality monitoring (Bentall 1990; Brébion et al. 2008; Dijkstra et al. 2022; Lau 2019; Lee et al. 2015; Rankin and O’Carroll 1995; Simons, Garrison, and Johnson 2017). I will argue that it depends on the type of hallucination. Some hallucinations are basic, others are more complex (Fénelon 2014) and it might make a difference regarding reality-monitoring.

4.2. Perception vs imagination

In particular, Dijkstra et al. (2022) argued that hallucinations reflect a failure to recognize the delineating properties between perception and imagination, thus confusing one with the other (see Table 2). For instance, one would confuse an imaginary content with a perceptual content in case one fails to recognize its poorly detailed precision, or fails to notice that imaginary contents are predictable.

	Perception	Imagination
Precision	Clear and richly detailed	Vague and less detailed
Cognitive control about phenomenal contents	No. Driven by external objects	Yes. Contents of imagination can be voluntarily determined and manipulated
Predictability	Eye-movements lead to predictable sensory change	Eye-movements do not lead to predictable change

Table 2. Summarized from Dijkstra et al. (2022). Distinguishing properties between perceptual and imaginary contents.

From table 2, it becomes clear that the reality-monitoring hypothesis of hallucinations is based on the assumption that hallucination-like contents do not share the same phenomenological properties as perceptual contents. And indeed, this hypothesis seems valid in cases where patients fail to properly investigate the incongruent properties of their singular percepts. However, there are multiple types of hallucinations varying in complexity and mechanisms (Fénelon 2014). Then, what if hallucinatory percepts were endowed with true perceptual properties? In other words, what if hallucinatory percepts were strictly indistinguishable from perceptual contents? In the next section, I would like to share anecdotal - yet puzzling phenomenological reports of hallucinations that are difficult to reconcile with the reality-monitoring hypothesis.

4.3. Some peculiar reports of hallucinations

Here, I would like to share accounts of hallucinations from Anna (pseudonym), who was a hospitalized patient at the time I met her. After having tried her best to perform the experimental tests described in chapter 2, I asked a few questions to Anna about the reasons why she was internalized. We rapidly reached the theme of hallucinations and it

certainly turned out to be one of the most fascinating reports I have ever heard, in such a way that it continues to make me think until this day. She told me affectionately about a 10-year old boy, called Stanley, who frequently visited her. He was wearing shorts and a beret. At first, he appeared to her in the street and engaged in conversation, then she saw him at the stairwell of her building, and finally he appeared in her room. Only at this time, she realized he was not a normal boy, because when her husband arrived home he walked through the body of Stanley without even noticing his presence, and subsequently the boy vanished. When I asked Anna about her reaction, and in particular whether she was afraid, she told me that she was used to these kinds of phenomena since she was a child. She could see animals like birds or cats, that anyone but she could see. The interesting thing was the rich interactions that Anna entertained with Stanley. They continued to meet on a regular basis. Sometimes he came upon her calling him, and sometimes he came spontaneously. He could tell a lot of details about himself, and revealed that he was shot and died during the first world war in 1914. Then I asked Anna how she dealt with these phenomena within her social life. She admitted that it could be inconvenient, and even embarrassing, mentioning cases where she started to speak with someone on the street, and soon realizing from the reactions of people around, that nobody but her could see the person she was speaking to. I was astonished, so I questioned her about the phenomenology of her percepts, wondering whether there was any clue she could rely upon to correctly identify her singular perceptions as such. She replied that it was impossible for her to know before a reality test was made - i.e. before engaging in an interaction and witnessing the following reactions of surrounding people. And to my amazement, Anna told me about an efficient strategy she naturally developed with her (real) daughter who acted as the secret-keeper, or as the reliable external witness. So, when Anna went out in town she could discreetly ask her daughter whether this person or that animal was real or not, and eventually, it would prevent her from being considered a freak.

Getting back to considerations about reality-monitoring, it seems that this specific account of hallucinations depicts a different picture from Dijkstra et al. (2022). Contrary to imaginary contents, Stanley had consistency in time, he had distinguishable perceptual traits, he acted as an independent agent with an idiosyncratic character and was not predictable. In other words, the hallucination-like percepts described by Anna seemed to share indistinguishable properties with true perceptions. In this sense, a perfectly functioning reality-monitoring ability would not - in principle - be sufficient to draw the line between these kinds of hallucinations and true percepts, or in other words to distinguish between percepts that are constitutive of a shared reality and those that are singular. Even if we consider perceptual reality-monitoring as conceived by Lau (2019), where a first-order representation

becomes conscious if it is deemed plausible by a second-order reality monitoring process, these types of hallucinatory percepts could sneak in consciousness by passing the reality-monitoring test.

In the spirit of neurophenomenology (Varela 1996; Varela, Thompson, and Rosch 2017), taking into account the first-person perspective from patients might help us to refine our theories, especially in cases where accounts are precise and (paradoxically) insightful. The next section deals with another way to conceive the specific kind of hallucinations reported by Anna, without involving a deficit in reality-monitoring. The reader should be informed that this part and the next are surely the most speculative.

4.4. SDT considerations: against the interpretation of “noisy” hallucinations

In SDT terms, a hallucination is conceived as a false alarm (i.e. the detection of an objectively absent stimulus). A false alarm is in turn the result of a piece of evidence sampled from a distribution of “noise” that exceeded the perceptual threshold, and thus erroneously categorized as “signal”. This conception of hallucinations in terms of “noisy” evidence implicitly contains the notion that the evidence is by nature scarce, volatile, not reliable and therefore it is deemed inappropriate to be confident in resulting detection errors. Indeed, when adopting a normative view opposing insightful healthy individuals with insightful patients, hallucinations might be interpreted as obvious failures to recognize or categorize something which is nonetheless easy to recognize. However, if the specific case of hallucination reported by Anna is taken seriously, then the source of evidence giving rise to false alarms can hardly be interpreted as “noisy” in its implicit sense (i.e. scarce, volatile, not reliable).

On the contrary, one might consider the complexity and coherence of some kinds of hallucinations to be comparable with dream-like perceptions (Waters, Barnby, and Blom 2021). In particular, “serotonergic hallucinations” - resulting from a 5-HT_{2A} receptor blockage -, such as those induced with psychedelic drugs tend to be more dream-like (Jacobs 1978). As mentioned in the introduction, patients that are non responsive to dopamine D₂ receptor antagonists are generally responsive to Clozapine, which is an antagonist of the 5-HT_{2A} receptor. Furthermore, the activity of the serotonergic 5-HT_{2A} receptor has been linked to the frightful hallucinations that one undergoes in sleep paralysis, a dream state often described as involving the felt or seen presence of a ghost-like individual and a panic-like fear reaction (Jalal 2018). Moreover, the comparison with dream-like experiences can be meaningful for

the understanding of hallucinations from a first-person perspective, as anybody already experienced an erroneous yet compelling sense of reality while dreaming.

What follows should be considered a thought experiment rather than a demonstration. Here, I would like to share the intuition I gained about some kinds of hallucinations from a specific lucid dream I had. One night I was dreaming, and suddenly I realized that everything that happened to me was nonsense. At this point, I began to question the substance of this strange world I was evolving in, and I willingly made reality-checks. I started to count my fingers and in bewilderment I realized I had 6. Now, it was clear to me, I knew I was dreaming. This metaphysical shift in my belief system produced a joyful wonderment, since this is the very moment I realized that perceived objects were not made of matter anymore, but mere projections of the mind instead, groundless appearances. Mind was the online creator of this whole complex and fascinating environment made of landscapes, objects and people. In this sense, the delineation between the self and the world faded away: there was no more ontological difference between myself as a perceiver, and the world that was perceived. Out of curiosity, I rushed into examining the texture of the dream, how rich was it? I started to stare at the palms of my hands and I could see my fingerprints. I got closer to a wall nearby and scrutinized its texture, it was rough to the touch and I could identify thousands of tiny bumps. I did not expect such a resolution. Then I told myself "Well, I am dreaming, so my body is not really substantial, it is not physical. As a consequence, I should be able to cross through my body with my arm." And with this hypothesis in mind, I made a slow circular movement with my right hand toward my left hip, expecting that my movement would not be stopped by the visible boundary of my body. And there, at the point of contact, my hypothesis was refuted. Not only did I hear the sound of the contact, exactly at the moment I saw it touching my body, but I could also feel it vividly. Multisensory integration was convincingly preserved. At this precise moment, with this evidence of congruent perception, I went through a second reversal in my metaphysical belief and entered a state of deep puzzlement: "Everything feels so real, how could I be sure anymore that I am dreaming?" And then a rational thought popped out: "My left hand had 6 fingers a few minutes ago, It's definitely a dream". I wanted to prove to myself that my body was not physical so I grabbed an ax and I was about to cut my arm, but before I did anything, another rational thought interrupted me: "I already checked and my body felt as real as possible. If I cut my arm, no doubt I will go through intolerable pain." And for a while I was stupefied, dumbstruck. My dream-like perceptions felt so real that I lost the initial certainty that I was dreaming. I couldn't distinguish anymore between what was real and what was not. And my dream ended up on this confusing note, in a state I will later refer to as "reality puzzlement".

Here is the core of my argument: hallucinations (or at least a specific kind of hallucination) are not mere mental imagery contents - as proposed by Dijkstra et al. 2022 -, they are like dream perceptions. They are convincing and hard to identify because upon investigation they have the same phenomenological properties as those of wakeful perceptions. Overlaps between dream states and hallucinations in terms of subjective descriptions and underlying mechanisms have been identified (Waters et al. 2016). Despite the need for further evidence, psychotic hallucinations might be understood as the intrusion of dream-like evidence within wakeful life, perfectly integrated within sensory evidence. Hallucinations would arise, for instance, from a parallel memory stream of evidence (e.g. resampling of memory contents actualized as presently lived) interpreted as sensory evidence.

Of course, in the absence of a physical stimulus, SDT cannot distinguish between a sample of evidence coming from a “noise” distribution (i.e. lacking internal coherence) and a sample of evidence coming from a distribution of more complex objects such as dream-like percepts (Figure 24.A.). As a consequence, the estimated sensory noise (i.e. the spread of the noise distribution) is inflated (Figure 24.B) resulting in a seemingly liberal criterion (Moritz et al. 2014) and a decreased sensitivity, which corresponds to what we observe among patients with schizophrenia.

To conclude on this section, the attempt to explain hallucinations solely in terms of SDT parameters might miss the variety of hallucination properties (Laroi 2006). In this sense, bayesian inference theories like the *circular inferences* (Jardri and Denève 2013) or other predictive coding accounts (Corlett et al. 2019) might be more promising approaches to explain levels of complexity of hallucinations. However, as we will see in the next section, we still have to explain the sense of reality of hallucinations, which might be distinct from confidence or conviction.

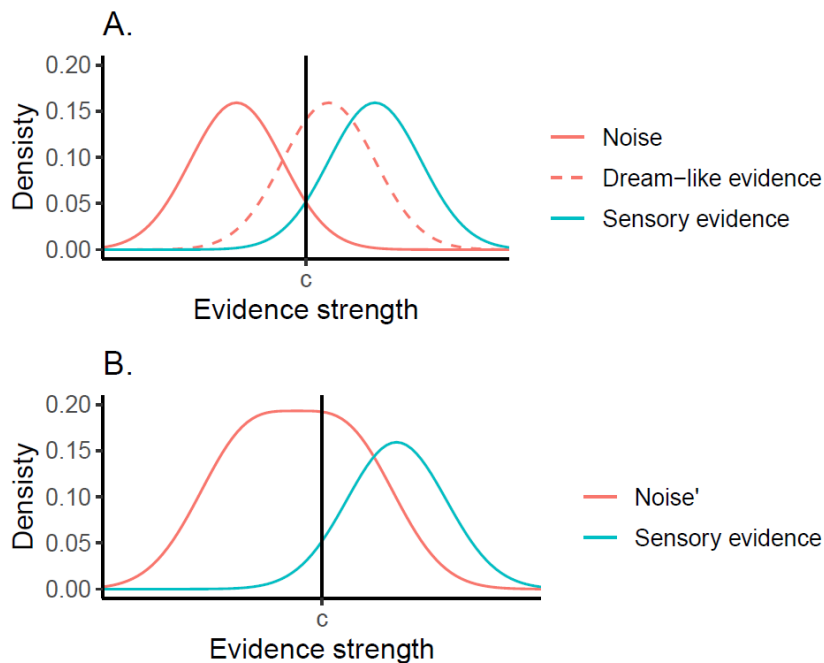


Figure 24. Distinguishing noise from dream-like evidence. A. Distributions of evidence, in the speculative case where hallucinations would originate from samples of “dream-like” evidence (red dashed curve), distinct from sensory noise (solid red curve) B. The red curve, called ‘Noise’, is the addition of two distributions: Noise and Dream-like evidence. Indeed, according to SDT categories, in the absence of a physical stimulus, “dream-like” samples of evidence fall under the label “noise”. Therefore, the noise distribution combines (or confounds) the two distributions.

4.5. Sense of reality

The analogy between hallucinations and dreams might be useful to elaborate different categories of hallucinations, in terms of degrees of insight and availability of reality-monitoring. Before I get into these analogies, a useful distinction to have in mind is the difference between the “sense of reality”, and “judgment of reality” (Fortier 2018). To get the flavor of the distinction, consider you are wearing a virtual reality head-mounted display, and you arrive at the top of a virtual cliff. Despite you “know” it is not real (reality judgment), and you “know” the ground around you is flat, you might experience some fear (sense of reality) to dare stepping into the seemingly vertiginous void.

With this distinction in mind, I will propose other thought experiments to try to have a better grasp on the phenomenological diversity of psychotic states, and the degrees of insight. I already introduced the “reality puzzlement” state (state A), which is endowed with a strong sense of reality combined with an impossibility to judge whether the experience is real

or not. I will now suggest that psychotic states may successively be conceived as a dream state (state B), a lucid nightmare state (C), and finally a lucid dream state (state D). As mentioned before, the power of the dream analogy lies in its graspable character, since anybody already felt an abnormal sense of reality within the dream realm.

State B: In a dream state, one hardly notices the perceptual inconsistencies before waking up. In Bayesian terms, it is a “weak prior” state: Nothing is surprising. The dreamer is immersed and believes the situations he finds himself into without questioning. There is complete adherence to this fantasy world. The dream state is well aligned with the hypothesis of a deficit of reality monitoring, since one fails to explicitly notice the peculiar phenomenal properties of non-lucid dream-like experiences compared to those lived in wakeful life. Often, after awakening from a dream, one recovers the ability to judge the implausibility of the situations we were immersed in. In other words, if one were lucid, one would have noticed bizarre features such as the low stability (volatility of environment) of the dream-like experience, or some other improbable situations (like finding ourselves suddenly naked on the street), and one would have stopped adhering to it like in the case of hallucinosis.

State C: In the case of a lucid nightmare, one knows one is dreaming but it does not prevent the dreamer from feeling endangered and scared by threatening events or agents. In a sense, the reality-monitoring ability that enables the dreamer to be lucid about the dream-like nature of the experience is not sufficient to recognize every aspect of the dream as the mere projection of the mind, hence dangerless experiences. It sounds like a functional split between a correct judgment of *ir*-reality, and a persistent sense of reality that is encapsulated¹³, i.e. resistant to any top-down strategies. State C is corroborated by several studies. For instance, auditory hallucination-related distress has been shown to increase with both the felt reality of hallucinations (Gaudio and Herbert 2006) and the degree to which the patients believed they were endangered (Hill et al. 2012), despite them recognizing these mental events as hallucinations. This result has been found also with visual hallucinations (Gauntlett-Gilbert and Kuipers 2005) where negative appraisals of hallucinations (in terms of bad outcome predictions) were correlated with hallucinatory distress. This “double-awareness state” (Sacks et al. 1974) where patients apparently *know* that they are hallucinating but nonetheless *feel* endangered suggest nested levels of metacognitive evaluations.

¹³ referring to Fodor’s modularity of the mind (1983), and corroborated by clinical considerations (Semerari et al. 2003)

Toward recovery: entering state D. At the moment of gaining full lucidity in a dream, the sense of reality falls apart simultaneously with the judgment of irreality. In the case of psychosis, reaching this state where threatening percepts lose their sense of reality (hence their distressful character) is part of the process of recovery (Sacks et al. 1974). In this process, the pernicious aspect of hallucinations might be unraveled by relying on high-level clues. One might start to query: Who can share these perceptions with me? Do these perceptions really have causal effects on myself and my physical environment? Another patient I met told me about the process of believing less and less in his singular perceptions, or at least becoming agnostic about their origin. He related the following event and commented on it:

One day, I was in my bedroom, and suddenly I could hear voices, endowed with a noticeable warmth, an enjoyable tone which sounded welcoming and kind. I wanted to know who they were so I simply asked "Who are you?". The voices replied that they were gods. "What kind of gods?" I asked in return. "Are you Hindu gods? Greek gods?" And still with their palpable kind presence, they replied that no matter the category, all I needed to know was that they were gods, and that it meant that they were powerful. In particular, they could destroy me if they wished. At this point I became very amused by this interaction, because these words, despite their threatening nature, were pronounced without any aggressivity. So, I started to tease them: "Hum, so you are so powerful that you can destroy me? Interesting... Well, let's try, I want you to destroy me!". And still with a confident and kind tone they replied "Well, it is not appropriate to destroy you right now." At this point, I laughed, it was not credible. And because it happened several times, voices saying things without any consequences on the physical reality, I started to become very skeptical about them. Fascination slowly decreased. To be honest, I don't really know what to do with these experiences. At first, I thought I was learning some deep truths about reality, but now I'm quite critical about them. Maybe they are real but useless, or maybe they are my own fantasies and so even more useless. It gave me hints that I should not pay so much attention to it.

This story was simply brilliant. So much critical thinking and resilience. It reminded me of the movie adaptation entitled "A beautiful mind" that I already mentioned in the introduction, from the life story of John Nash, the famous mathematician from Princeton who suffered from schizophrenia. In the original biography which inspired the movie, he was described as "slightly cold, a bit superior, somewhat secretive" (Nasar 1998), a brilliant and solitary guy gifted with a very singular way of thinking. John Nash progressively learned how

to recognize the deceitful character of his delusions, and not to pay too much attention to it, just as in a lucid dream.

Of note, since John Nash himself criticized the visual hallucinations depicted in the movie as alien to him¹⁴, I was skeptical that such veridical hallucinations could even be possible. However, my skepticism has decreased after I met Anna. The hallucinated characters she met were not questioned as unreal because of their complexity and their internal coherence: they had seemingly a life of their own. Their veridical properties made it difficult for doubt to arise, like in a dream state. Dream agents become convincing alterities because of the convincing way they behave as external agents. Therefore, the movie adaption is still relevant because it provides us with a flavor of “what it is like”¹⁵ (Nagel 1974) to have a realistic hallucination.

The following table (Table 3) summarizes the mental states evoked in this part, providing phenomenological distinctions for the categorization of hallucinations. The advantage is to propose different types of hallucinatory mental states, originating from different processes: Dream-state hallucinations result from a full deficit of reality monitoring; Lucid-nightmare hallucinations stem from a partial deficit of reality monitoring; Reality-puzzlement hallucinations (like in the case of Anna) are neither a first-order deficit nor a second-order deficit, but rather a first-order peculiarity that takes dream-like evidence as sensory evidence. And finally, lucid dreaming which is ideally the recovery state resulting from a high-level metacognitive training.

		Sense of reality	
		Yes	No
Judgment of reality	Yes	Dream state (state B)	(Derealization)
	No	Lucid nightmare (state C)	Lucid dreaming (state D), Hallucinosiis
	Guess	Reality-puzzlement (state A)	

Table 3. Categorizations of dream states (except for derealization), depending on the felt sense of reality and the judgment of reality.

¹⁴ <https://www.youtube.com/watch?v=UiWBWwCa1E0>

¹⁵ Of course we don't have access to the qualia of John Nash, but we do gain an insight of what an hallucination might look like from a first-person perspective.

5. Limitations

5.1. Theoretical limitations

The claim that local metacognition is relatively preserved among patients (schizophrenia in our meta-analysis, but also in FCD patients in Bhome et al. 2022) essentially comes from studies using the meta-d' framework. However, there are some limitations inherent to this measure.

First, despite the elegance of the metacognitive efficiency index (M-ratio) which theoretically controls first-order performance, its formal expression nevertheless contains a dependence on first-order performance that is counter-intuitive. As mentioned in Box 2, the formal expression of M-ratio is the following:

$$Mratio = \frac{\sigma_{sensory}}{\sqrt{\sigma_{sensory}^2 + \sigma_{meta}^2}}$$

where $\sigma_{sensory}$ refers to sensory noise,

and σ_{meta} refers to metacognitive noise.

Now, if we plot M-ratios as a function of $\sigma_{sensory}$, for fixed values of σ_{meta} , we can see that metacognitive efficiency increases as a function of sensory noise (Figure 25). In other words, the formal expression of the metacognitive efficiency index inherently contains the contradictory prediction that a decreased type 1 sensitivity (i.e. higher sensory noise, which means lower task-performance) leads to an increased metacognitive efficiency, given a constant level of metacognitive noise. This result has been observed with experimental data (Bang et al. 2017).

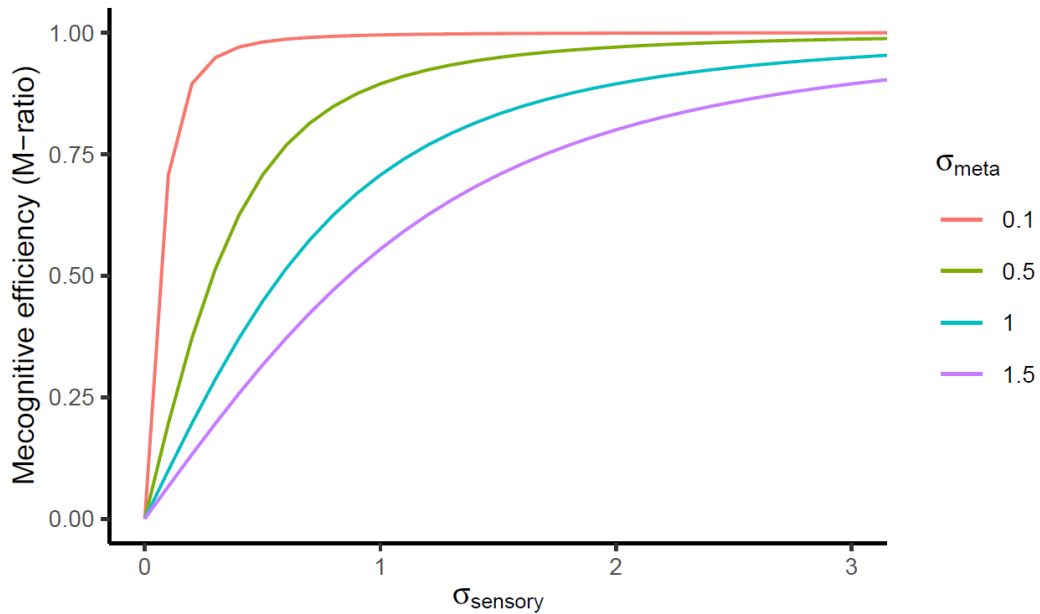


Figure 25. Metacognitive efficiency as a function of sensory noise, for different values of metacognitive noise.

The problem in the context of assessing metacognition among individuals with schizophrenia is obvious: given that these patients suffer from first-order deficits compared to healthy controls, the metacognitive efficiency index might have artificially reduced the magnitude of the metacognitive deficit among this population.

Second, it has been shown that the metacognitive efficiency index is not perfectly independent from metacognitive bias, in the sense that higher confidence ratings lead to higher values of M-ratios (Xue, Shekhar, and Rahnev 2021). Again, it is problematic in the context of the evaluation of metacognition within a population of patients that are known to be overconfident (Moritz et al. 2014; Hoven et al. 2019). An overall bias of overconfidence might have also contributed to reducing the magnitude of the metacognitive deficit among patients with schizophrenia.

Another critical point about M-ratio is its dependency on some parameters of decision dynamics (Desender, Ridderinkhof, and Murphy 2022). When quantifying metacognitive efficiency within an evidence accumulation framework, i.e. when taking into account the dynamics of the decisions, while manipulating the response caution (or the speed-accuracy trade-off), it has been shown that M-ratio increases with decreasing decision boundaries (Figure 26).

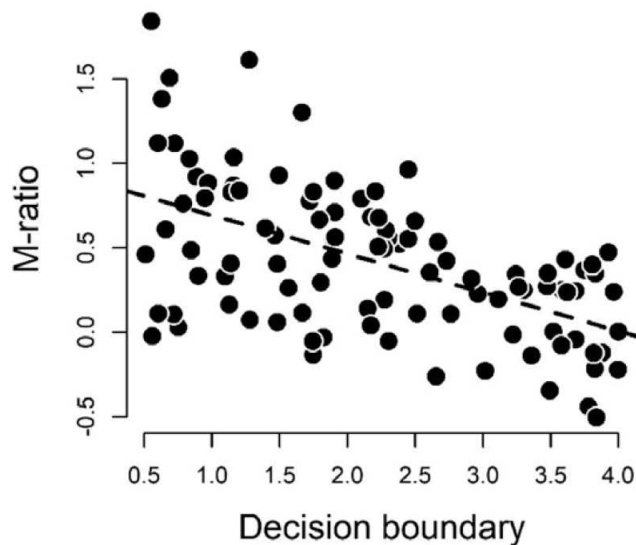


Figure 26. Correlation between M-ratio and decision boundary. Simulated data showing a negative correlation between M-ratio and decision boundary. Reprint from Desender et al. (2022)

This feature of M-ratio might again be problematic when assessing metacognition within a population of patients that is known to have “Jump to conclusion” (JTC) behaviors (for a meta-analysis, see Dudley et al. 2016), i.e. committing early in a decision in spite of a poor amount of evidence. Within an evidence accumulation framework, JTC can be understood either as an elevated drift rate, an elevated decision bias, or a reduced decision boundary. All these types of parameter changes would lead to accumulating less evidence before committing to a decision. Experimentally, the JTC has been linked to an elevated decision bias in schizophrenia (Limongi et al. 2018), which could in turn be involved in an artificial elevation of M-ratios (even though not discussed in Desender et al. 2022).

However (and fortunately), under controlled and standardized experimental conditions with low-level perceptual stimuli (e.g. random-dot kinematogram), none of the decision parameters including decision boundary were found to differ between patients with schizophrenia and healthy controls (Faivre et al. 2021), which tends to rule out the possibility that M-ratios were artificially elevated among patients with schizophrenia.

Last but not least, M-ratio might not even reflect a metacognitive process. Here, we should return to epistemological considerations about the metacognitive M-ratio measure. Indeed, I have already mentioned that a confidence rating is strictly speaking a type 2 task rather than a type 2 process. Therefore, measuring metacognition with a type 2 task assumes that a type 2 process underlies the type 2 task. However, the initial SDT model for type 2 responses is a single-process model of confidence, i.e. it does not specify a specific

processing stream for metacognitive evaluations, and thus it is in principle impossible to disentangle between the contributions of first-order and second-order processes to the type 2 responses. In these conditions, it has been conceptually demonstrated that controlling for first-order performance is not sufficient to remove the confound between second-order sensitivity from first-order sensitivity (Paulewicz, Siedlecka, and Koculak 2020). However, this criticism might be attenuated by modeling efforts that have shown that a hierarchical process model of confidence (which is the model presented in Box 2 and discussed in part 2.3.1., with serial sources of noise) outperforms the single-process model (Maniscalco and Lau 2016). And in particular, the confidence efficiency framework (Mamassian and de Gardelle 2021) providing a detailed generative model of confidence should make the confidence efficiency index immune to this criticism.

5.2. Clinical limitations

5.2.1. Samples of patients with schizophrenia are heterogeneous

Most studies included in our meta-analysis made use of the DSM IV (1994-2000) for the diagnosis of schizophrenia. In the DSM IV, a patient received the diagnosis schizophrenia in case he or she had at least two symptoms among the following:

1. delusions
2. hallucinations
3. disorganized speech
4. disorganized or catatonic behavior
5. negative symptoms (not expressing any feelings or emotions)

By means of combinatorics, 27 different combinations can be obtained upon these criteria, involving fairly distinct profiles. For instance, an individual having hallucinations and delusions is equally diagnosed with schizophrenia as an individual with disorganized speech and negative symptoms. The DSM IV diagnosis itself can easily explain the great heterogeneity of patients diagnosed with schizophrenia.

More homogeneity has been gained with the DSM-V (2013) diagnosis, which requires that at least one of those symptoms is among the following 3:

- delusions
- hallucinations
- disorganized speech

From this new criterion the number of combinations drops to 12. Yet, the previous remark about the variety of profiles that fall under the diagnosis of schizophrenia is still valid.

5.2.2. Toward a transdiagnostic approach

As already discussed in the introduction, diagnoses based on a collection of symptoms might mix distinct etiologies, and in turn the resulting clusters of co-existing symptoms might blur distinct sources of variation. Rouault and al. (2018) used a transdiagnostic approach with three self-reported psychiatric dimensions (Anxious-depression, compulsivity, and social withdrawal) on a large dataset among the general population (N = 995). They have shown that the variance in local metacognition was better explained using these psychiatric dimensions compared to any psychiatric diagnosis. Interestingly, Anxious-depression was found to lower confidence levels and heighten metacognitive efficiency, whereas the opposite pattern was found for compulsive behaviors. Since anxious-depression and compulsive behaviors (including intrusive thought in particular) might coexist within the diagnosis of schizophrenia. Therefore, the specific combinations of symptoms constituting the diagnosis itself might have masked the relevant factors contributing to specific metacognitive alterations.

Furthermore, the strength of the transdiagnostic approach to psychosis has already been demonstrated and replicated with the B-SNIP (Bipolar-Schizophrenia Network for Intermediate Phenotypes) consortium (Clementz et al. 2016, 2022). Three “biotypes”, i.e. specific patterns of behavioral and electrophysiological markers have emerged from machine learning clustering applied to a large dataset, irrespective of the diagnosis (Bipolar, schizophrenia, schizo-affective). In contrast, the DSM diagnoses explained less variance in the data. However, some drawbacks regarding the potential application of this method to diagnosis might be considered. First, assigning a biotype to a patient will be costly and time-consuming (and sometimes not applicable since it also requires testing first-order relatives), and second, it might deeply alter the social role of the clinician. Indeed, if the diagnosis 2.0 relies only on biomarkers, at the expense of clinically observable symptoms, the first-person perspective of patients will be informationally irrelevant. The resulting communication gap in the patient-clinician might need some readjustment to ensure treatment observance.

6. Metacognitive training

Finally, in chapter 4 we re-assessed the efficiency of a metacognitive training that was potentially transferable as a remediation strategy for patients, but we concluded the

inefficiency of the training once confounding factors related to instructions and incentives were controlled for.

6.1. The explanatory gap

Even though we have provided evidence that local metacognitive training as proposed by Carpenter and colleagues (2019) was not efficient, we may question the relevance of such an approach as a potential clinical remediation. Under the framework of Seow et al. (2021), the metacognitive training we assessed is tapping into the building blocks of metacognition. Yet, there is a concern from the clinical side (Schnakenberg Martin and Lysaker 2022) that this approach to metacognition - that is focused only on the monitoring of the quality of first-order contents (whatever the domain) - might be too narrow to reach clinically relevant dimensions. It is interesting to note from the correspondence between clinicians (Schnakenberg Martin and Lysaker 2022) and cognitivists (Seow et al. 2022) the gap in the way metacognition is conceived from both sides. The cognitivist definition of metacognition is the one outlined in this manuscript, referring to the ability to monitor the correctness of our decisions in two-choice tasks. From the clinical side, the definition includes the notions of self-reflexivity, understanding of others, one's place within a larger community, and mastery (i.e. the adaptive control of behavior on the basis of the knowledge about the 3 previous dimensions), which are quantified with the Metacognition Assessment Scale (MAS, Semerari et al. 2003). As noted by Seow et al. (2022), the core of the misunderstanding between the two approaches certainly lies in what we consider a reliable and valid measure. Standardized behavioral tasks have been developed as a complementary approach to subjective scales and questionnaires that are amenable to confabulation (Nisbett and Wilson 1977) and to demand characteristics effect (Orne 1962) to a larger extent. However, in line with Schnakenberg and Lysaker, I would agree that a low-level metacognitive training in terms of confidence calibration in two-choice tasks, even if working, would certainly miss the higher-level metacognitive ingredient needed for recovery (e.g. broader self representations integrating various aspects of one's life under one coherent narrative, which should be distinguished from considerations in terms of domain-generality and task-performance). Of course, the attempt to address high-level metacognitive problems by tapping into low-level mechanisms is highly relevant under the hierarchical framework proposed by Seow et al (2022), but dissociations between the two levels (Bhome et al. 2022) are important clues that theories need to be refined (which is actually acknowledged by the whole field of visual metacognition, Rahnev et al. 2021, see Figure 27) and that the present cognitivist approach might miss important leverage elements for the purpose of clinical recovery.

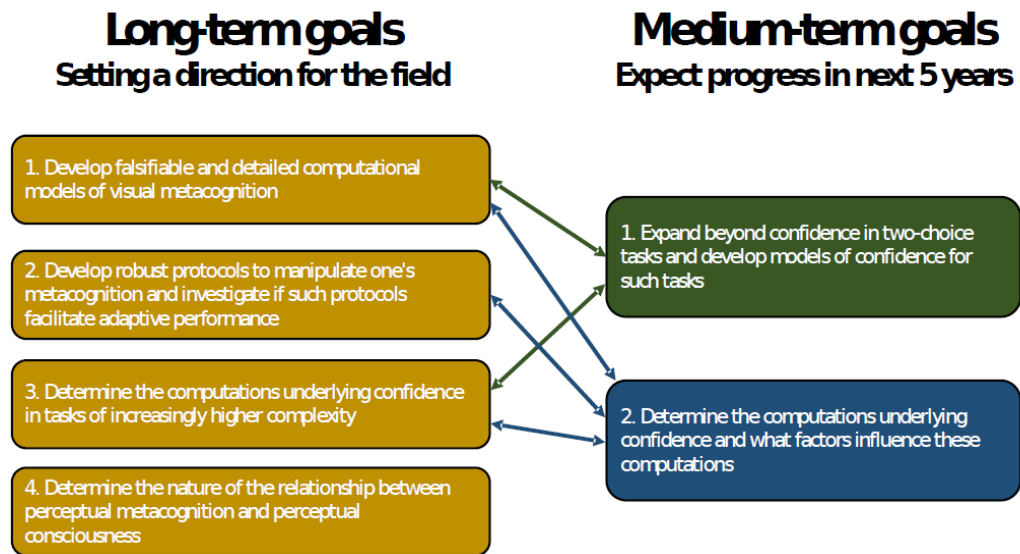


Figure 27. Medium- and long-term consensus goals in the field of visual metacognition. In particular, these objectives include the assessment of more complex metacognitive evaluations beyond confidence in two-choice tasks and the understanding of the underlying computations. Reprint from Ranhev et al. (2021)

To conclude on this part, despite no evidence for local metacognition impairment, high-level metacognitive training tapping directly into explicit beliefs like MERIT and MCT might still be highly relevant. And it is also interesting to note, regarding the different types of hallucinations reported above, that recovery might have a very different meaning from one patient to another. Considering a category of patients who would agree with John Nash that delusions and hallucinations are “essentially a hopeless waste of intellectual effort”, recovery would certainly mean receiving an antipsychotic treatment efficiently reducing positive symptoms (given a reasonable amount of side-effects). However, patients like Anna who developed affective bonds with hallucinatory characters might have a different clinical goal. Anna told me with a sad tone that from the moment she took antipsychotic medication, Stanley was not visiting her anymore. She confessed that she missed him a lot.

It was also interesting to note that most patients I asked the reason why they were medicated answered were either elusive: e.g. “I had anxiety problems so that I could not take care of myself anymore”, or explicitly avoided mentioning the schizophrenia category, e.g. “My psychiatrist thinks I have a schizophrenia disorder, but he’s wrong. Rather, I think I have a narcissistic personality.”, or “I had drinking problems, I have bipolar disorder, but it does not mean I am mad”. Should we conclude that these patients suffer from poor insight into their illness? It is interesting to consider the case of Charles Bonnet syndrome, which refers to a kind of hallucination with insights occurring in non-psychiatric but visually

impaired people. It has been reviewed that these hallucinations cause distress, and that people having this syndrome do not talk easily about these symptoms, fearing being stigmatized as mad (Menon 2005). Among patients with schizophrenia, the fear of stigma also exists (Fénelon 2014) and has been associated with the “insight paradox” (Davis et al. 2020): insight is negatively correlated with the quality of life, i.e. gaining insight into their illness, the quality of life of patients with schizophrenia decreases. Thus, recovery might have several dimensions that are conflicting with one another, and weighing their importance in patients’ daily lives is not a trivial issue.

6.2. The need for larger collaborations

The introduction gave only a glimpse of the varieties of existing theories within the field of research on schizophrenia. Many levels of observation are investigated, from genes to behaviors, and resulting theories are sometimes difficult to articulate with one another. And even among the researchers working on the same level, the varieties of methods and constructs lead to contradicting results and conclusions. Ultimately, the disagreement mentioned above about what should be called metacognition, how it should be measured, and most importantly what is the relevant metacognitive element for recovery, reminded me about considerations from the philosopher of science Thomas Kuhn, in his masterpiece “The structure of scientific revolutions” (1970). Regarding the history of science, Kuhn identified a repetitive structure and outlined four stages within each cycle: the pre-paradigmatic phase, “normal” science, crisis, and revolution.

The pre-paradigmatic phase refers to a phase where there are no shared theories, concepts and methods. Scientists have different background assumptions, and do different kinds of measurements. As a consequence, they never really know whether they are all talking about the same thing, and it makes collaborations very difficult. Rather, everyone is pursuing their own ideas. Eventually, agreement is reached upon a paradigm (theories, concepts, and methods) that enable substantial progress. At some point, scientists are not critical anymore about the paradigm itself and even take it for granted: this is the sign that we have reached the “normal” science phase. Every scientist trusts the paradigm and tries to solve scientific questions within the rules of the paradigm. Anecdotal anomalies consisting of incongruences between predictions and observations are dismissed. But if the paradigm fails to give a proper account for a growing number of anomalies, then scientists become critical about it, which is the signature of a scientific “crisis”. In the end, a scientific “revolution” occurs when a new paradigm endowed with more explanatory power is found.

Regarding this structure, we might best describe the current state of scientific research on metacognition in schizophrenia as “pre-paradigmatic”. In order to conceive an integrative theoretical framework and to reach common agreement upon concepts, methods, and goals, we need to foster interdisciplinary collaborations combining the first-person perspective of patients and the third-person perspectives of both clinicians and cognitivists, much like the “Hearing The Voice” project did (<https://www.dur.ac.uk/hearingthevoice/>). In this regard, we might also be inspired by the recent initiatives from Dobromir Rahnev - e.g. the creation of a large database for confidence experiments (Rahnev et al. 2020), and the call for an agreement upon consensus goals (Rahnev et al. 2021) - who provided substantial efforts to make the field of metacognition reach the status of *normal* science.

Conclusion

In this thesis, we have provided evidence that local metacognition, construed as the ability to monitor one's decisions' correctness in two-choice tasks on a trial-by-trial basis, was relatively preserved in the perceptual and the memory domains among patients with schizophrenia. In particular, we have highlighted the importance of controlling for first-order performance to assess the specificity of metacognitive abilities. We discussed the possibility that overconfidence in errors might result from first-order deficits, much like we would expect in the case of a Dunning-Kruger effect. On the basis of these results together with phenomenological reports of hallucinations, I also discussed the paradoxical possibility that source- and reality-monitoring might be preserved in some very specific cases of hallucinations. As already suggested 30 years ago, different metacognitive deficits might lead to different types of hallucinations (Bentall 1990), and a model integrating this hierarchy of metacognitive evaluations is still missing. Theoretical advances are needed to better understand how the hierarchy of metacognitive processes might interact with one another, ultimately explaining the varieties of hallucinations and improving treatment research.

References

- Allen, S.J., Bharadwaj R., Hyde, T.M., and Kleinman J.E. 2020. GENETIC NEUROPATHOLOGY REVISITED, GENE EXPRESSION IN PSYCHOSIS *Psychotic Disorders: Comprehensive Conceptualization and Treatments*, 163.
- Amador, X. F., & Strauss, S. A. 1993. Scale to Assess Unawareness of Mental Disorders. *Human Sciences*.
- Amador, X. F., Strauss, D. H., Yale, S. A., Flaum, M. M., Endicott, J., & Gorman, J. M. 1993. Assessment of insight in psychosis. *American Journal of Psychiatry*, 150, 873-873.
- Amador, X. F., and David, S. A. eds. 1998. *Insight and Psychosis*. New York: Oxford University Press.
- Amador, X. F., and David, S. A. eds. 2004. *Insight and Psychosis*. 2nd ed. Oxford ; New York: Oxford University Press.
- Analayo. 2003. *Satipatthana: The Direct Path to Realization*. Windhorse.
- Andreou, C., Bozikas, V. P., Luedtke, T., & Moritz, S. 2015. Associations between visual perception accuracy and confidence in a dopaminergic manipulation study. *Frontiers in psychology*, 6, 414.
- Arango-Muñoz, Santiago. 2011. Two Levels of Metacognition. *Philosophia* 39(1):71–82. doi: 10.1007/s11406-010-9279-0.
- Aynsworth, C., Nemat, N., Collerton, D., Smailes, D., & Dudley, R. 2017. Reality monitoring performance and the role of visual imagery in visual hallucinations. *Behaviour Research and Therapy*, 97, 115-122.
- Babashova, S. I. 2020. Comparative analysis of testing heteroscedasticity in nonlinear regression. *Current and Historical Debates in Social Sciences: Field Studies and Analysis*, 211.
- Bäckman, L., & Lipinska, B. 1993. Monitoring of general knowledge: Evidence for preservation in early Alzheimer's disease. *Neuropsychologia*, 31(4), 335-345.
- Bahrami, Bahador, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. 2010. Optimally Interacting Minds. *Science* 329(5995):1081–85. doi: 10.1126/science.1185718.
- Bang, D., Kishida, K. T., Lohrenz, T., White, J. P., Laxton, A. W., Tatter, S. B., ... & Montague, P. R. 2020. Sub-second dopamine and serotonin signaling in human striatum during perceptual decision-making. *Neuron*, 108(5), 999-1010.
- Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General*, 148(3), 437.

- Beck, A. 2004. A New Instrument for Measuring Insight: The Beck Cognitive Insight Scale. *Schizophrenia Research* 68(2–3):319–29. doi: 10.1016/S0920-9964(03)00189-0.
- Bentall, R. P. 1990. The Illusion of Reality: A Review and Integration of Psychological Research on Hallucinations. *Psychological Bulletin* 107(1):82–95. doi: 10.1037/0033-2909.107.1.82.
- Bentall, R. P., and P. D. Slade. 1985. Reality Testing and Auditory Hallucinations: A Signal Detection Analysis. *British Journal of Clinical Psychology* 24(3):159–69. doi: 10.1111/j.2044-8260.1985.tb01331.x.
- van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. 2016. A common mechanism underlies changes of mind about decisions and confidence. *Elife*, 5, e12192.
- Berkovitch, L., Dehaene, S., and Gaillard, R. 2017. Disruption of Conscious Access in Schizophrenia. *Trends in Cognitive Sciences* 21(11):878–92. doi: 10.1016/j.tics.2017.08.006.
- Berridge, K. C., & Robinson, T. E. 1998. What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Salience? *Brain Research Reviews* 28(3):309–69. doi: 10.1016/S0165-0173(98)00019-8.
- Berrios, G. E. 2018. Obsessional disorders during the nineteenth century: terminological and classificatory issues. In *The anatomy of madness* (pp. 166-187). Routledge.
- Bhome, R., McWilliams, A., Huntley, J. D., Fleming, S. M., & Howard, R. J. 2022. Metacognition in Functional Cognitive Disorder. *Brain Communications* 4(2):fcac041. doi: 10.1093/braincomms/fcac041.
- Biéder, J. 2011. Eugène Billod (1818–1886). *Annales Médico-psychologiques, revue psychiatrique* 169(5):332–36. doi: 10.1016/j.amp.2011.04.001.
- Blakemore, S. J., Smith, J., Steel, R., Johnstone, E. C., & Frith, C. D. 2000. The Perception of Self-Produced Sensory Stimuli in Patients with Auditory Hallucinations and Passivity Experiences: Evidence for a Breakdown in Self-Monitoring. *Psychological Medicine* 30(5):1131–39. doi: 10.1017/S0033291799002676.
- Bloomfield, M. A. P., & Howes, O. D. 2020. DOPAMINERGIC MECHANISMS UNDERLYING PSYCHOSIS. *Psychotic Disorders: Comprehensive Conceptualization and Treatments*, 277.
- Boldt, A., De Gardelle, V., & Yeung, N. 2017. The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. *Journal of Experimental Psychology: Human Perception and Performance* 43(8):1520–31. doi: 10.1037/xhp0000404.
- Brébion, G., Amador, X., David, A., Malaspina, D., Sharif, Z., & Gorman, J. M. 2000. Positive Symptomatology and Source-Monitoring Failure in Schizophrenia — an Analysis of Symptom-Specific Effects. *Psychiatry Research* 95(2):119–31. doi: 10.1016/S0165-1781(00)00174-8.
- Brébion, G., Ohlsen, R. I., Pilowsky, L. S., & David, A. S. 2008. Visual Hallucinations in

Schizophrenia: Confusion between Imagination and Perception. *Neuropsychology* 22(3):383–89. doi: 10.1037/0894-4105.22.3.383.

Brown, T. M. 2018. Descartes, dualism, and psychosomatic medicine. In *The anatomy of madness* (pp. 40-62). Routledge.

Carota, A., & Bogousslavsky, J. 2019. Neurology versus Psychiatry? Hallucinations, Delusions, and Confabulations. Pp. 127–40 in *Frontiers of Neurology and Neuroscience*. Vol. 44, edited by J. Bogousslavsky, F. Boller, and M. Iwata. S. Karger AG.

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. doi: 10.1037/xge0000505.

Clementz, B. A., Parker, D. A., Trotti, R. L., McDowell, J. E., Keedy, S. K., Keshavan, M. S., ... & Tamminga, C. A. 2022. Psychosis Biotypes: Replication and Validation from the B-SNIP Consortium. *Schizophrenia Bulletin* 48(1):56–68. doi: 10.1093/schbul/sbab090.

Clementz, B. A., Sweeney, J. A., Hamm, J. P., Ivleva, E. I., Ethridge, L. E., Pearlson, G. D., ... & Tamminga, C. A. 2016. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *American Journal of Psychiatry* 173(4):373–84. doi: 10.1176/appi.ajp.2015.14091200.

Connell, P. H. 1957. Amphetamine psychosis. *British medical journal*, 1(5018), 582.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers III, A. R. 2019. Hallucinations and Strong Priors. *Trends in Cognitive Sciences* 23(2):114–27. doi: 10.1016/j.tics.2018.12.001.

David, A. S. 1990. Insight and Psychosis. *British Journal of Psychiatry* 156(6):798–808. doi: 10.1192/bjp.156.6.798.

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. 2012. Failures of Metacognition and Lack of Insight in Neuropsychiatric Disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1594):1379–90. doi: 10.1098/rstb.2012.0002.

Davis, B. J., Lysaker, P. H., Salyers, M. P., & Minor, K. S. 2020. The Insight Paradox in Schizophrenia: A Meta-Analysis of the Relationship between Clinical Insight and Quality of Life. *Schizophrenia Research* 223:9–17. doi: 10.1016/j.schres.2020.07.017.

Davis, K. L., Kahn, R. S., Ko, G., & Davidson, M. 1991. Dopamine in Schizophrenia: A Review and Reconceptualization. *American Journal of Psychiatry* 148(11):1474–86. doi: 10.1176/ajp.148.11.1474.

Delay, J., & Deniker, P. 1955. Neuroleptic effects of chlorpromazine in therapeutics of neuropsychiatry. *Journal of clinical and experimental psychopathology*, 16(2), 104-112.

Delion, P. 2014. La psychothérapie institutionnelle : d'où vient-elle et où va-t-elle ? *Empan* 96(4):104. doi: 10.3917/empa.096.0104.

- Demjaha, A., Murray, R. M., McGuire, P. K., Kapur, S., & Howes, O. D. 2012. Dopamine Synthesis Capacity in Patients With Treatment-Resistant Schizophrenia. *American Journal of Psychiatry* 169(11):1203–10. doi: 10.1176/appi.ajp.2012.12010144.
- Descartes, R., & Rodis-Lewis, G. 1994. Les passions de l'âme. Vrin.
- Descartes, R. 2020. Discours de la méthode [suivi de] Méditations métaphysiques. Paris: BoD-Books on demand.
- Desender, K., Ridderinkhof, K. R., & Murphy, P. R. 2021. Understanding Neural Signals of Post-Decisional Performance Monitoring: An Integrative Review. *eLife* 10:e67556. doi: 10.7554/eLife.67556.
- Dietrichkeit, M., Grzella, K., Nagel, M., & Moritz, S. 2020. Using Virtual Reality to Explore Differences in Memory Biases and Cognitive Insight in People with Psychosis and Healthy Controls. *Psychiatry Research* 285:112787–112787. doi: 10.1016/j.psychres.2020.112787.
- Dijkstra, N., Kok, P., & Fleming, S. M. 2022. Perceptual Reality Monitoring: Neural Mechanisms Dissociating Imagination from Reality. *Neuroscience & Biobehavioral Reviews* 135:104557. doi: 10.1016/j.neubiorev.2022.104557.
- Dollfus, S., Mach, C., & Morello, R. 2016. Self-Evaluation of Negative Symptoms: A Novel Tool to Assess Negative Symptoms. *Schizophrenia Bulletin* 42(3):571–78. doi: 10.1093/schbul/sbv161.
- Dondé, C., Avissar, M., Weber, M. M., & Javitt, D. C. 2019. A Century of Sensory Processing Dysfunction in Schizophrenia. *European Psychiatry* 59:77–79. doi: 10.1016/j.eurpsy.2019.04.006.
- Dubreucq, J. 2020. Auto-stigmatisation dans les troubles psychiques sévères et persistants (Doctoral dissertation, Université de Lyon)
- Dudley, R., Taylor, P., Wickham, S., & Hutton, P. 2016. Psychosis, Delusions and the “Jumping to Conclusions” Reasoning Bias: A Systematic Review and Meta-Analysis. *Schizophrenia Bulletin* 42(3):652–65. doi: 10.1093/schbul/sbv150.
- Dunne, J. D., Thompson, E., & Schooler, J. 2019. Mindful Meta-Awareness: Sustained and Non-Propositional. *Current Opinion in Psychology* 28:307–11. doi: 10.1016/j.copsyc.2019.07.003.
- Eack, S. M., Wojtalik, J. A., Keshavan, M. S., & Minshew, N. J. 2017. Social-Cognitive Brain Function and Connectivity during Visual Perspective-Taking in Autism and Schizophrenia. *Schizophrenia Research* 183:102–9. doi: 10.1016/j.schres.2017.03.009.
- Esquirol, E. 1838. Des maladies mentales considérées sous les rapports médical, hygiénique et médico-légal (Vol. 1). chez JB Baillière.
- Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. 2018. Behavioral, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition. *The Journal of Neuroscience* 38(2):263–77. doi: 10.1523/JNEUROSCI.0322-17.2017.

- Fénelon, G. 2014. Les hallucinations en neurologie. *Pratique Neurologique - FMC* 5(4):277–86. doi: 10.1016/j.praneu.2014.10.001.
- Flavell, J. H. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *American Psychologist* 34(10):906–11. doi: 10.1037/0003-066X.34.10.906.
- Fleming, S. M. 2021. *Know Thyself: How the New Science of Self-Awareness*. London: John Murray.
- 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):1280–86. doi: 10.1098/rstb.2012.0021.
- Fleming, S. M., & Lau, H. C. 2014. How to Measure Metacognition. *Frontiers in Human Neuroscience* 8. doi: 10.3389/fnhum.2014.00443.
- Fletcher, P. C., and Frith, C. D. 2009. Perceiving Is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia. *Nature Reviews Neuroscience* 10(1):48–58. doi: 10.1038/nrn2536.
- Fodor, J. A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass: MIT Press.
- Fortier, M. 2018. Sense of reality, metacognition, and culture in schizophrenic and drug-induced hallucinations: An interdisciplinary approach. In J. Proust & M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach* (pp. 343–378). Oxford University Press.
- Franck, N.. 2021. Principes et outils de la réhabilitation psychosociale. *Annales Médico-psychologiques, revue psychiatrique* 179(10):953–58. doi: 10.1016/j.amp.2021.10.002.
- Frith, C. D., and Frith, U. 2012. Mechanisms of Social Cognition. *Annual Review of Psychology* 63(1):287–313. doi: 10.1146/annurev-psych-120710-100449.
- Fusar-Poli, P., & Meyer-Lindenberg, A. 2013. Striatal Presynaptic Dopamine in Schizophrenia, Part II: Meta-Analysis of [18F/11C]-DOPA PET Studies. *Schizophrenia Bulletin* 39(1):33–42. doi: 10.1093/schbul/sbr180.
- Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. 2003. Type 2 Tasks in the Theory of Signal Detectability: Discrimination between Correct and Incorrect Decisions. *Psychonomic Bulletin & Review* 10(4):843–76. doi: 10.3758/BF03196546.
- de Gardelle, V., and Mamassian, P. 2014. Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychological Science* 25(6):1286–88. doi: 10.1177/0956797614528956.
- Gaudio, B. A., & Herbert, J. D. 2006. Believability of Hallucinations as a Potential Mediator of Their Frequency and Associated Distress in Psychotic Inpatients. *Behavioural and*

Cognitive Psychotherapy 34(04):497. doi: 10.1017/S1352465806003080.

Gauntlett-Gilbert, J., and Kuipers, E. 2005. Visual Hallucinations in Psychiatric Conditions: Appraisals and Their Relationship to Distress. *British Journal of Clinical Psychology* 44(1):77–87. doi: 10.1348/014466504x19451.

Gaweda, L., and Moritz, S. 2019. The Role of Expectancies and Emotional Load in False Auditory Perceptions among Patients with Schizophrenia Spectrum Disorders. *European Archives of Psychiatry and Clinical Neuroscience*. doi: 10.1007/s00406-019-01065-2.

Gignac, G. E., and Zajenkowski, M. 2020. The Dunning-Kruger Effect Is (Mostly) a Statistical Artefact: Valid Approaches to Testing the Hypothesis with Individual Differences Data. *Intelligence* 80:101449. doi: 10.1016/j.intell.2020.101449.

Goodman, L. A., and Kruskal, W. H. 1979. Measures of Association for Cross Classifications. Pp. 2–34 in *Measures of Association for Cross Classifications*, Springer Series in Statistics. New York, NY: Springer New York.

Gopal, Y. V., and Variend, H. 2005. First-Episode Schizophrenia: Review of Cognitive Deficits and Cognitive Remediation. *Advances in Psychiatric Treatment* 11(1):38–44. doi: 10.1192/apt.11.1.38.

Green, D. M., & Swets, J. A. 1966. Signal detection theory and psychophysics (Vol. 1, pp. 1969-2012). New York: Wiley.

Gur, R. E., Roalf, D. R., Alexander-Bloch, A., McDonald-McGinn, D. M., & Gur, R. C. 2021. Pathways to Understanding Psychosis through Rare – 22q11.2DS - and Common Variants. *Current Opinion in Genetics & Development* 68:35–40. doi: 10.1016/j.gde.2021.01.007.

Hart, J. T. 1965. Memory and the Feeling-of-Knowing Experience. *Journal of Educational Psychology* 56(4):208–16. doi: 10.1037/h0022263.

Heinrichs, R. W., and Zakzanis, K. K. 1998. Neurocognitive Deficit in Schizophrenia: A Quantitative Review of the Evidence. *Neuropsychology* 12(3):426–45. doi: 10.1037/0894-4105.12.3.426.

Henriksen, M. G., and Parnas, J. 2014. Self-Disorders and Schizophrenia: A Phenomenological Reappraisal of Poor Insight and Noncompliance. *Schizophrenia Bulletin* 40(3):542–47. doi: 10.1093/schbul/sbt087.

Hill, K., Varese, F., Jackson, M., & Linden, D. E. 2012. The Relationship between Metacognitive Beliefs, Auditory Hallucinations, and Hallucination-Related Distress in Clinical and Non-Clinical Voice-Hearers: *Metacognitive Beliefs, Voices, and Distress*. *British Journal of Clinical Psychology* 51(4):434–47. doi: 10.1111/j.2044-8260.2012.02039.x.

Hill, S. K., Keefe, R. S., & Sweeney, J. A. 2020. COGNITIVE BIOMARKERS OF PSYCHOSIS. *Psychotic Disorders: Comprehensive Conceptualization and Treatments*, 195.

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. doi: 10.1038/s41398-019-0602-7.

Howes, O. D., Kambeitz, J., Kim, E., Stahl, D., Slifstein, M., Abi-Dargham, A., & Kapur, S. 2012. The Nature of Dopamine Dysfunction in Schizophrenia and What This Means for Treatment: Meta-Analysis of Imaging Studies. *Archives of General Psychiatry* 69(8). doi: 10.1001/archgenpsychiatry.2012.169.

Howes, O. D., McCutcheon, R., Owen, M. J., & Murray, R. M. 2017. The Role of Genes, Stress, and Dopamine in the Development of Schizophrenia. *Biological Psychiatry* 81(1):9–20. doi: 10.1016/j.biopsych.2016.07.014.

Hu, X., Yang, C., & Luo, L. (2022). Are the Contributions of Processing Experience and Prior Beliefs to Confidence Ratings Domain-general or Domain-specific?.

Jacobs, B. L. 1978. Dreams and Hallucinations: A Common Neurochemical Mechanism Mediating Their Phenomenological Similarities. *Neuroscience & Biobehavioral Reviews* 2(1):59–69. doi: 10.1016/0149-7634(78)90007-6.

Jalal, B. 2018. The Neuropharmacology of Sleep Paralysis Hallucinations: Serotonin 2A Activation and a Novel Therapeutic Drug. *Psychopharmacology* 235(11):3083–91. doi: 10.1007/s00213-018-5042-1.

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. 2021. A Rational Model of the Dunning–Kruger Effect Supports Insensitivity to Evidence in Low Performers. *Nature Human Behaviour* 5(6):756–63. doi: 10.1038/s41562-021-01057-0.

Jardri, R., and Denève, S. 2013. Circular Inferences in Schizophrenia. *Brain* 136(11):3227–41. doi: 10.1093/brain/awt257.

de Jong, S., Van Donkersgoed, R. J. M., Timmerman, M. E., Aan Het Rot, M., Wunderink, L., Arends, J., ... & Pijnenborg, G. H. M. 2019. Metacognitive Reflection and Insight Therapy (MERIT) for Patients with Schizophrenia. *Psychological Medicine* 49(2):303–13. doi: 10.1017/S0033291718000855.

Kapur, S. 2003. Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia. *American Journal of Psychiatry* 160(1):13–23. doi: 10.1176/appi.ajp.160.1.13.

Keefe, R. S., Arnold, M. C., Bayen, U. J., & Harvey, P. D. 1999. Source Monitoring Deficits in Patients with Schizophrenia; a Multinomial Modelling Analysis. *Psychological Medicine* 29(4):903–14. doi: 10.1017/S0033291799008673.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. 2008. Neural Correlates, Computation and Behavioural Impact of Decision Confidence. *Nature* 455(7210):227–31. doi: 10.1038/nature07200.

Keshavan, M. S., Clementz, B. A., Pearlson, G. D., Sweeney, J. A., & Tamminga, C. A. 2013. Reimagining Psychoses: An Agnostic Approach to Diagnosis. *Schizophrenia Research* 146(1–3):10–16. doi: 10.1016/j.schres.2013.02.022.

Keshavan, M. S., Torous, J., & Tandon, R. 2020. CONCEPTUALIZATION OF PSYCHOSIS IN PSYCHIATRIC NOSOLOGY. *Psychotic Disorders: Comprehensive Conceptualization and Treatments*, 3.

Kiani, R., and Shadlen, M. N. 2009. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science* 324(5928):759–64. doi: 10.1126/science.1169405.

Ko, Y., and Lau, H. 2012. A Detection Theoretic Explanation of Blindsight Suggests a Link between Conscious Perception and Metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1594):1401–11. doi: 10.1098/rstb.2011.0380.

Koob, G. F., & Moal, M. L. 1997. Drug Abuse: Hedonic Homeostatic Dysregulation. *Science* 278(5335):52–58. doi: 10.1126/science.278.5335.52.

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. doi: 10.1016/j.psychres.2012.03.004.

Kreps, D. M. 1989. Nash Equilibrium. Pp. 167–77 in *Game Theory*, edited by J. Eatwell, M. Milgate, and P. Newman. London: Palgrave Macmillan UK.

Kronbichler, L., Stelzig-Schöler, R., Pearce, B. G., Tschernegg, M., Said-Yürekli, S., Crone, J. S., ... & Kronbichler, M. 2019. Reduced Spontaneous Perspective Taking in Schizophrenia. *Psychiatry Research: Neuroimaging* 292:5–12. doi: 10.1016/j.psychresns.2019.08.007.

Krueger, J., and Mueller, R., A. 2002. Unskilled, Unaware, or Both? The Better-than-Average Heuristic and Statistical Regression Predict Errors in Estimates of Own Performance. *Journal of Personality and Social Psychology* 82(2):180–88. doi: 10.1037/0022-3514.82.2.180.

Kruger, J., and Dunning, D. 1999. Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology* 77(6):1121–34. doi: 10.1037/0022-3514.77.6.1121.

Krugwasser, A. R., Stern, Y., Faivre, N., Harel, E. V., & Salomon, R. 2022. Impaired Sense of Agency and Associated Confidence in Psychosis. *Schizophrenia* 8(1):32. doi: 10.1038/s41537-022-00212-4.

Kuhn, T. S. 1970. The structure of scientific revolutions (Vol. 111). University of Chicago Press: Chicago.

Larøi, F. 2006. The Phenomenological Diversity of Hallucinations: Some Theoretical and Clinical Implications. *Psychologica Belgica* 46(1–2):163. doi: 10.5334/pb-46-1-2-163.

Lau, H. C., and Passingham, R. E. 2006. Relative Blindsight in Normal Observers and the Neural Correlate of Visual Consciousness. *Proceedings of the National Academy of Sciences* 103(49):18763–68. doi: 10.1073/pnas.0607716103.

- Lau, H. 2019. Consciousness, Metacognition, & Perceptual Reality Monitoring. *preprint*. PsyArXiv. doi: 10.31234/osf.io/ckbyf.
- Lee, J. S., Chun, J. W., Lee, S. H., Kim, E., Lee, S. K., & Kim, J. J. 2015. Altered Neural Basis of the Reality Processing and Its Relation to Cognitive Insight in Schizophrenia. edited by S. Lui. *PLOS ONE* 10(3):e0120478. doi: 10.1371/journal.pone.0120478.
- Lefebvre, P. 1988. Le traité des maladies mentales d'Esquirol: cent cinquante ans après. *Histoire des sciences médicales*, 22(2), 169-174.
- Leonesio, R. J., and Nelson, T. O. 1990. Do Different Metamemory Judgments Tap the Same Underlying Aspects of Memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(3):464–70. doi: 10.1037/0278-7393.16.3.464.
- Lepoutre, T., and Dening, T. 2012. “De La Non-Existence de La Monomanie”, by Jean-Pierre Falret (1854): Introduction and Translation (Part 1). *History of Psychiatry* 23(3):356–70. doi: 10.1177/0957154X12445421.
- Levitt, H. 1971. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America* 49(2B):467–77. doi: 10.1121/1.1912375.
- Liddle, P. F. 1987. The Symptoms of Chronic Schizophrenia: A Re-Examination of the Positive-Negative Dichotomy. *British Journal of Psychiatry* 151(2):145–51. doi: 10.1192/bjp.151.2.145.
- Limongi, R., Bohaterewicz, B., Nowicka, M., Plewka, A., & Friston, K. J. 2018. Knowing When to Stop: Aberrant Precision and Evidence Accumulation in Schizophrenia. *Schizophrenia Research* 197:386–91. doi: 10.1016/j.schres.2017.12.018.
- Lou, H. C., Skewes, J. C., Thomsen, K. R., Overgaard, M., Lau, H. C., Mouridsen, K., & Roepstorff, A. 2011. Dopaminergic Stimulation Enhances Confidence and Accuracy in Seeing Rapidly Presented Words. *Journal of Vision* 11(2):15–15. doi: 10.1167/11.2.15.
- Lysaker, P. H., Vohs, J., Minor, K. S., Irrarázaval, L., Leonhardt, B., Hamm, J. A., ... & Dimaggio, G. 2015. Metacognitive Deficits in Schizophrenia: Presence and Associations With Psychosocial Outcomes. *The Journal of Nervous and Mental Disease* 203(7):530–36. doi: 10.1097/NMD.0000000000000323.
- Mamassian, P., and de Gardelle, V. 2021. Modeling Perceptual Confidence and the Confidence Forced-Choice Paradigm. *Psychological Review*. doi: 10.1037/rev0000312.
- Manford, M., and Andermann, F. 1998. Complex Visual Hallucinations. Clinical and Neurobiological Insights. *Brain* 121(10):1819–40. doi: 10.1093/brain/121.10.1819.
- Maniscalco, B., and Lau, H. 2014. Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-D', Response-Specific Meta-D', and the Unequal Variance SDT Model. Pp. 25–66 in *The Cognitive Neuroscience of Metacognition*, edited by S. M. Fleming and C. D. Frith. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Maniscalco, B., and Lau, H. 2016. The Signal Processing Architecture Underlying Subjective

- Reports of Sensory Awareness. *Neuroscience of Consciousness* 2016(1). doi: 10.1093/nc/niw002.
- Martinez, M. E. 2006. What Is Metacognition? *Phi Delta Kappan* 87(9):696–99. doi: 10.1177/003172170608700916.
- Mattila, T., Koeter, M., Wohlfarth, T., Storosum, J., van den Brink, W., de Haan, L., ... & Denys, D. 2015. Impact of DSM-5 Changes on the Diagnosis and Acute Treatment of Schizophrenia. *Schizophrenia Bulletin* 41(3):637–43. doi: 10.1093/schbul/sbu172.
- Mayer, J. S., and Park, S. 2012. Working Memory Encoding and False Memory in Schizophrenia and Bipolar Disorder in a Spatial Delayed Response Task. *Journal of Abnormal Psychology* 121(3):784–94. doi: 10.1037/a0028836.
- Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. 2020. Is There a G Factor for Metacognition? Correlations in Retrospective Metacognitive Sensitivity across Tasks. *Journal of Experimental Psychology: General* 149(9):1788–99. doi: 10.1037/xge0000746.
- Mazor, M., Friston, K. J., & Fleming, S. M. 2020. Distinct Neural Contributions to Metacognition for Detecting, but Not Discriminating Visual Stimuli. *eLife* 9:e53900. doi: 10.7554/eLife.53900.
- McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. 2013. Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. *Journal of Neuroscience* 33(5):1897–1906. doi: 10.1523/JNEUROSCI.1890-12.2013.
- McDonald-McGinn, D. M., Sullivan, K. E., Marino, B., Philip, N., Swillen, A., Vorstman, J. A., ... & Bassett, A. S. 2015. 22q11.2 Deletion Syndrome. *Nature Reviews Disease Primers* 1(1):15071. doi: 10.1038/nrdp.2015.71.
- Menon, G. J.. 2005. Complex Visual Hallucinations in the Visually Impaired: A Structured History-Taking Approach. *Archives of Ophthalmology* 123(3):349. doi: 10.1001/archophth.123.3.349.
- Mondino, M., Dondé, C., Lavallé, L., Haesebaert, F., & Brunelin, J. 2019. Reality-Monitoring Deficits and Visual Hallucinations in Schizophrenia. *European Psychiatry* 62:10–14. doi: 10.1016/j.eurpsy.2019.08.010.
- Mora, G. 2008. Renaissance conceptions and treatments of madness. In *History of psychiatry and medical psychology* (pp. 227-254). Springer, Boston, MA.
- Morales, J., Lau, H., and Fleming, S. M. 2018. Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *Journal of Neuroscience* 38(14):3534–46. doi: 10.1523/JNEUROSCI.2360-17.2018.
- Moritz, S. 2012. A Bias-Oriented Treatment Approach: The Metacognitive Training for Schizophrenia Patients (MCT). *A Bias-Oriented Treatment Approach: The Metacognitive Training for Schizophrenia Patients (MCT)* 14–18.

- Moritz, S., Ramdani, N., Klass, H., Andreou, C., Jungclaussen, D., Eifler, S., ... & Zink, M. 2014. Overconfidence in Incorrect Perceptual Judgments in Patients with Schizophrenia. *Schizophrenia Research: Cognition* 1(4):165–70. doi: 10.1016/j.scog.2014.09.003.
- Moritz, S., and Woodward, T. S. 2002. Memory Confidence and False Memories in Schizophrenia. *Journal of Nervous and Mental Disease* 190(9):641–43. doi: 10.1097/00005053-200209000-00012.
- Moritz, S., and Woodward, T. S. 2006. The Contribution of Metamemory Deficits to Schizophrenia. *Journal of Abnormal Psychology* 115(1):15–25. doi: 10.1037/0021-843X.115.1.15.
- Moritz, S., Woodward, T. S., Cuttler, C., Whitman, J. C., and Watson, J. M. 2004. False Memories in Schizophrenia. *Neuropsychology* 18(2):276–83. doi: 10.1037/0894-4105.18.2.276.
- Moritz, S., Woodward, T. S. and Ruff, C. C. 2003. Source Monitoring and Memory Confidence in Schizophrenia. *Psychological Medicine* 33(1):131–39. doi: 10.1017/S0033291702006852.
- Moritz, S., Woznica, A., Andreou, C., and Köther, U. 2012. Response Confidence for Emotion Perception in Schizophrenia Using a Continuous Facial Sequence Task. *Psychiatry Research* 200(2–3):202–7. doi: 10.1016/j.psychres.2012.07.007.
- Moritz, S., Andreou, C., Schneider, B. C., Wittekind, C. E., Menon, M., Balzan, R. P., & Woodward, T. S. 2014. Sowing the Seeds of Doubt: A Narrative Review on Metacognitive Training in Schizophrenia. *Clinical Psychology Review* 34(4):358–66. doi: 10.1016/j.cpr.2014.04.004.
- Moritz, S., Pfuhl, G., Lüdtke, T., Menon, M., Balzan, R. P., & 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. doi: 10.1016/j.jbtep.2016.07.004.
- Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review* 83(4):435. doi: 10.2307/2183914.
- Nasar, S. 1998. *A Beautiful Mind: A Biography of John Forbes Nash, Jr., Winner of the Nobel Prize in Economics, 1994*. New York, NY: Simon & Schuster.
- Nassar, M. R., Waltz, J. A., Albrecht, M. A., Gold, J. M., & Frank, M. J. 2021. All or Nothing Belief Updating in Patients with Schizophrenia Reduces Precision and Flexibility of Beliefs. *Brain* 144(3):1013–29. doi: 10.1093/brain/awaa453.
- Nelson, T. O., & Narens, L. 1990. Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 1–45). New York, NY: Academic Press.
- Nelson, T. O. 1996. Consciousness and metacognition. *American psychologist*, 51(2), 102.

- Nichols, D. E. 2004. Hallucinogens. *Pharmacology & Therapeutics* 101(2):131–81. doi: 10.1016/j.pharmthera.2003.11.002.
- Nisbett, R. E., and Wilson T. D. 1977. Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84(3):231–59. doi: 10.1037/0033-295X.84.3.231.
- O’Callaghan, C., Hall, J. M., Tomassini, A., Muller, A. J., Walpola, I. C., Moustafa, A. A., ... & Lewis, S. J. 2017. Visual hallucinations are characterized by impaired sensory evidence accumulation: insights from hierarchical drift diffusion modeling in Parkinson’s disease. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(8), 680-688.
- O’Donnell, P. 2011. Adolescent Onset of Cortical Disinhibition in Schizophrenia: Insights From Animal Models. *Schizophrenia Bulletin* 37(3):484–92. doi: 10.1093/schbul/sbr028.
- Orne, M. T. 2009. Demand characteristics and the concept of quasi-controls. *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow’s classic books, 110*, 110-137.
- Van Os, J., Kenis, G., & Rutten, B. P. 2010. The Environment and Schizophrenia. *Nature* 468(7321):203–12. doi: 10.1038/nature09563.
- Overgaard, M. 2008. Introspection. *Scholarpedia* 3(5):4953. doi: 10.4249/scholarpedia.4953.
- Overgaard, S. 2015. How to Do Things with Brackets: The Epoché Explained. *Continental Philosophy Review* 48(2):179–95. doi: 10.1007/s11007-015-9322-8.
- Pang, L. 2016. Hallucinations Experienced by Visually Impaired: Charles Bonnet Syndrome. *Optometry and Vision Science* 93(12):1466–78. doi: 10.1097/OPX.0000000000000959.
- Paulewicz, B., Siedlecka, M., & Koculak, M. 2020. Confounding in Studies on Metacognition: A Preliminary Causal Analysis Framework. *Frontiers in Psychology* 11:1933. doi: 10.3389/fpsyg.2020.01933.
- Penfield, W. 1974. The mind and the highest brain-mechanism. *The American Scholar*, 237-246.
- Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., ... & Faivre, N. 2021. Evidence Accumulation Relates to Perceptual Consciousness and Monitoring. *Nature Communications* 12(1):3261. doi: 10.1038/s41467-021-23540-y.
- Pinel, P. 1809. *Traité médico-philosophique sur l’aliénation mentale*. J. Ant. Brosson.
- Phillips, I. 2021. Blindsight Is Qualitatively Degraded Conscious Vision. *Psychological Review* 128(3):558–84. doi: 10.1037/rev0000254.
- Pleskac, T. J., and Busemeyer, J. R. 2010. Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence. *Psychological Review* 117(3):864–901. doi: 10.1037/a0019737.
- Porter, R. 2002. *Madness: A Brief History*. Oxford ; New York: Oxford University Press.

Powers, A. R., Mathys, C., and Corlett, P. R. 2017. Pavlovian Conditioning–Induced Hallucinations Result from Overweighting of Perceptual Priors. *Science* 357(6351):596–600. doi: 10.1126/science.aan3458.

Rahnev, D. A., Maniscalco, B., Luber, B., Lau, H., & Lisanby, S. H. 2012. Direct Injection of Noise to the Visual Cortex Decreases Accuracy but Increases Decision Confidence. *Journal of Neurophysiology* 107(6):1556–63. doi: 10.1152/jn.00985.2011.

Rahnev, D., Balsdon, T., Charles, L., De Gardelle, V., Denison, R., Desender, K., ... & Zylberberg, A. 2021. *Consensus Goals in the Field of Visual Metacognition*. preprint. PsyArXiv. doi: 10.31234/osf.io/z8v5x.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., ... & Zylberberg, A. The Confidence Database. *Nature Human Behaviour* 4(3):317–25. doi: 10.1038/s41562-019-0813-1.

Ramachandiraiah, C. T., Subramaniam, N., & Tancer, M. 2009. The story of antipsychotics: Past and present. *Indian journal of psychiatry*, 51(4), 324.

Rankin, P. M., & O'Carroll, P. J. 1995. Reality Discrimination, Reality Monitoring and Disposition towards Hallucination. *British Journal of Clinical Psychology* 34(4):517–28. doi: 10.1111/j.2044-8260.1995.tb01486.x.

Ratcliff, R., and McKoon, G. 2008. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation* 20(4):873–922. doi: 10.1162/neco.2008.12-06-420.

Ratcliff, R., and Rouder, J. N. 1998. Modeling Response Times for Two-Choice Decisions. *Psychological Science* 9(5):347–56. doi: 10.1111/1467-9280.00067.

Raven, J., and Raven, J. 2003. Raven Progressive Matrices. Pp. 223–37 in *Handbook of Nonverbal Assessment*, edited by R. S. McCallum. Boston, MA: Springer US.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. 2009. Changes of Mind in Decision-Making. *Nature* 461(7261):263–66. doi: 10.1038/nature08275.

Ribot, T. 2002. Les maladies de la volonté. *Les maladies de la volonté*, 1-180.

Rouault, M., and Stephen M. Fleming, S. M. 2020. Formation of Global Self-Beliefs in the Human Brain. *Proceedings of the National Academy of Sciences* 117(44):27268–76. doi: 10.1073/pnas.2003094117.

Rouault, M., McWilliams, A., Allen, M. G., and Fleming, S. M. 2018. Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience* 1:e17. doi: 10.1017/pen.2018.16.

Rouault, M., Seow, T., Gillan, C. M., and Fleming, S. M. 2018. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry* 84(6):443–51. doi: 10.1016/j.biopsych.2017.12.017.

- Sacks, M. H., Carpenter, W. T., & Strauss, J. S. 1974. Recovery From Delusions: Three Phases Documented by Patient's Interpretation of Research Procedures. *Archives of General Psychiatry* 30(1):117. doi: 10.1001/archpsyc.1974.01760070093015.
- Sackur, J. 2009. L'introspection en psychologie expérimentale. *Revue d'histoire des sciences* 62(2):349. doi: 10.3917/rhs.622.0349.
- Salomon, R., Kannape, O. A., Debarba, H. G., Kaliuzhna, M., Schneider, M., Faivre, N., ... & Blanke, O. 2022. Agency Deficits in a Human Genetic Model of Schizophrenia: Insights From 22q11DS Patients. *Schizophrenia Bulletin* 48(2):495–504. doi: 10.1093/schbul/sbab143.
- Sanz, M., Constable, G., Lopez-Ibor, I., Kemp, R., & David, A. S. 1998. A Comparative Study of Insight Scales and Their Relationship to Psychopathological and Clinical Variables. *Psychological Medicine* 28(2):437–46. doi: 10.1017/S0033291797006296.
- Schaefer, J., Giangrande, E., Weinberger, D. R., & Dickinson, D. 2013. The Global Cognitive Impairment in Schizophrenia: Consistent over Decades and around the World. *Schizophrenia Research* 150(1):42–50. doi: 10.1016/j.schres.2013.07.009.
- Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. 2021. Striatal Dopamine Mediates Hallucination-like Perception in Mice. *Science* 372(6537):eabf4740. doi: 10.1126/science.abf4740.
- Schnakenberg Martin, A. M., and Lysaker, P.H. 2022. Metacognition, Adaptation, and Mental Health. *Biological Psychiatry* 91(8):e31–32. doi: 10.1016/j.biopsych.2021.09.028.
- Schneider, K. 1959. *Clinical psychopathology*. Grune & Stratton.
- Schneider, M., Debbané, M., Bassett, A. S., Chow, E. W., Fung, W. L. A., Van Den Bree, M. B., ... & International Consortium on Brain and Behavior in 22q11. 2 Deletion Syndrome. 2014. Psychiatric Disorders From Childhood to Adulthood in 22q11.2 Deletion Syndrome: Results From the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. *American Journal of Psychiatry* 171(6):627–39. doi: 10.1176/appi.ajp.2013.13070864.
- Schooler JW, Smallwood J 2009. Metacognition. In *Oxford Companion to Consciousness*. Edited by Bayne T, Cleeremans A, Wilken P. Oxford University Press; 2009:438-442.
- Seeman, P., Lee, T., Chau-Wong, M., and Wong. K. 1976. Antipsychotic Drug Doses and Neuroleptic/Dopamine Receptors. *Nature* 261(5562):717–19. doi: 10.1038/261717a0.
- Semerari, A., Carcione, A., Dimaggio, G., Falcone, M., Nicolo, G., Procacci, M., & Alleva, G. 2003. How to Evaluate Metacognitive Functioning in Psychotherapy? The Metacognition Assessment Scale and Its Applications: Assessing Metacognitive Functions in Psychotherapy. *Clinical Psychology & Psychotherapy* 10(4):238–61. doi: 10.1002/cpp.362.
- Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. 2021. How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry* 90(7):436–46. doi: 10.1016/j.biopsych.2021.05.013.

- Seow, T. X., Rouault, M., Gillan, C. M., & Fleming, S. M. 2022. Reply to: Metacognition, Adaptation, and Mental Health. *Biological Psychiatry* 91(8):e33–34. doi: 10.1016/j.biopsych.2021.11.005.
- Shea, N., and Frith, C. D. 2019. The Global Workspace Needs Metacognition. *Trends in Cognitive Sciences* 23(7):560–71. doi: 10.1016/j.tics.2019.04.007.
- Shergill, S. S., Samson, G., Bays, P. M., Frith, C. D., & Wolpert, D. M. 2005. Evidence for Sensory Prediction Deficits in Schizophrenia. *American Journal of Psychiatry* 162(12):2384–86. doi: 10.1176/appi.ajp.162.12.2384.
- Simons, J. S., Garrison, J. R., & Johnson, M. K. 2017. Brain Mechanisms of Reality Monitoring. *Trends in Cognitive Sciences* 21(6):462–73. doi: 10.1016/j.tics.2017.03.012.
- Souchay, C. 2007. Metamemory in Alzheimer's Disease. *Cortex* 43(7):987–1003. doi: 10.1016/S0010-9452(08)70696-8.
- Stone, M. H. 2008. A Brief History of Psychiatry. Pp. 203–42 in *Psychiatry*, edited by A. Tasman, J. Kay, J. A. Lieberman, M. B. First, and M. Maj. Chichester, UK: John Wiley & Sons, Ltd.
- Strauss, J. S., Carpenter, W. T. and Bartko, J. J. 1974. Part III. Speculations on the Processes That Underlie Schizophrenic Symptoms and Signs. *Schizophrenia Bulletin* 1(11):61–69. doi: 10.1093/schbul/1.11.61.
- Taine, H. 1870. *De l'intelligence*, Librairie Hachette et Cie
- Tandon, R. 2012. The Nosology of Schizophrenia. *Psychiatric Clinics of North America* 35(3):557–69. doi: 10.1016/j.psc.2012.06.001.
- Tversky, A., & Kahneman, D. 1986. Judgment under uncertainty: Heuristics and biases. In H. R. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 38–55). Cambridge University Press. (This chapter originally appeared in "Science," 1974, 185, 1124-1131)
- Vaccaro, A. G., and Fleming, S. M. 2018. Thinking about Thinking: A Coordinate-Based Meta-Analysis of Neuroimaging Studies of Metacognitive Judgements. *Brain and Neuroscience Advances* 2:239821281881059. doi: 10.1177/2398212818810591.
- Van Camp, L. S. C., Sabbe, B. G. C., & Oldenburg, J. F. E. 2017. Cognitive Insight: A Systematic Review. *Clinical Psychology Review* 55:12–24. doi: 10.1016/j.cpr.2017.04.011.
- Varela, F. J. 1996. Neurophenomenology: A methodological remedy for the hard problem. *Journal of consciousness studies*, 3(4), 330-349.
- Varela, F. J., Thompson, E., and Rosch, E. 2016. *The Embodied Mind: Cognitive Science and Human Experience*. revised edition. Cambridge, Massachusetts ; London England: MIT Press.

Verdoux, H. 2003. Psychiatry in France. *International Journal of Social Psychiatry*, 49(2), 83-86.

Vinogradov, S., Luks, T. L., Schulman, B. J., & Simpson, G. V. 2008. Deficit in a Neural Correlate of Reality Monitoring in Schizophrenia Patients. *Cerebral Cortex (New York, N.Y. : 1991)* 18(11):2532–39. doi: 10.1093/cercor/bhn028.

Von Helmholtz, H. 1911. Handbuch der physiologischen. *Optik*, 2.

Waters, F., Barnby, J. M., & Blom, J. D. 2021. Hallucination, Imagery, Dreaming: Reassembling Stimulus-Independent Perceptions Based on Edmund Parish's Classic Misperception Framework. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1817):20190701. doi: 10.1098/rstb.2019.0701.

Waters, F., Blom, J. D., Dang-Vu, T. T., Cheyne, A. J., Alderson-Day, B., Woodruff, P., & Collerton, D. 2016. What Is the Link Between Hallucinations, Dreams, and Hypnagogic–Hypnopompic Experiences? *Schizophrenia Bulletin* 42(5):1098–1109. doi: 10.1093/schbul/sbw076.

Wright, A., Nelson, B., Fowler, D., & Greenwood, K. 2020. Perceptual Biases and Metacognition and Their Association with Anomalous Self Experiences in First Episode Psychosis. *Consciousness and Cognition* 77:102847–102847. doi: 10.1016/j.concog.2019.102847.

Yeung, N., and Summerfield, C. 2012. Metacognition in Human Decision-Making: Confidence and Error Monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1594):1310–21. doi: 10.1098/rstb.2011.0416.

Xue, K., Shekhar, M., & Rahnev, D. 2021. The shape of metacognitive noise confounds metacognitive efficiency with confidence bias. *Journal of Vision*, 21(9), 2794-2794.

Zahavi, D. 2003. Husserl's Phenomenology. Stanford, Calif: Stanford University Press.

Appendices

1. Supplementary information for project 1 (Meta-analysis)

Supplementary methods

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.

Search strategy: In Scopus, we retrieved results from both “documents” and “secondary documents” tabs.

Data selection: to ensure mutual agreement, MR and PS iteratively performed consistency checks on 20 randomly selected studies prior to the selection process, until they reached 90% of agreement. The selection process consisted in the application of the inclusion criteria from the PICO scheme following a two-step process, first on the basis of title and abstract, and then on full-text. The raters were blind to each other’s decisions. The first step resulted in a total of 192 disagreements, which were resolved on the basis of information from the abstracts. The raters commented and double-checked each conflicting judgement until an agreement was reached. The same strategy was used for resolving the disagreements in the full-text selection.

Data extraction: when scores of interest were reported separately across variables irrelevant for our purpose, we computed weighted means and pooled standard deviations. In studies reporting several experiments from the same sample, we computed mean scores and pooled standard deviations, resulting in one score for each variable of interest per study. On the contrary, studies reporting scores from different population samples were considered as if they were independent experiments.

Clinical scores: Brief Psychiatric Rating Scale (BPRS), Scale for the Assessment of Positive Symptoms (SAPS) and Scale for the Assessment of Negative Symptoms (SANS) scores were transformed into PANSS equivalent scores according to the recommended formulae (Leucht et al., 2013; van Erp et al., 2014). We had further pre-registered the extraction of confidence bias

(mean confidence), depression, insight, psychosocial functioning scores, and brief psychosis episode, but did not proceed due to too few studies reporting them (N = 7, N = 4, N = 6, N = 5, and N = 7, respectively).

Formulation of the meta-analytic model:

Let ES be the observed effect sizes:

$$ES_i \sim \text{Normal}(\mu_i, \sigma_i)$$

$$\mu_i = \alpha + \alpha_{study[i]}$$

$$\alpha \sim \text{Normal}(-0.3, 1)$$

$$\alpha_{study[i]} \sim \text{Normal}(0, \tau)$$

$$\tau \sim \text{Half-Cauchy}(0.1)$$

Where μ_i indicates the effect size of study i , and σ_i^2 is the known variance of the effect in study i . α is the intercept parameter of the model (the average effect size in the population). We chose a random-effect model rather than a fixed-effect model as the distributions of effect sizes are expected to be heterogeneous for metacognition in schizophrenia. Because metacognitive deficit in schizophrenia is commonly described, we specified a mildly informative prior corresponding to a metacognitive deficit with small to medium effect-size, and τ^2 is the between-studies variance, provided with a mildly informative prior. Because studies with multiple experiments were rare (4 in total), we could not add experiment in addition to study as a random factor in our model.

Computation of the effect sizes (Hedge' g):

Effect sizes were computed as Hedge's G with the R package `esc` (Lüdtke, 2018) using the procedure given in Borenstein et al. (2010) as follows:

$$g = J \times \frac{mean_p - mean_c}{sd_{pooled}}$$

where J is the correction factor to achieve an unbiased estimator, defined as:

$$J = 1 - \frac{3}{4(n_p + n_c - 2) - 1}$$

$mean_p$ is the average meta-performance reported for the patient group of size n_p , and $mean_c$ is the average meta-performance reported for the control group of size n_c .

sd_{pooled} is the pooled standard deviation within both groups:

$$sd_{pooled} = \sqrt{\frac{(n_p - 1)sd_p^2 + (n_c - 1)sd_c^2}{n_p + n_c - 2}}$$

with sd_p the standard deviation of the meta-performance reported for the patient group, and sd_c the standard deviation for the control group.

Finally, the variance of g is defined as:

$$var(g) = j^2 \cdot \left[\frac{n_p + n_c}{n_p n_c} + \frac{1}{2(n_p + n_c)} \left(\frac{mean_p - mean_c}{sd_{pooled}} \right)^2 \right]$$

To assess the extent to which metacognitive performance was contaminated by first-order performance in studies which did not control for it, we fitted a meta-regression model identical to M1 with the z-scores of first-order performance as an additional continuous regressor. Based on the literature (Faivre et al., 2020), we specified an informative Gaussian prior ($m = 0.56$, $sd = 0.24$) for the slope of the meta-regression.

Risk of bias: The risk of bias regarding selection, comparability and outcome was assessed in parallel by two raters (MR and PS) and intraclass correlation (ICC) scores of agreement were computed with the R package *irr* (Gamer et al., 2012)

Supplementary results

We assessed whether our selection of studies contained any extreme effect size values via a leave-one-out sensitivity analysis, which computes the effect sizes for each fold of $n-1$ studies, with n the total number of selected studies. This analysis revealed a strongly deviant study driving the overall effect size 4 standard deviations above the mean (Fig S1). This study was therefore excluded.

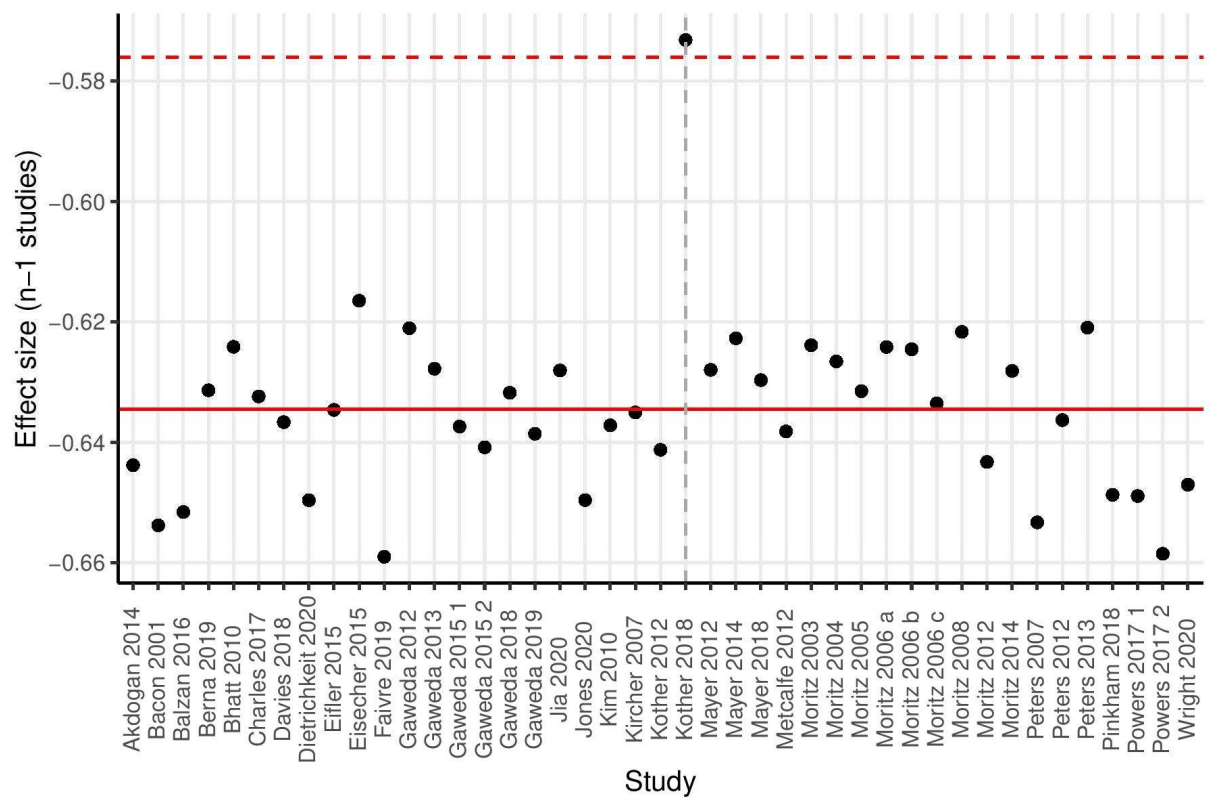


Figure S1: Effect size for each fold of n-1 studies. The horizontal red solid line indicates the average effect size, the dashed red line is four standard deviations above the mean. The vertical grey dashed line points to the deviant article, which has been excluded from our analysis.

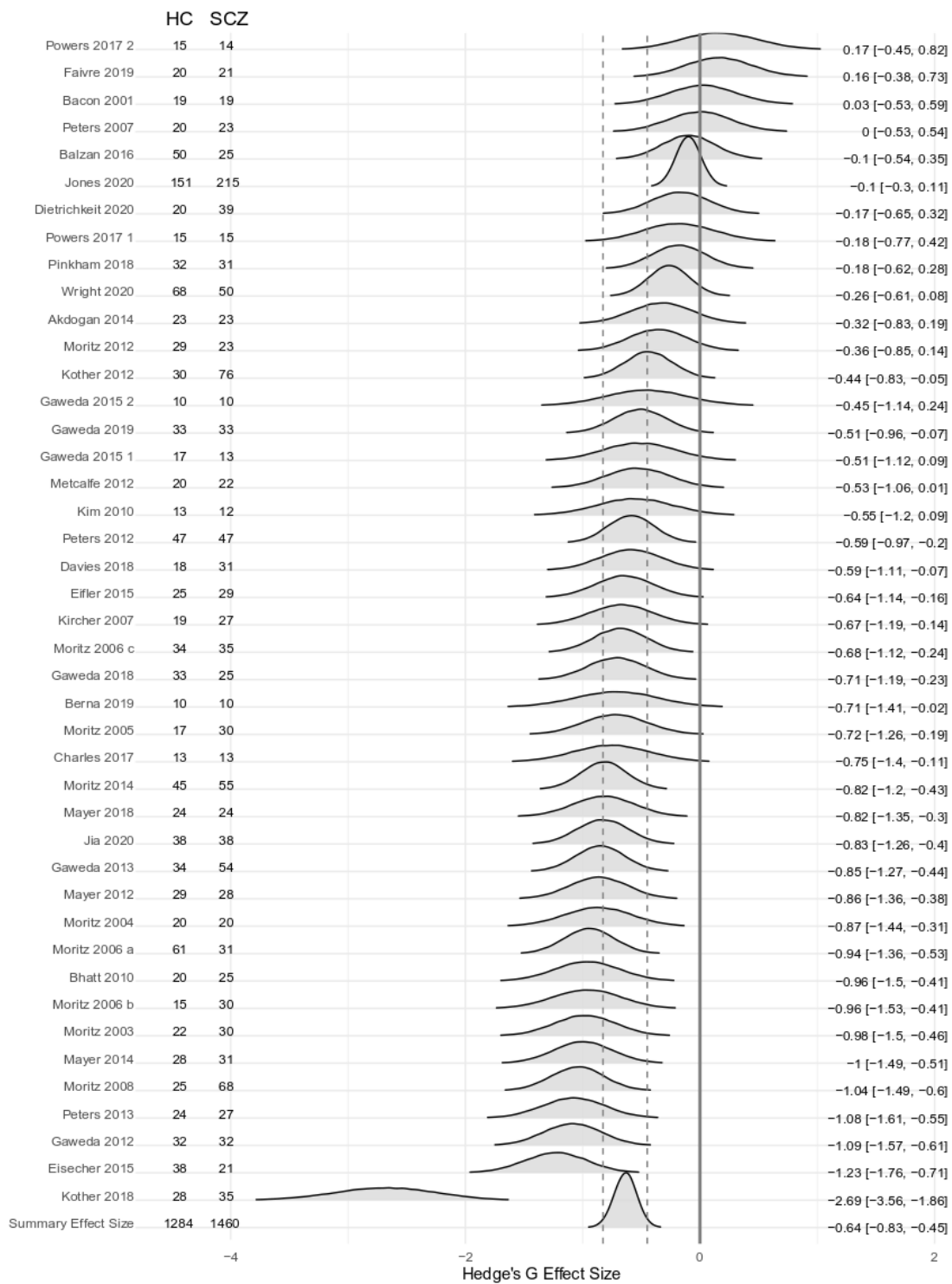


Figure S2: Forest plot of the metacognitive deficit in schizophrenia including the outlier study. Left: Authors with publication year; Middle: posterior distribution of the effect size; Right: mean and 95% CrI of the posterior distribution. The summary effect size is displayed on the last row: the

solid vertical grey line is centred on zero (i.e., equivalent metacognitive performance between groups), and the dashed vertical lines depict the boundaries of the 95% CrI.

Correspondence between effect sizes and metacognitive measures

To get a sense of how effect sizes translate into differences in measures of metacognitive performance, we provide the correspondences between the two in table S1.

Metric	sample	Group Difference (patients – controls)	g Effect Size
confidence gap (%)	8	-15.64	-0.61
KCI (%)	9	8.06	0.60*
M-ratio	5	-0.04	-0.05

Table S1 : Values of the principal measures of metacognition (i.e. M-ratio, confidence gap, KCI : Knowledge Corruption Index) and the corresponding g effect sizes.

* By construction, a metacognitive deficit in patients results in a positive KCI, whereas other metrics such as Confidence gap and M-ratio return negative values. Therefore, we homogenized these metrics by reversing the sign of KCI.

Robustness analysis

To assess the influence of our choice of prior on effect size estimates, we re-ran the model M1 with a set of different priors α varying in mean and SD (Fig.S3). Except for very informative priors (SD = 0.1), results were robust to prior variations, in support of our main findings.

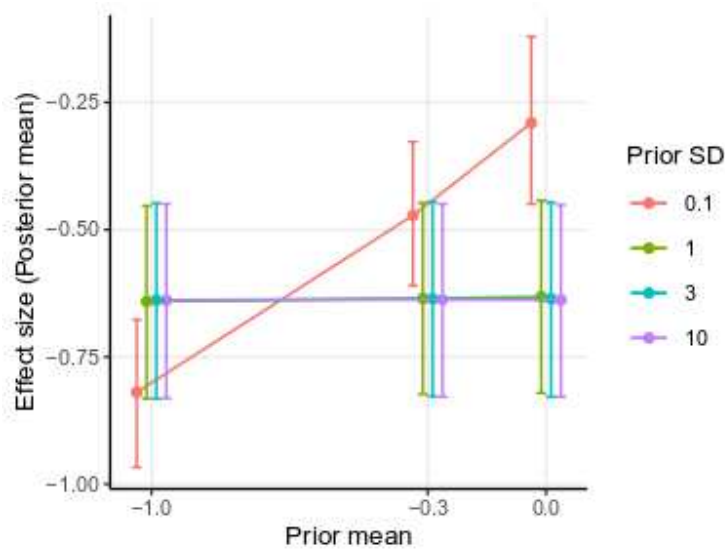


Figure S3: Metacognitive deficit effect size estimated by M1 with priors α varying in means (x-axis) and SDs (0.1: red; 1: green; 3: cyan; 10: purple).

The analysis of the moderating role of first-order performance was motivated by a significant moderation factor ($Q_{\text{between}} = 8.41$, $df = 1$, $p = .0004$), which means that the effect size was related to the control of first-order performance. Regarding subgroup analyses, I^2 for the non-controlled and controlled sub-groups were 0.65 and 0.65, corresponding to a reduction of 1.52% compared to I^2 across all studies. Q-statistic remained significant for the non-controlled and controlled sub-groups. We assessed the influence of performance-matching with a model identical to M1 including performance-matching as an additional binary predictor, with a Gaussian prior centered on -0.3 (SD = 1).

To examine the correlation between cognitive and metacognitive deficits among studies which did not control for first-order performance between groups, we performed a meta-regression by adding the standardized cognitive deficit as a continuous predictor to the model M1. On the basis of a previous study (Faivre et al., 2020), we specified a Normal prior with mean = 0.56 and SD = 0.24 for the slope parameter (the value of 0.24 corresponded to 0.1 before standardization). The mean slope value was $b = 0.15$, 95% CrI [-0.003, 0.30], with 97.3% of the slope estimates above 0, and very strong evidence in support of our hypothesis for a positive relationship between cognitive and metacognitive deficits ($BF_{10} = 36$). Although a prior with SD = 0.24 is quite informative, the robustness analysis revealed stable patterns for a prior with mean = 0.56 (Fig.S5).

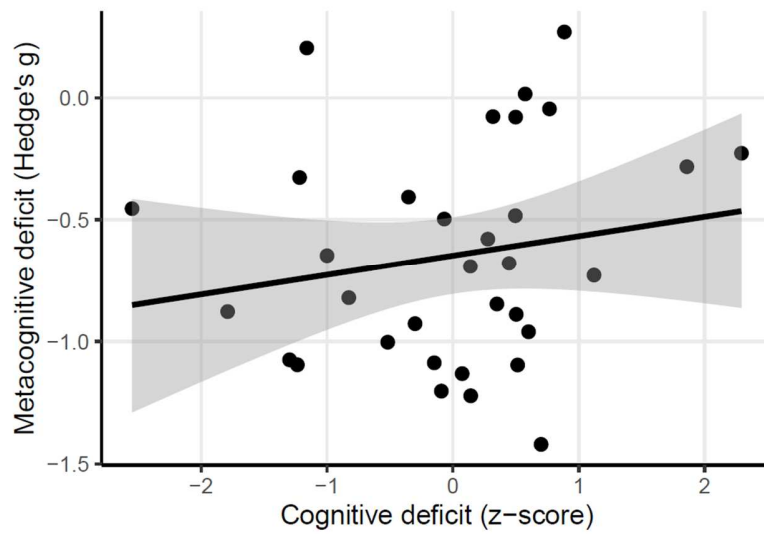


Figure S4: Meta-regression of the metacognitive deficit as a function of the cognitive deficit for studies which do not control for first-order performance. Each data point corresponds to one study (N = 33).

Robustness analysis

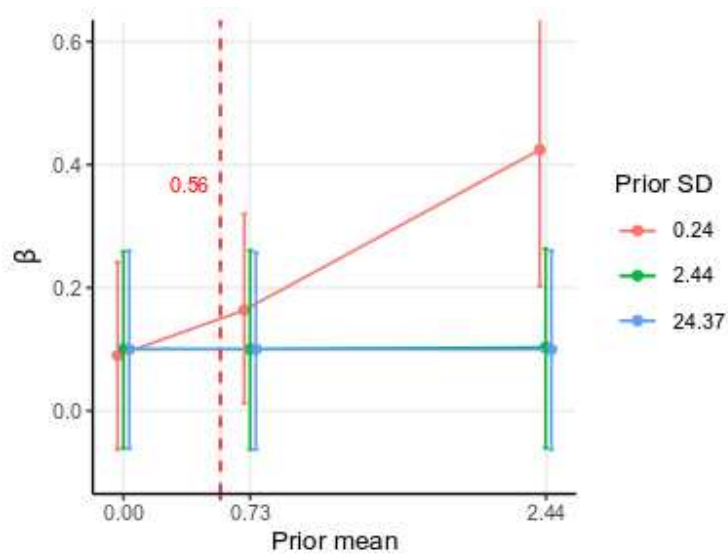


Figure S5: Slope estimates for the relationship between cognitive and metacognitive deficits under priors varying in means (x-axis, 0.73 and 2.44 corresponding to 0.3 and 1, respectively, after standardisation) and SDs (0.24: red; 2.44: green; 24.37: blue, corresponding to 0.1, 1, 10, respectively, after standardisation). The vertical red dashed line indicates the prior's mean value specified in our analysis.

Regarding metacognitive deficits across cognitive domains, I^2 for the memory, perception and other domain sub-groups were 0.52, 0.64 and 0 respectively, corresponding to a reduction of

21%, 3%, and 100% compared to the global analysis. Q-statistic remained significant for the memory and perception sub-groups. We performed a sub-group analysis with a weakly informative prior with mean = -0.3, and SD = 1 for the effect of cognitive domains. We found extremely strong evidence supporting a greater influence of memory vs. perception studies on the metacognitive deficit ($m = -0.38$, 95% CrI [-0.66, -0.09], $BF_{10} = 203$). This pattern was even more pronounced when memory was compared with other domains (social and agency; $m = -0.49$, 95% CrI [-0.83, -0.12], $BF_{10} = 182$).

Finally, we performed meta-regressions between the metacognitive deficits and clinical variables, with a prior of mean 0 and SD = 1. Contrary to what we had predicted, none of these meta-regressions revealed conclusive evidence (Figure S6)

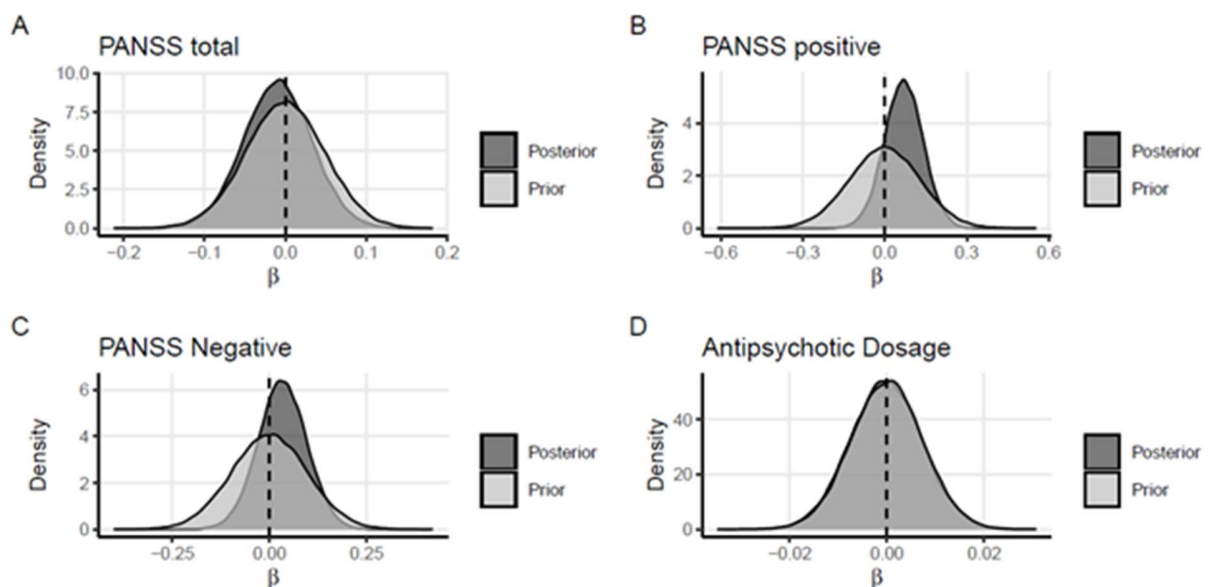


Figure S6: Meta-regressions of the metacognitive deficit with PANSS total scores (A), PANSS positive scores (B), PANSS negative scores (C), and antipsychotic dosage (D). The x-axes represent the posterior estimates for the slope parameter. Posterior and prior distributions are depicted in dark gray and light gray, respectively.

Similar results were found between first-order cognitive deficits and clinical features (Fig. S7) : $BF_{01} = 1.08$ for PANSS total scores ($N = 35$), $BF_{01} = 1.55$ for PANSS positive scores ($N = 32$), $BF_{01} = 1.20$ for PANSS negative scores ($N = 33$), $BF_{01} = 0.98$ for pharmacological treatment ($N = 20$).

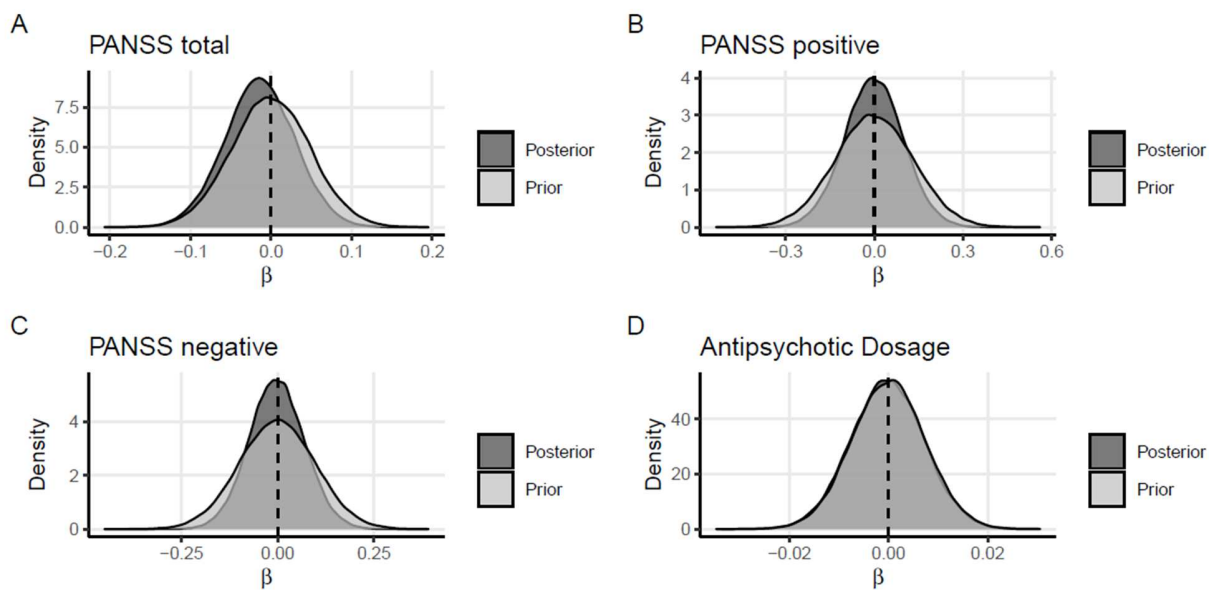


Figure S7: Meta-regressions of the cognitive deficit with PANSS total scores (A), PANSS positive scores (B), PANSS negative scores (C), and antipsychotic dosage (D). The x-axes represent the posterior estimates for the slope parameter. Posterior and prior distributions are depicted in dark gray and light gray, respectively.

Effect size of confidence bias

Although we had preregistered to compare average confidence per group, the scarcity of these data in our selection prevented us from performing these analyses. However, confidence bias is worth considering given the large literature on error detection in schizophrenia. The best analysis we could do with the current dataset was to compare the average confidence estimations of patients versus controls in a subset of 7 studies. At odds with the literature, we found a confidence bias taking the form of underconfidence in patients, though with a small effect size ($g = -0.14$, 95% CI [-0.4, 0.01], $BF_{10} = 8.05$) (see Figure S8). Of note, this effect may be explained by a lower first-order performance among patients, which mitigates this finding. We now mention it in the SI.

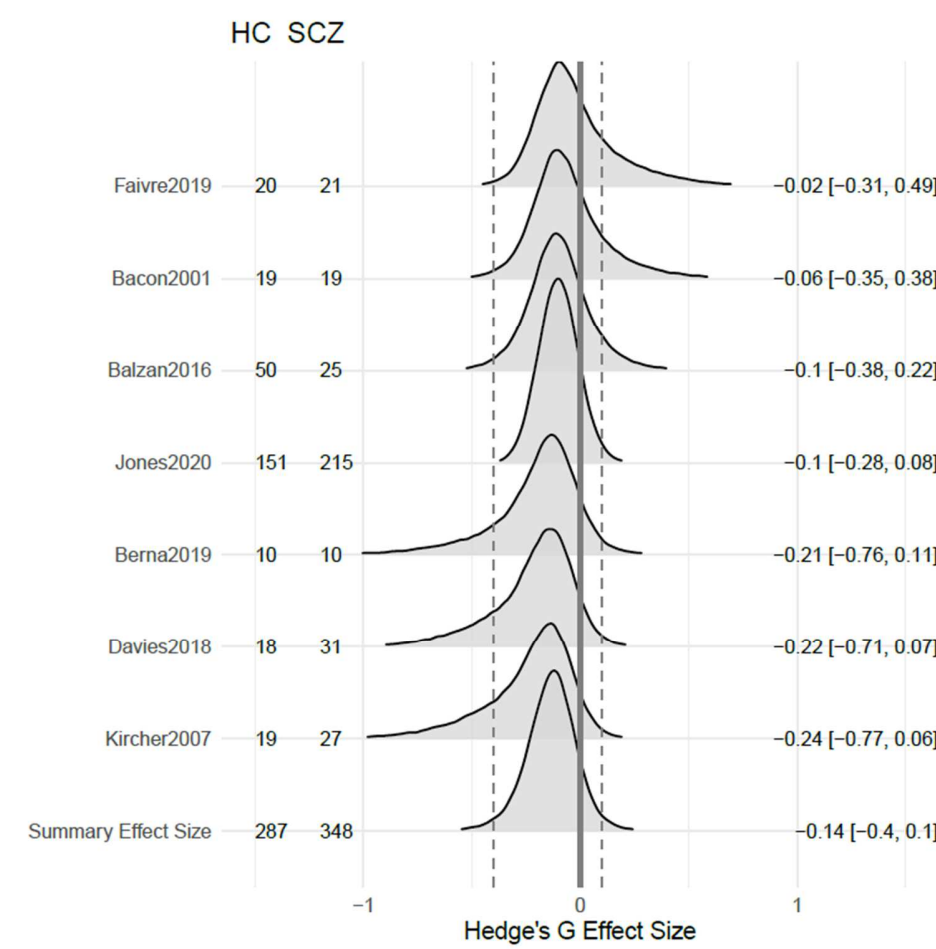


Figure S8: Forest plot of the confidence bias in schizophrenia. Left: Authors with publication year and sample sizes; Middle: posterior distribution of the effect size; Right: mean and 95% CrI of the posterior distribution. The summary effect size is displayed on the last row: the solid vertical grey line is centred on zero (i.e., equivalent metacognitive performance between groups), and the dashed vertical lines depict the boundaries of the 95% CrI.

Risk of bias

The risk of bias was assessed using the Newcastle Ottawa Scale. The total ICC score (two way model, agreement type, single unit) revealed an average agreement between the two raters (MR and PS) of 0.55 according to interpretation schemes given by Koo & Li (2016). We then targeted the studies for which there were more than two divergences out of nine between the two raters. The six studies which reached this criterion were assessed again by two others raters (NF and PR). For these six studies, the final NOS score was obtained by averaging the scores given MR, PS, NF and PR.

About half of the studies included in this meta-analysis were rated as poor according to the interpretation scheme (Table S2) provided by the Newcastle Ottawa Scale (Fig. S9-10).

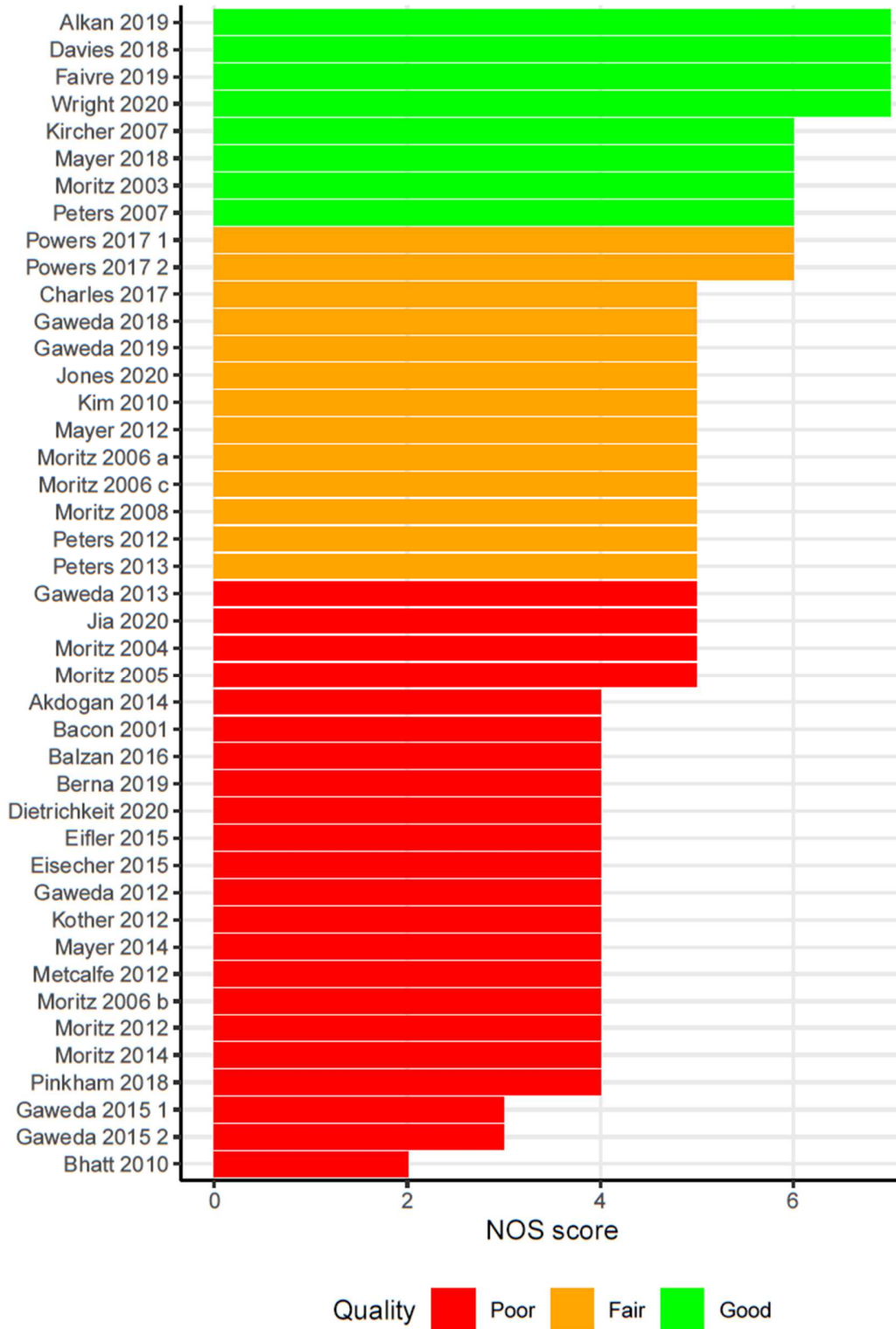


Figure S9: NOS quotation for each article included in the present meta-analysis

Quality \ Domain	Selection	Comparability	Outcome
Good	3 or 4 *	1 or 2 *	2 or 3 *
Fair	2 *	1 or 2 *	2 or 3 *
Poor	0 or 1 *	0 *	0 or 1*

Table S2 : Correspondence between the NOS score and quality rating (Poor, Fair, Good), according to the NOS official recommendations

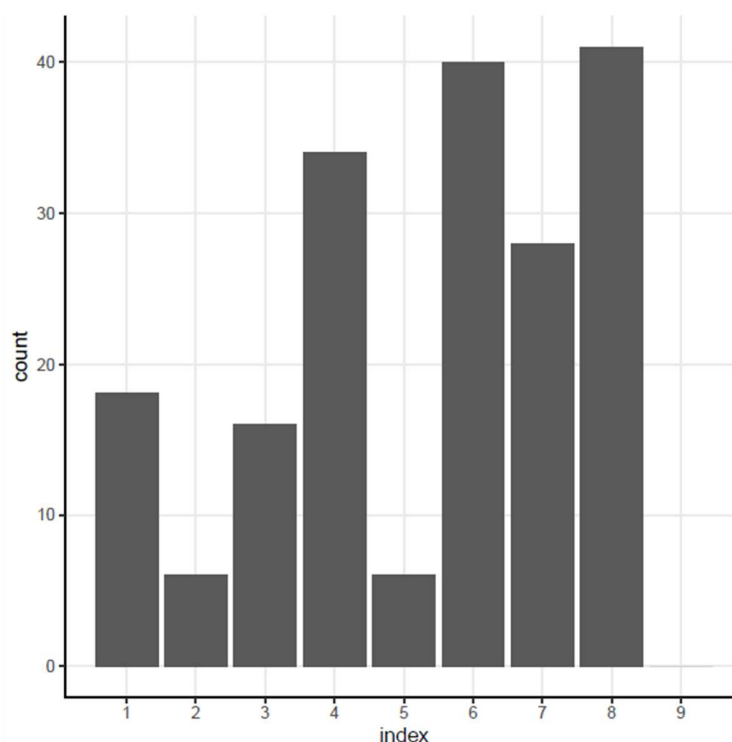


Figure S10 : The nine NOS items are presented on the x-axis: 1: Case definition adequacy, 2: Case representativeness, 3: Control selection, 4: Definition of controls, 5: Control for first-order performance between groups, 6: Control for Age/QI between groups , 7: Computerized protocol, 8: Same protocol for both groups, 9: Non-response rate. The y-axis represents the total number of articles which were granted a point for each NOS item.

Supplementary Discussion

The same research group (hereafter Group A) contributed a large number of selected studies ($n = 19$). To assess whether this could have influenced our results, we ran an exploratory analysis comparing the amplitude of the metacognitive deficit between studies led by Group A vs. other groups. As Group A's studies are all non-controlled studies regarding first-order performance, we restricted the analysis to this category to get a meaningful comparison. The results reveal that metacognitive deficits estimated by Group A are of larger magnitudes ($BF_{10} = 48.71$, see Figure S11).

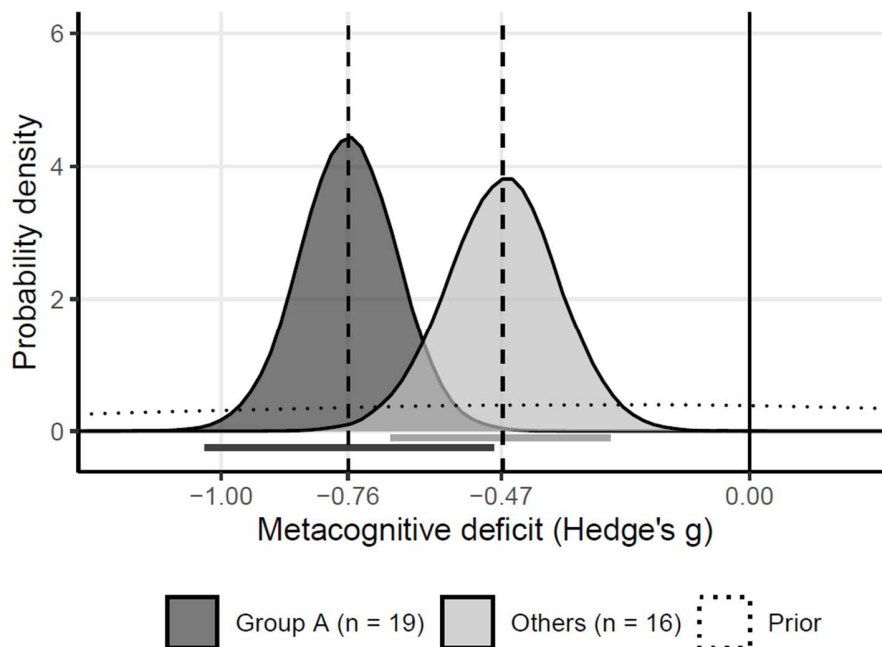


Figure S11: Posterior distributions of the metacognitive deficit among non-controlled studies. Dark gray: Group A, Light gray: Other research groups. The dotted line represents the prior distribution, vertical dashed lines the mean posterior values, and the horizontal bars the 95% CrI.

Study	PANSS (total)	PANSS (positive)	PANSS (negative)	Chlorpromazine (mg)
Akdogan2014	69 ± 27.8	14.9 ± 4.6	18.8 ± 10.5	Not reported
Bacon2001	67.625 ± 11.5	44.7568 ± 16	54.48318 ± 20.3	348 ± 273
Balzan2016	11.08 ± 3.12	Not reported	Not reported	Not reported
Berna2019	75.13 ± 19.68	20.5 ± 5.04	17.88 ± 6.66	Not reported
Bhatt2010	Not reported	Not reported	Not reported	Not reported
Charles2017	Not reported	11.3 ± 3.4	15.5 ± 4.4	224.3 ± 103
Davies2018	Not reported	5.43 ± 2.44	5.4 ± 2.8	279.4 ± 150.2
Dietrichkeit2020	51.69 ± 13.02	Not reported	Not reported	Not reported
Eifler2015	67.41 ± 16.01	15.75 ± 5.34	16.59 ± 6.54	425.96 ± 183.68
Eisecher2015	85.67 ± 18.62	22.95 ± 5.04	18.26 ± 7.02	Not reported
Faivre2019	78.5 ± 12.79	17.2 ± 5.85	20.5 ± 6.44	439.7 ± 27.08
Gaweda2012	72 ± 13.21	13.2 ± 5.36	19.54 ± 6.89	Not reported
Gaweda2013	62.57 ± 11.37	16.81 ± 5.43	17.81 ± 6.25	Not reported
Gaweda2015_1	66 ± 17.1	20.15 ± 8.07	18.23 ± 4.56	Not reported
Gaweda2015_2	79.55 ± 16	23.88 ± 6.97	19.33 ± 10.13	Not reported
Gaweda2018	87.645 ± 14.58	11.5 ± 4.3	3.72 ± 1.18	405.98 ± 175.04
Gaweda2019	46.81 ± 12.27	12 ± 4.9	13 ± 5.22	Not reported
Jia2020	54.6 ± 12.9	13.2 ± 5.6	14.1 ± 6.4	369 ± 397
Jones2020	Not reported	Not reported	Not reported	Not reported
Kim2010	22.5625 ± 6.6	20.3984 ± 7.8	22.69635 ± 10.1	Not reported
Kircher2007	66.4 ± 21.4	17.3 ± 7.2	17.2 ± 6.5	411 ± 257
Kother2012	58.28 ± 15.82	15.2 ± 8.21	14.14 ± 6.34	Not reported
Mayer2012	20.795 ± 5.74	24.173952 ± 10.23	31.100829 ± 13.81	361.12 ± 381.78
Mayer2014	Not reported	25.32544 ± 11.12	32.530722 ± 14.95	329.23 ± 299.36
Mayer2018	31.29 ± 7.92	14.92 ± 4.27	16.17 ± 6.31	501.13 ± 265.57
Metcalfe2012	46.1 ± 5.8	Not reported	36.89241 ± 12.2	Not reported
Moritz2003	63.3025 ± 12.3	8.86 ± 4.4	7.62 ± 3	253.04 ± 189.7
Moritz2004	13.84 ± 4.51	Not reported	Not reported	676.59 ± 523.86
Moritz2005	17.33 ± 7.35	Not reported	Not reported	671.6 ± 494.4
Moritz2006_a	68.45 ± 17.48	Not reported	Not reported	687.88 ± 864.69
Moritz2006_b	Not reported	9.67 ± 4.2	12.77 ± 6.17	Not reported
Moritz2006_c	66.31 ± 16.62	Not reported	Not reported	Not reported
Moritz2008	62.24 ± 18.37	Not reported	Not reported	Not reported
Moritz2012	60.04 ± 11.11	Not reported	Not reported	Not reported
Moritz2014	Not reported	1.8 ± 0.39	2.32 ± 0.54	Not reported
Peters2007	55.6375 ± 8.52	9.17 ± 4.29	5.3 ± 2.12	Not reported
Peters2012	54.7 ± 13.27	14.81 ± 6.28	11.7 ± 4.94	Not reported
Peters2013	49.29 ± 9.75	12.66 ± 4.76	11.59 ± 4.63	368.87 ± 260.63
Pinkham2018	30.48 ± 6.44	17.54 ± 5.68	12.23 ± 3.2	452.26 ± 416.14
Powers2017_1	61.2 ± 7.94	19.67 ± 4.14	15.2 ± 4.8	431.46 ± 90.92
Powers2017_2	54.43 ± 10.33	14.93 ± 3.26	14.14 ± 4.19	330.43 ± 66.28
Wright2020	Not reported	12.4 ± 4.7	11.5 ± 4	Not reported

Table S3: Summary of clinical data for each study (PANSS total, PANSS positive, PANSS negative, chlorpromazine equivalent and diagnostic tool).

References

- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (Eds.). (2010). *Introduction to meta-analysis* (Reprinted). Wiley.
- Faivre, N., Roger, M., Pereira, M., de Gardelle, V., Vergnaud, J.-C., Passerieux, C., & Roux, P. (2020). Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *Psychiatry and Clinical Neurosciences*.
<https://doi.org/10.1503/jpn.200022>
- Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr'. *Various coefficients of interrater reliability and agreement*. R package version 0.84.1.
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163.
<https://doi.org/10.1016/j.jcm.2016.02.012>
- Leucht, S., Rothe, P., Davis, J. M., & Engel, R. R. (2013). Equipercntile linking of the BPRS and the PANSS. *European Neuropsychopharmacology, 23*(8), 956–959.
<https://doi.org/10.1016/j.euroneuro.2012.11.004>
- Lüdecke, D. (2018). *Esc: Effect Size Computation For Meta Analysis*. Zenodo.
<https://doi.org/10.5281/ZENODO.1249218>
- van Erp, T. G. M., Preda, A., Nguyen, D., Faziola, L., Turner, J., Bustillo, J., Belger, A., Lim, K. O., McEwen, S., Voyvodic, J., Mathalon, D. H., Ford, J., Potkin, S. G., & Fbirt. (2014). Converting positive and negative symptom scores between PANSS and SAPS/SANS. *Schizophrenia Research, 152*(1), 289–294. <https://doi.org/10.1016/j.schres.2013.11.013>
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57.
<https://doi.org/10.3758/s13423-017-1343-3>

2. Supplementary information for project 2 (Assessment of metaperceptual and metamemory abilities)

Supplementary Information

Preprint on medrxiv :

Martin Rouy, Michael Pereira, Pauline Saliou, Remi Sanchez, Wassila el Mardi, Hanna Sebban, Eugenie Baque, Perrine Porte, Childeric Dezier, Vincent de Gardelle, Pascal Mamassian, Chris Moulin, Clement Donde, Paul Roux, and Nathan Faivre. (2023). Confidence in visual detection, familiarity and recollection judgements is preserved in schizophrenia spectrum disorder. <https://doi.org/10.1101/2023.03.28.23287851>

Methods

Participants

Optional bayesian stopping rule:

As per our preregistered plan, we sought to include 50 patients and 50 healthy controls, or to stop the recruitment whenever moderate evidence for the presence ($BF > 3$) or absence (< 0.33) of a specific metacognitive impairment among individuals with schizophrenia indicated by an interaction between group and task-domain on metaperformance. It turned out that our preregistered evidence thresholds were already exceeded when we fully inspected the data for the first time (Figure S1).

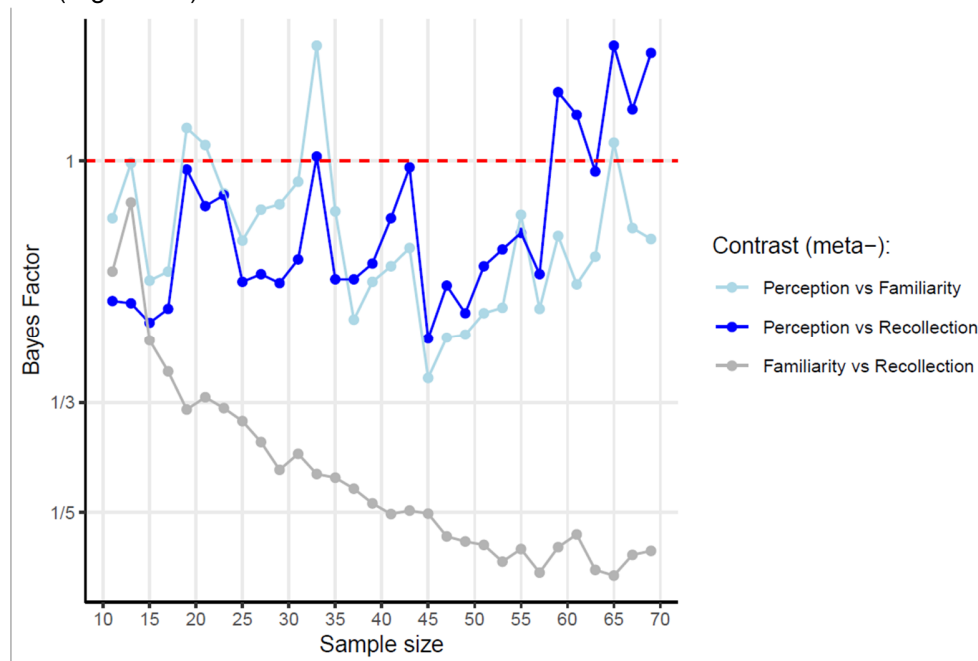


Figure S1. Bayesian sequential analysis of the interaction between task and group on metaperformance for each contrast: Visual detection vs Familiarity in blue, Visual detection vs Recollection in light blue, Familiarity vs Recollection in gray. Bayes Factors $< \frac{1}{3}$ for the contrast between meta-familiarity and meta-recollection are evidence for an absence of a specific deficit in memory tasks.

Exclusions

Visual inspection led to participant exclusion in the following situations: 1) a pattern of first-order performance resulting from a strong response bias, revealed by high (resp. low) and stable first-order performance across all tasks for all non-null stimulus strength, and low (resp. high) performance across all tasks in catch trials (i.e. null evidence), strongly suggesting that those participants did not do the task properly; 2) no variability in confidence judgments (leading to the impossibility to compute indices of metacognitive performance). Accordingly, we excluded 3 patients with extreme values of criterion as well as weak sensitivities across all tasks; 3 control participants with a ceiling effect on confidence ratings (i.e. no variance in their responses); one patient due to the misuse of the confidence scale, revealed by a bimodal distribution of confidence ratings.

Recruitment procedure

Patients were recruited from community mental health centers and outpatient clinics in Versailles and Grenoble and were included if they met the criteria for a diagnosis of schizophrenia or schizoaffective disorders according to the DSM-5 during a diagnostic interview. Healthy volunteers between 18 and 65 years old were recruited from the general population, matched to the patients for age, gender, and education. All participants had normal or corrected-to-normal vision. Exclusion criteria for both groups comprised an estimated IQ (from Wechsler Adult Intelligence Scale IV matrix subtest¹) strictly lower than 2 standard deviations below the mean of the general population; substances or alcohol dependence within the past 6 months and current; or prior history of untreated significant medical illness or of neurological illness. The control group was screened for current psychiatric illness during a diagnostic interview and participants were excluded in case they met criteria for any mental disorders according to the DSM-V.

Clinical and neuropsychological evaluation

We used the French National Adult Reading Test² to assess premorbid IQ, and the matrix reasoning subtest from the WAIS-IV to exclude participants scoring lower than two standard deviations below the mean of the general population.

Stimuli

Using a 2-D Fourier Transform, the phase of each face image was randomized to create random noise backgrounds with spatial frequencies identically distributed. This noise background was

grayscaled (familiarity and visual detection task) or presented in blue or red to provide contextual information (recollection task). Luminosity was balanced between blue and red noise backgrounds using the following formula: $L = 0.30 \cdot R + 0.59 \cdot G + 0.11 \cdot B$. (where R, G and B stand for the blue, green and red channels).

Randomization

Within each trial, the sequence of faces during the encoding phase was pseudo-randomized regarding gender and background color to get 2 male faces and 2 women faces (hence 6 combinations possible), 2 blue and 2 red backgrounds (6 combinations), totalizing 36 possible combinations. Participants were asked to provide confidence judgments only in session 2.

Structure of the experiment

The rationale behind the split into two sessions is the following: a pilot study had previously demonstrated that individual performances were similar between familiarity and recollection tasks. Thus, the first session of the experiment started with the assessment of familiarity and recollection performance for each lag. In order to match perceptual performance on memory performance, we determined a visual psychometric curve from which we selected perceptual intensities corresponding to the memory performances obtained previously (see Figure S2 D.).

The first session had three parts, in the following order: 1) 5 blocks of 40 memory trials (familiarity and recollection condition randomly interleaved), 2) 35 trials in the visual detection task with a 1up/2down staircase procedure to determine the 71% detection threshold, and 3) 3 blocks of 80 trials in the visual detection condition, with 10 contrast levels relative to the detection threshold (relative intensities: [0, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5]). A psychometric curve was fitted for each participant (see Figure S2 A., B, C.).

To match perceptual performance with memory performance, we used the psychometric curve obtained from the end of session 1 as follows: for each perceptual trial in session 2, we selected the stimulus intensities corresponding numerically to the minimum and maximum of memory performance. Adding two equidistant performance levels between the minimum and maximum, we obtained four contrast levels to map with the four memory stimulus strengths (or lags), in terms of resulting performance (see Figure S2 D.).

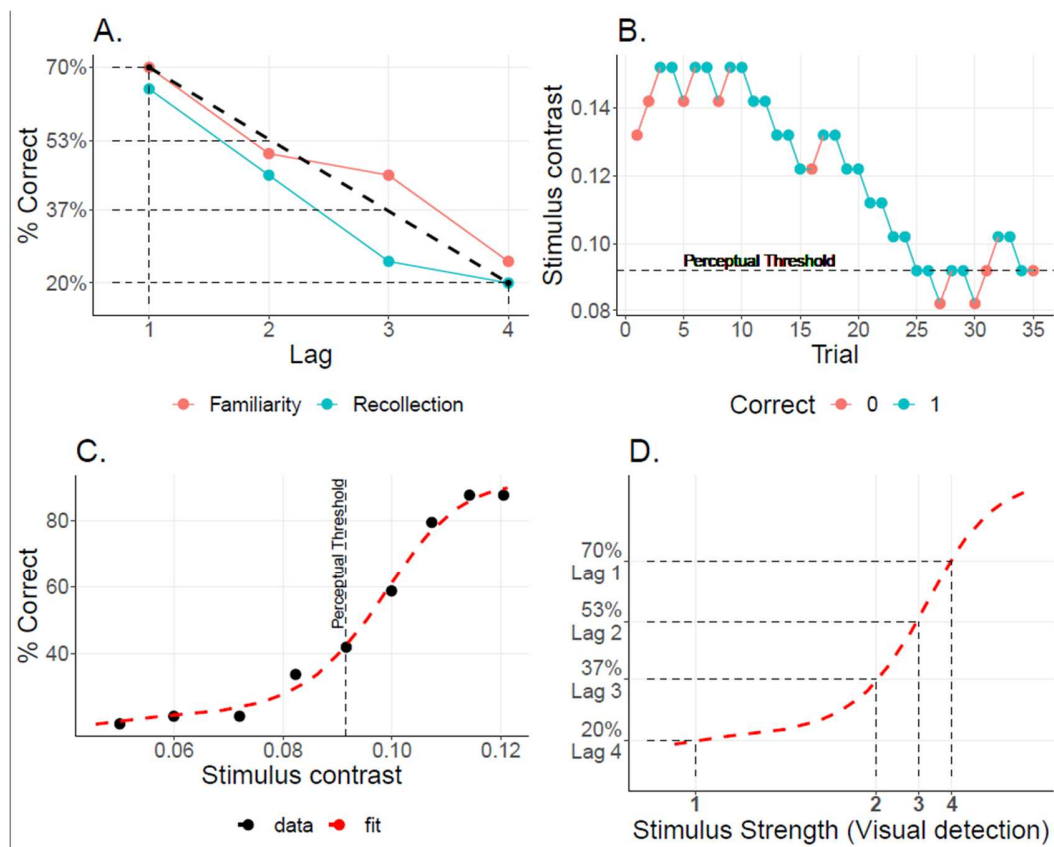


Figure S2. Experimental procedure for performance-matching. A. Memory performance obtained from one participant, during session 1. Memory performances to be matched by perceptual performance were computed at this stage for each stimulus strength. To ensure a wide range of performance, we took the minimum performance at Lag 4, the maximum performance at Lag 1, and equidistant values for Lag 2 and Lag 3. B: Visual detection staircase from one participant, during session 1. C: Psychometric curve obtained from one participant, during session 1. D. Illustration of the performance-matching procedure for one subject. Perceptual intensities were determined from memory performance measured in session 1 projected onto the individual psychometric curve, and updated online during session 2.

Bayesian analysis: prior specification

Priors were defined as follows:

We defined weakly informative Gaussian priors in the direction of a lower task performance for patients compared to controls (mean = -0.5, SD = 1); equivalent accuracy across the three tasks for patients (mean = 0, SD = 1) as well as for healthy controls (mean = 0, SD = 1); increased accuracy for higher stimulus strength (mean = 0.5, SD = 1); a positive regression slope of accuracy as a function of confidence for control participants, which served as baseline metacognitive performance (mean = 0.5, SD = 1), a metacognitive deficit among individuals with schizophrenia (lower regression slopes for patients compared to healthy controls: interaction confidence * group, mean = -0.1, SD = 1). Due to the absence of evidence for a specific

metacognitive impairment among individuals with schizophrenia (studies controlling for first-order performance across multiple domains), the prior for the double interaction confidence * group * task was centered on 0 (mean = 0, SD = 1).

Response times

Response times were log-transformed and modeled with a bayesian linear mixed-effects regression, with accuracy (binary categorical variable: correct or incorrect), standardized confidence (continuous variable), group (binary categorical variable: controls vs patients), evidence (ordinal variable with 4 levels, i.e. a common scale for memory lag levels and perceptual contrast levels. Stimulus strength is symmetrical to difficulty), task (categorical variable: perception, familiarity, recollection) as fixed effects, and a full random effect structure.

Formula:

$$\log(\text{RT}) \sim \text{accuracy} * \text{confidence} * \text{group} * \text{task} * \text{evidence} \\ + (\text{confidence} * \text{task} * \text{evidence} \mid \text{participant}) \quad (2)$$

Based on Faivre and colleagues³, we expected to replicate the following effects on response times: longer response times for patients compared to healthy control participants, shorter response times for correct vs. incorrect responses, a negative correlation between response times and confidence, a lower link between response times and response correctness among individuals with schizophrenia compared to healthy controls, a lower link between response times and confidence ratings among individuals with schizophrenia compared to healthy controls. According to these predictions, we defined the following Gaussian priors: shorter response times for correct vs. incorrect responses (mean = -0.1, SD = 1), longer response times for patients compared to healthy control participants (mean = 0.1, SD = 1); similar response times across the three tasks for patients (mean = 0, SD = 1) as well as for healthy controls (mean = 0, SD = 1); shorter response times for higher stimulus strength (mean = -0.1, SD = 1); a negative correlation between response times and confidence ratings (mean = -0.1, SD = 1), a lower link between response times and accuracy among individuals with schizophrenia compared to healthy controls (mean = 0.1, SD = 1), and a lower link between response times and confidence ratings among individuals with schizophrenia compared to healthy controls (mean = 0.1, SD = 1).

Domain-generality of metacognition

To assess whether metacognitive performance correlated across tasks while avoiding spurious correlations due to group-level shrinkage in hierarchical models, independent generalized mixed models were conducted for each subject and each task (i.e. 3 models per subject) as follows:

$$\text{accuracy} \sim \text{confidence} * \text{evidence} \quad (3)$$

Under the assumption of a domain-general architecture of metacognition⁴, we expected pairwise task metaperformance correlations. According to the disconnection hypothesis in schizophrenia⁵,

we expected lower pairwise task-metaperformance correlations among individuals with schizophrenia compared to healthy controls.

Correlation with clinical scores

Robust linear regressions were performed to explore the correlations between individual metacognitive performance (indicated by regression slopes between accuracy and confidence) and 1) demographic and neuropsychological scores (age, education level, premorbid IQ, depression (CDS), and WAIS matrix subtest standardized score), 2) clinical scores (only for patients: PANSS positive, negative and disorganization scores, cognitive insight (BCIS) and subjective cognitive functioning (SSTICS total and working memory scores).

Results

First-order performance

Contrast	Estimate [95% CrI]	Bayes Factor
Familiarity - Visual detection (Control)	0.13 [-0.31, 0.56]	0.26
Recollection - Visual detection (Control)	-0.2 [-0.62, 0.21]	0.32
Recollection - Familiarity (Control)	-0.33 [-0.78, 0.12]	0.52
Familiarity - Visual detection (Schizophrenia)	0.14 [-0.31, 0.58]	0.19
Recollection - Visual detection (Schizophrenia)	0 [-0.44, 0.43]	0.16
Recollection - Familiarity (Schizophrenia)	-0.14 [-0.59, 0.31]	0.14
Stim. strength (Control, Visual detection)	1.02 [0.88, 1.16]	>1000
Stim. strength (Control, Familiarity)	0.73 [0.58, 0.88]	>1000
Stim. strength (Control, Recollection)	0.47 [0.33, 0.61]	>1000
Stim. strength (Schizophrenia, Visual detection)	1.03 [0.91, 1.16]	>1000
Stim. strength (Schizophrenia, Familiarity)	0.74 [0.52, 0.97]	>1000
Stim. strength (Schizophrenia, Recollection)	0.48 [0.25, 0.71]	83.03
Stim. strength * group (Familiarity - Visual detection)	-0.12 [-0.38, 0.14]	0.19
Stim. strength * group (Recollection - Visual detection)	0.05 [-0.21, 0.31]	0.14
Stim. strength * group (Recollection - Familiarity)	0.17 [-0.09, 0.44]	0.21

Table S1.

Plausible explanation of imperfect intra-individual performance matching:

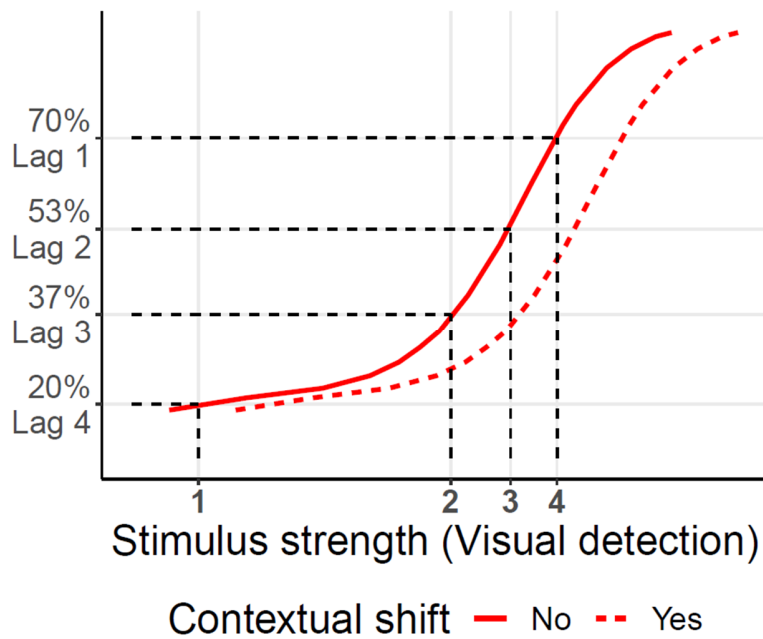


Figure S3. Reproduction of Figure S2 D., with an additional dashed psychometric curve rightward shifted compared to the solid curve, that would result from a contextual effect of high-contrast memory stimuli embedded within low-contrast visual detection stimuli during session 2.

Logistic regressions

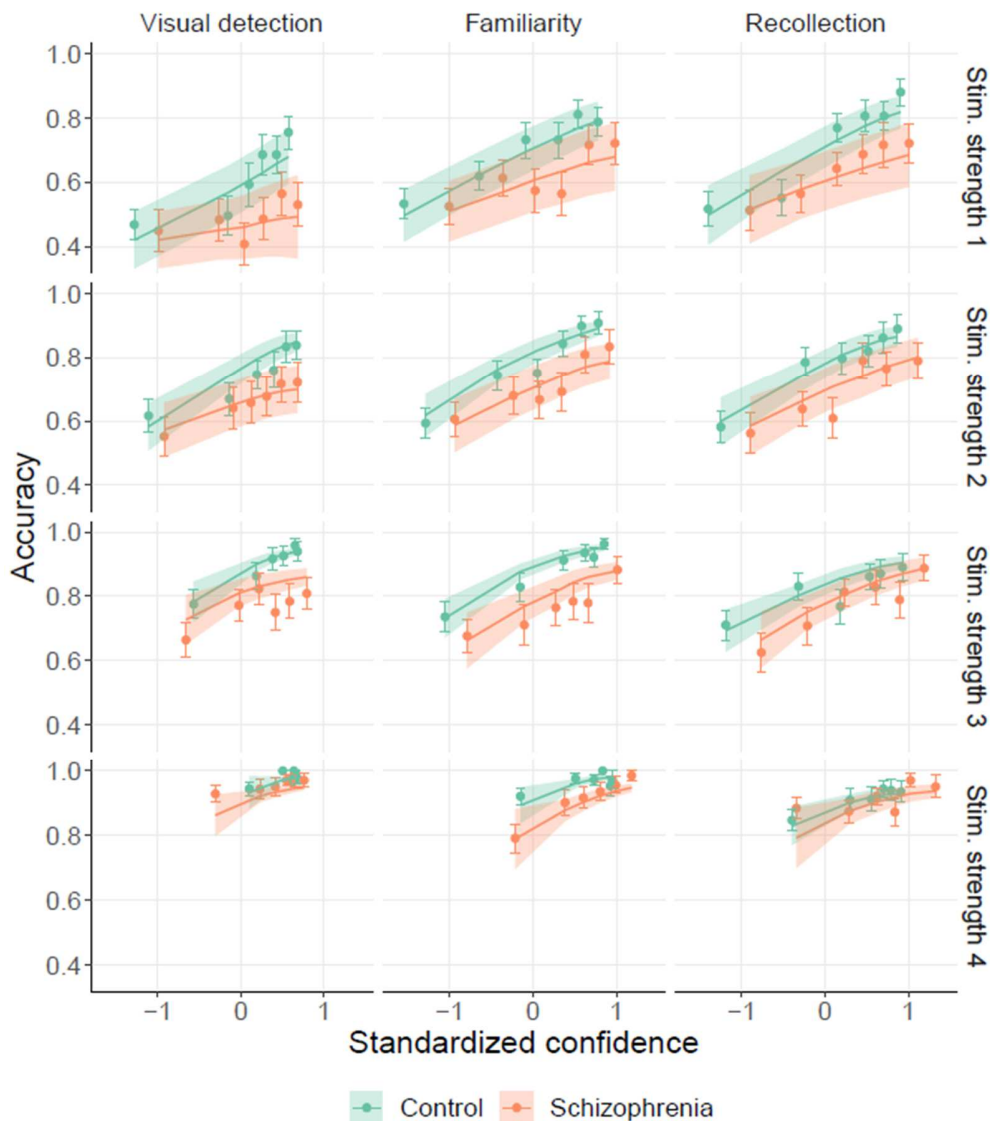


Figure S4: Accuracy as a function of standardized confidence, by group, task (columns), and stimulus strength (rows). Points and error bars indicate average accuracy and standard error of the mean, respectively, computed for each confidence bin. Solid lines and shaded areas represent model fit mean and 95% confidence interval, respectively.

Response times

On average, patients were slower to respond compared to healthy controls (0.19 [0.07, 0.31], BF = 257). Response times were shorter for correct responses compared to incorrect responses among individuals with schizophrenia (-0.08 [-0.10, -0.05], BF = 8000) and control participants (-0.23 [-0.26, -0.20], BF = 8000). However, correctness was less predictive of response times among patients compared to healthy controls, as indicated by the very strong evidence for the 'correctness * group' interaction (0.15 [0.11, 0.19], BF = 8000).

Response times were negatively correlated with confidence in both groups (patients: -0.19 [-0.24 ; -0.14], $BF = 8000$; controls: -0.21 [-0.24 , -0.18], $BF = 8000$), meaning that participants were longer to respond when less confident, regardless of the group (as indicated by the weak and inconclusive ‘confidence * group’ interaction: 0.02 [-0.04 , 0.07], $BF = 2.44$). However, the strength of the link between response times and confidence was higher among controls compared to individuals with schizophrenia for correct responses, but not for incorrect responses (Figure S5, ‘confidence * group * correctness’ double interaction: 0.08 [0.04 , 0.13], $BF = 614$).

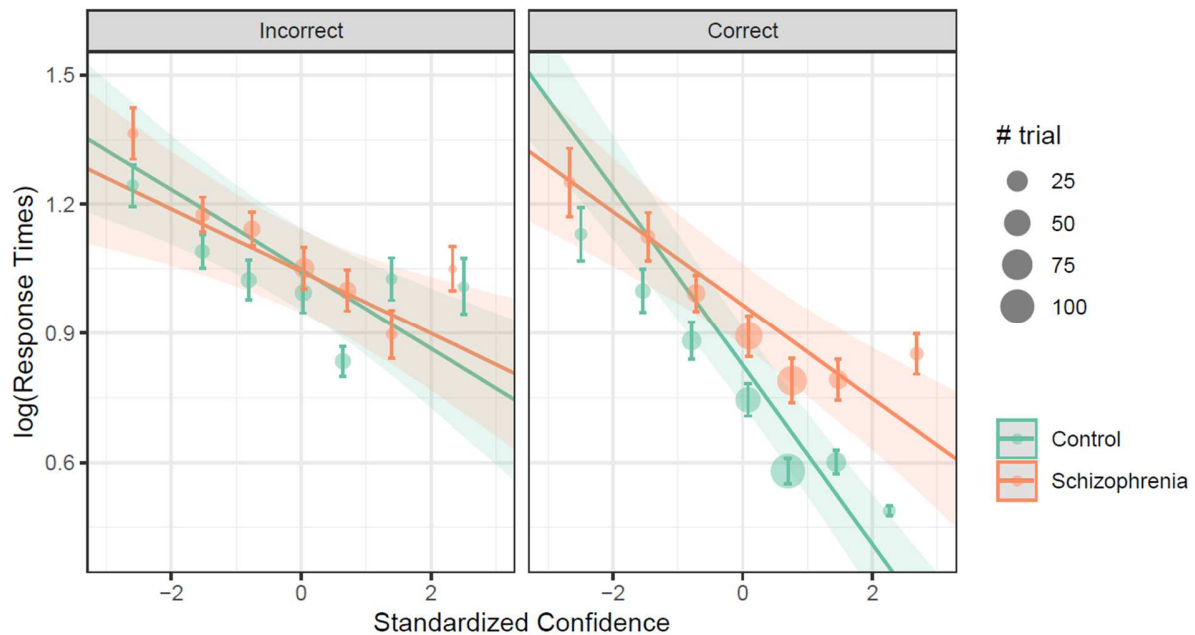


Figure S5: Response times (log-transformed) regressed on standardized confidence for incorrect responses (left panel) and correct responses (right panel). Points, size of points, and error bars indicate means, averaged number of individual trials, and standard errors, resp.; solid lines and shaded areas represent model fit means and 95% CrI, resp.

In both groups of participants, response times for correct responses were longer in the visual detection task compared to the familiarity task (Patients: -0.18 [-0.25 , -0.12], $BF = 8000$, Controls: -0.16 [-0.21 , -0.10], $BF = 8000$) and compared to the recollection task (Patients: -0.05 [-0.11 , 0.02], $BF = 6.71$, Controls: -0.04 [-0.10 , 0.02], $BF = 6.71$)(Figure S6).

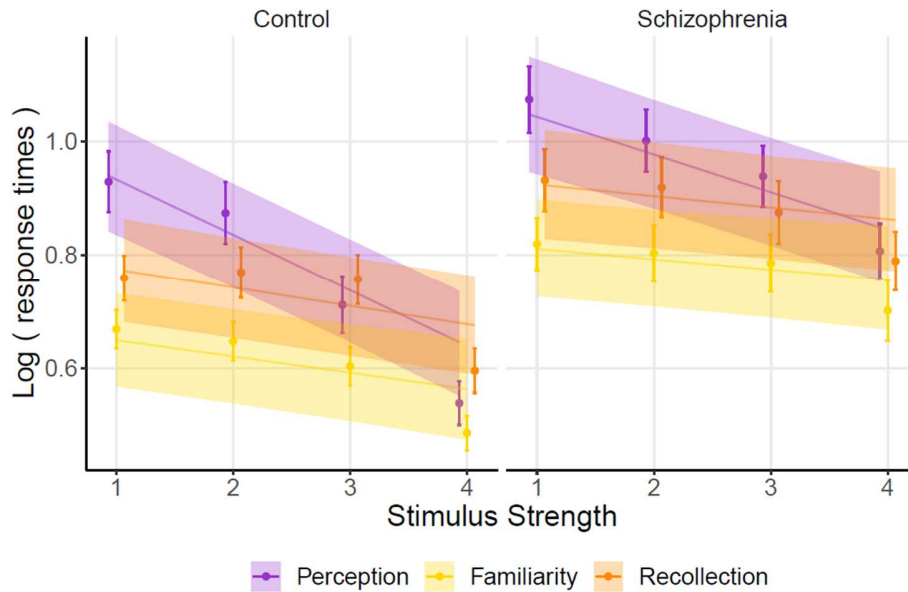


Figure S6: Log-Response times (only for correct responses) as a function of stimulus strength, for each task and group. Points and error bars indicate average accuracy and standard error of the mean, respectively; solid lines and shaded areas represent model fit mean and 95% CrI, respectively.

Domain-generality

Interestingly, contrary to the notion that metacognition obeys domain-general rules, we found no pairwise correlations between indices of metacognitive sensitivity across tasks (Figure S7; Perception - Familiarity: estimate = -0.16, std err. = 0.37, statistic = -0.45, $p = 0.65$, BF = 0.36; Recollection - Familiarity: estimate = 0.08, std err. = 0.11, statistic = 0.69, $p = 0.49$, BF = 0.40; Recollection - Perception: estimate = -0.07, std err. = 0.27, statistic = -0.26, $p = 0.79$, BF = 0.33)).

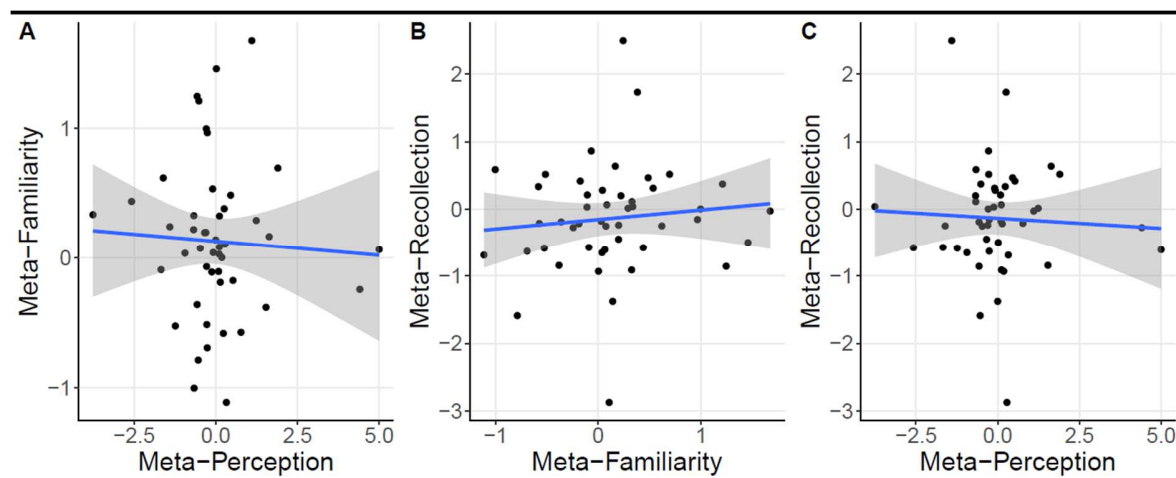


Figure S7: Pairwise task-metaperformance correlations. Dots represent individual regressions slopes (accuracy regressed on confidence). Solid lines and shaded areas represent the correlational fits, and standard errors, resp.

Correlation with clinical scores

Among patients, metacognitive performance tended to be positively correlated with cognitive insight scores (estimate = 0.03, std err. = 0.02, $t = 1.75$, $p = 0.08$, $BF = 1.08$), and was negatively correlated - yet with inconclusive evidence - with PANSS disorganized symptoms (estimate = -0.05, std err. = 0.02, $t = -2.43$, $p < 0.05$, $BF = 1.09$) and with PANSS negative symptoms (estimate = -0.04, std err. = 0.02, $t = -2.23$, $p < 0.05$, $BF = 0.72$). Other clinical scores - PANSS positive symptoms and subjective cognitive functioning (SSTICS) - were not correlated with metacognitive performance (Figure S8).

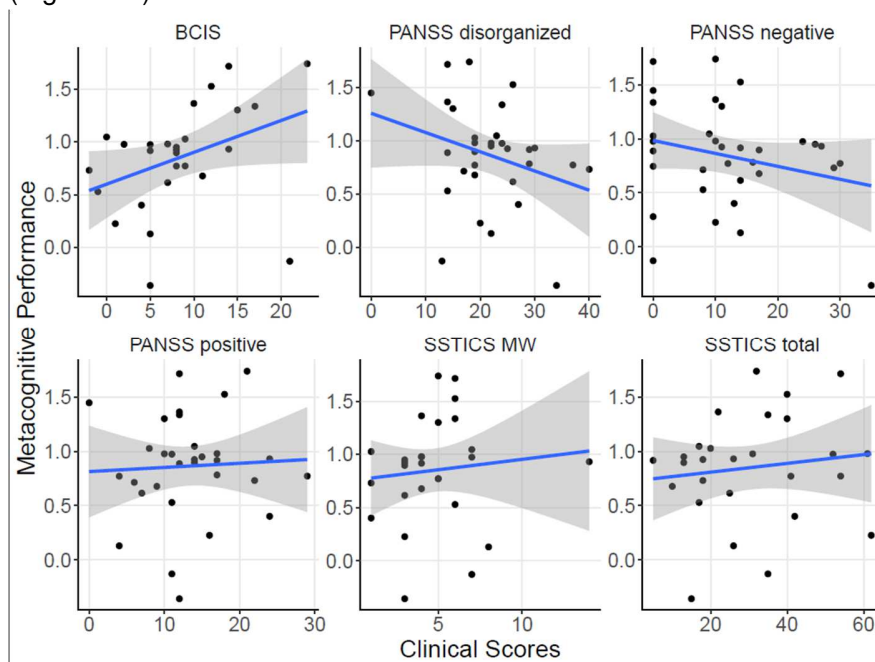


Figure S8: Metacognitive scores regressed on clinical scores among patients. Dots are individual regression slopes averaged across the three tasks. BCIS: Beck Cognitive Insight Scale; PANSS: Positive And Negative Symptoms in Schizophrenia, SSTICS WM: Subjective Scale To Investigate Cognition in Schizophrenia: Working Memory score.

Among demographic and neuropsychological variables, only the WAIS matrix subtest scores were correlated with metacognitive performance (Figure S9, estimate = 0.09, std err. = 0.03, $t = 3.13$, $p < 0.01$, $BF = 415$).

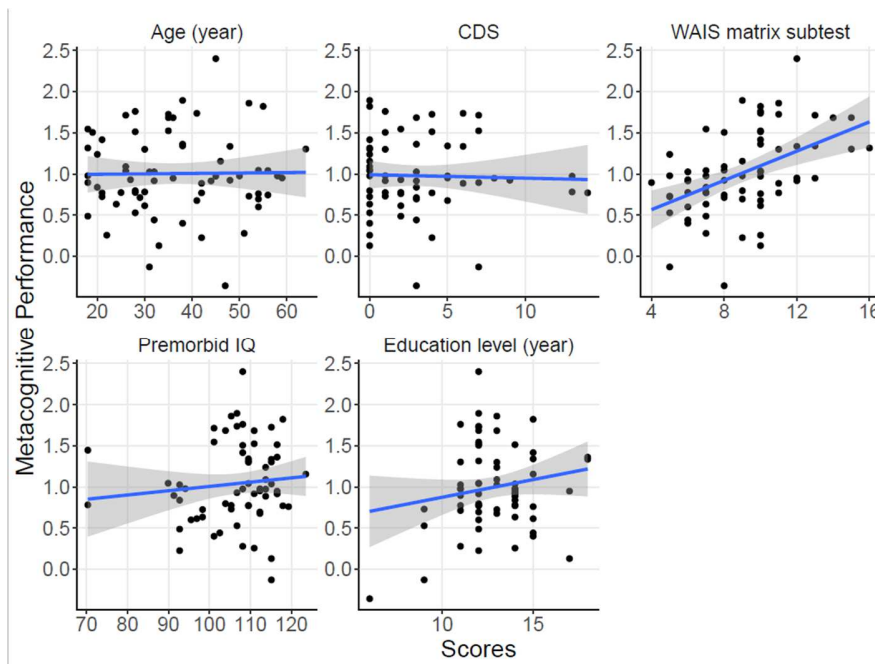


Figure S9: Metacognitive scores regressed on demographic data and neuropsychological scores (all participants). Dots are individual regression slopes averaged across the three tasks.

References

1. Wechsler D. Wechsler Adult Intelligence Scale--Fourth Edition. November 2012. doi:10.1037/t15169-000
2. Mackinnon A, Mulligan R. Estimation de l'intelligence prémorbide chez les francophones. *L'Encéphale*. 2005;31(1):31-43. doi:10.1016/S0013-7006(05)82370-X
3. Faivre N, Roger M, Pereira M, et al. Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *J Psychiatry Neurosci*. 2021;46(1):E65-E73. doi:10.1503/jpn.200022
4. Mazancieux A, Fleming SM, Souchay C, Moulin CJA. Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *J Exp Psychol Gen*. 2020;149(9):1788-1799. doi:10.1037/xge0000746
5. Friston KJ. The disconnection hypothesis. *Schizophr Res*. 1998;30(2):115-125. doi:10.1016/S0920-9964(97)00140-0

3. Supplementary information for project 3 (Electrophysiological markers of confidence)

Supplementary information

Rouy, M., Roger, M., Goueytes, D., Pereira, M., Roux, P., & Faivre, N. (2023). Preserved electrophysiological markers of confidence in schizophrenia spectrum disorder. *Schizophrenia*, 9(1), 12.

1. Relationship between confidence and confidence history

Consistently with Zheng and colleagues (2022), we applied the following model:

$\text{confidence} \sim \text{accuracy} * \text{group} * (\text{RT} + \text{confidence history}) + (\text{accuracy} + \text{RT} + \text{conf history} | \text{subj})$

where confidence history is the confidence averaged over the five trials prior to the current decision.

We found a main effect of RT (Estimate = -0.29, 95%CI [-0.36, -0.23], evidence ratio = 16000), and a main effect of confidence history (Estimate = 0.25, 95%CI [0.18, 0.32], evidence ratio = 16000). We still found an interaction effect between RT and group on confidence level (Estimate = 0.11, 95%CI [0.01, 0.21], evidence ratio = 28.8) indicating that confidence was less correlated with response times among patients compared to control participants. However, there was no interaction between history of confidence and group on confidence (Estimate = 0.04, 95%CI [-0.09, 0.18], $BF_{01} = 11.8$) indicating that the result obtained by Zheng and colleagues did not extend to our perceptual task (Figure S1).

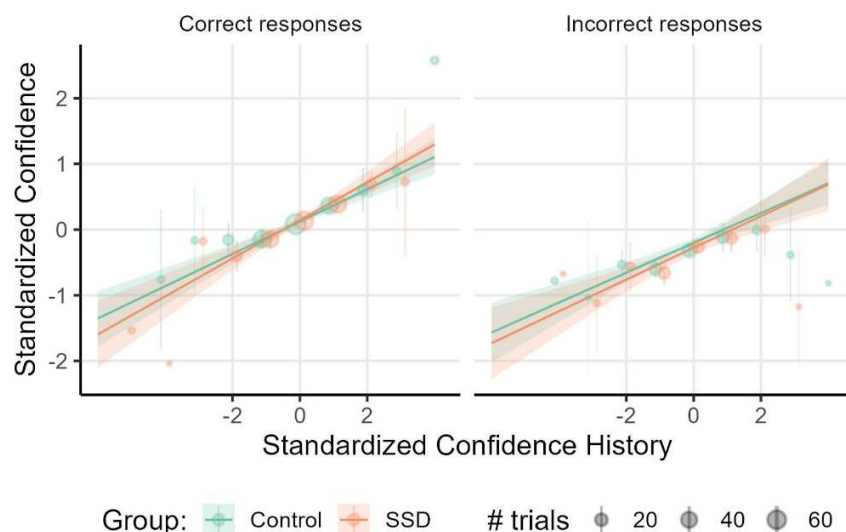


Figure S1: Standardized confidence as a function of standardized confidence history. Dots represent averaged data and lines are regression fits. Control participants are depicted in green, and patients with SSD in red. Error bars represent 95%CI.

2. Time-Frequency analysis

Time-frequency analyses were conducted with the EEGLAB toolbox (v2021.0, EEGLAB, Delorme and Makeig 2004). We used a wavelet decomposition of 165 linearly spaced complex-valued Morlet wavelets ranging from 4 Hz (3 cycles) to 45 Hz (16.875 cycles). For every trial, the EEG signal between -500 ms and 1000 ms after the movement onset was convolved with each Morlet wavelet. We then compared the average magnitude of each condition in the log-domain.

Analysis of confidence for correct trials

Below, we show the time-frequency representation of the confidence contrast between high versus low confidence in correct responses (where high and low categories are determined by a median split of confidence for each participant) for control participants and patients with SSD. We then conducted t-tests to compare the power of each frequency at each time sample between the two groups, while applying False-discovery rate (fdr-) correction for multiple comparisons. No effect of group resisted this correction with a corrected alpha level of 0.05.

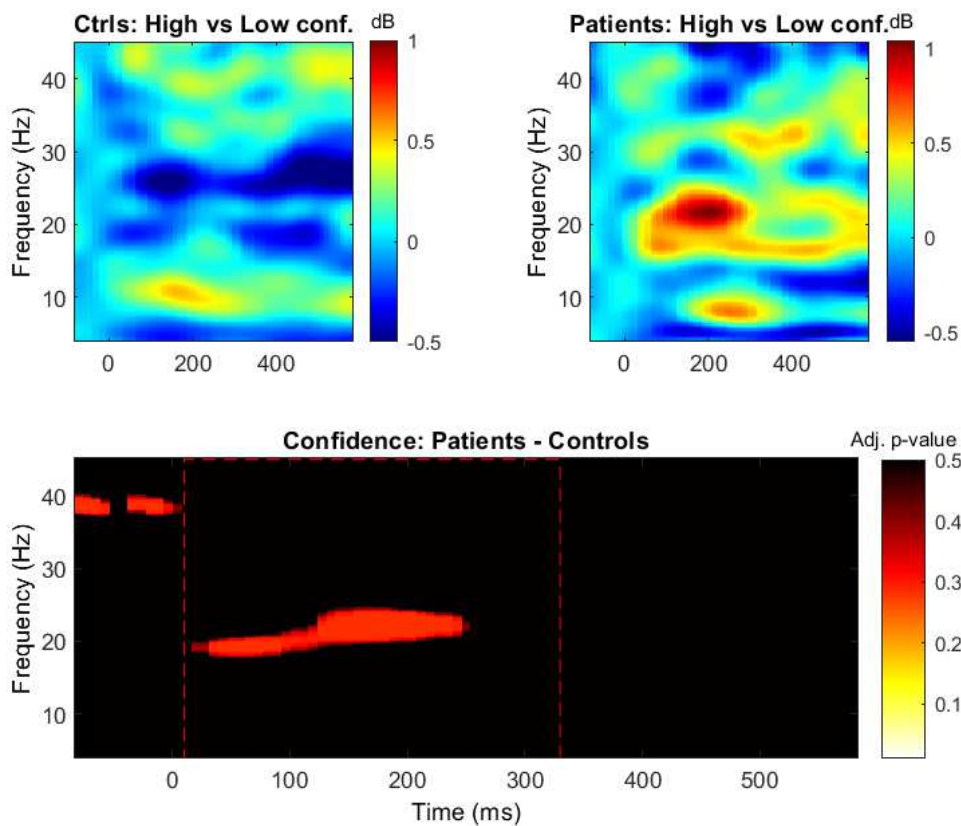


Figure S2: Time-frequency representations. Confidence contrasts (high versus low) for controls (upper left panel) and patients (upper right panel). Diagram of p-values adjusted for multiple comparisons (fdr-correction) between controls and patients (bottom panel). All adjusted p-values are > 0.05 . Dashed red lines delimitate the time window where a significant main effect of confidence on EEG amplitude was found in the cluster analysis reported in the manuscript.

Analysis of correctness

We conducted the same time-frequency analysis for the contrast between correct and incorrect responses for control participants and patients with SSD but again found no significant differences after correcting for multiple comparisons.

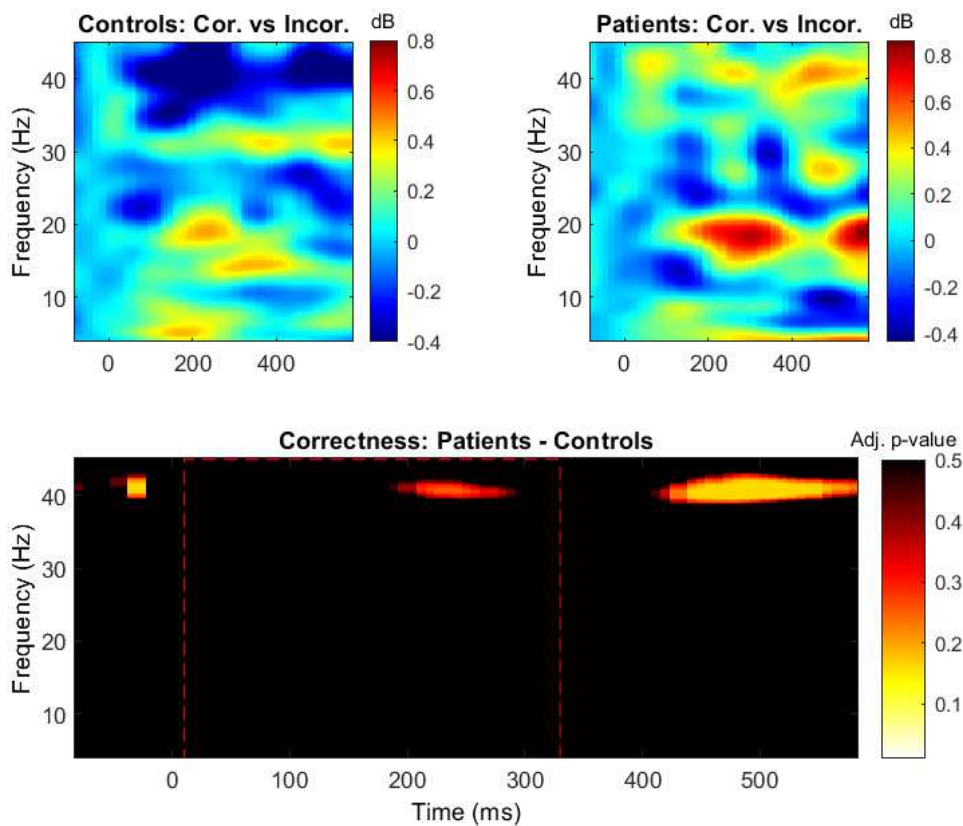


Figure S3: Time-frequency representations. Accuracy contrasts (correct versus incorrect) for controls (upper left panel) and patients (upper right panel). Diagram of adjusted p-values adjusted for multiple comparisons (fdr-correction) between controls and patients (bottom panel). All adjusted p-values are > 0.05 . Dashed red lines delimitate the time window where a significant main effect of correctness was found in the cluster analysis reported in the manuscript.

4. The Dunning-Kruger effect is not a statistical artifact

Krueger and Muller (2002) proposed an alternative interpretation to the Dunning-Kruger effect: the same pattern of results is obtained from the combination of a regression to the mean effect (see Box 3 for an illustration), together with a better-than-average effect.

To arbitrate between the two explanations (a true deficit versus a statistical artifact), Gignac and Zajenkowski (2020) proposed to assess the data in terms of heteroscedasticity and non-linearity.

Nonlinearity: If the Dunning-Kruger effect depends on task-performance, it involves that the magnitude of the correlation between self-assessed ability and objectively measured ability increases with task-performance. Therefore, we should predict a non-linear relationship between task performance and self-assessments, see Figure A1 for an simulated illustration of the expected pattern).

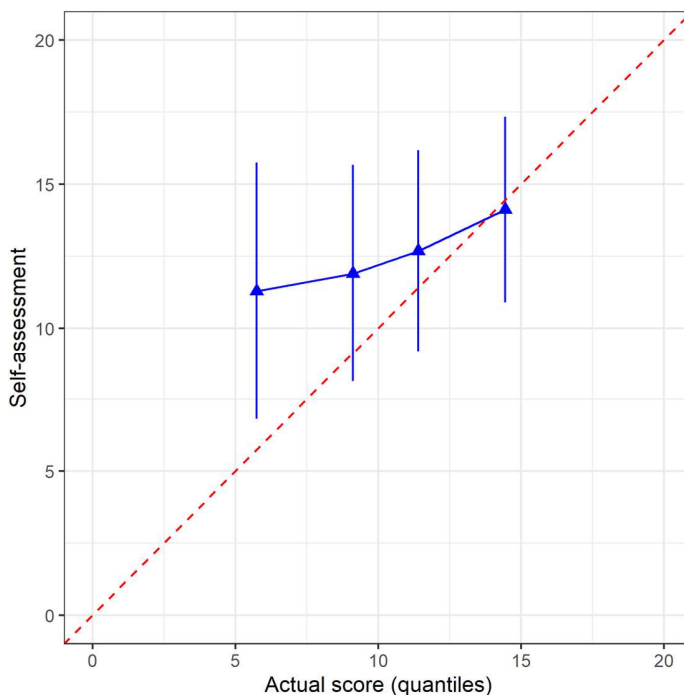


Figure A1. Simulation of self-assessments with a nonlinear relationship with actual scores. If the Dunning-Kruger effect depends on task-performance, the gain in accuracy of self-assessments is expected to be non-linear.

Heteroscedasticity: If the Dunning-Kruger effect really depends on task-performance, individuals with lower task-performance should have more difficulty to judge their ability, which means that the dispersion of the regression residuals (i.e. degree of misprediction) should decrease as a function of task-performance. It leads to the prediction that the residual variance of the regression should be significantly heteroscedastic (see Figure A2 for an simulated illustration).

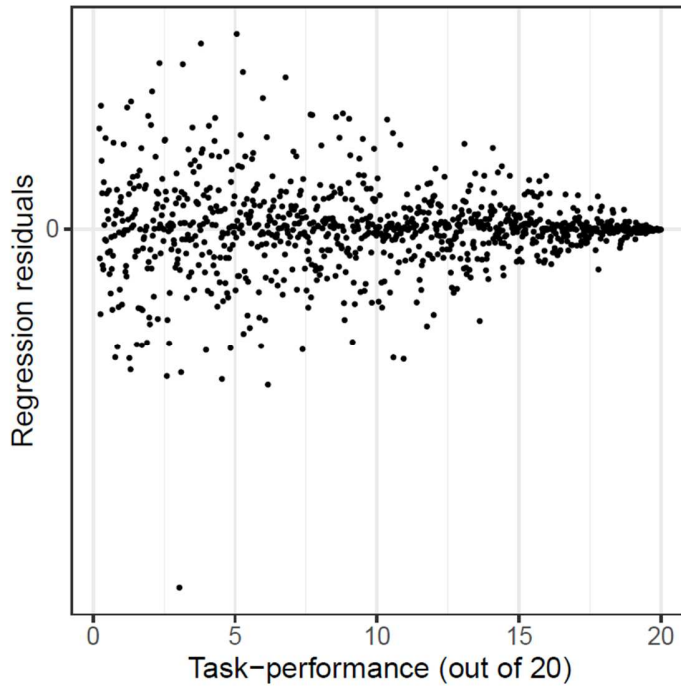


Figure A2. Expected heteroscedastic residuals (i.e. decreasing dispersion as a function of task-performance) in case the Dunning Kruger effect depends on task-performance.

Gignac and Zajenkowski (2020) tested these two hypotheses on a large sample (N = 929). Participants' objective intelligence was assessed with the Advanced Progressive Matrices test (APM; Raven and Raven 2003) and they were then asked to provide a self-assessment of intelligence on a scale ranging from 1 to 25 (Figure A3 provided in the supplementary information)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Very low					Low					Average					High					Very high				

Figure A3. The measure of self-estimated intelligence (SEI) used in the study from Gignac and Zajenkowski (2020)

The authors have shown that the residuals were not heteroscedastic, and that the relationship between objective scores of intelligence and self-assessments of intelligence were not nonlinear, hence congruently converging toward the conclusion that the Dunning-Kruger effect was a statistical artifact.

However, given the operationalization proposed by Gignac and collaborators, I reasoned that it was not clear whether the participants would rely on their performance on the objective test of intelligence to provide their self-assessment of intelligence. To gain in validity I applied the same statistical tests on the large dataset made available online by Jansen et al. (2021). In this study, participants performed a grammar task (20 grammar questions, close to the original questions from Dunning and Kruger, N = 3515) and a logical reasoning task (20 logical questions, close to the original questions from Dunning and Kruger, N = 3543) (see the original article for

details about the tasks), and were then asked to estimate how many responses were correct, out of 20, for each task.

Methods:

Regarding the nonlinearity hypothesis, I compared the following models for each task:

a linear model: $\text{self-assessment} \sim \text{score}$

a nonlinear model: $\text{self-assessment} \sim \text{poly}(\text{score}, 2)$

Regarding the heteroscedasticity hypothesis, I applied the White Test, which is convenient for large datasets and nonlinear data (Babashova 2020)

Results:

Model comparisons revealed a better nonlinear fit for both tasks (Grammar: $F(1, 3512) = 34.25$, $p < 0.001$; Logical reasoning: $F(1, 3540) = 56.87$, $p < 0.001$)(Figure A4).

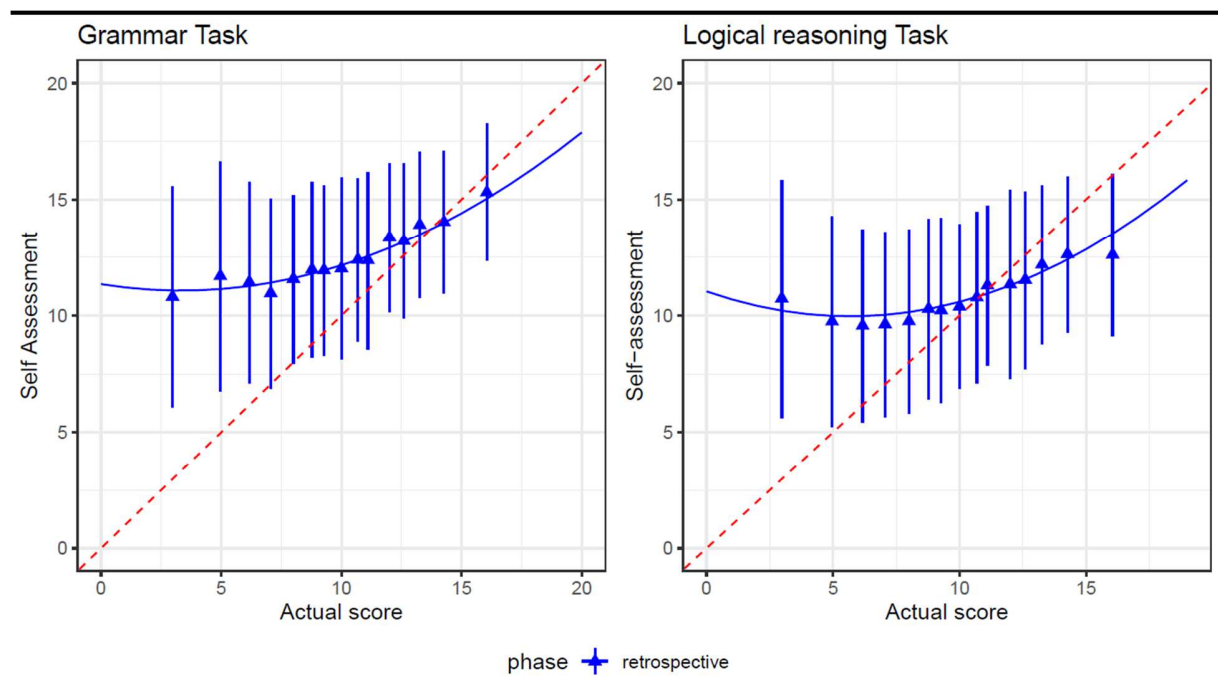


Figure A4: Self-assessed scores regressed on actual scores with a non linear regression model (solid blue curves). Actual scores were binned into deciles. Triangles indicate average scores and error bars indicate standard deviations.

Residuals from the previous nonlinear models were found to be heteroscedastic in both tasks (grammar: White Test statistic = 165, $p < 0.001$; logical reasoning: White Test statistic = 134, $p < 0.001$)(Figure A5)

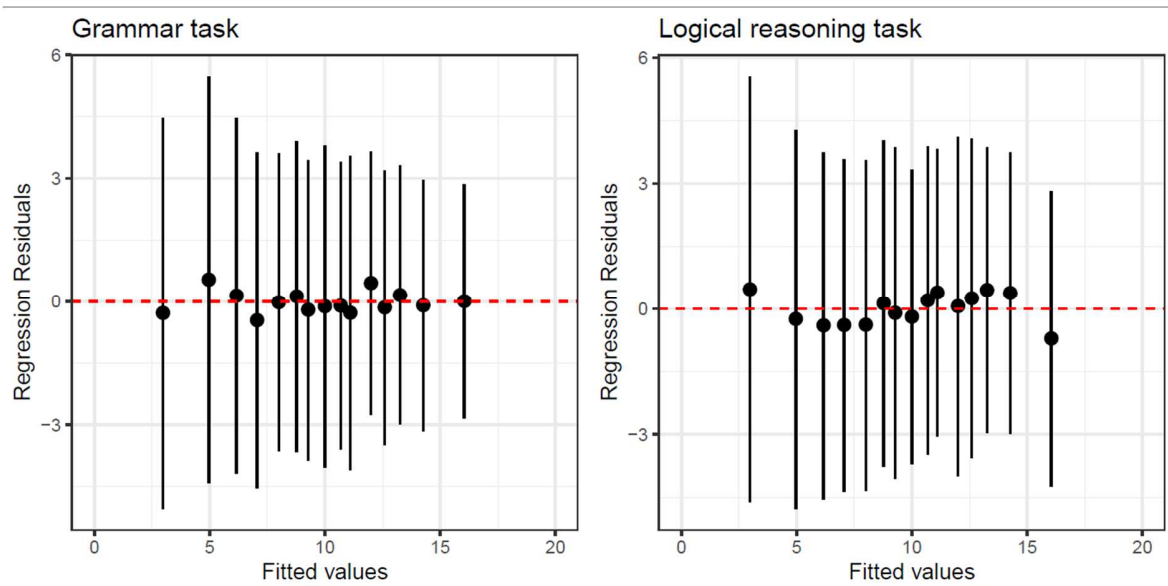


Figure A5: Residuals of the nonlinear models plotted against the fitted values. Points indicate average residuals for deciles of fitted values, and error bars indicate standard deviations.

Conclusion:

In order to determine whether the Dunning Kruger effect (1999) genuinely depended on task-performance rather than resulted from statistical artifacts (regression to the mean and better than average effect), I applied the statistical tests of nonlinearity and heteroscedasticity (recommended by Gignac and Zajenkowski 2020) on a large dataset ($N \sim 3500$) where self-assessments of performance were explicitly related to the tasks. Upon this operationalization, contrary to Gignac et al who demonstrated that the Dunning Kruger effect was a statistical artifact, I reached the opposite conclusion: the Dunning Kruger effect depends on task-performance.