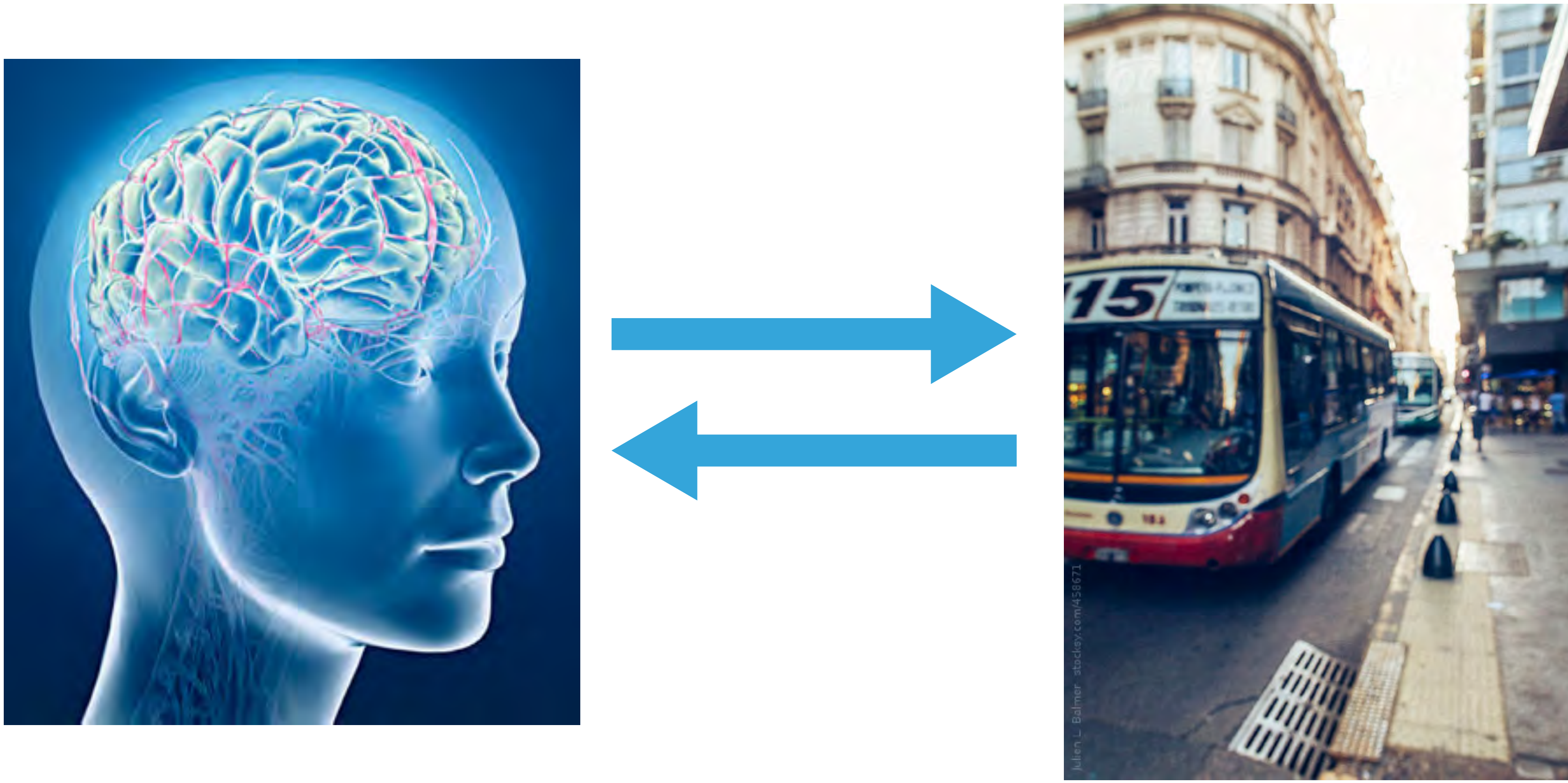
Why metacognition matters for (computational) psychiatry

Steve Fleming
Wellcome Centre for Human Neuroimaging, UCL
stephen.fleming@ucl.ac.uk

metacoglab.org

MAX PLANCK UCL CENTRE
for Computational Psychiatry and Ageing Research

Learning models of the world



WE KNOW AN INCREASING AMOUNT ABOUT HOW WE LEARN AND DECIDE ABOUT THE EXTERNAL WORLD (STATES, REWARDS)

Learning models of ourselves

DID I MAKE A MISTAKE?
IS MY MEMORY ACCURATE?
AM I GIVING A GOOD TALK?

WE KNOW LESS ABOUT HOW PEOPLE FORM BELIEFS ABOUT THEMSELVES, AND HOW SELF-KNOWLEDGE GUIDES BEHAVIOUR

Defining metacognition

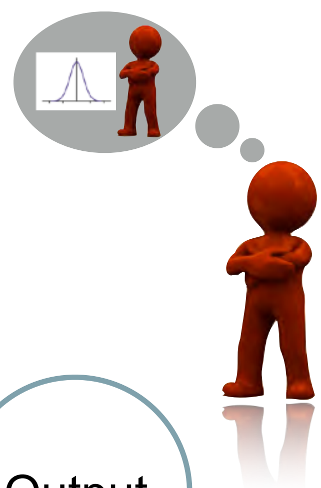
- "cognition about cognitive phenomena..." **(Flavell, 1979)**

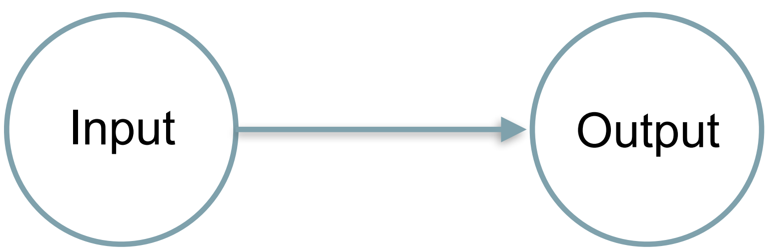
Self-reflection

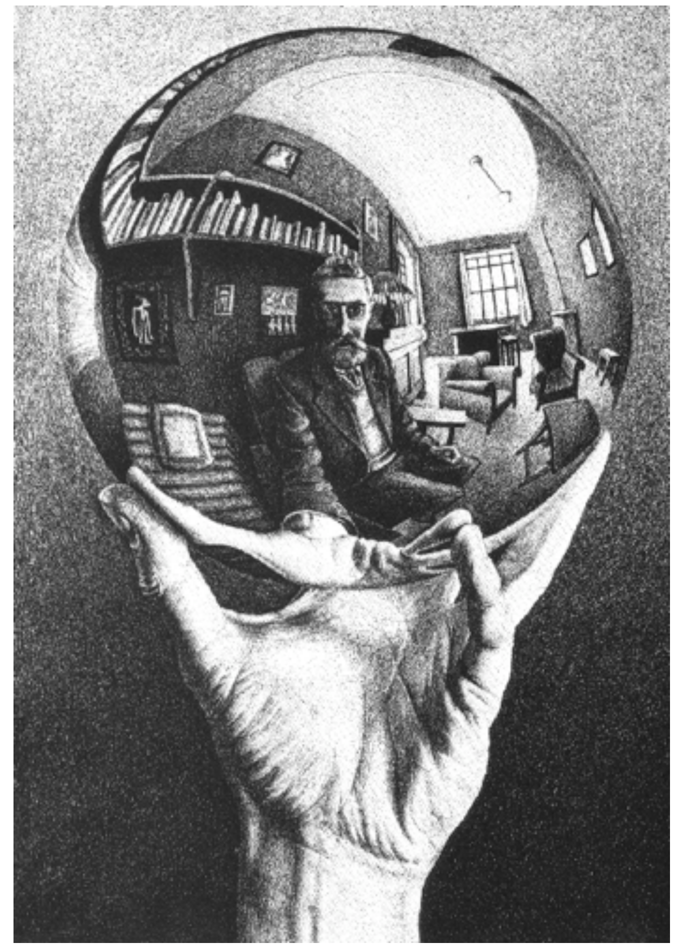
Recursive thought

Introspection

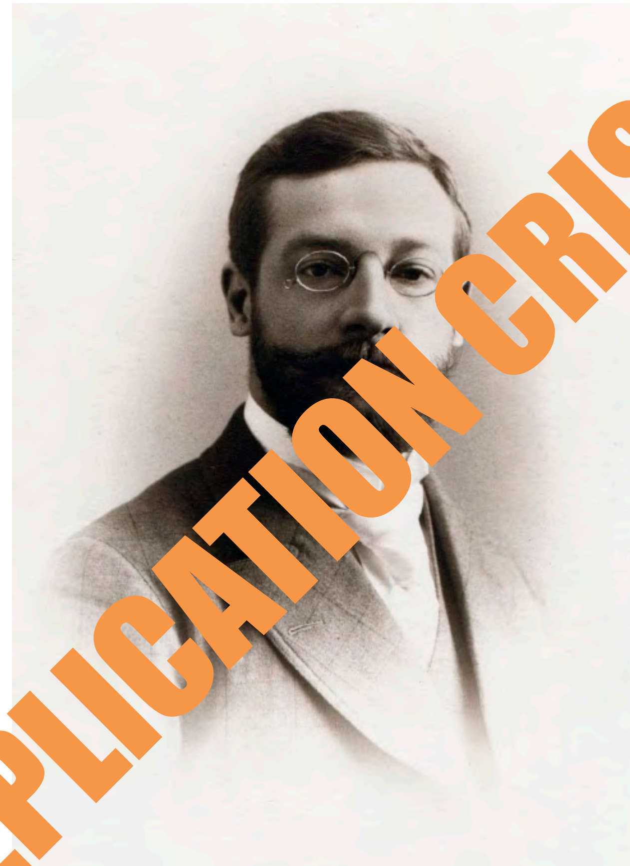
etc…



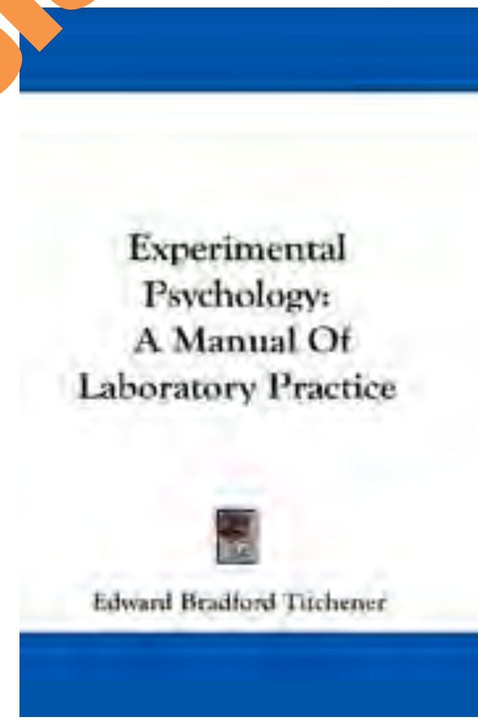




Wilhelm Wundt



Edward B. Titchener



REPLICATION CRISIS!

Introspection, self-report

Behaviorism / psychophysics

Ericsson & Simon (1984)

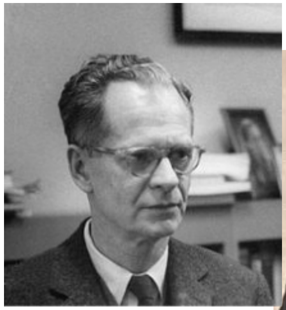
Metacognition (1980s-)

"skeptical introspectionists"

Nisbett & Wilson (1977)

W. Wundt

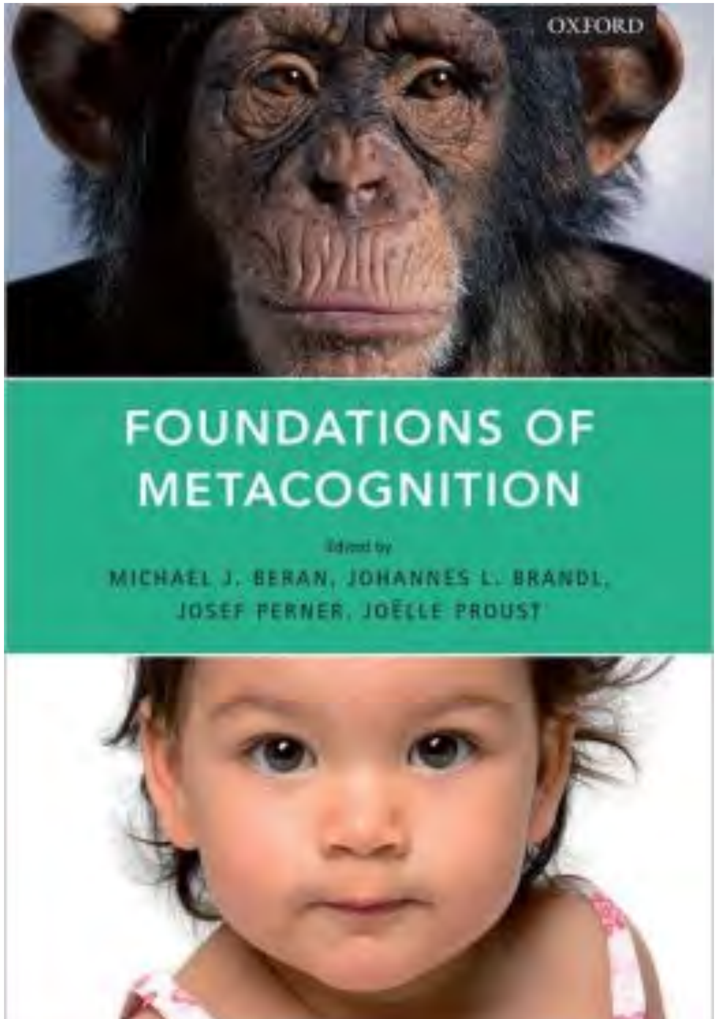
E. Titchener



B. F. Skinner

G. Fechner

A primer on measuring metacognition



BEHAVIOUR ↔ **SECOND-ORDER REPORT**

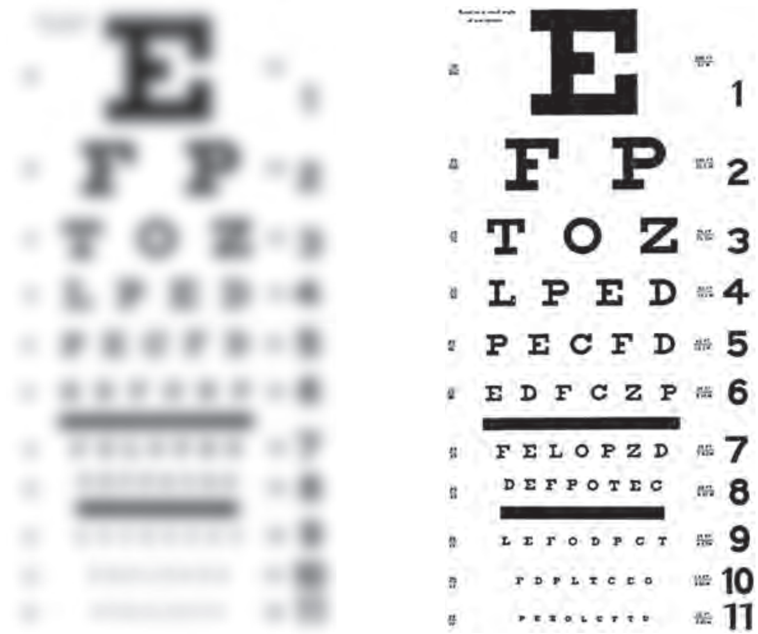
E.g. answer to exam question; response in a psychophysics experiment

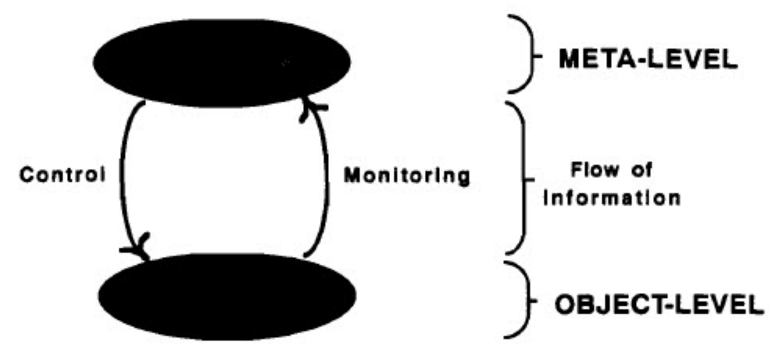
E.g. **confidence** in getting the answer right

Confidence as core variable of interest for metacognition

How does metacognition work?
= hard question…

Which processes support the formation of confidence in perception/action/cognition?
= easier question





Nelson & Narens (1990)

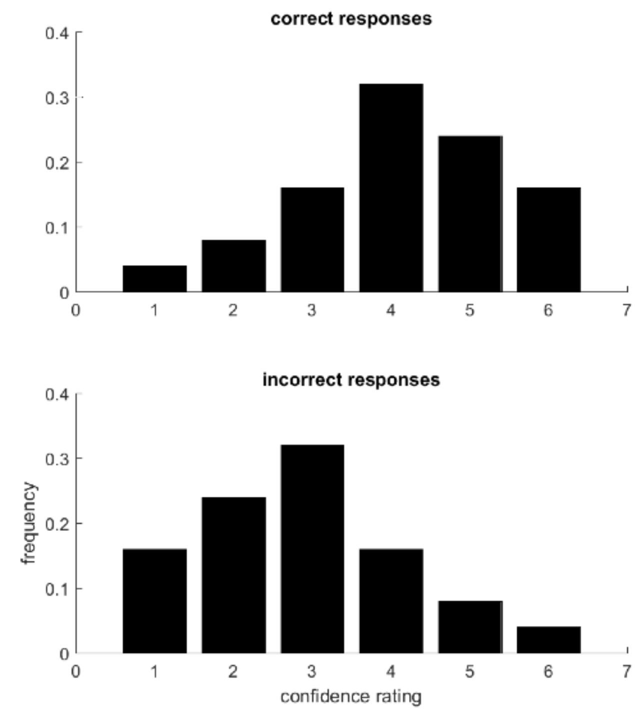
Quantifying metacognition - type 2 ROCs

Two Types of ROC Curves and Definitions of Parameters*

F. R. Clarke, T. G. Birdsall, and W. P. Tanner, Jr.
Electronic Defense Group, University of Michigan, Ann Arbor, Michigan
(Received February 26, 1959)

Type 2 receiver operating characteristic curves are a compact representation of the *quality* of confidence ratings

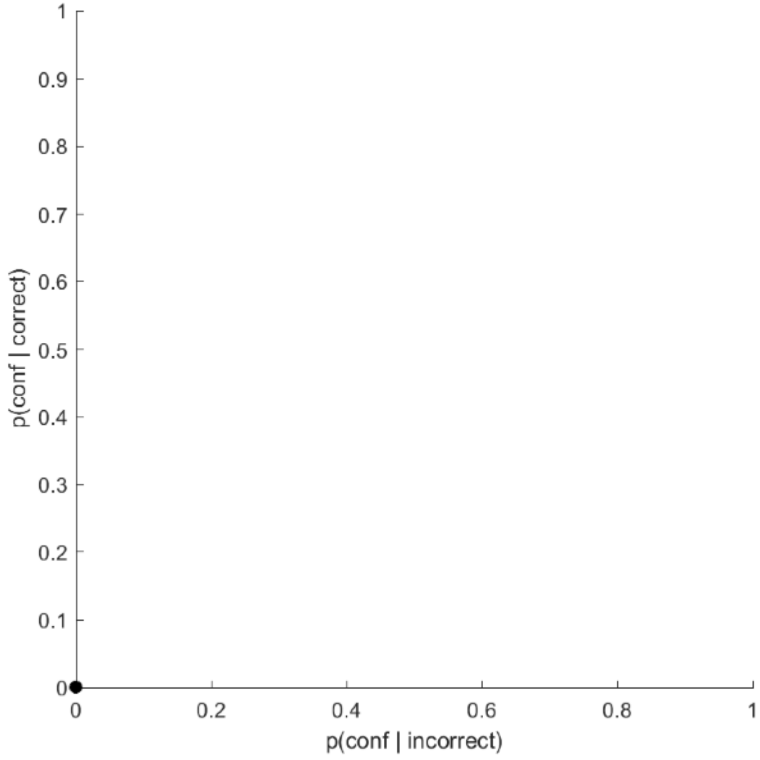
In general, the more different the confidence distributions for correct and for incorrect responses are, the more insight one has into the quality of individual decisions



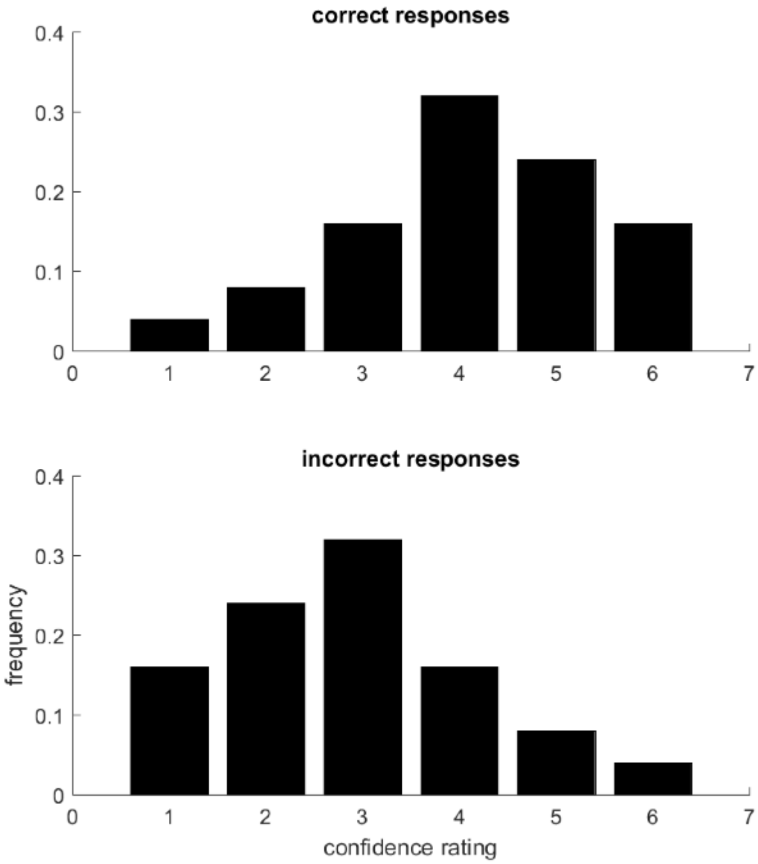
with thanks to Matan Mazor

Clarke, Birdsall & Tanner (1959) *J Acoust Soc Am*

Type 2 ROCs

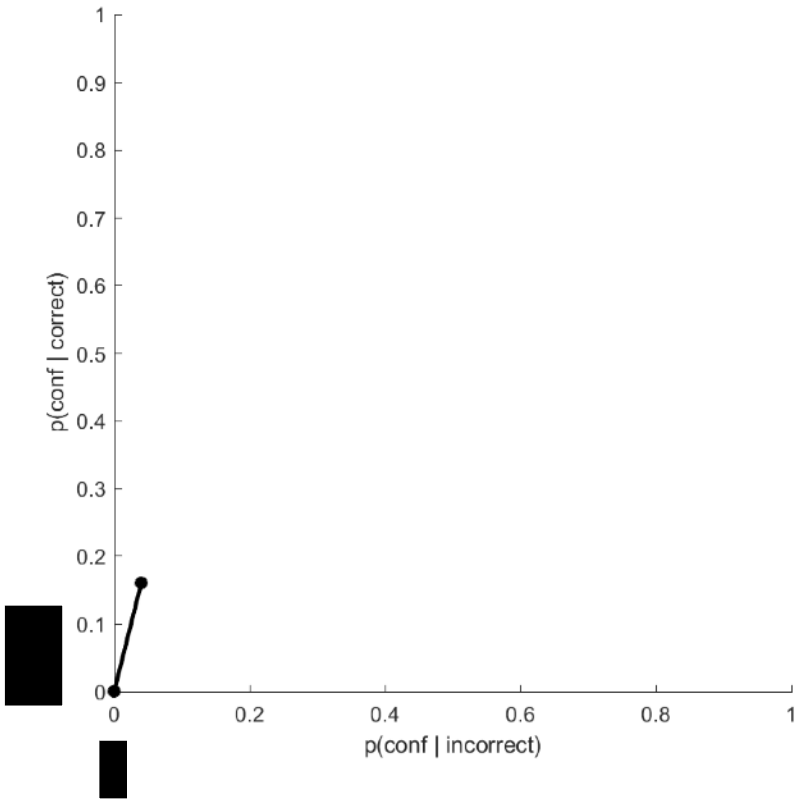


p(conf | correct)
p(conf | incorrect)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8

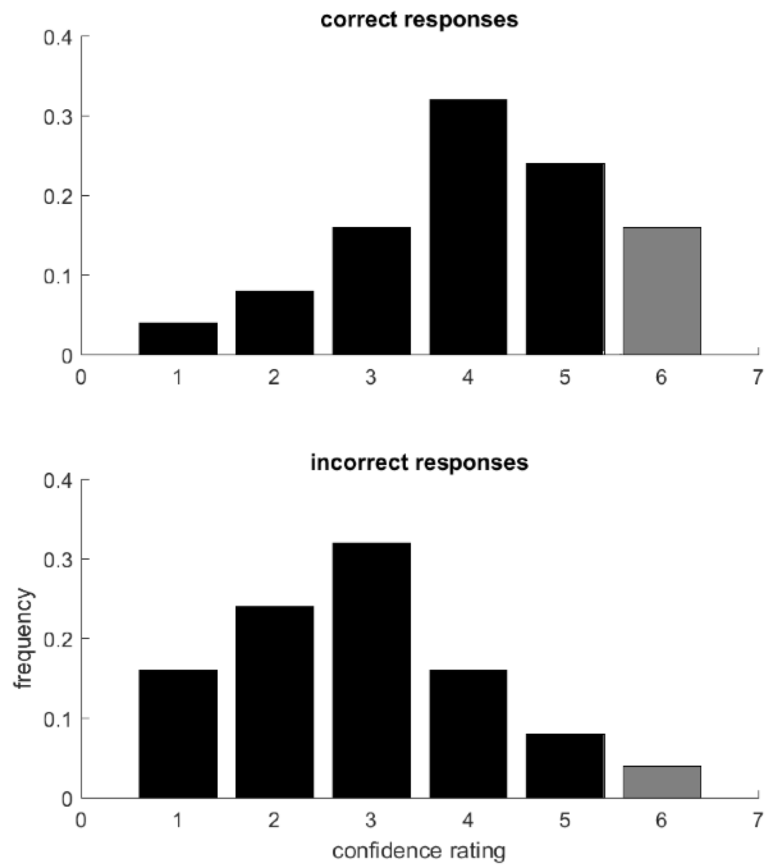


correct responses
incorrect responses
frequency
confidence rating
0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7

Type 2 ROCs

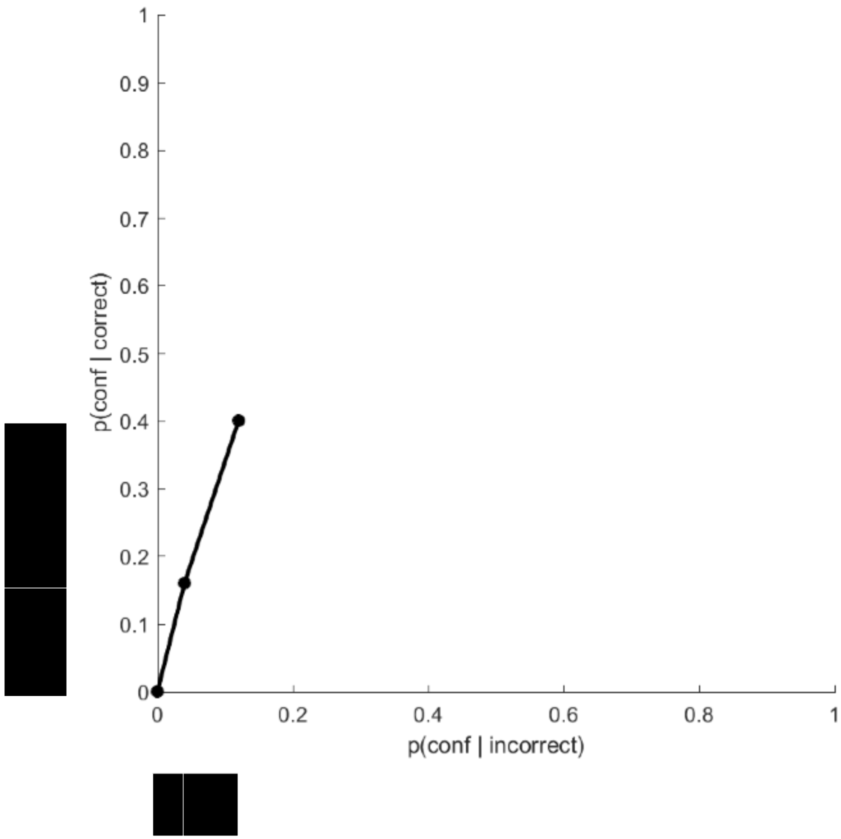


p(conf | correct)
p(conf | incorrect)

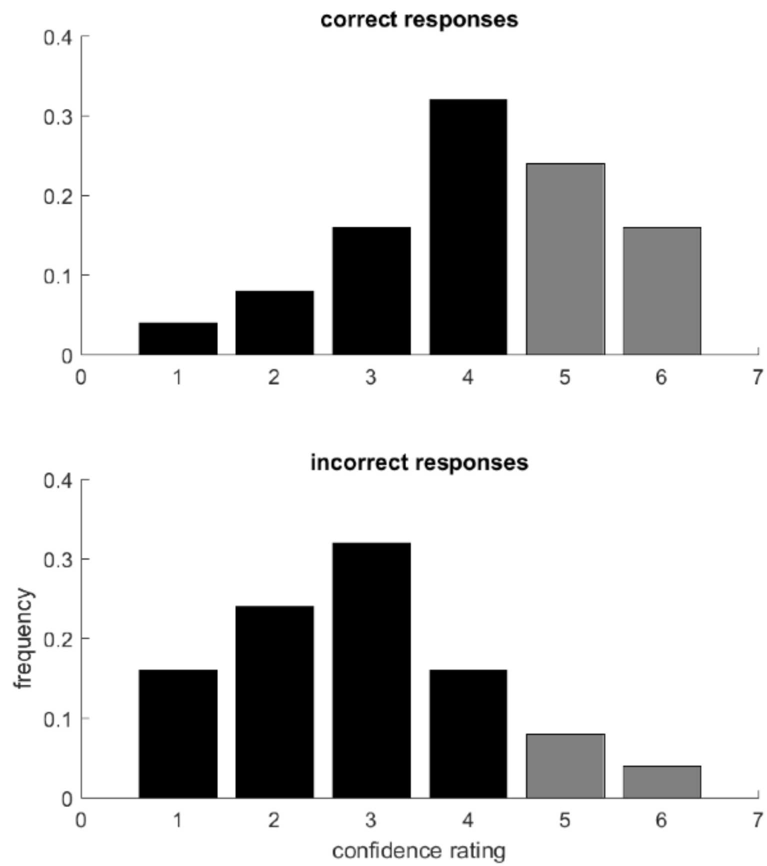


correct responses
incorrect responses
frequency
confidence rating

Type 2 ROCs

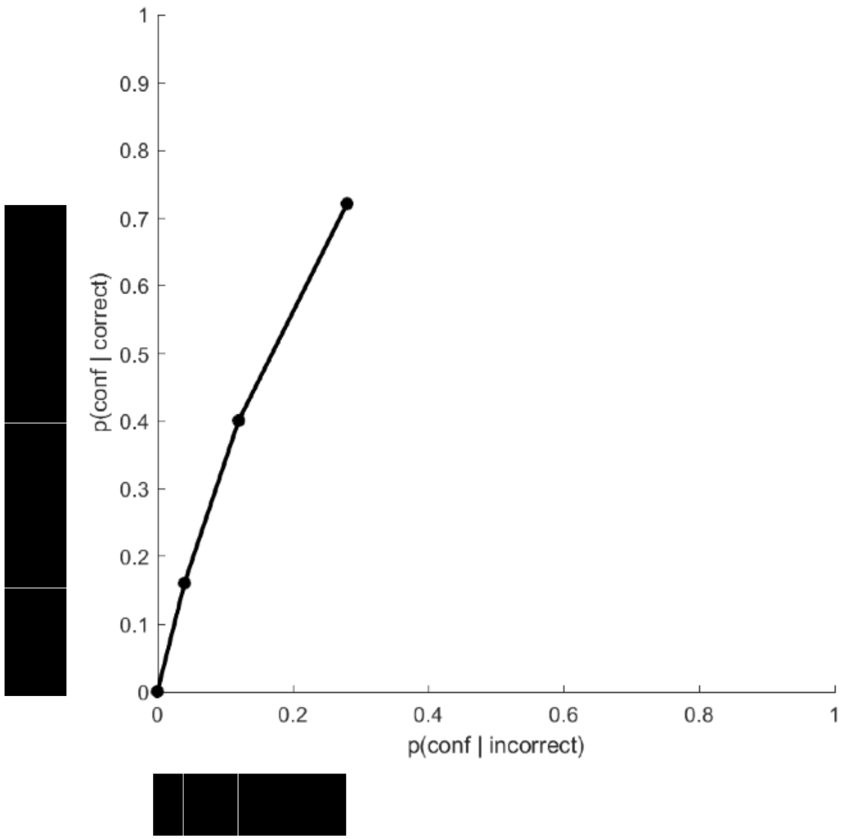


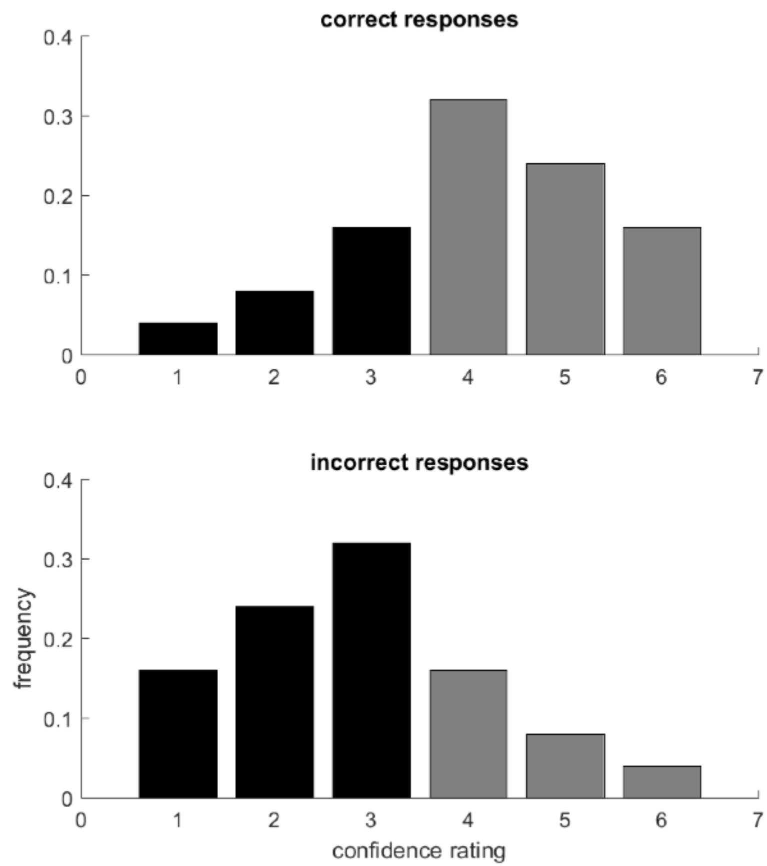
p(conf | correct)
p(conf | incorrect)



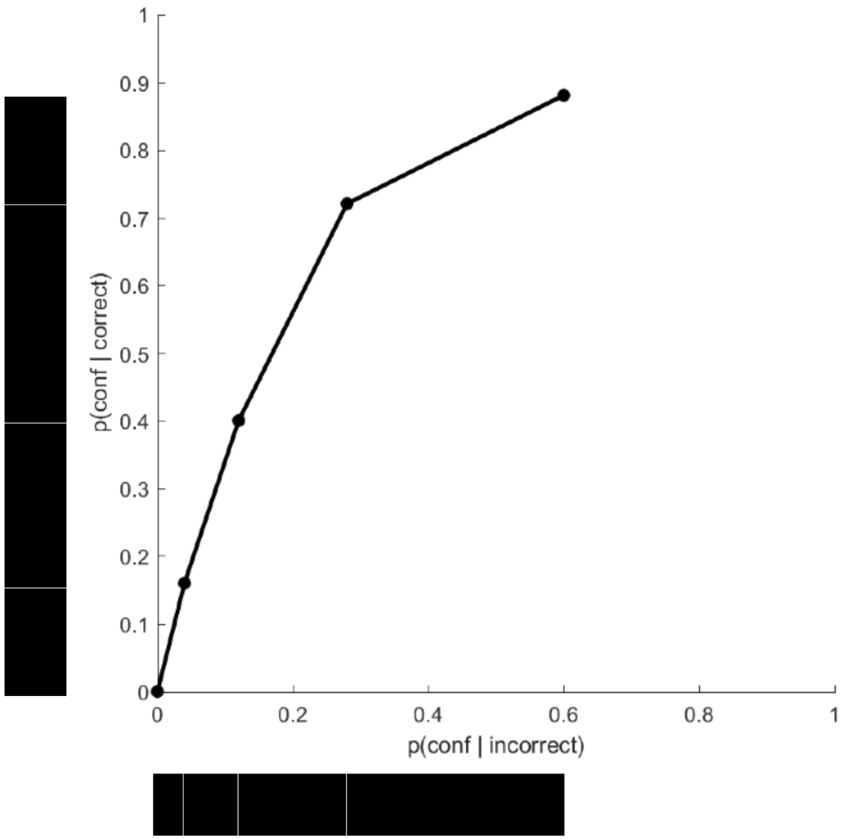
correct responses
incorrect responses
frequency
confidence rating

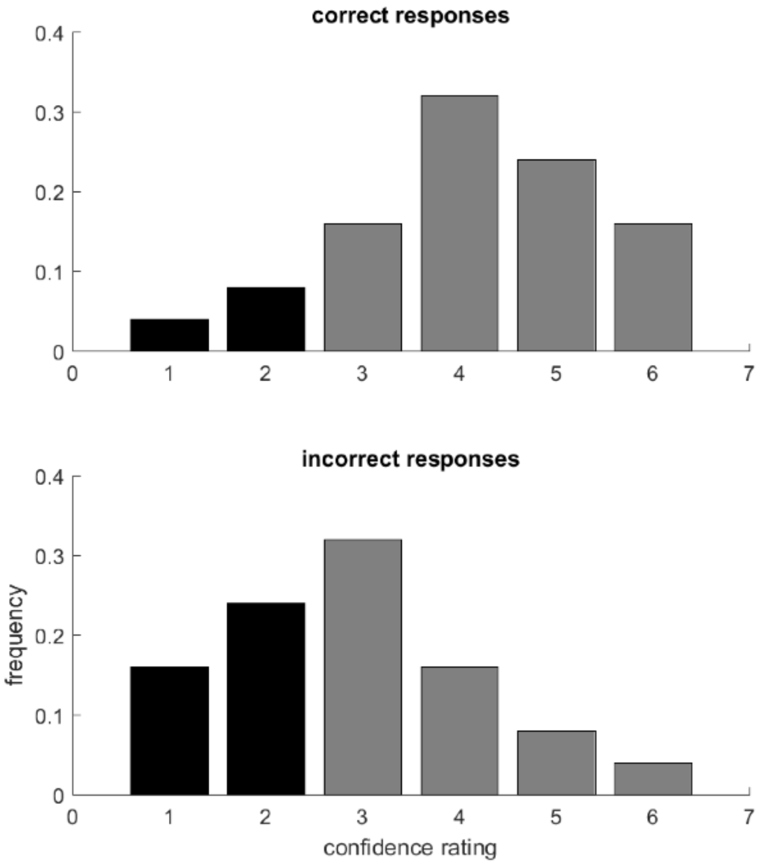
Type 2 ROCs


p(conf | correct)
p(conf | incorrect)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8

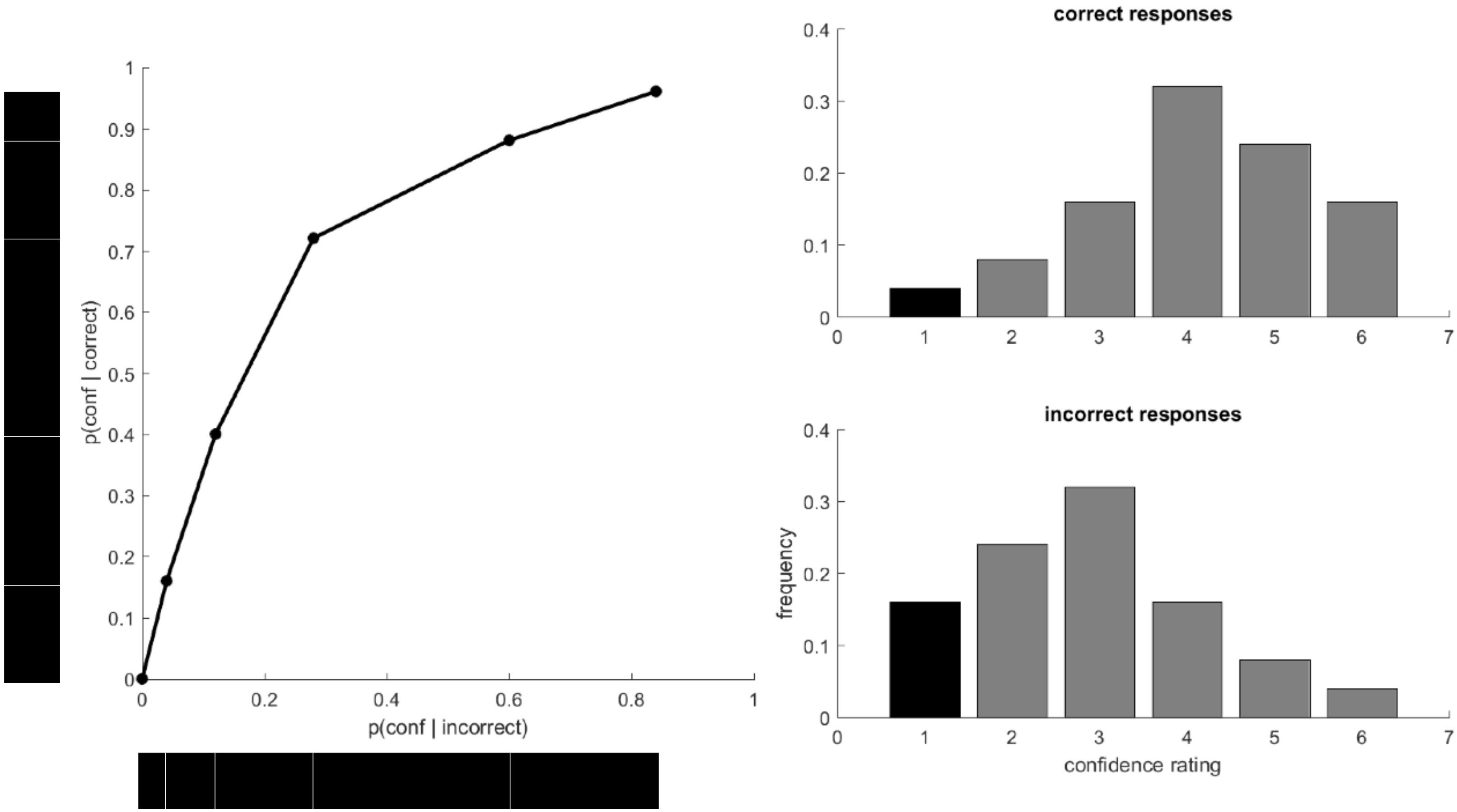

correct responses
incorrect responses
frequency
confidence rating
0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7

Type 2 ROCs

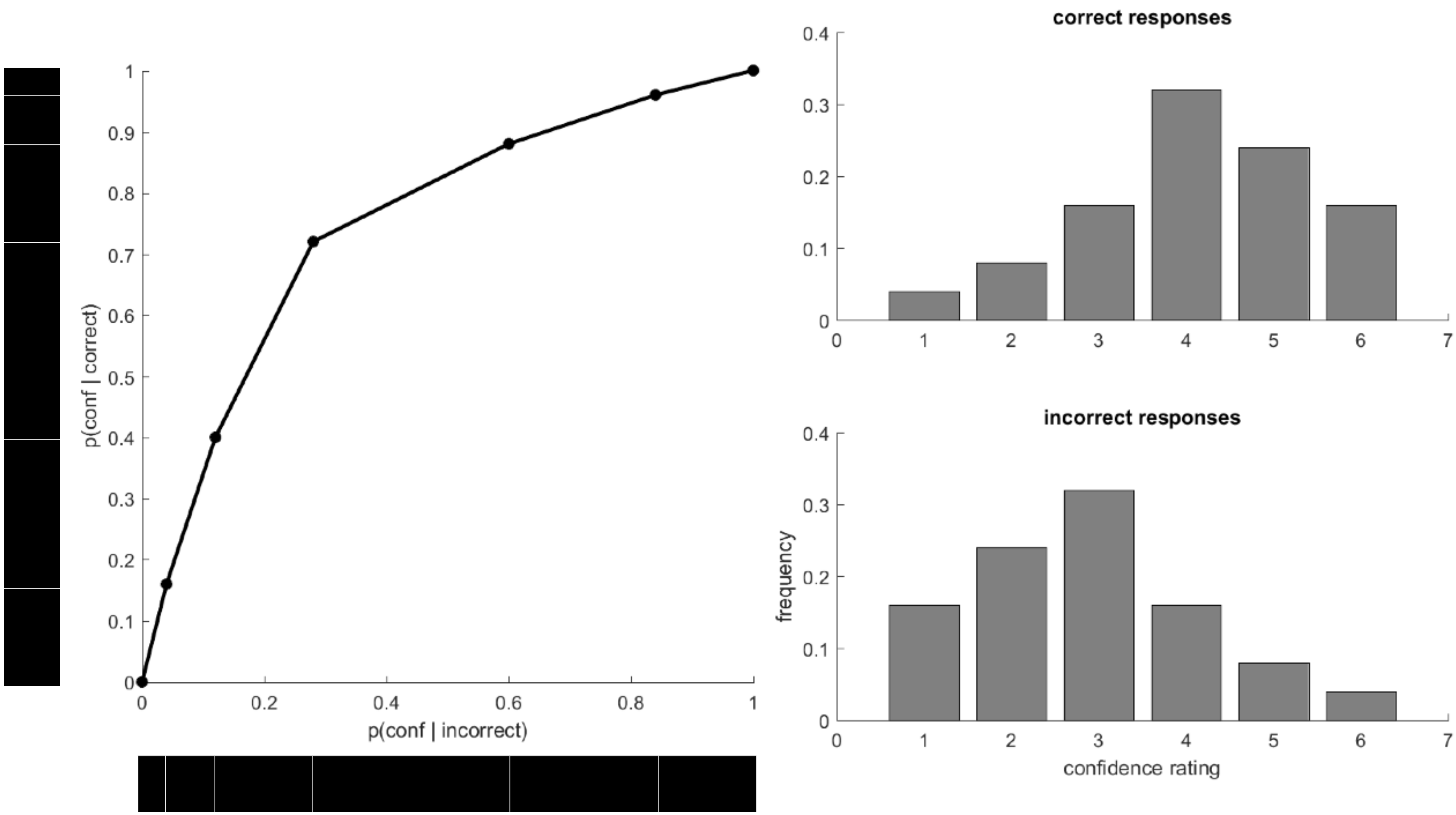

p(conf | correct)
p(conf | incorrect)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


correct responses
incorrect responses
frequency
confidence rating
0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7

Type 2 ROCs

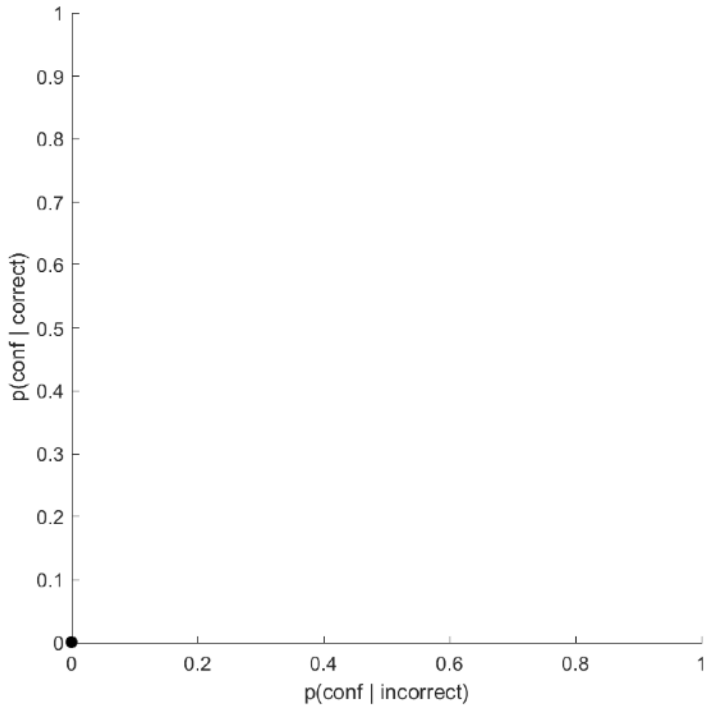

p(conf | correct)
p(conf | incorrect)
correct responses
incorrect responses
frequency
confidence rating

Type 2 ROCs

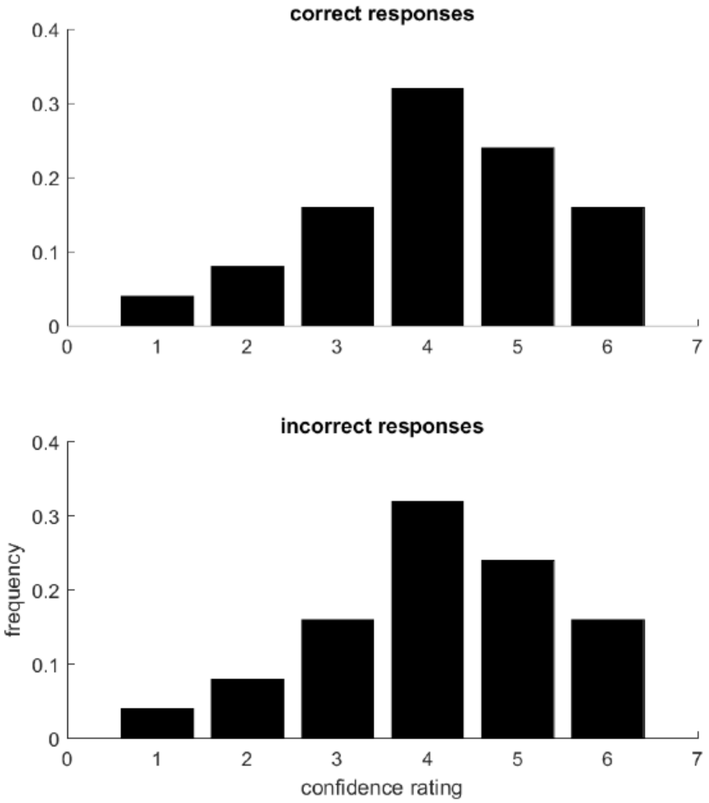


p(conf | correct)
p(conf | incorrect)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
correct responses
incorrect responses
frequency
confidence rating
0
1
2
3
4
5
6
7

Type 2 ROCs

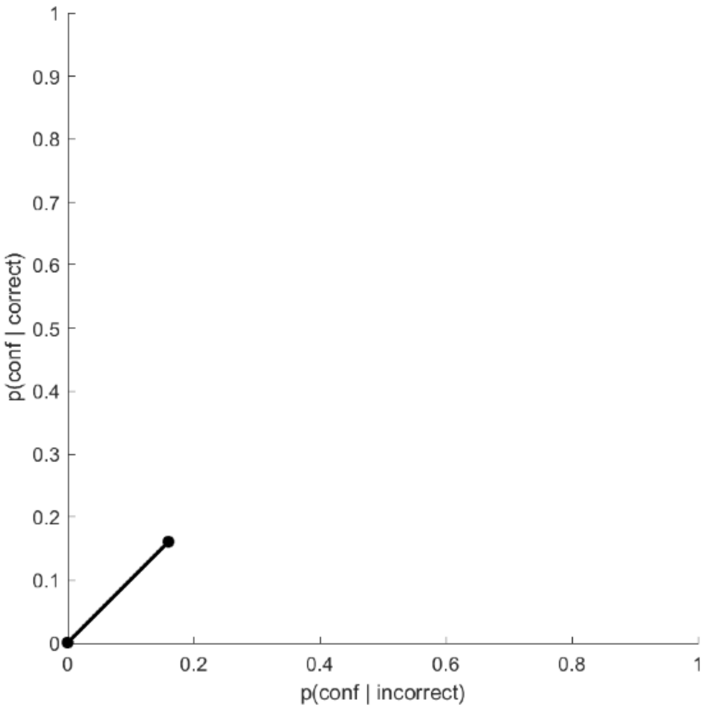


No metacognition:



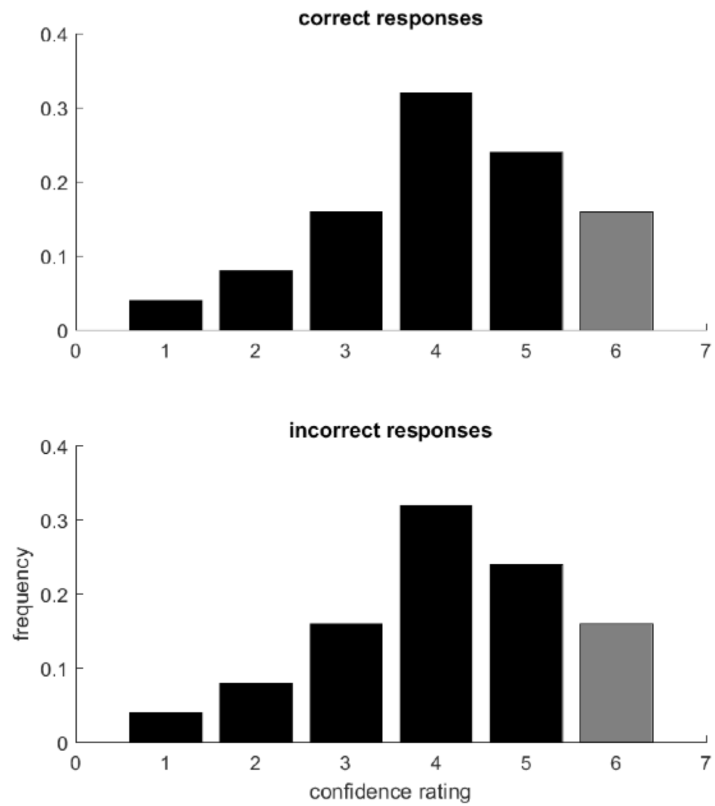
Type 2 ROCs

No metacognition:

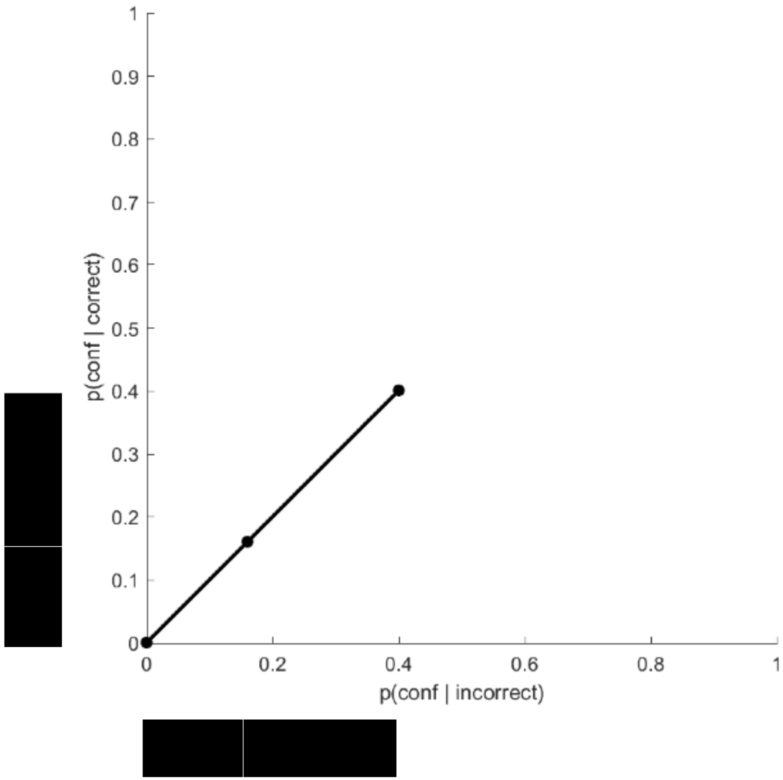


[redacted]

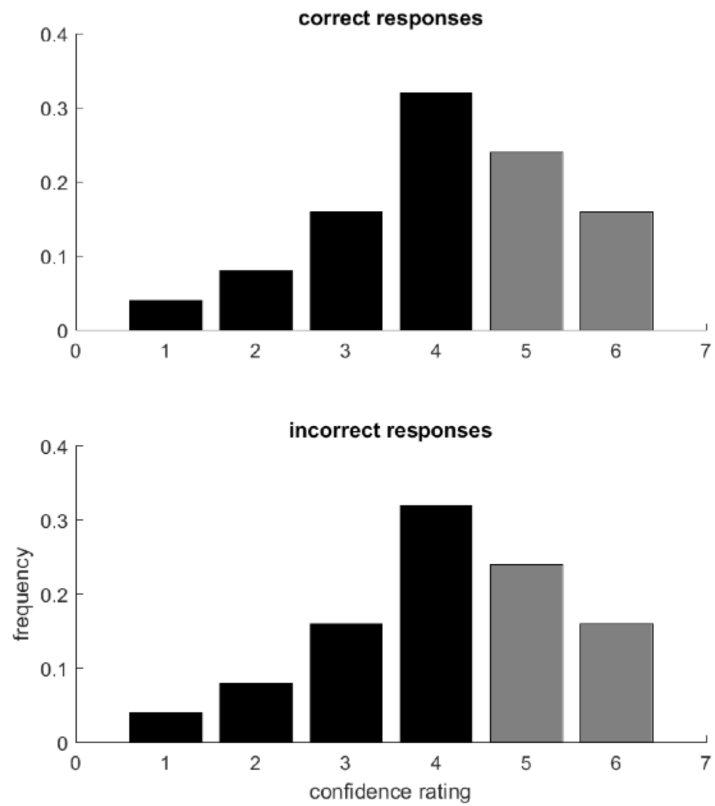
[redacted]



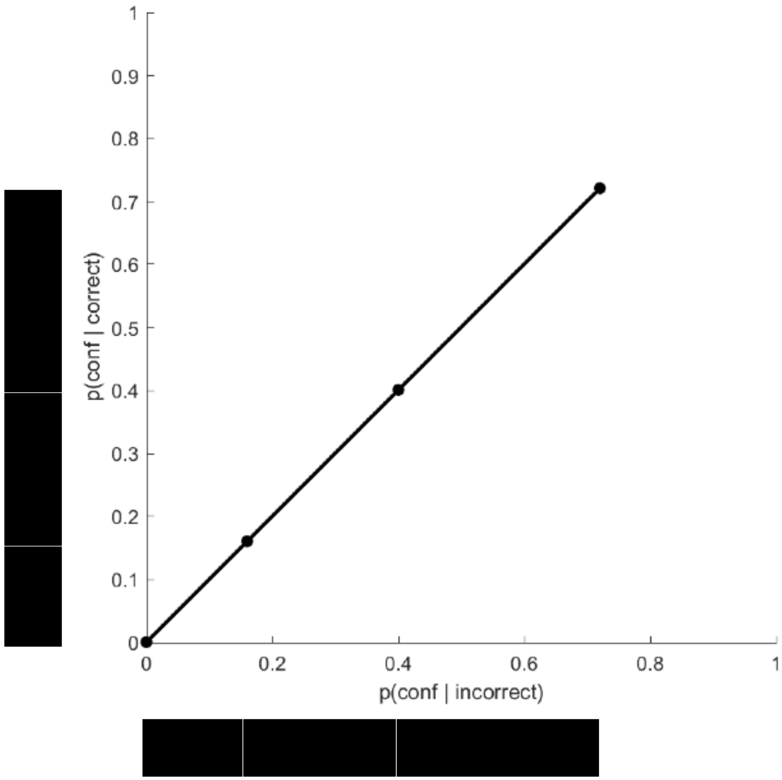
Type 2 ROCs



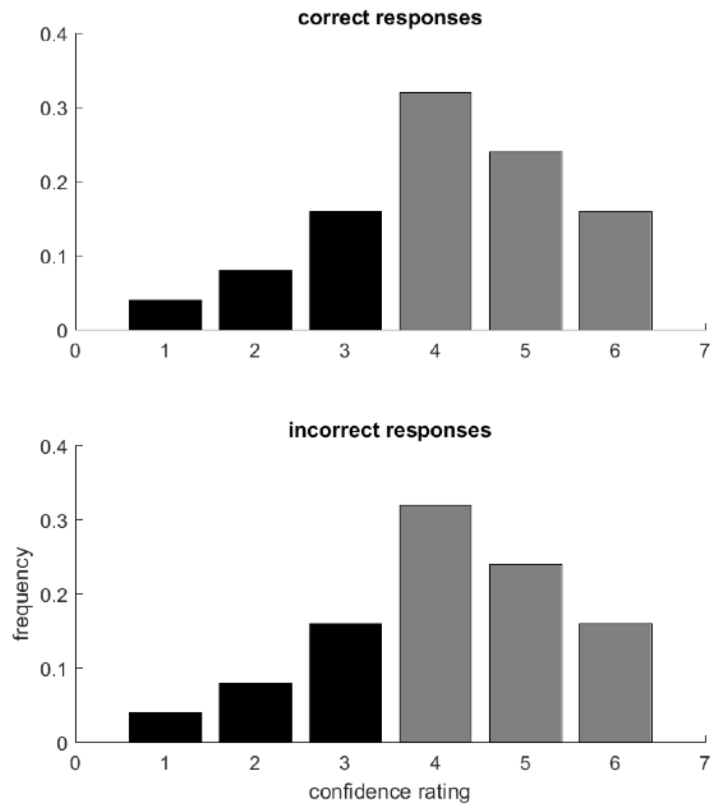
No metacognition:



Type 2 ROCs

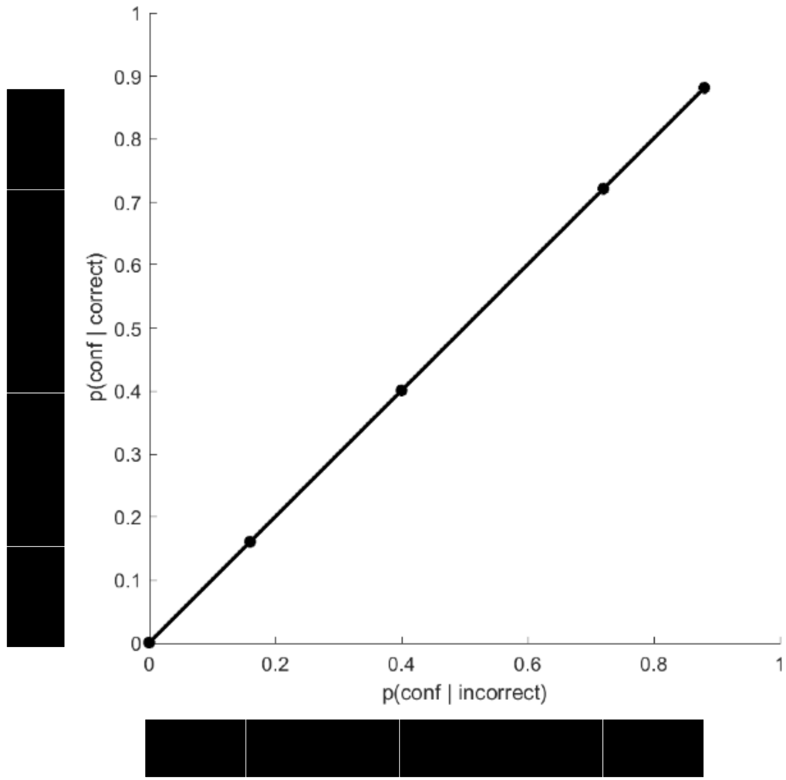


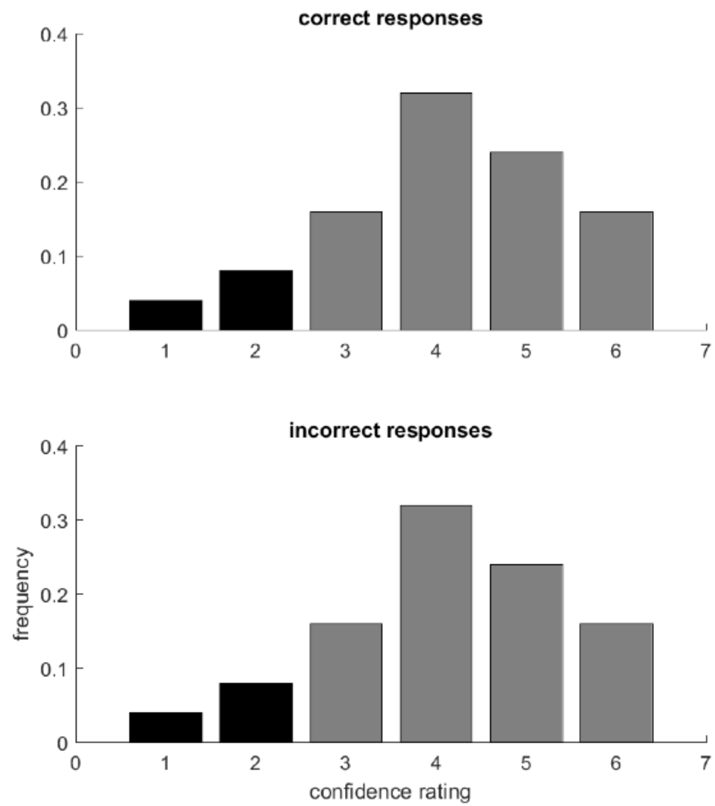
No metacognition:



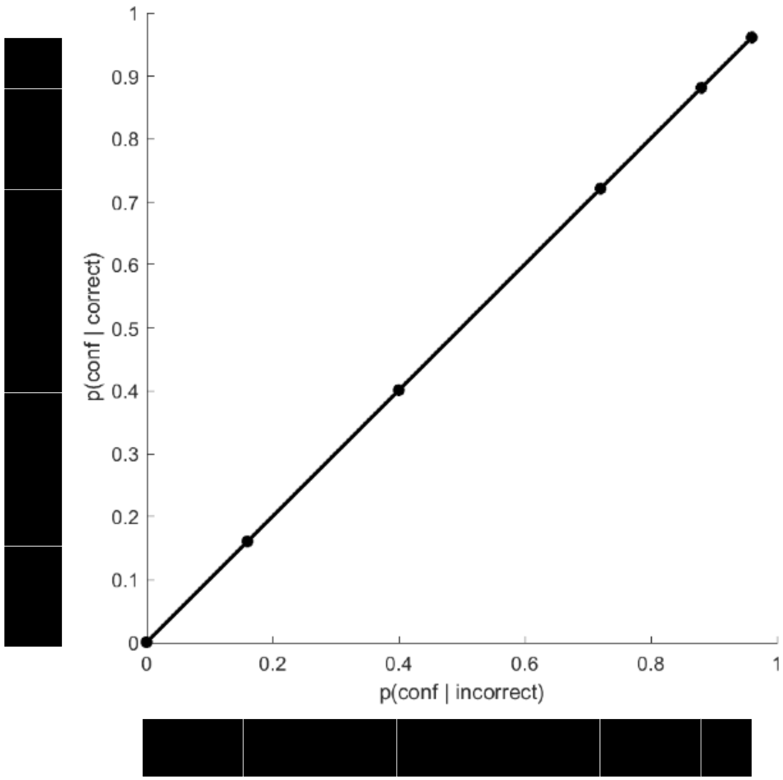
Type 2 ROCs

No metacognition:

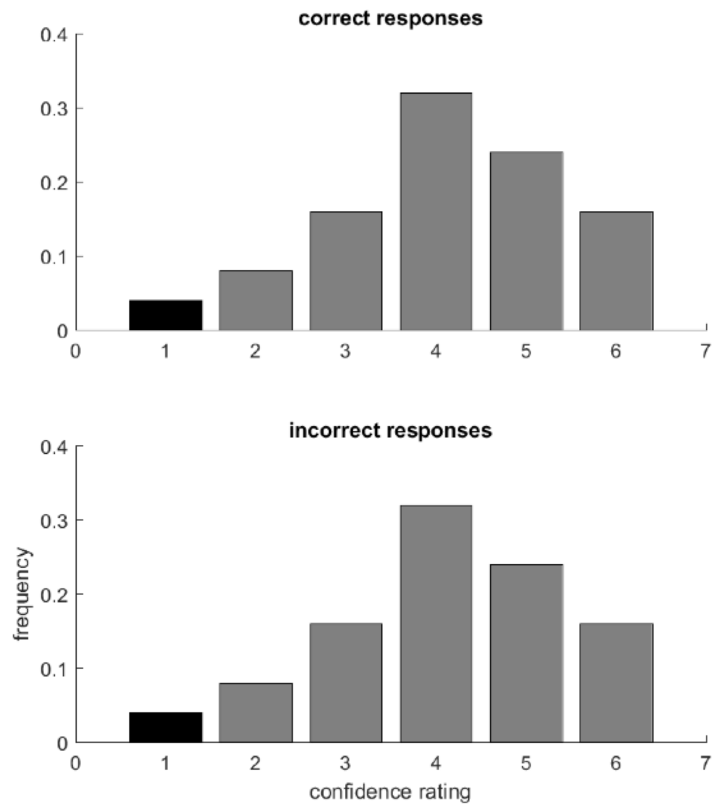




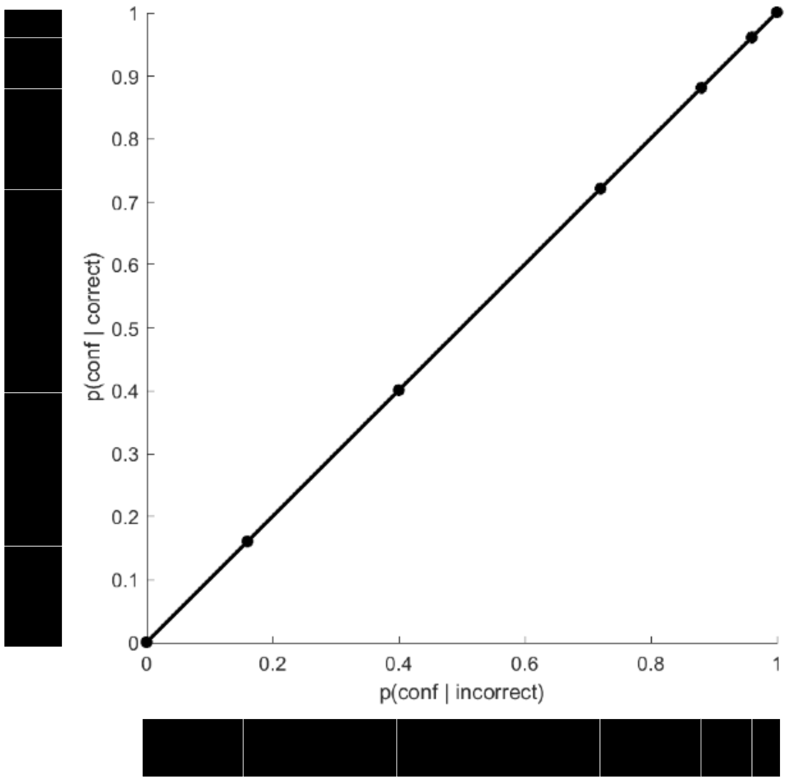
Type 2 ROCs



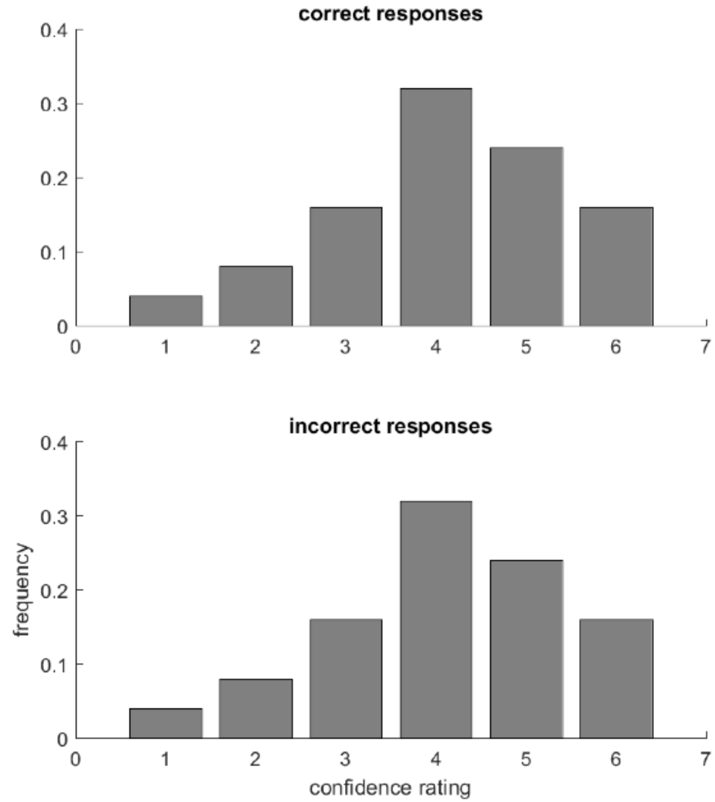
No metacognition:



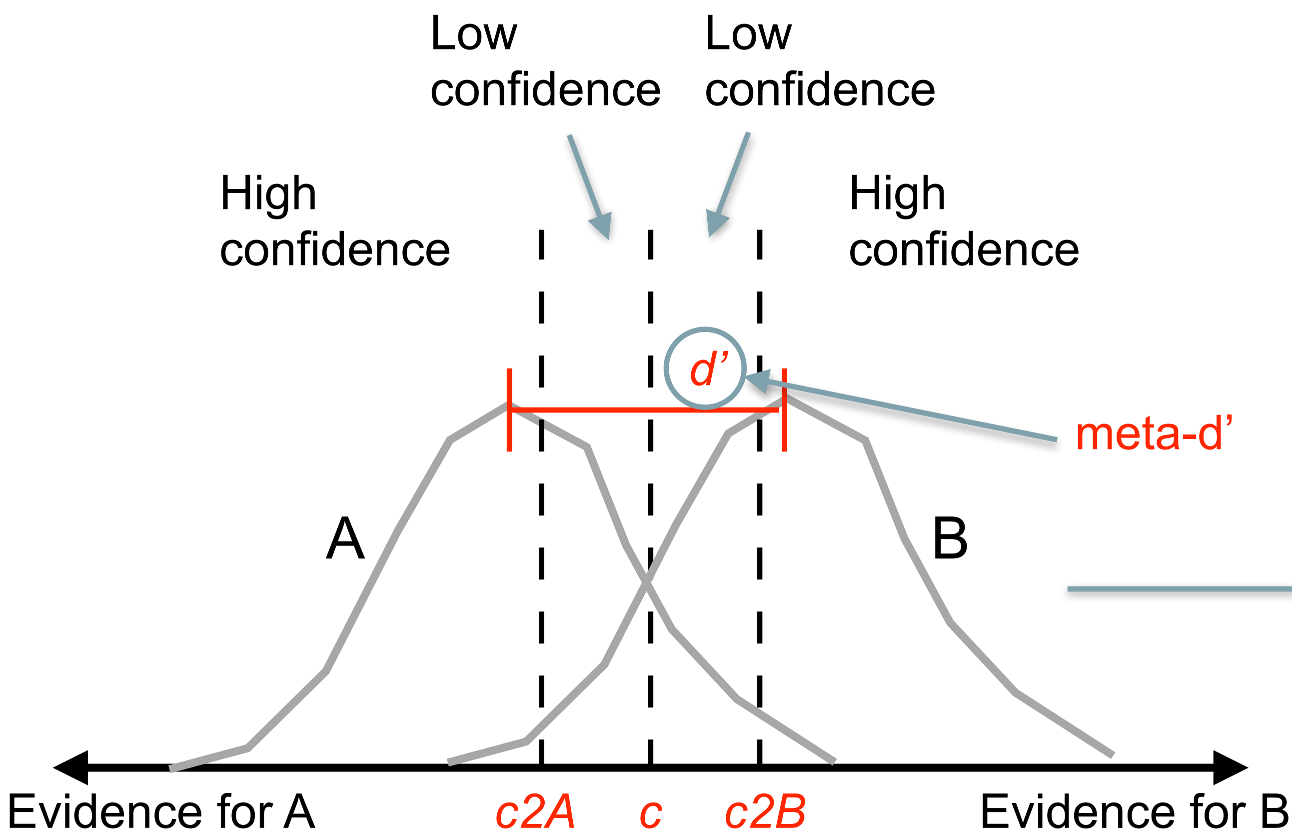
Type 2 ROCs

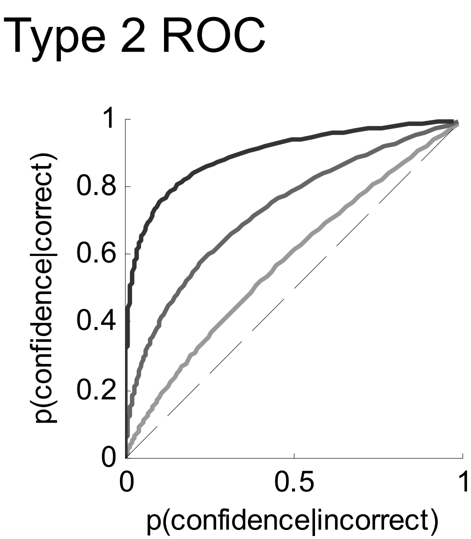


No metacognition:



Generative model for confidence





Find parameter set that best fits subjects' type 2 ROC

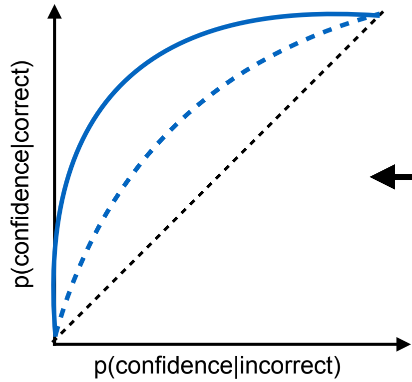
Maniscalco & Lau (2012) *Consciousness and Cognition*

Generative model for confidence

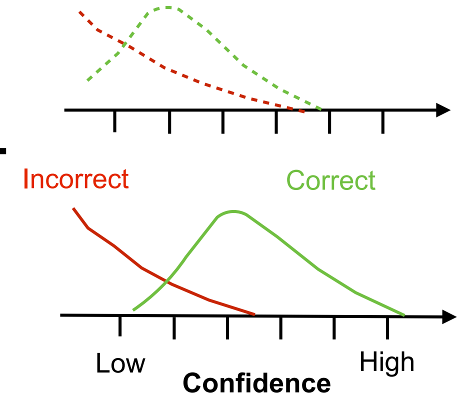
Type 1 SDT parameters

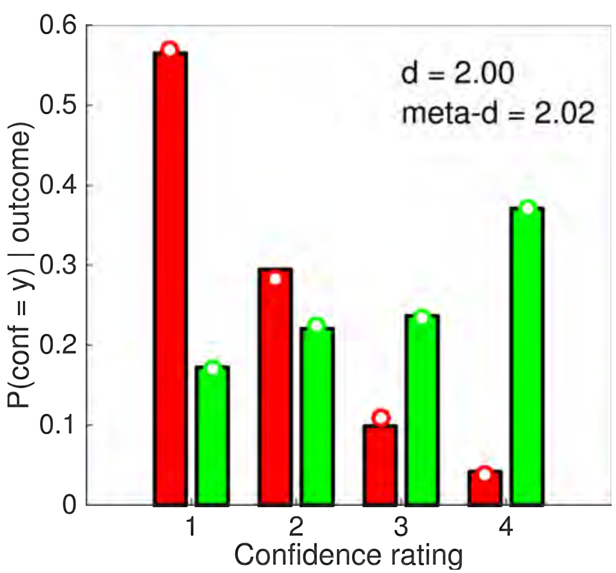
meta-*d’* (fitted to type 2 ROC) compared to observed *d’*

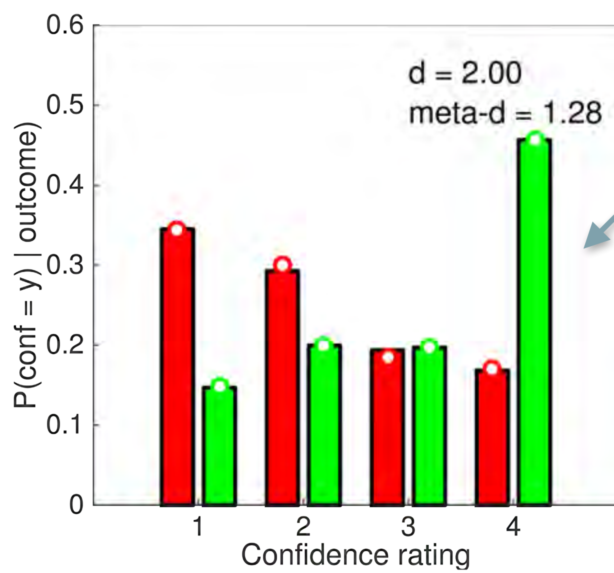
Type 2 ROC



Observed confidence distributions







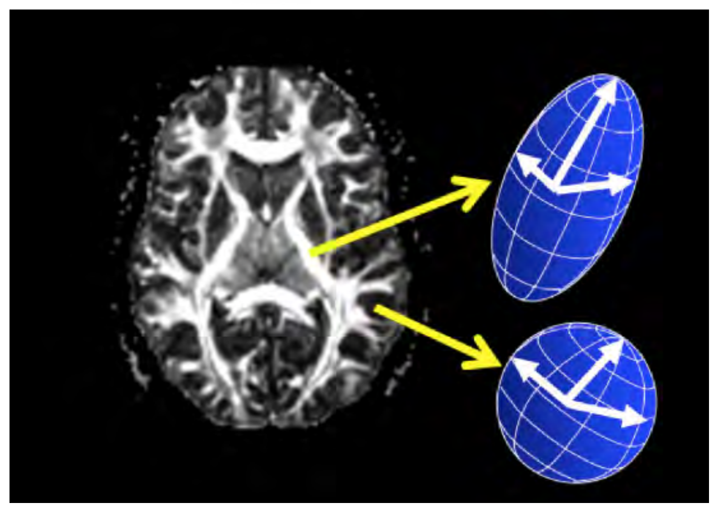
Gaussian noise added to confidence ratings

meta-d’/d’ = metacognitive efficiency

Maniscalco & Lau (2012) *Consciousness & Cognition*; Fleming (2017) *Neuroscience of Consciousness*

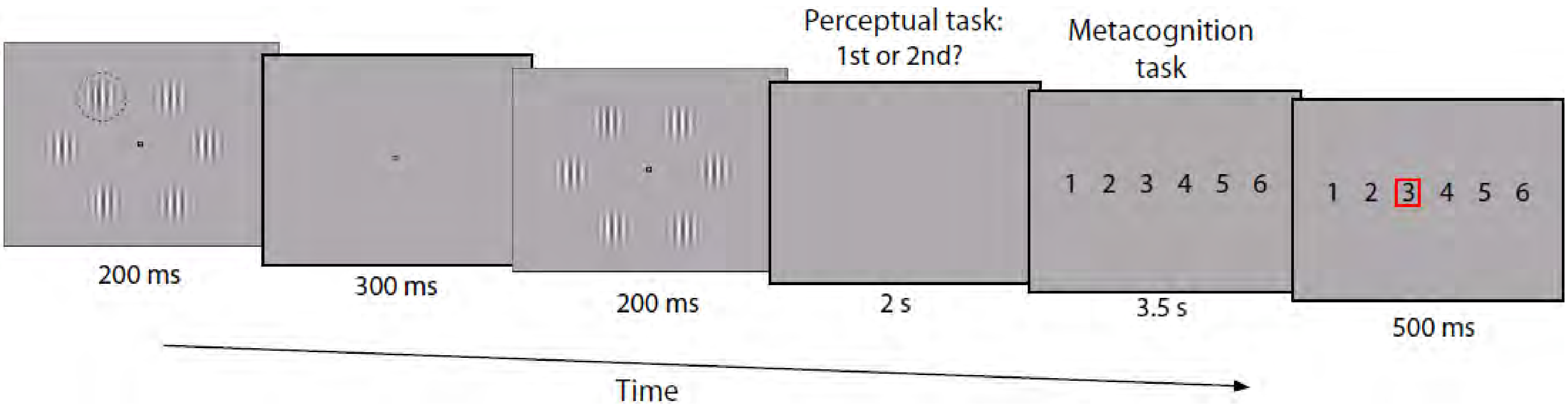
Individual differences in metacognition

Grey matter volume



White matter integrity

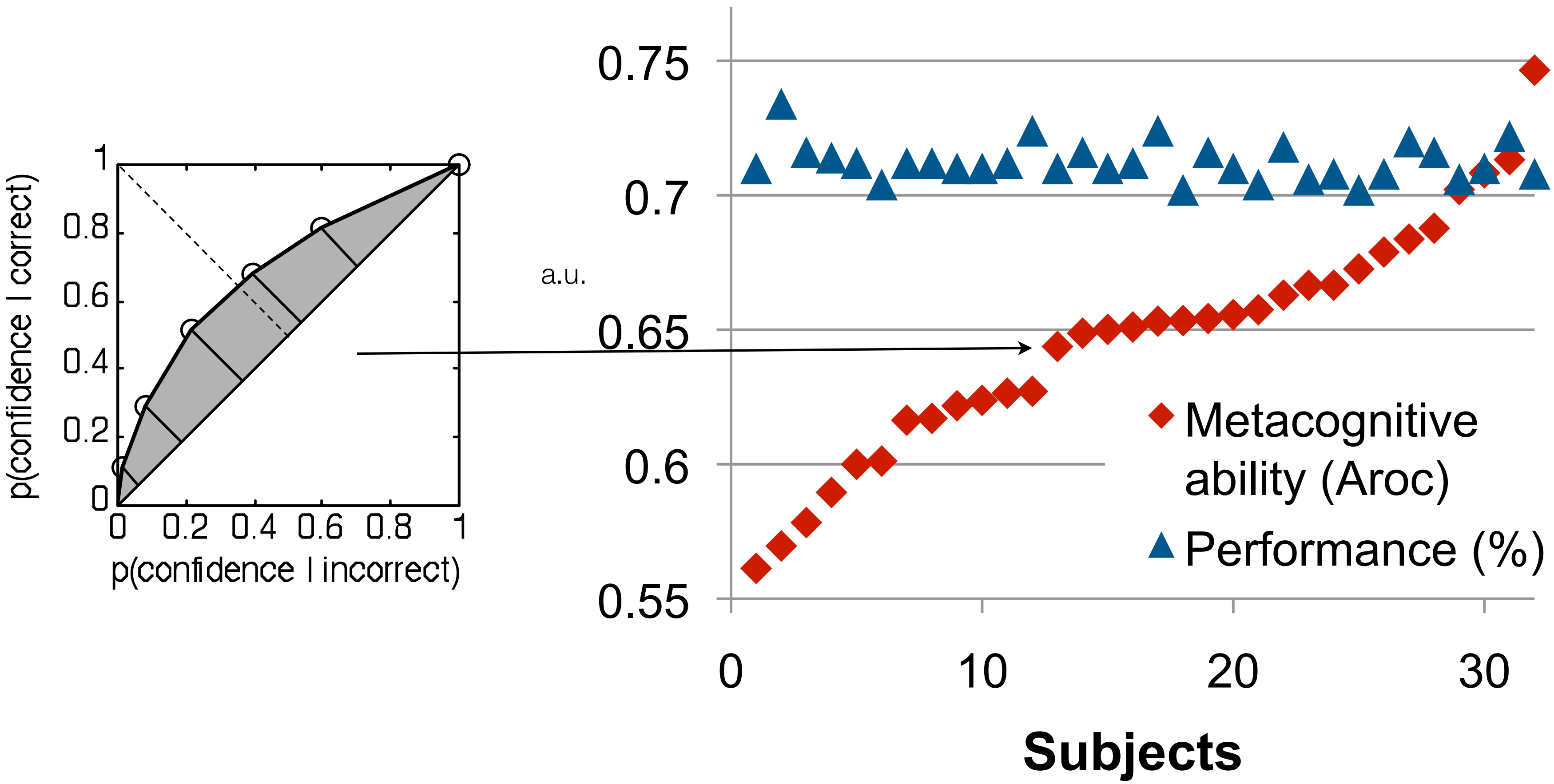
Isolating metacognition from performance



Performance titrated using a 2-down 1-up staircase
32 participants
600 trials per participant

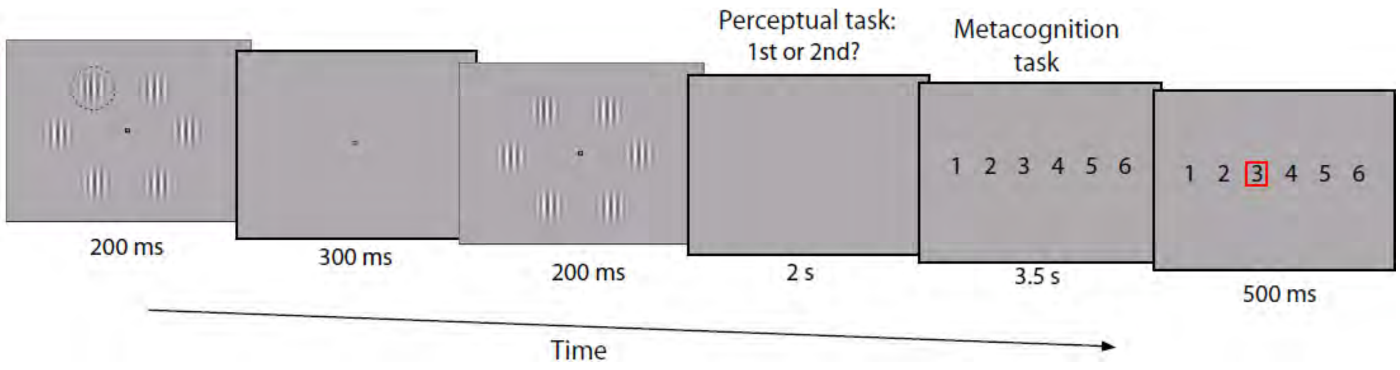
Fleming et al. (2010) *Science*

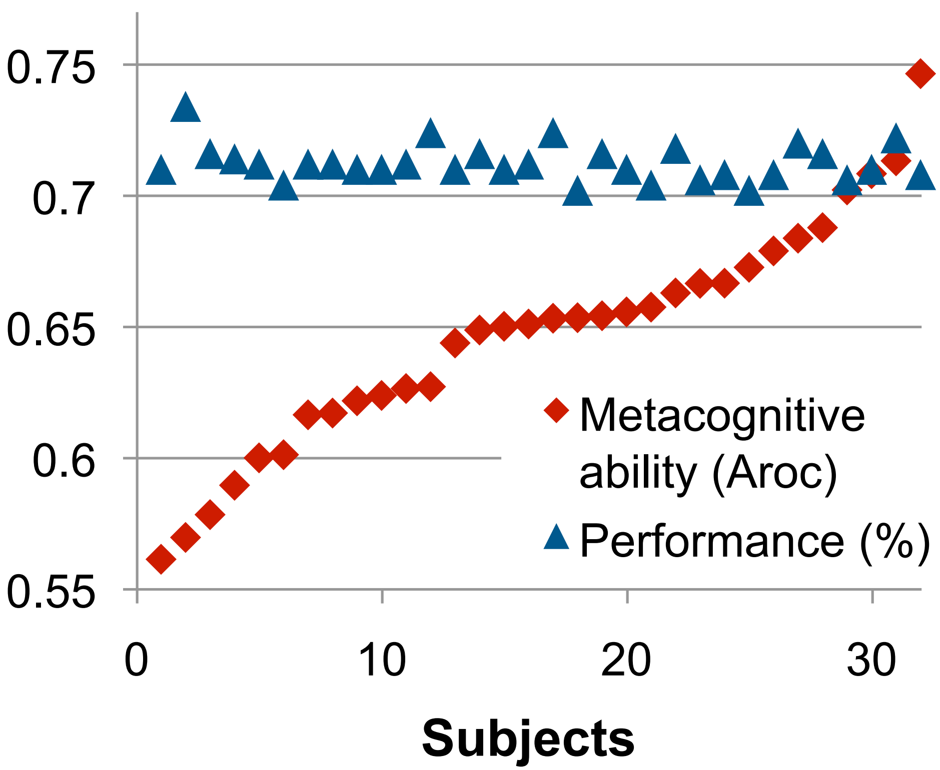
Isolating metacognition from performance

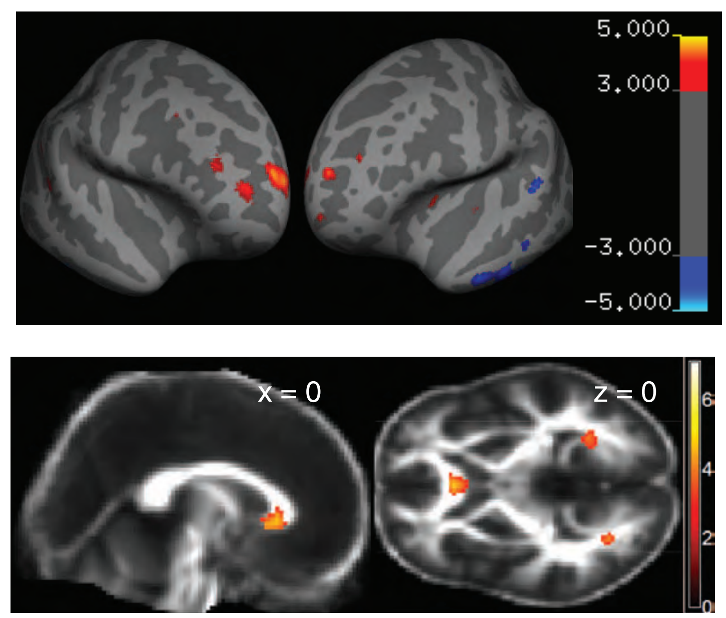


Fleming et al. (2010) *Science*

Isolating metacognition from performance

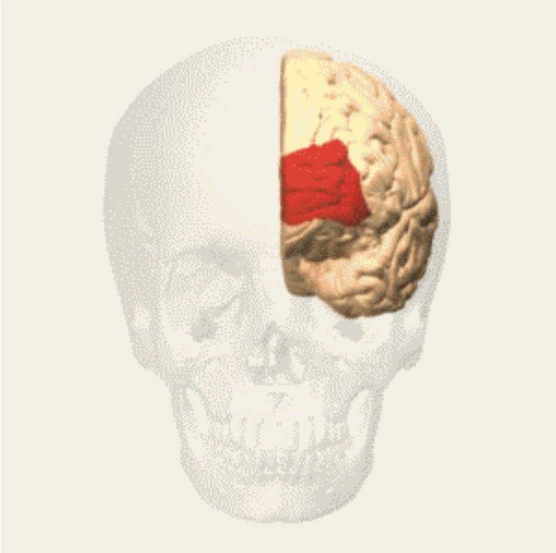






Fleming et al. (2010) *Science*

Metacognitive sensitivity and aPFC



Relating Introspective Accuracy to Individual Differences in Brain Structure

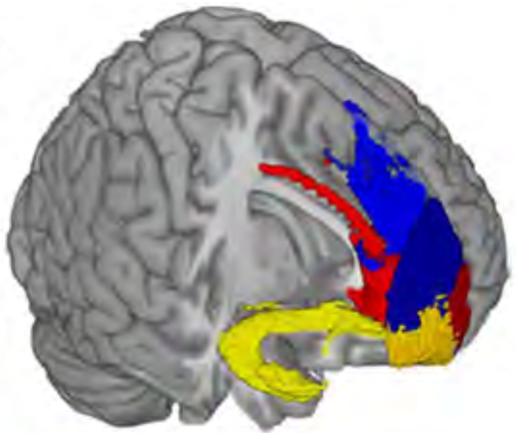
Stephen M. Fleming,[1]*† Rimona S. Weil,[1,2]* Zoltan Nagy,[1] Raymond J. Dolan,[1] Geraint Rees[1,2]

Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance

Osamu Yokoyama[a,b,c], Naoki Miura[c,d], Jobu Watanabe[c,d,e], Atsushi Takemoto[b,c], Shinya Uchida[d,f], Motoaki Sugiura[g], Kaoru Horie[e,h], Shigeru Sato[e,h], Ryuta Kawashima[c,d,e,f], Katsuki Nakamura[b,c,*]

Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception

Li Yan McCurdy,[1] Brian Maniscalco,[1] Janet Metcalfe,[1] Ka Yuet Liu,[2] Floris P. de Lange,[3] and Hakwan Lau[1,3]
[1]Department of Psychology, Columbia University, New York, New York 10027, [2]Department of Sociology, University of California, Los Angeles, Los Angeles, California 90095, and [3]Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behaviour, 6500 HE Nijmegen, The Netherlands



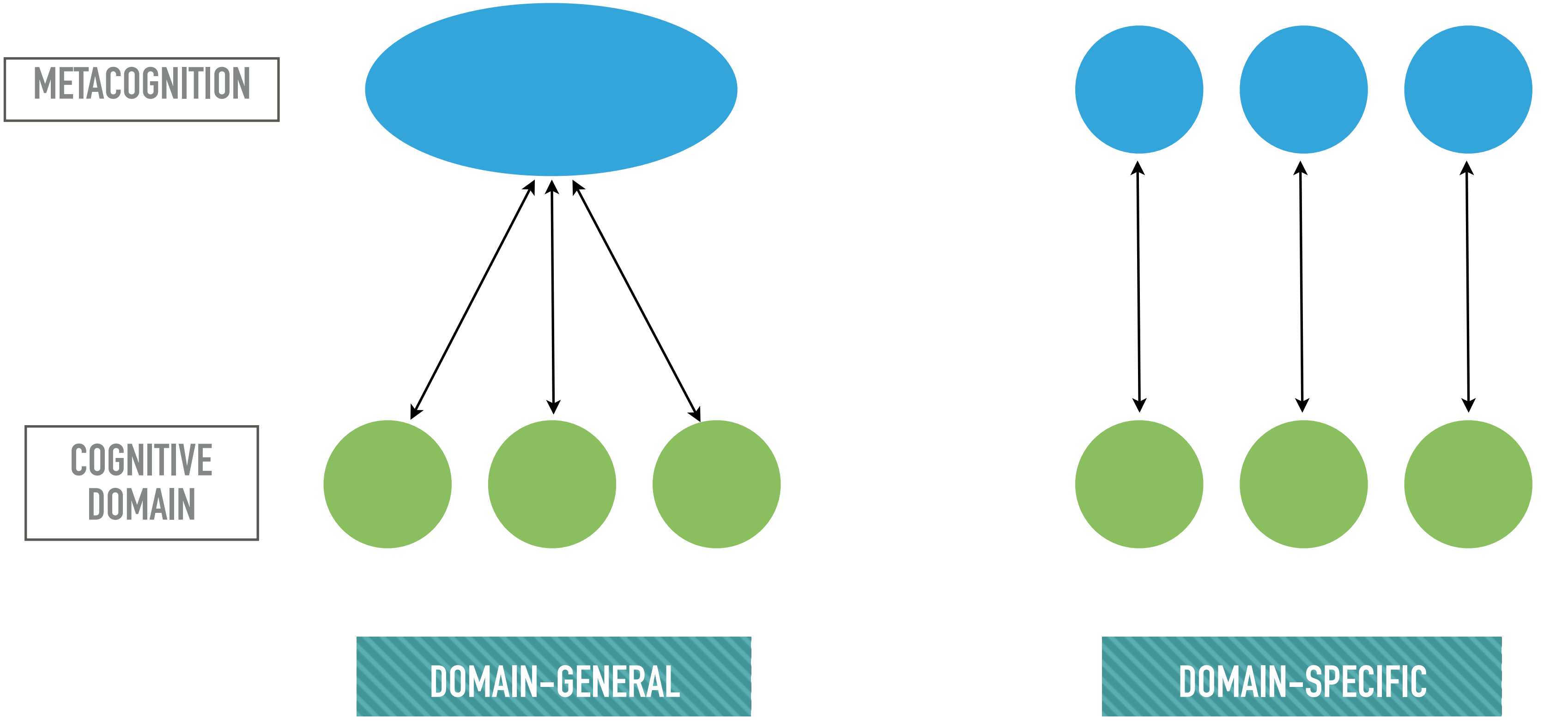
Metacognitive ability correlates with hippocampal and prefrontal microstructure

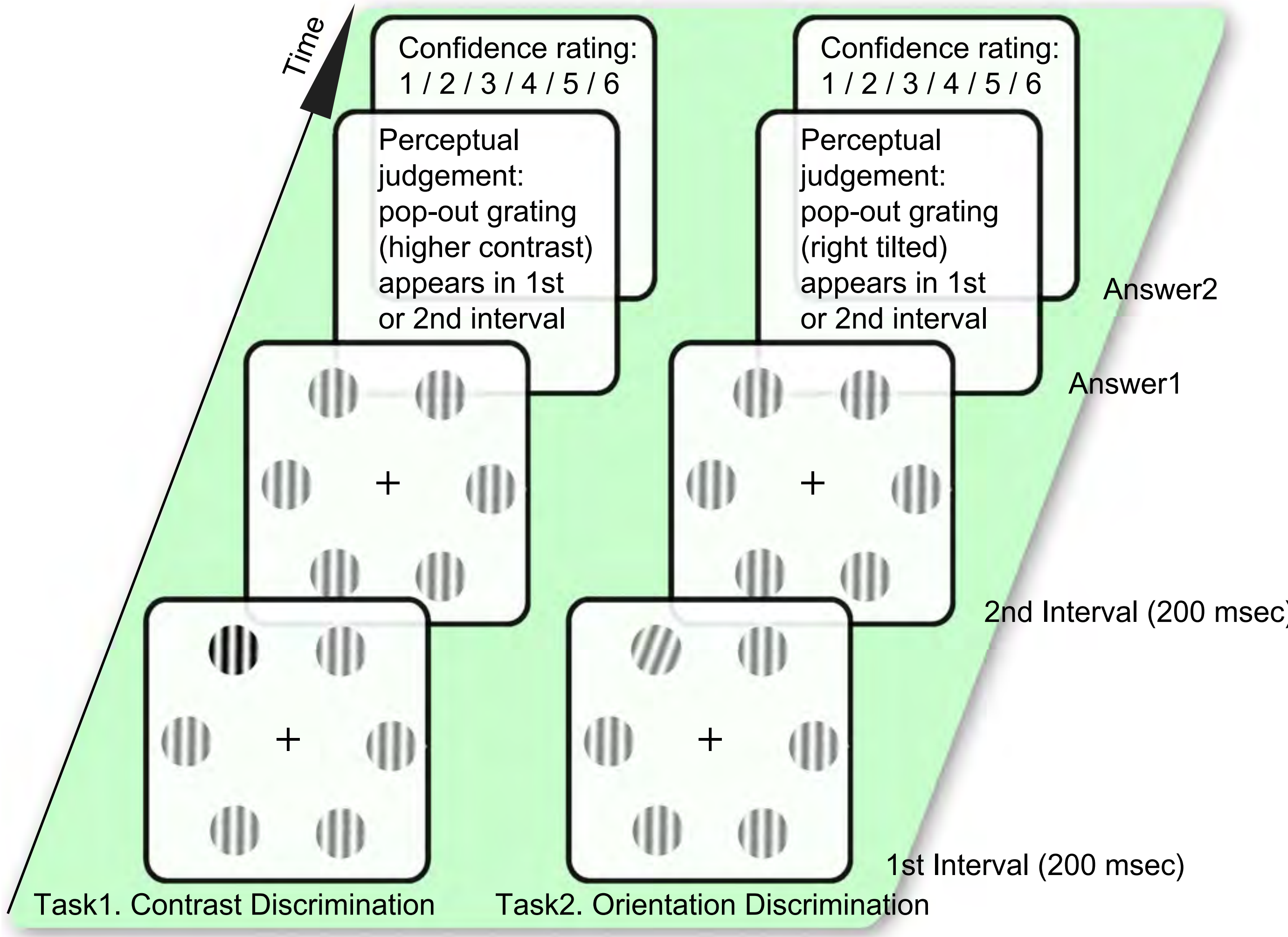
Micah Allen[a,b,*], James C. Glen[a], Daniel Müllensiefen[c], Dietrich Samuel Schwarzkopf[a,d], Francesca Fardo[a,e,f], Darya Frank[g], Martina F. Callaghan[b], Geraint Rees[a,b]

Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception

Benjamin Baird,[1] Jonathan Smallwood,[2] Krzysztof J. Gorgolewski,[3] and Daniel S. Margulies[3]
[1]Department of Psychological and Brain Sciences, University of California, Santa Barbara, California 93106, [2]Department of Psychology, University of York, North Yorkshire YO10 5DD, United Kingdom, and [3]Max Planck Research Group: Neuroanatomy & Connectivity, Max Planck Institute for Human Cognitive and Brain Sciences, 04103, Leipzig, Germany

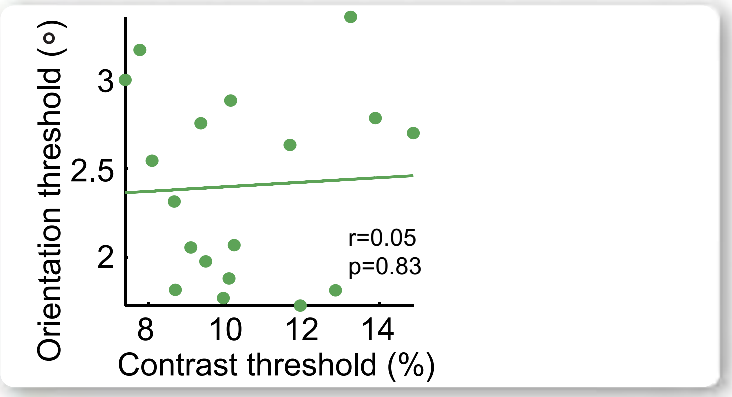
Domain-general or domain-specific?





Does having good metacognition on task 1 predict good metacognition on task 2?

Song, Kanai, Fleming et al. (2011) *Consciousness & Cognition*

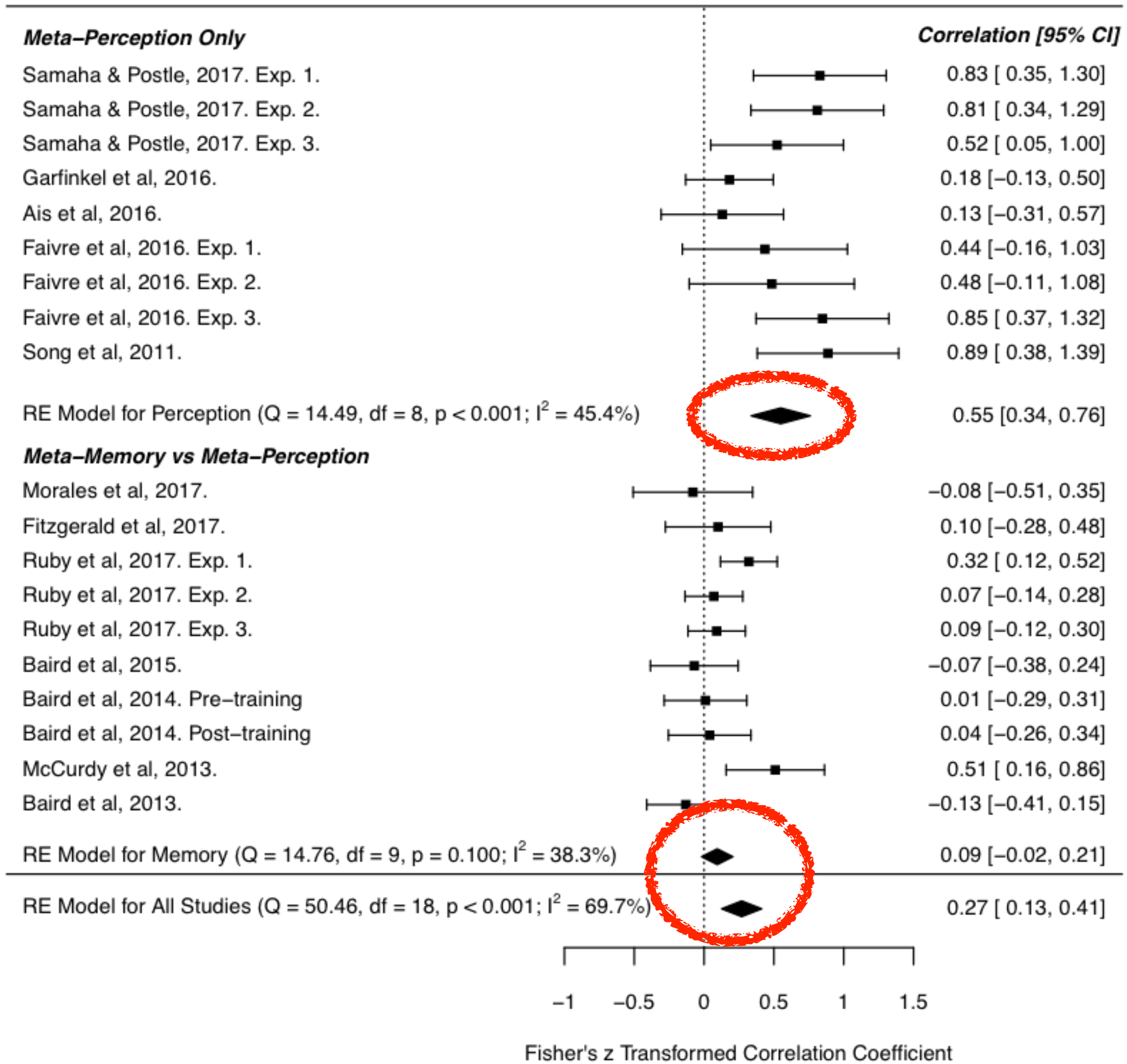


Performance thresholds uncorrelated

Song, Kanai, Fleming et al. (2011) *Consciousness & Cognition*

Meta-analysis of metacognitive sensitivity correlations

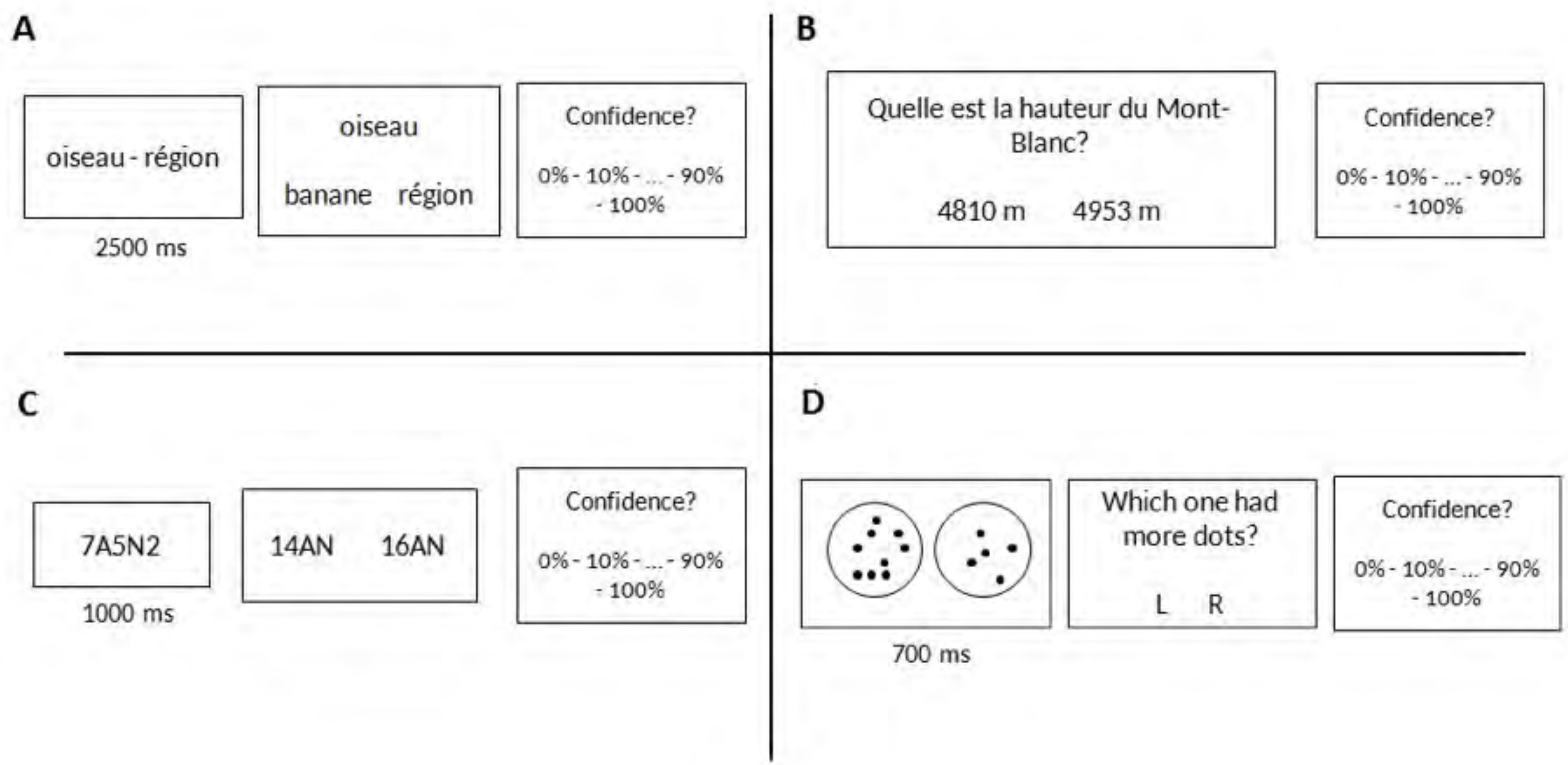
Some evidence for domain-generality, but lack of consistency across task designs / low power



Rouault, McWilliams, Allen & Fleming (2018) *Personality Neuroscience*

Assessing domain-generality of metacognition

N=181, hierarchical modelling of covariance in **metacognitive efficiency** across 4 distinct 2AFC tasks



Mazancieux, Fleming, Souchay & Moulin *in prep*

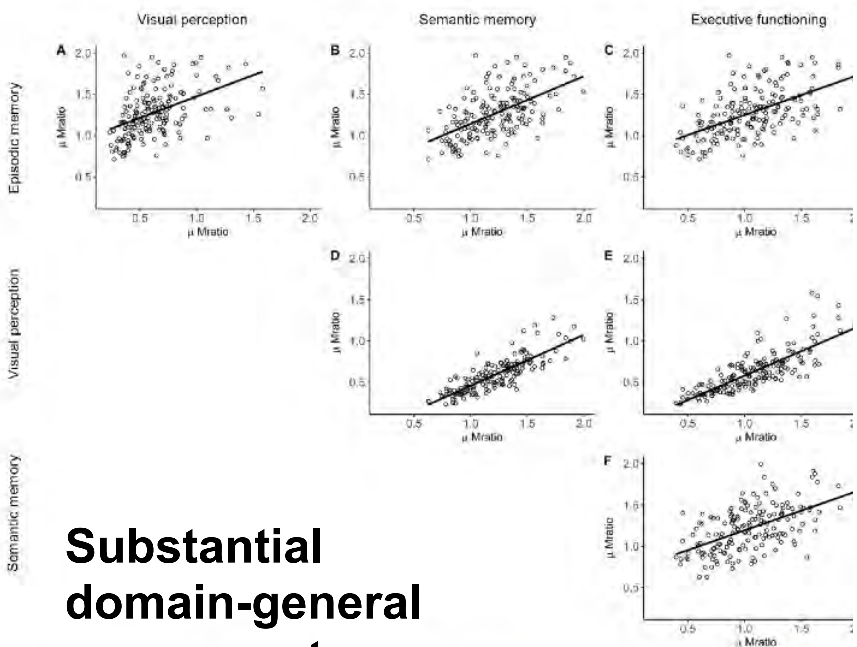
$$[\log(M1_s)\log(M2_s)\log(M3_s)\log(M4_s)] \sim N\left(\begin{bmatrix}\mu_{M1}\\ \mu_{M2}\\ \mu_{M3}\\ \mu_{M4}\end{bmatrix}, \begin{bmatrix}\sigma^2{}_{M1} & \rho_{M1M2}\sigma_{M1}\sigma_{M2} & \rho_{M1M3}\sigma_{M1}\sigma_{M3} & \rho_{M1M4}\sigma_{M1}\sigma_{M4}\\ \rho_{M1M2}\sigma_{M1}\sigma_{M2} & \sigma^2{}_{M2} & \rho_{M2M3}\sigma_{M2}\sigma_{M3} & \rho_{M2M4}\sigma_{M2}\sigma_{M4}\\ \rho_{M1M3}\sigma_{M1}\sigma_{M3} & \rho_{M2M3}\sigma_{M2}\sigma_{M3} & \sigma^2{}_{M3} & \rho_{M3M4}\sigma_{M3}\sigma_{M4}\\ \rho_{M1M4}\sigma_{M1}\sigma_{M4} & \rho_{M2M4}\sigma_{M2}\sigma_{M4} & \rho_{M3M4}\sigma_{M3}\sigma_{M4} & \sigma^2{}_{M4}\end{bmatrix}\right)$$

Priors were specified as follows:

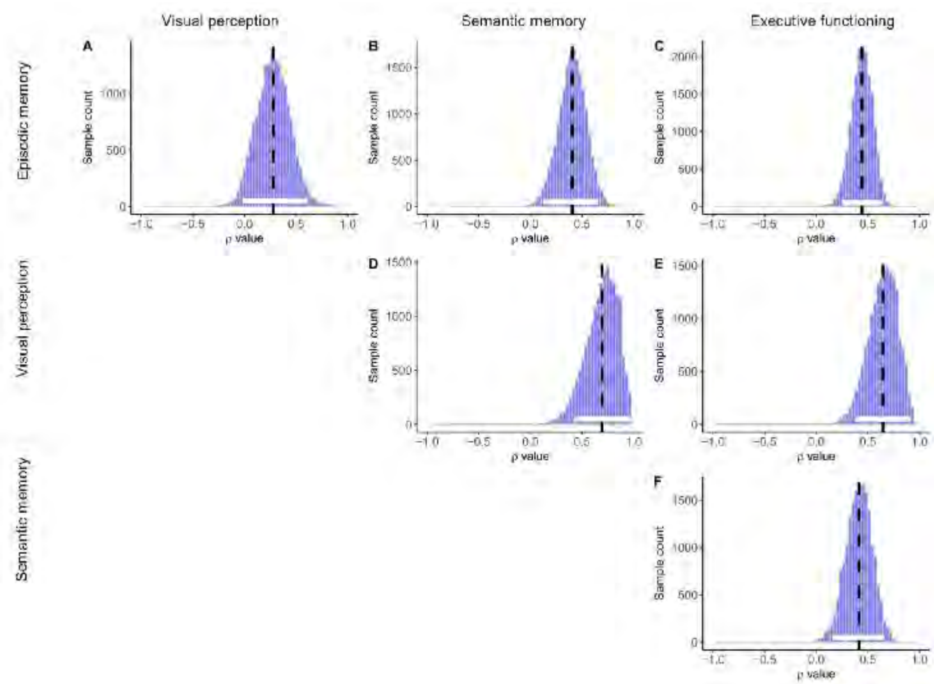
$$\mu_{M1}, \mu_{M2}, \mu_{M3}, \mu_{M4} \sim N(0,1)$$

$$\sigma_{M1}, \sigma_{M2}, \sigma_{M3}, \sigma_{M4} \sim \mathrm{InvSqrtGamma}(0.001, 0.001)$$

$$\rho_{M1M2}, \rho_{M1M3}, \rho_{M1M4}, \rho_{M2M3}, \rho_{M2M4}, \rho_{M3M4} \sim \mathrm{Uniform}(-1,1)$$



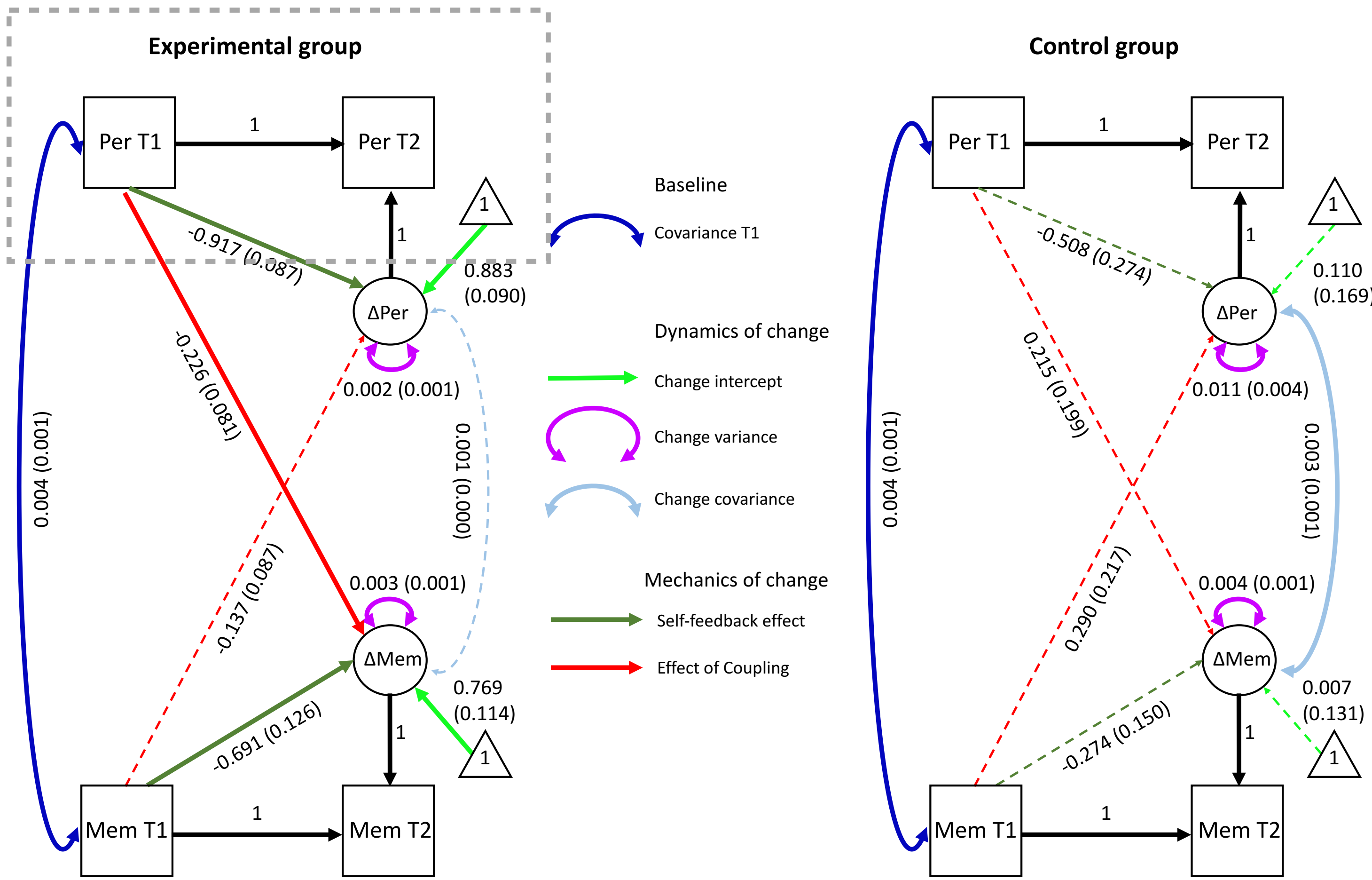
Substantial domain-general component



Mazancieux, Fleming, Souchay & Moulin *in prep*

Training perceptual metacognition leads to generalised improvements

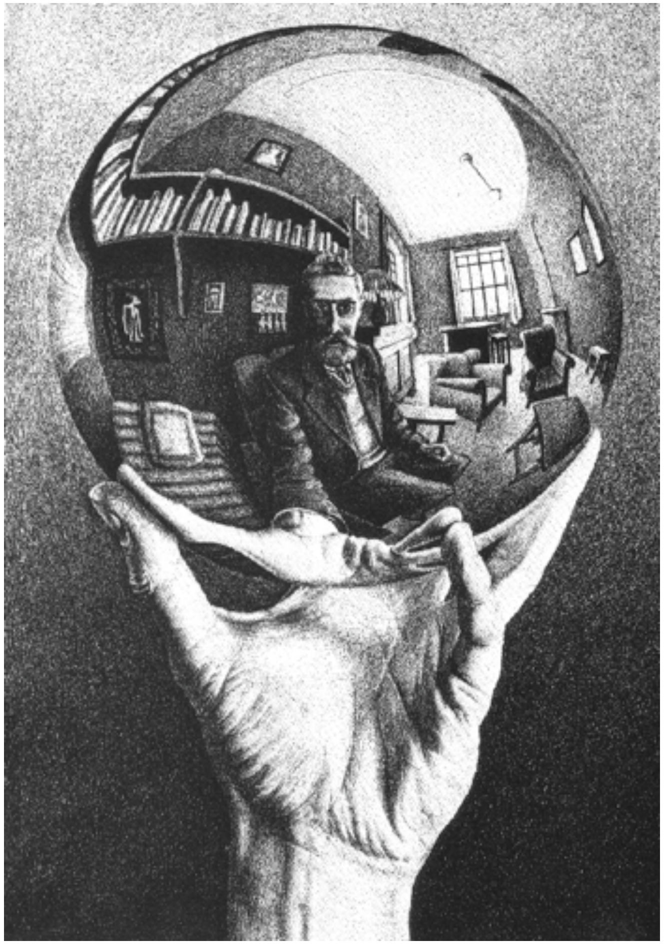
Trained only on perceptual metacognition



Carpenter, Sherman, Kievit, Seth, Lau & Fleming (in press) *JEP:General*

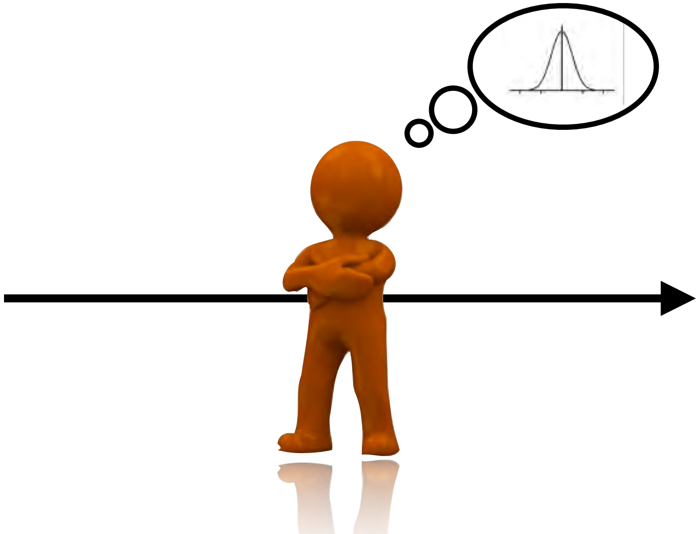
Outline

- **Metacognition as a domain-general resource**
- **PFC as a hub for confidence formation**
- Altered metacognitive beliefs in psychopathology



Individual differences in metacognition

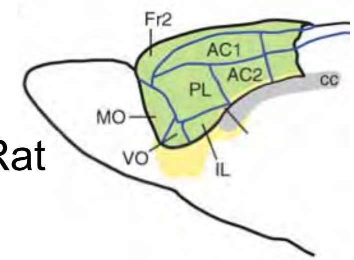
Within-subject construction of confidence

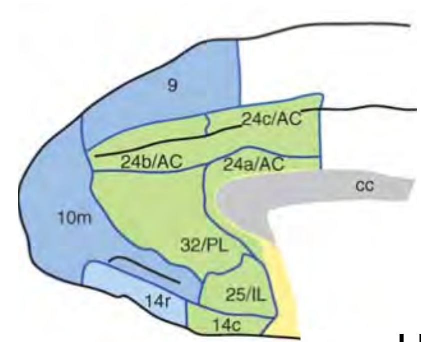


The confusing anatomy of confidence

- **Rodents**
  - **OFC** (Kepecs et al., 2008; Lak et al., 2014)
- **Monkeys**
  - **supplementary eye field** (Middlebrooks & Sommer, 2012)
  - **LIP** (Kiani & Shadlen, 2009)
  - **thalamic pulivinar** (Komura et al., 2013)
- **Humans**
  - **vmPFC** (De Martino, Fleming et al., 2013; Lebreton et al., 2015)
  - **ventral striatum** (Hebart et al., 2016)
  - **dACC/pre-SMA** (Fleck et al., 2006; Fleming et al., 2012)
  - **rostrolateral prefrontal cortex** (De Martino, Fleming et al., 2013; Fleming et al., 2010; 2012; 2014; Hilgenstock et al., 2014)

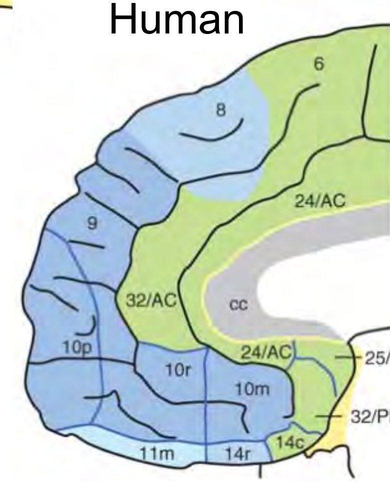
Rat





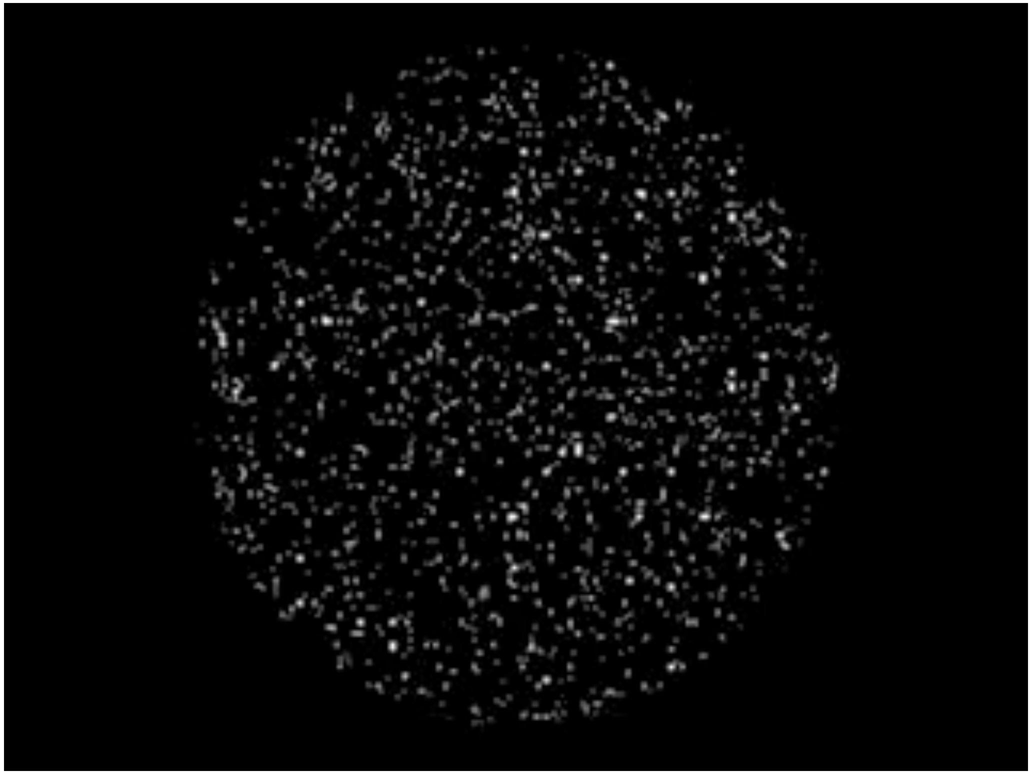
Macaque

Human



Wallis (2011) *Nat Neuro*

Confounding multiple computations?



From
Newsome lab

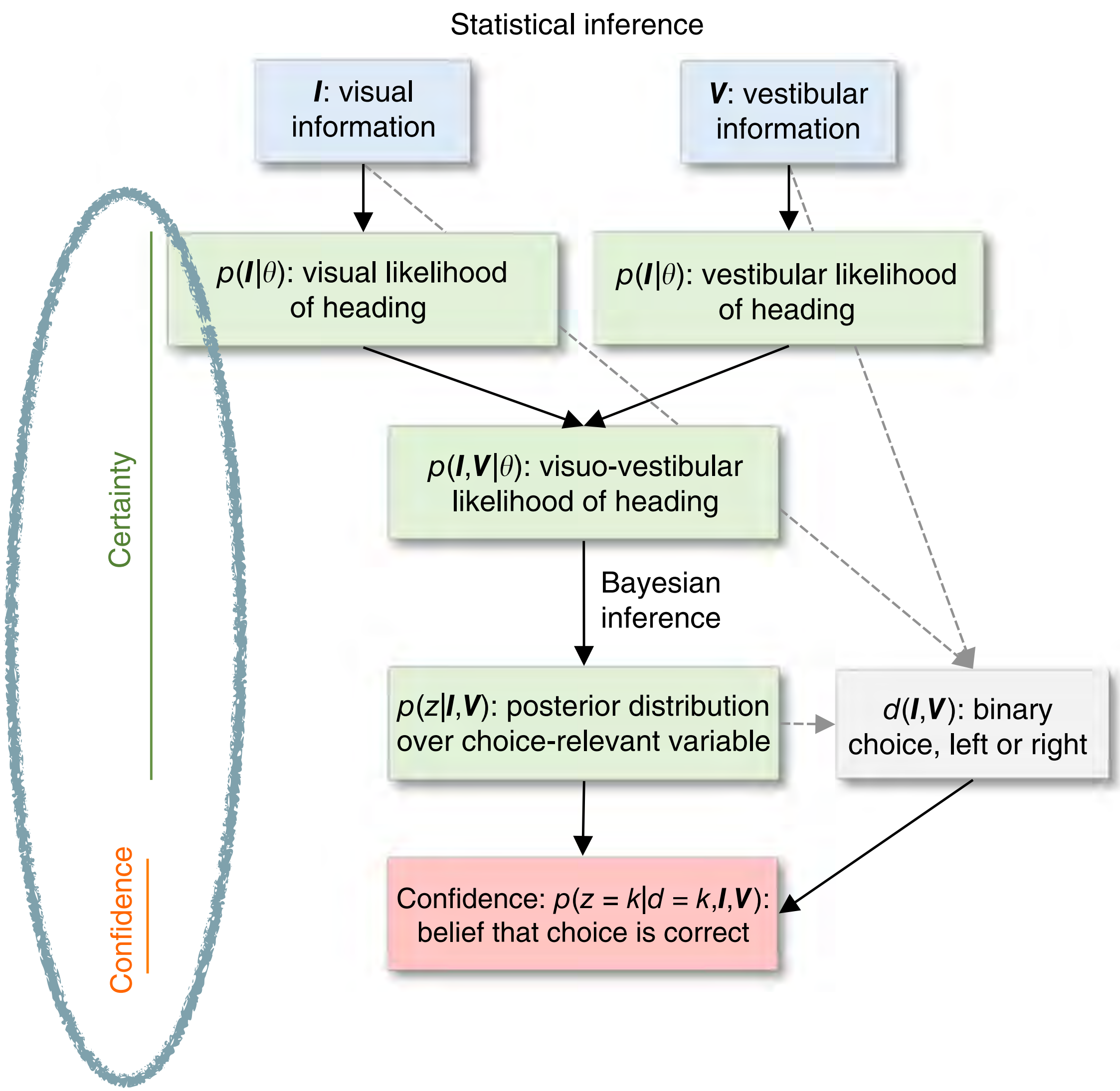
Left or right?

Greater certainty about motion direction = greater decision confidence
Sensory certainty and confidence are confounded…

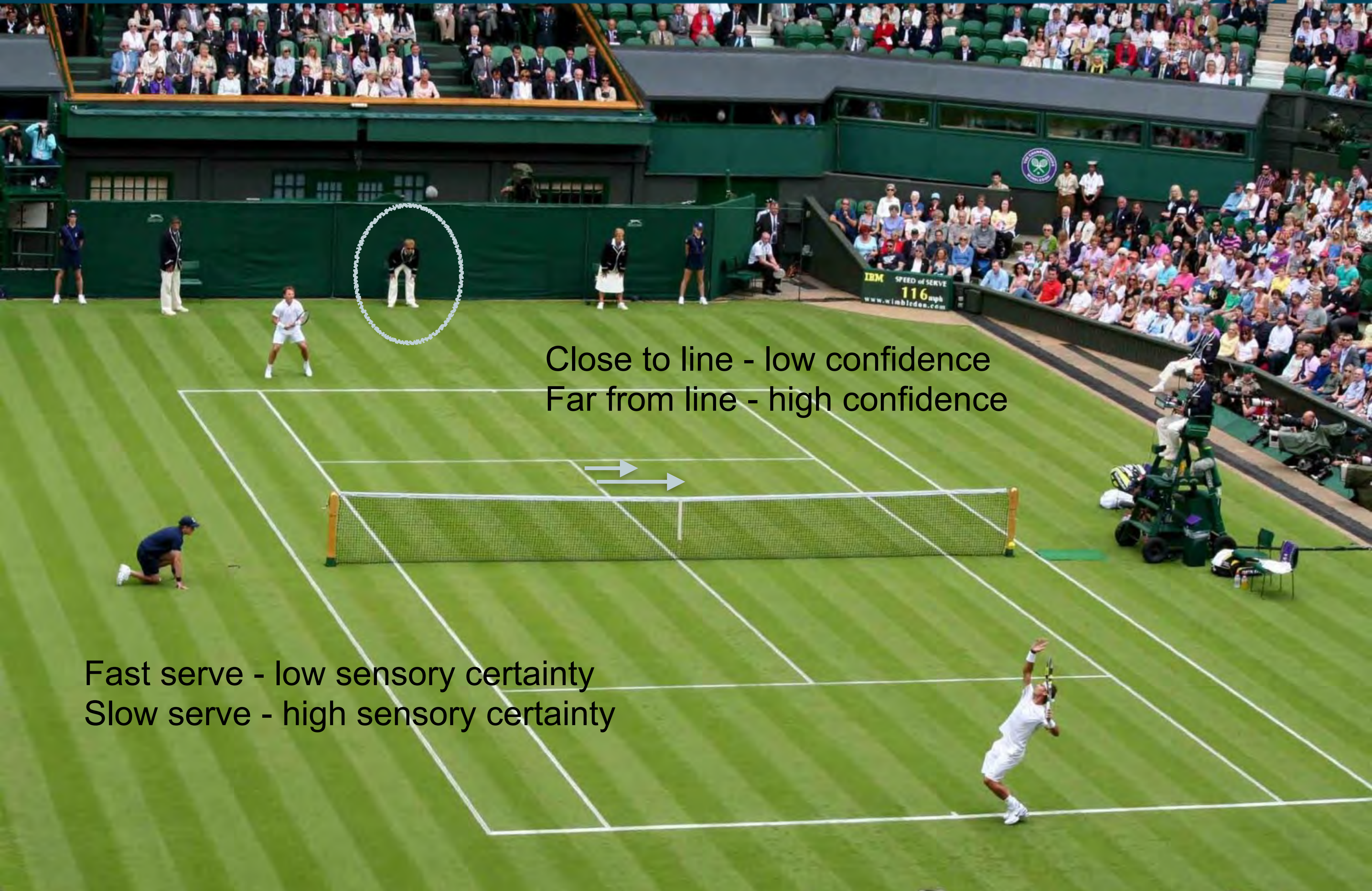
Confounding multiple computations?

“First-order” uncertainty about properties the world

Metacognitive uncertainty about your decision about the world



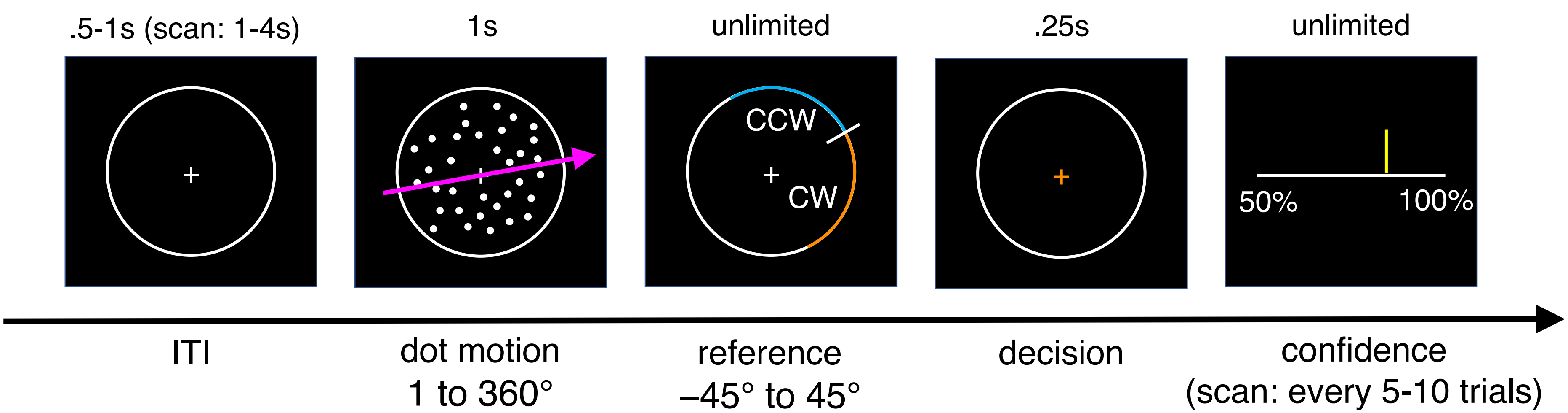
Pouget et al. (2016) *Nat Neuro*



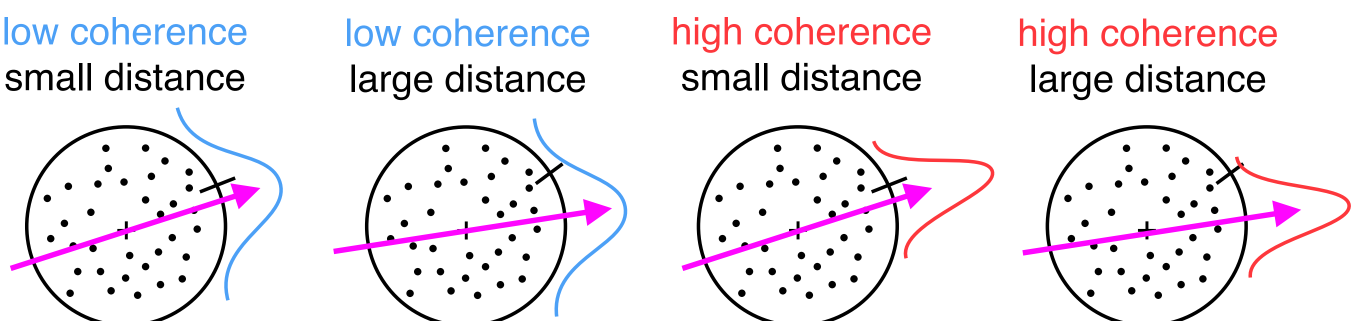
Close to line - low confidence
Far from line - high confidence

Fast serve - low sensory certainty
Slow serve - high sensory certainty

Task outline - scan session

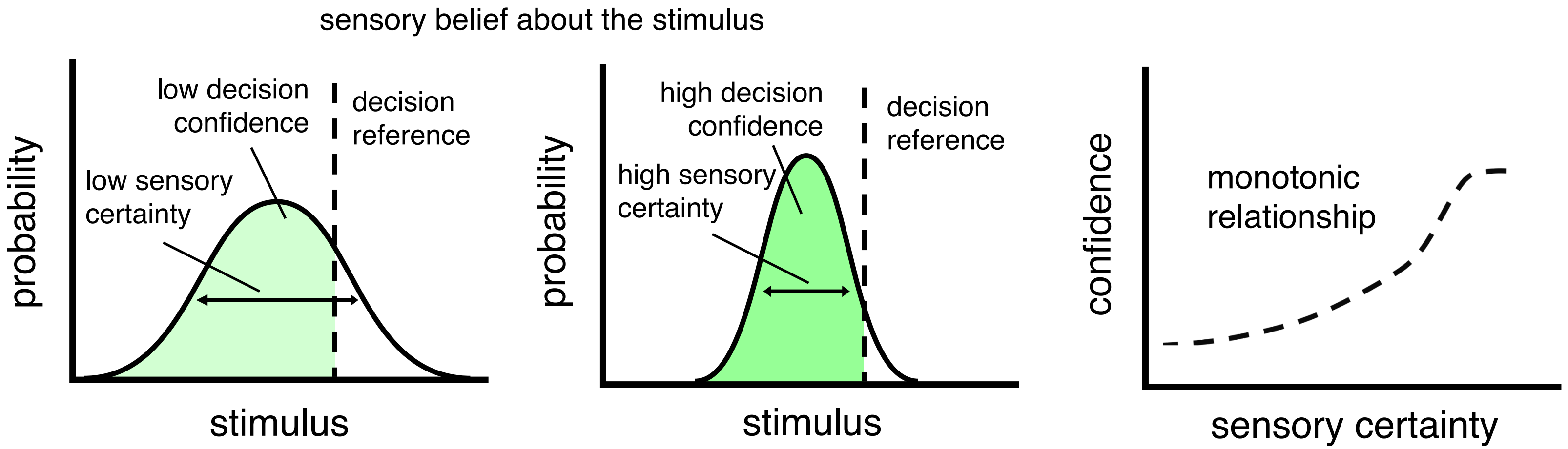


2x2 design used
in scan session

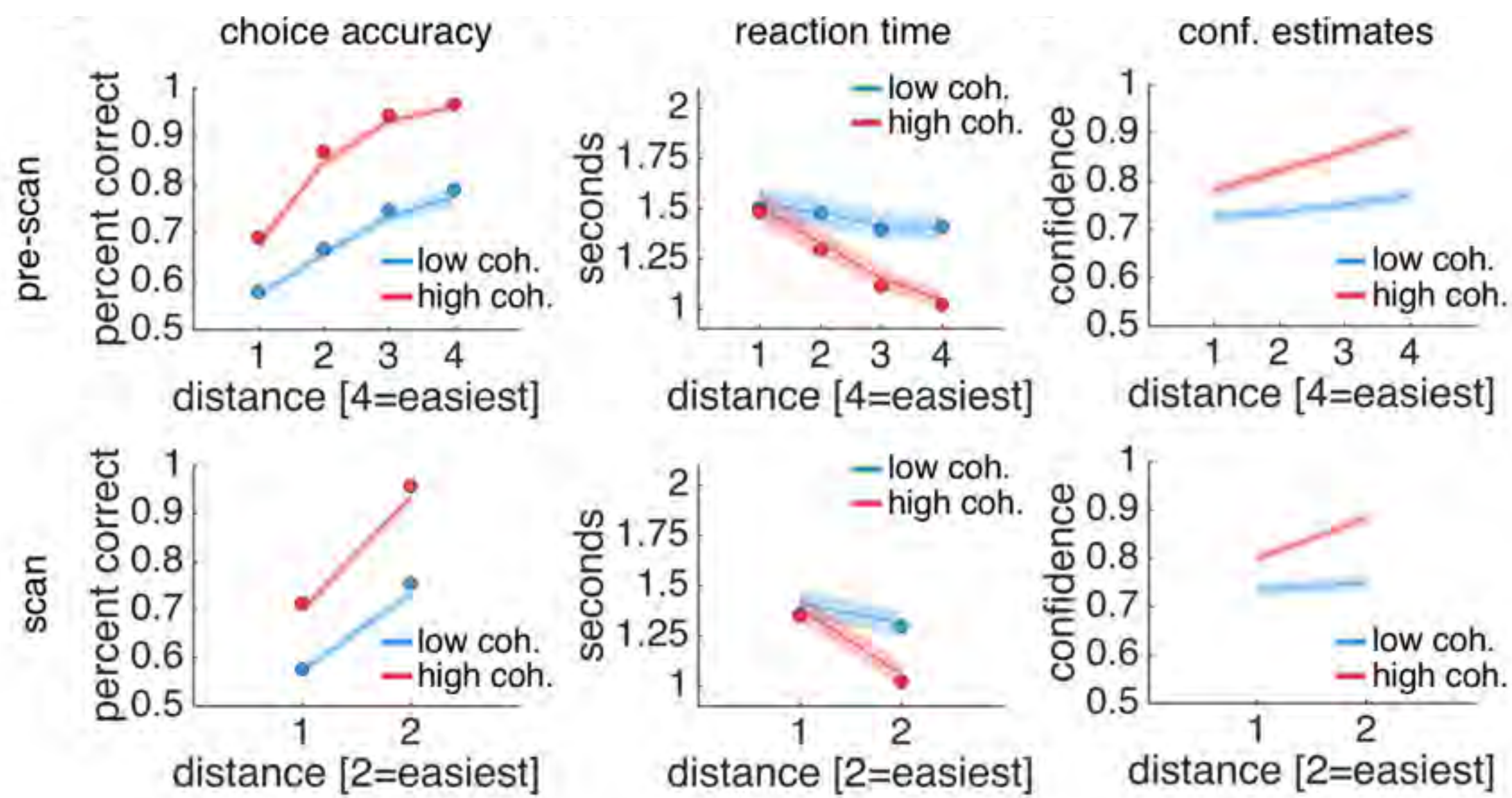


Bang & Fleming (2018) *PNAS*

Typical perceptual decision task:



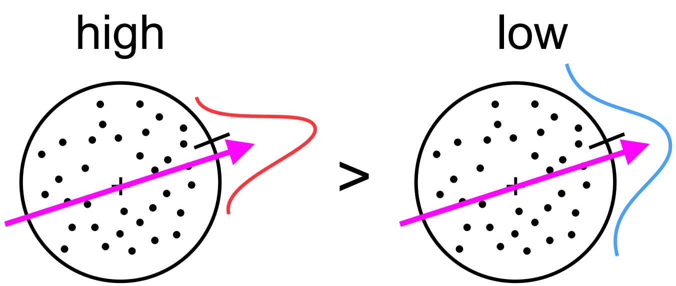
Construction of confidence

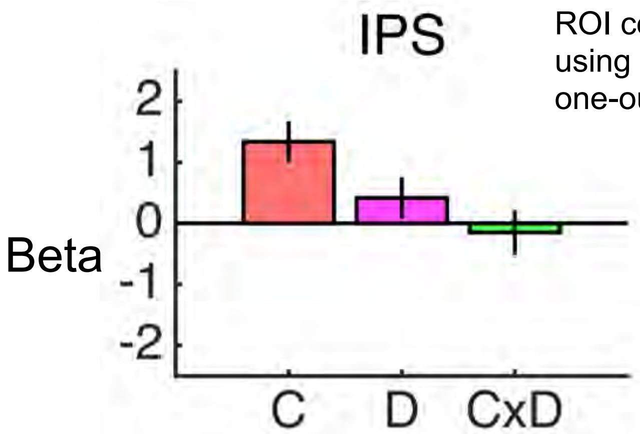


Confidence estimation relies on tracking an **interaction** between sensory certainty and boundary distance

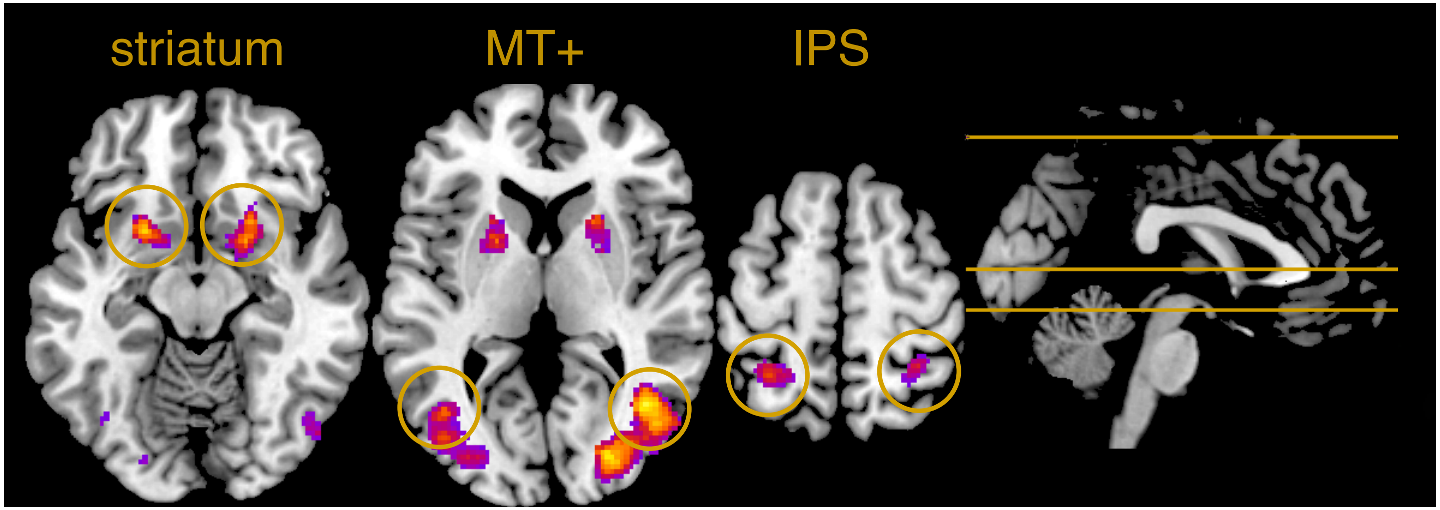
Bang & Fleming (2018) *PNAS*

Components of confidence - sensory certainty





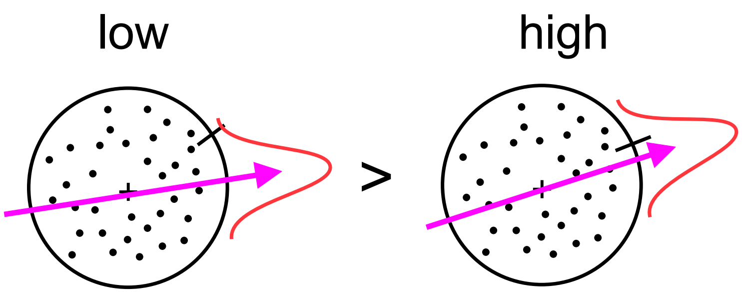
ROI computed using leave-one-out

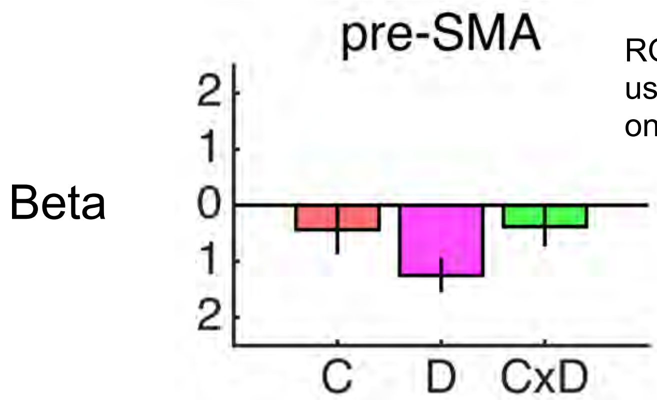


cluster-defining threshold: p < .001; pFWE < .05 corrected; correct trials

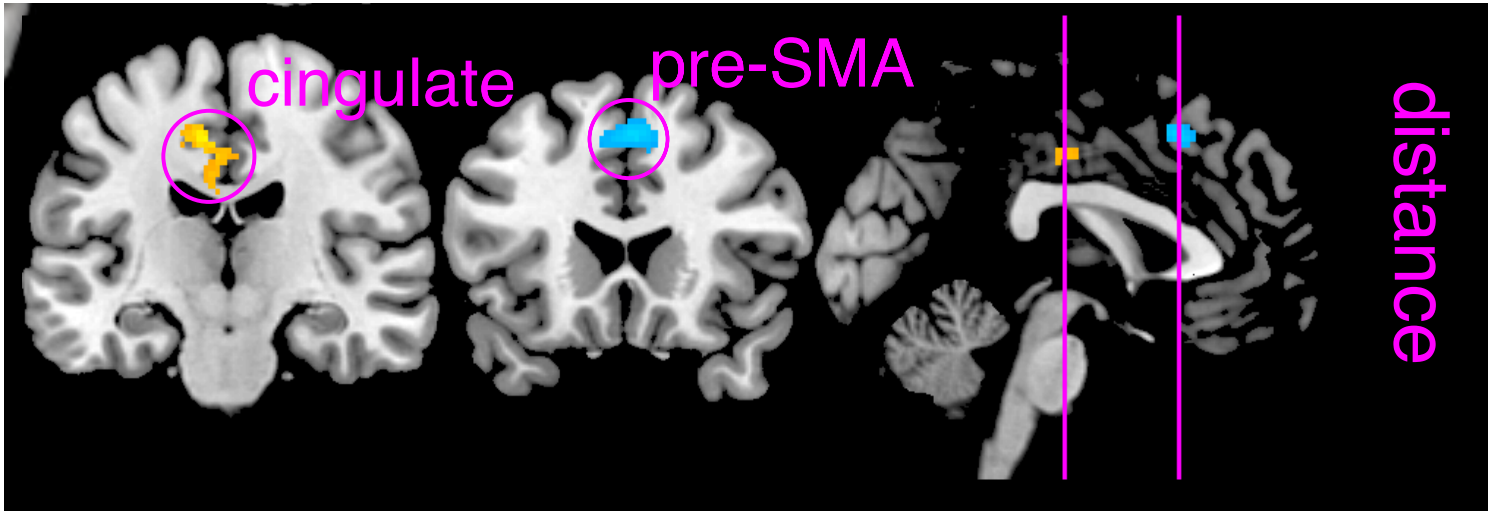
Bang & Fleming (2018) *PNAS*

Components of confidence - boundary distance





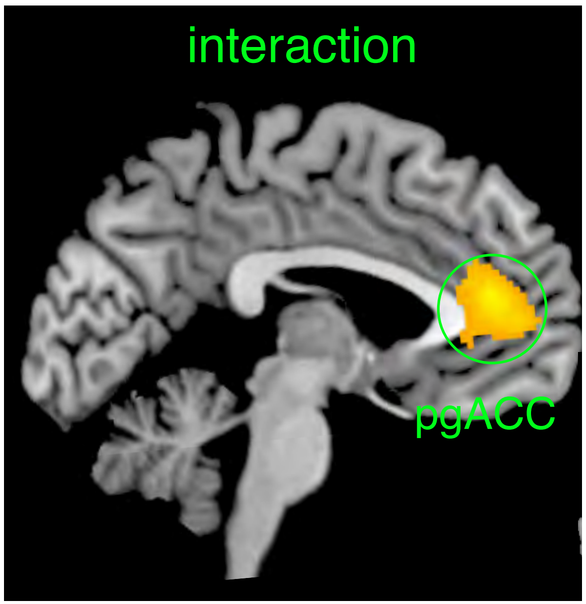
ROI computed using leave-one-out

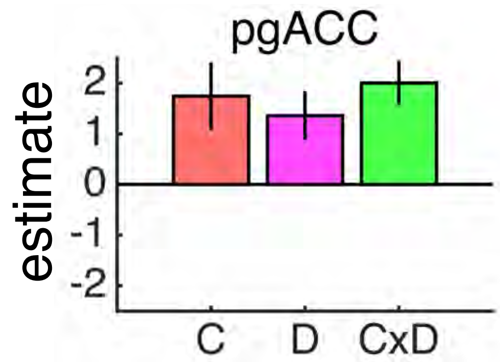


cluster-defining threshold: p < .001; pFWE < .05 corrected; correct trials

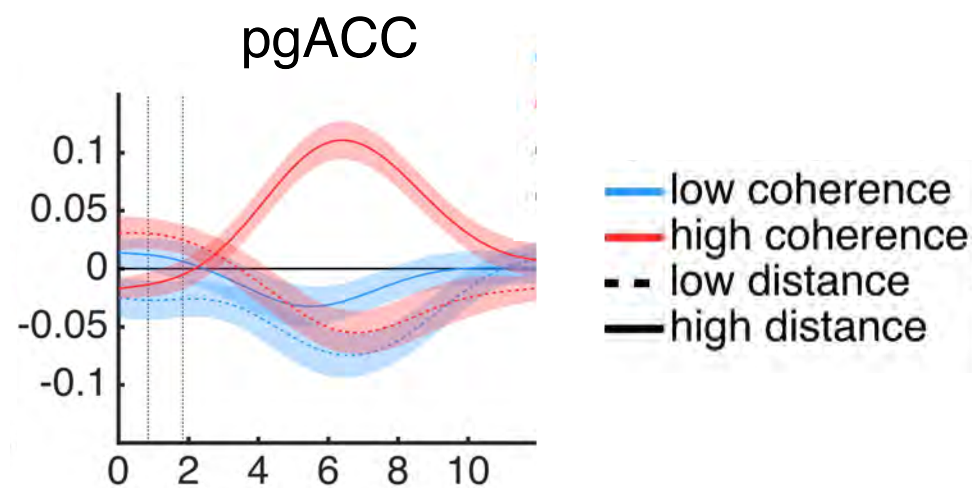
Bang & Fleming (2018) *PNAS*

Interaction of certainty and difficulty





ROI computed using leave-one-out



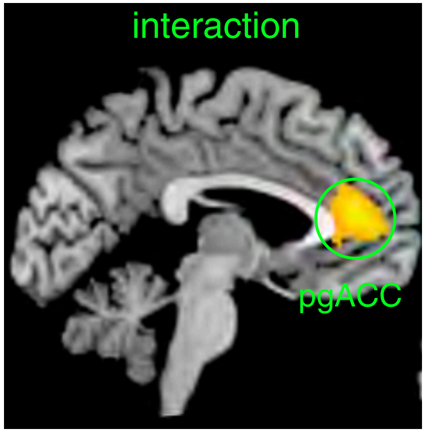
Bang & Fleming (2018) *PNAS*

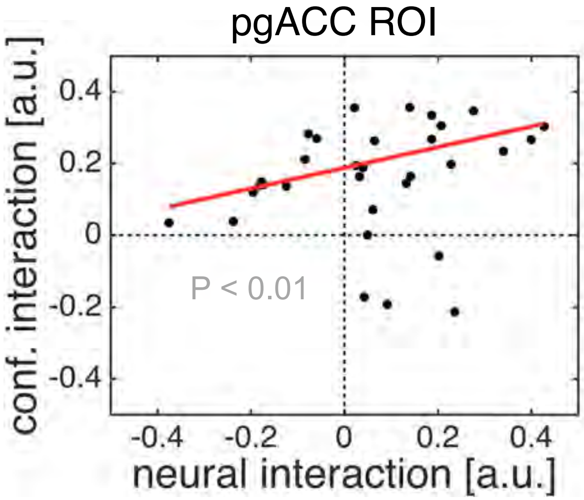
Individual differences in confidence

pgACC activation tracks degree of behavioural integration of certainty and distance in confidence reports

behavioural certainty x distance interaction

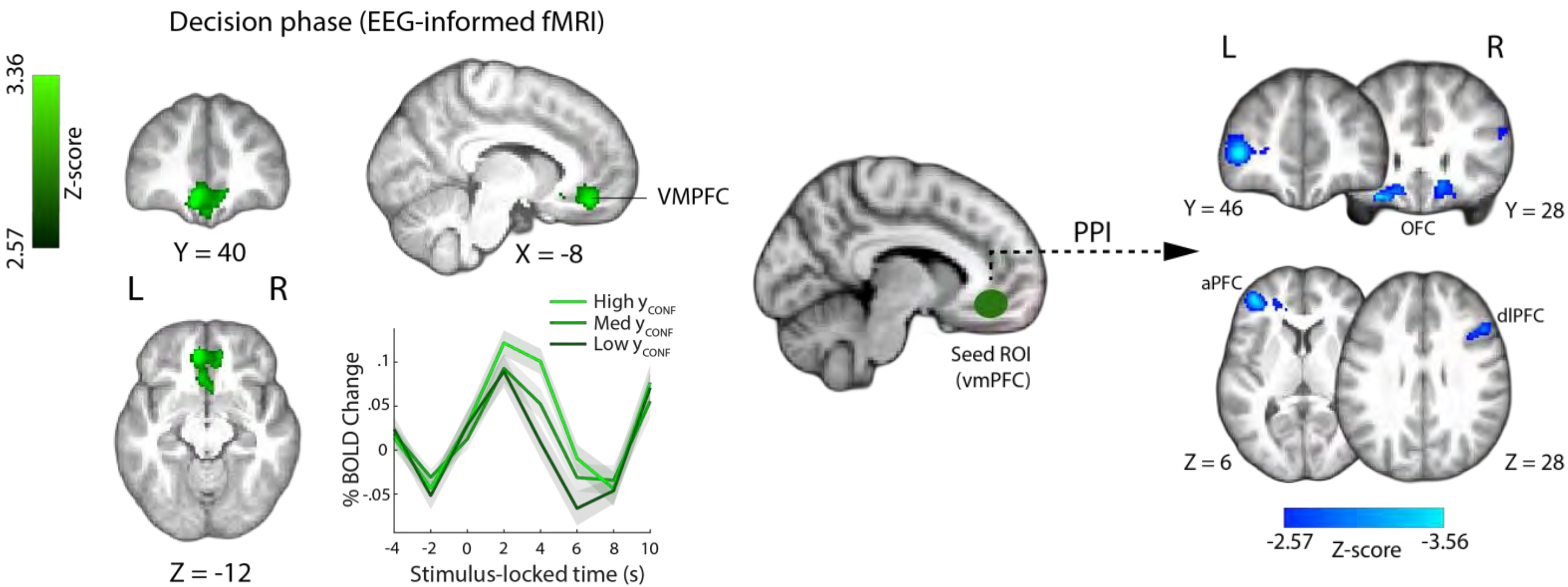
interaction contrast





Bang & Fleming (2018) *PNAS*

mPFC carries early confidence signals

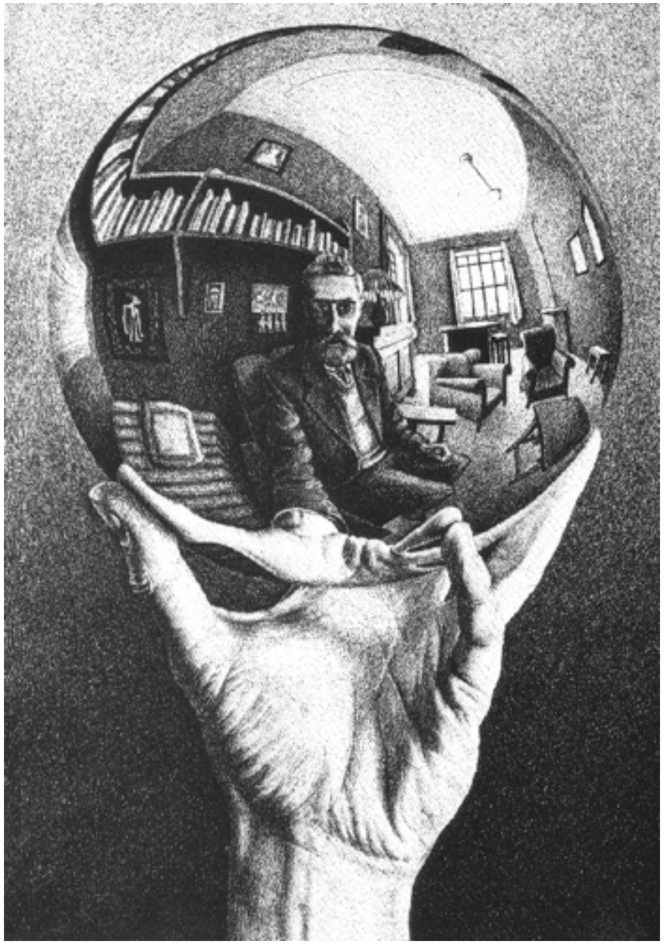


mPFC tracks early decision confidence; subsequently coupled with lateral aPFC to support metacognitive control / communication

Gherman & Philiastedes (in press) *eLife*

Outline

- Metacognition as a domain-general resource
- PFC as a hub for confidence formation
- Altered metacognitive beliefs in psychopathology



Metacognition and computational psychiatry

- Disorders of mental health are *subjectively* - **introspectively** - distressing

- One source of distressing beliefs may be **metacognitive distortions** - e.g. believing we are performing poorly when we are doing well; misinterpreting cognitive/emotional states

- Often **global**, applying to the *self* rather than any one particular task/measure/ aspect of life (e.g. depression, GAD)

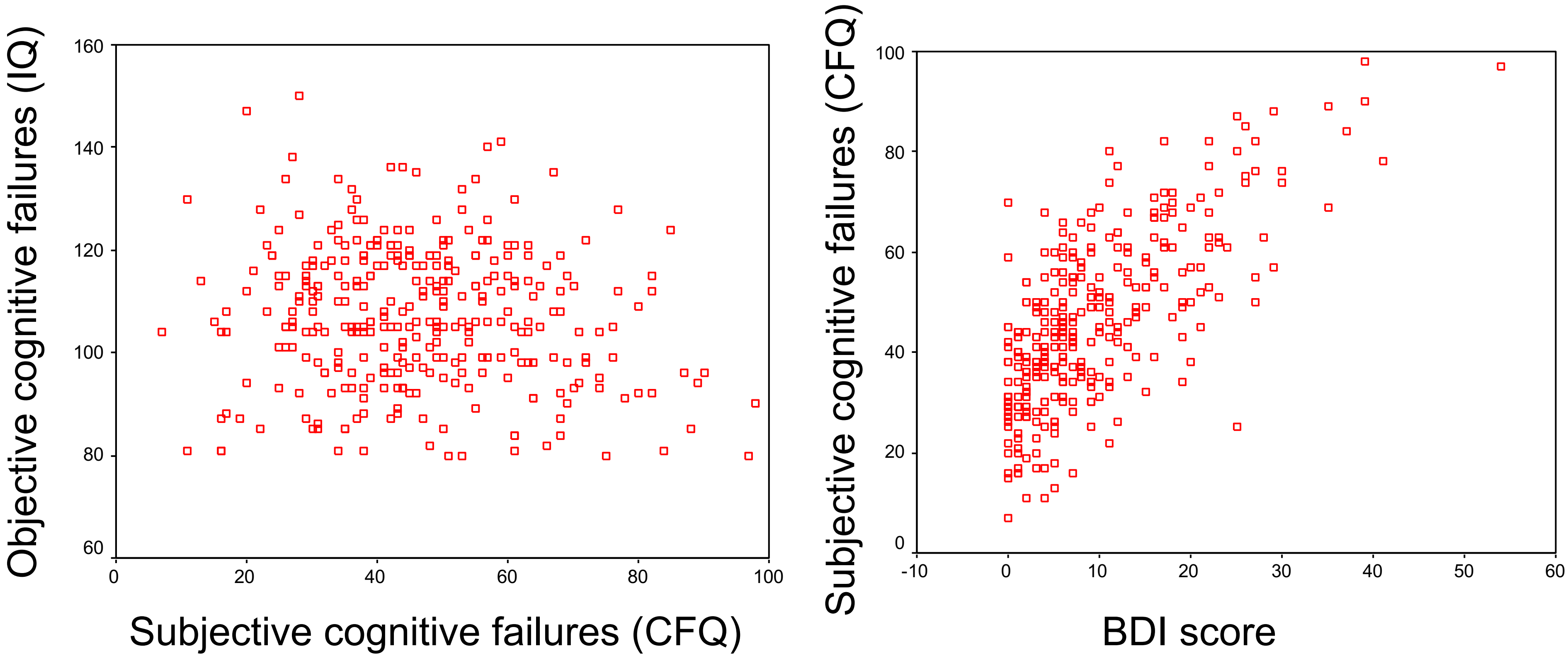
- In extreme cases, deficits in metacognition may lead to lack of insight / anosagnosia

David et al. (2012) *Phil Trans B*

Psychological Medicine, 2002, **32**, 1357–1370. © 2002 Cambridge University Press
DOI: 10.1017/S0033291702006359 Printed in the United Kingdom

Cognitive functioning and disturbances of mood in UK veterans of the Persian Gulf War: a comparative study

A. S. DAVID,[1] L. FARRIN, L. HULL, C. UNWIN, S. WESSELY AND T. WYKES
From the Gulf War Illnesses Research Unit, Guy's, King's and St Thomas' School of Medicine, King's College and Institute of Psychiatry, London



David et al. (2002) *Psych Med*

Metacognition and psychopathology

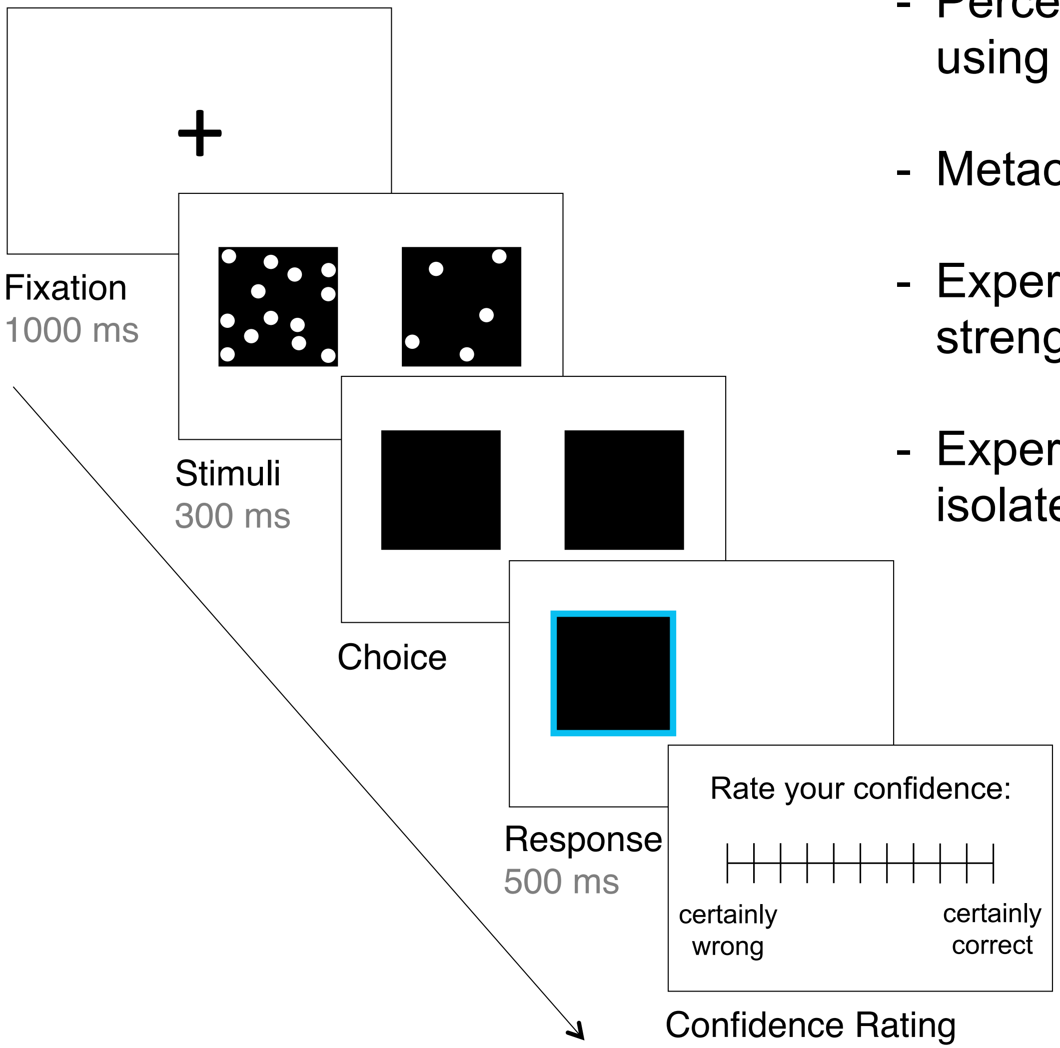
Experiment 1 - N=498 participants
Experiment 2 - N=497 participants

amazon mechanical turk

Perceptual decision-making task → Self-reported symptom questionnaires

Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Decision-making + confidence task



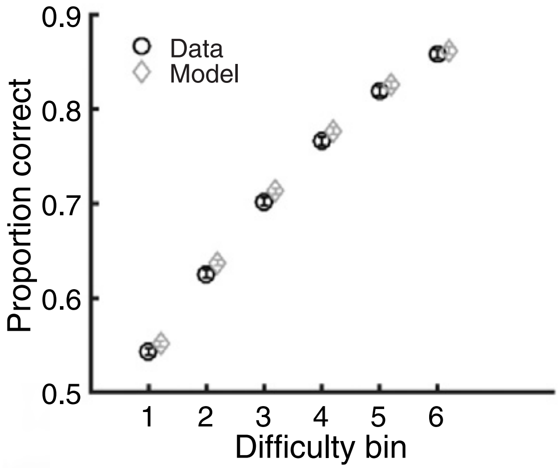
- Perceptual decision-making quantified using SDT and drift-diffusion modelling
- Metacognition quantified using meta-d'
- Experiment 1 = variable stimulus strength
- Experiment 2 = staircase used to isolate of metacognitive variability

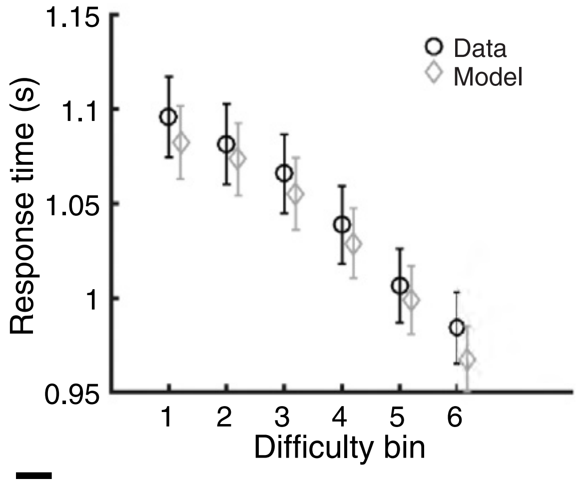
Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Decision-making + confidence task

Confidence ratings

Perceptual decision performance

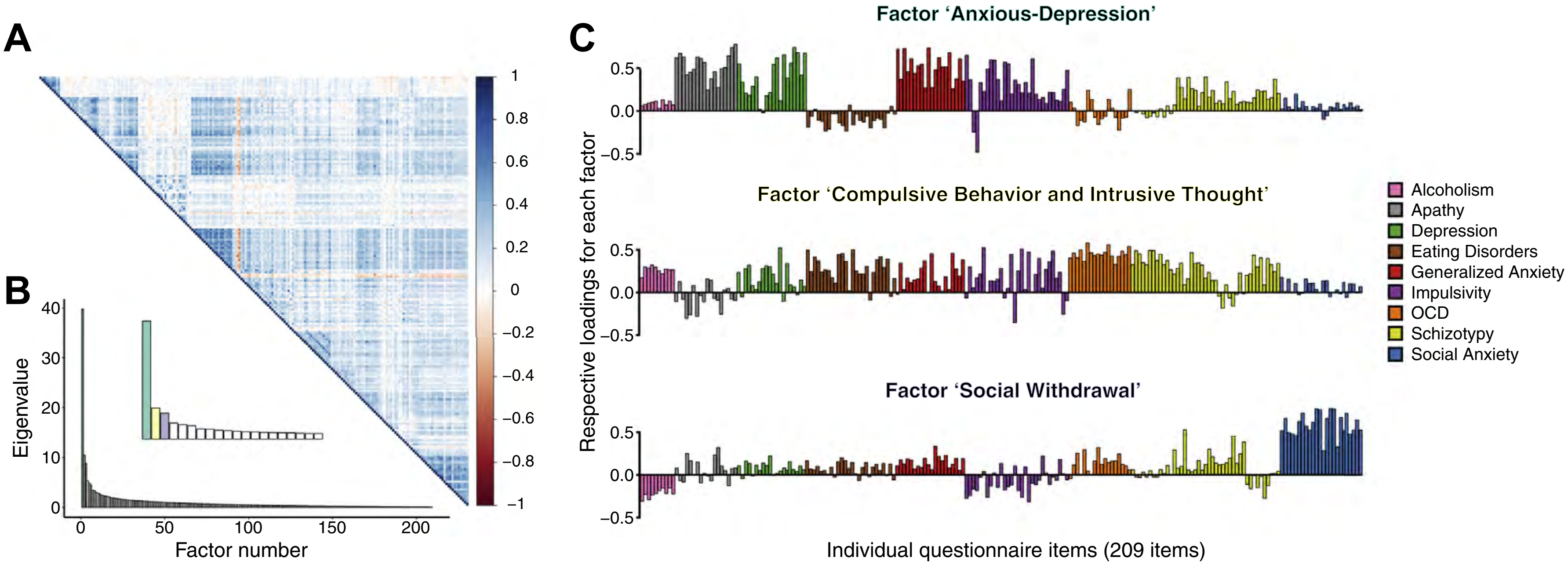




Experiment 1, N=498

Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Identifying latent transdiagnostic dimensions



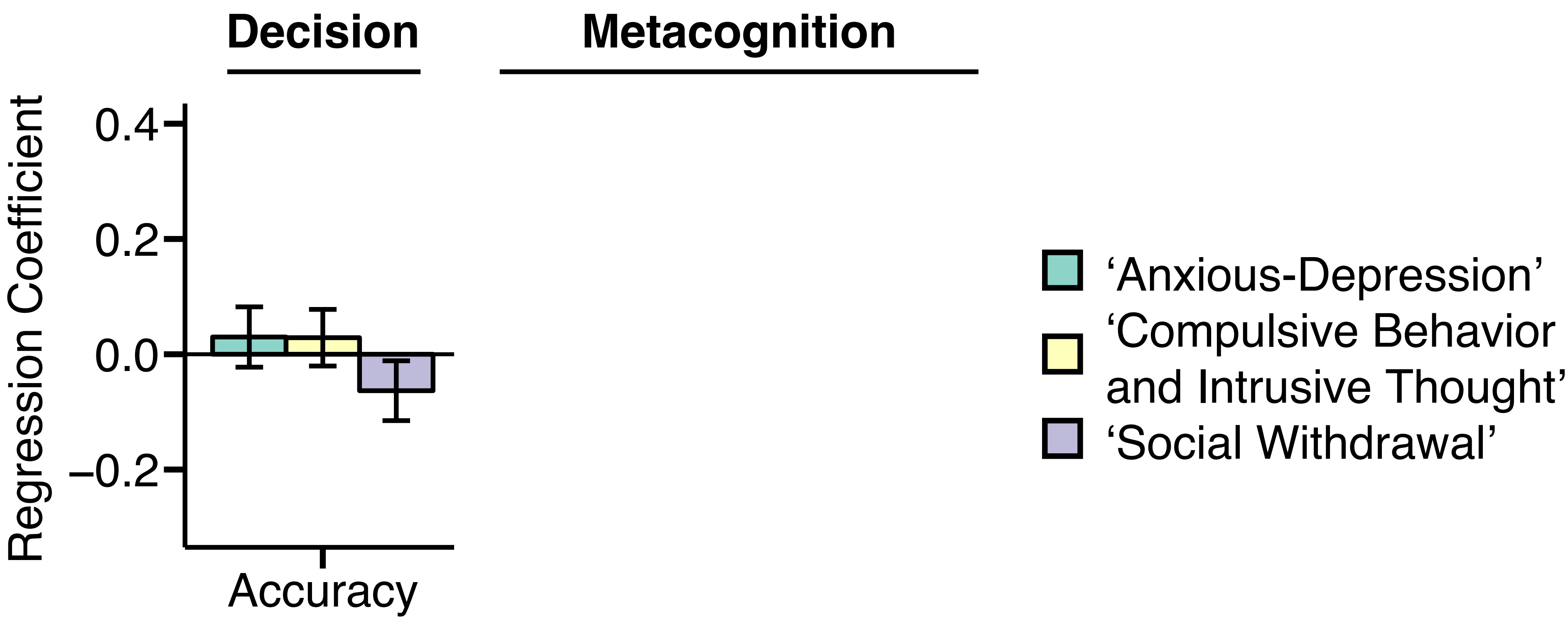
See also Gillan et al. (2016) eLife

Anxious/ Depression

Compulsive Behavior and Intrusive Thought

Social Withdrawal

Metacognition (but not decision performance) is associated with latent symptom dimensions



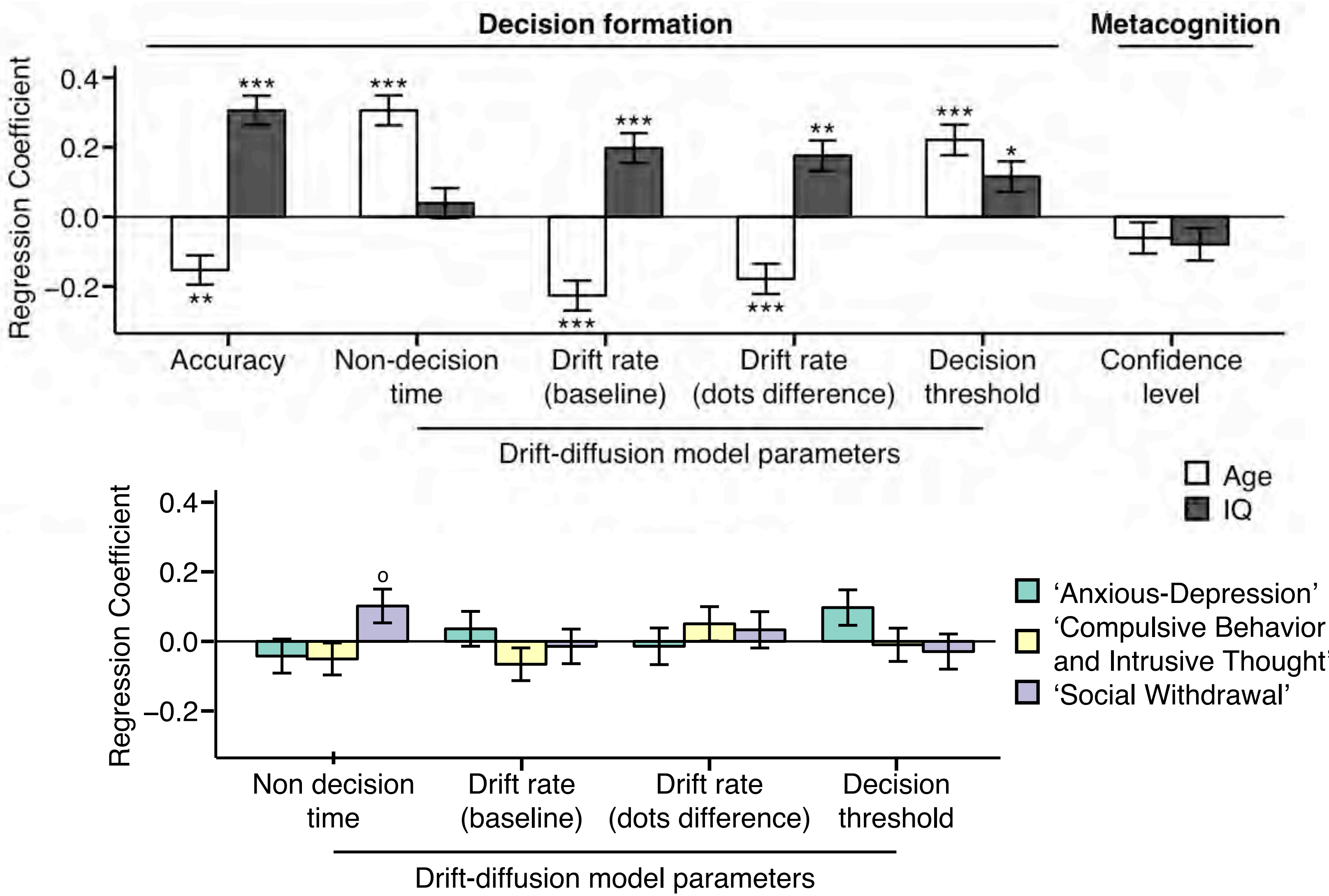
Experiment 2, N=497

Controlled for age, IQ, gender

**p*<.05 uncorr, ****p*<.001, corrected for multiple comparisons over number of dependent variables

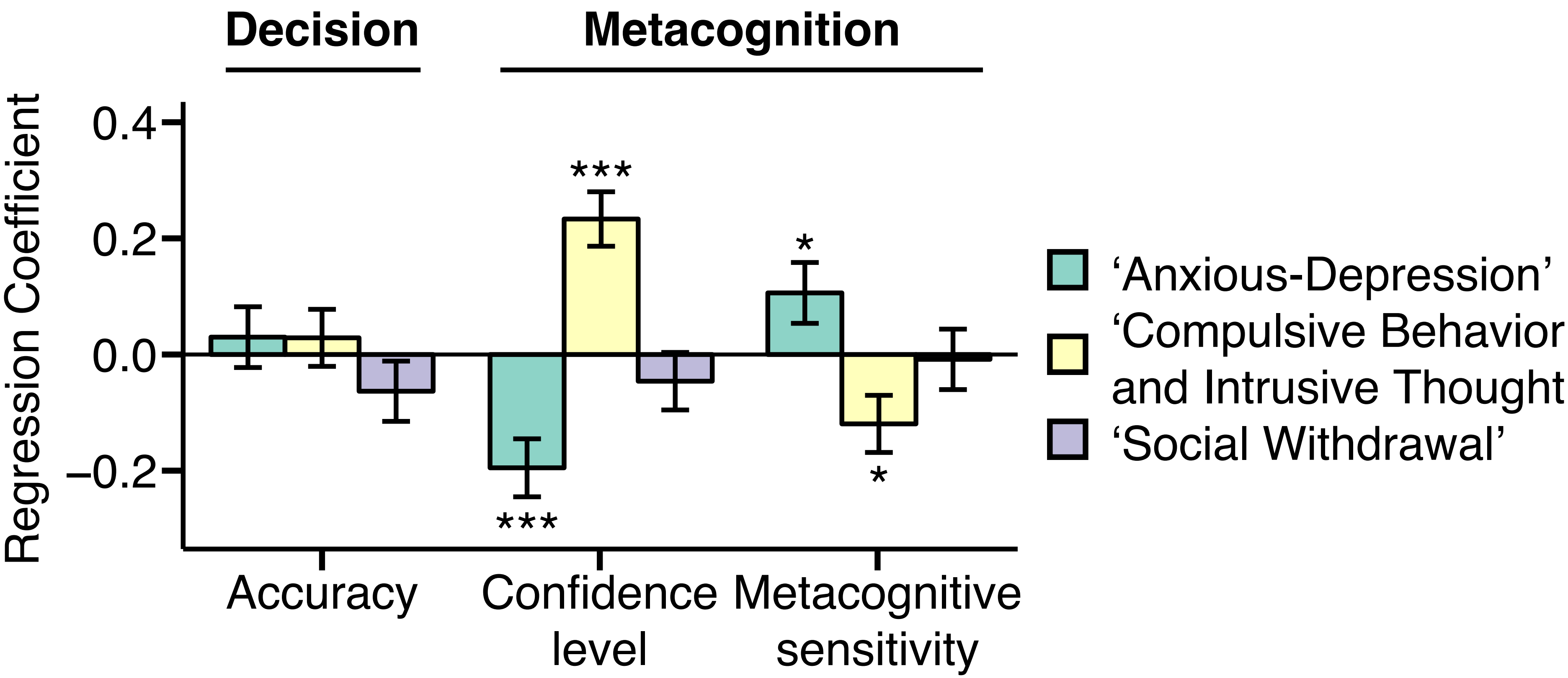
Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Dissociating metacognition and decision performance



Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Metacognition (but not decision performance) is associated with latent symptom dimensions



Experiment 2, N=497

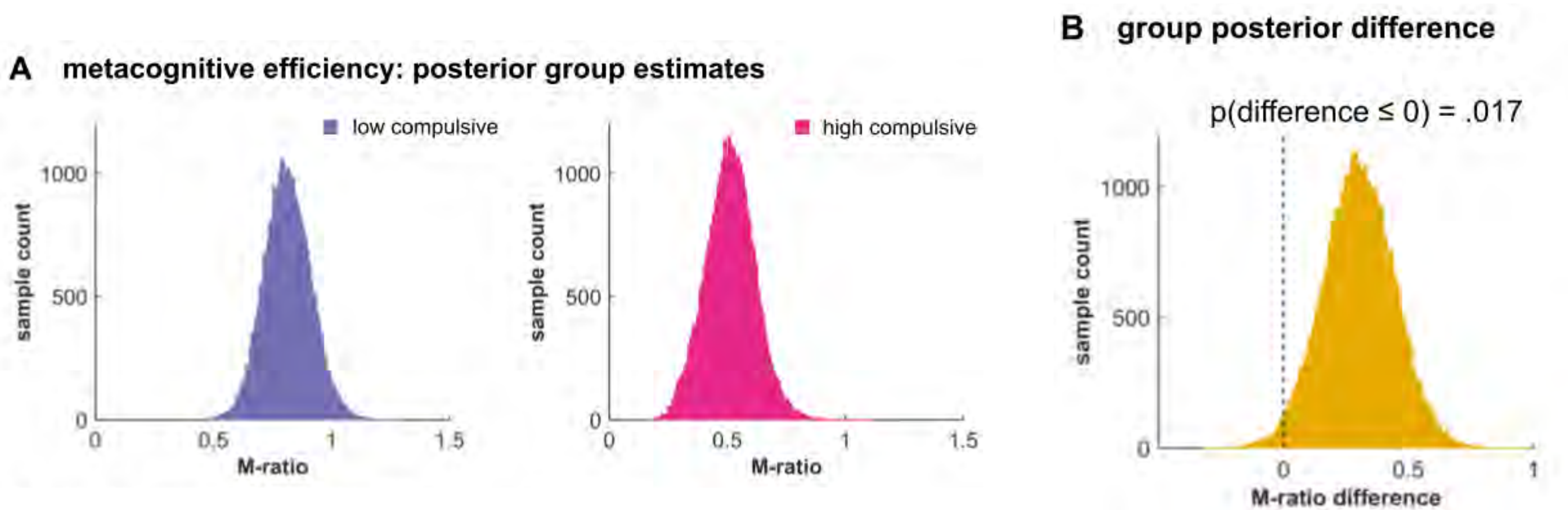
Controlled for age, IQ, gender

*p<.05 uncorr, ***p<.001, corrected for multiple comparisons over number of dependent variables

Rouault*, Seow*, Gillan & Fleming (2018) *Biological Psychiatry*

Metacognitive impairments extend perceptual decision making weaknesses in compulsivity

Tobias U. Hauser[1,2], Micah Allen[1,3], NSPN Consortium*, Geraint Rees[1,3] &
Raymond J. Dolan[1,2]

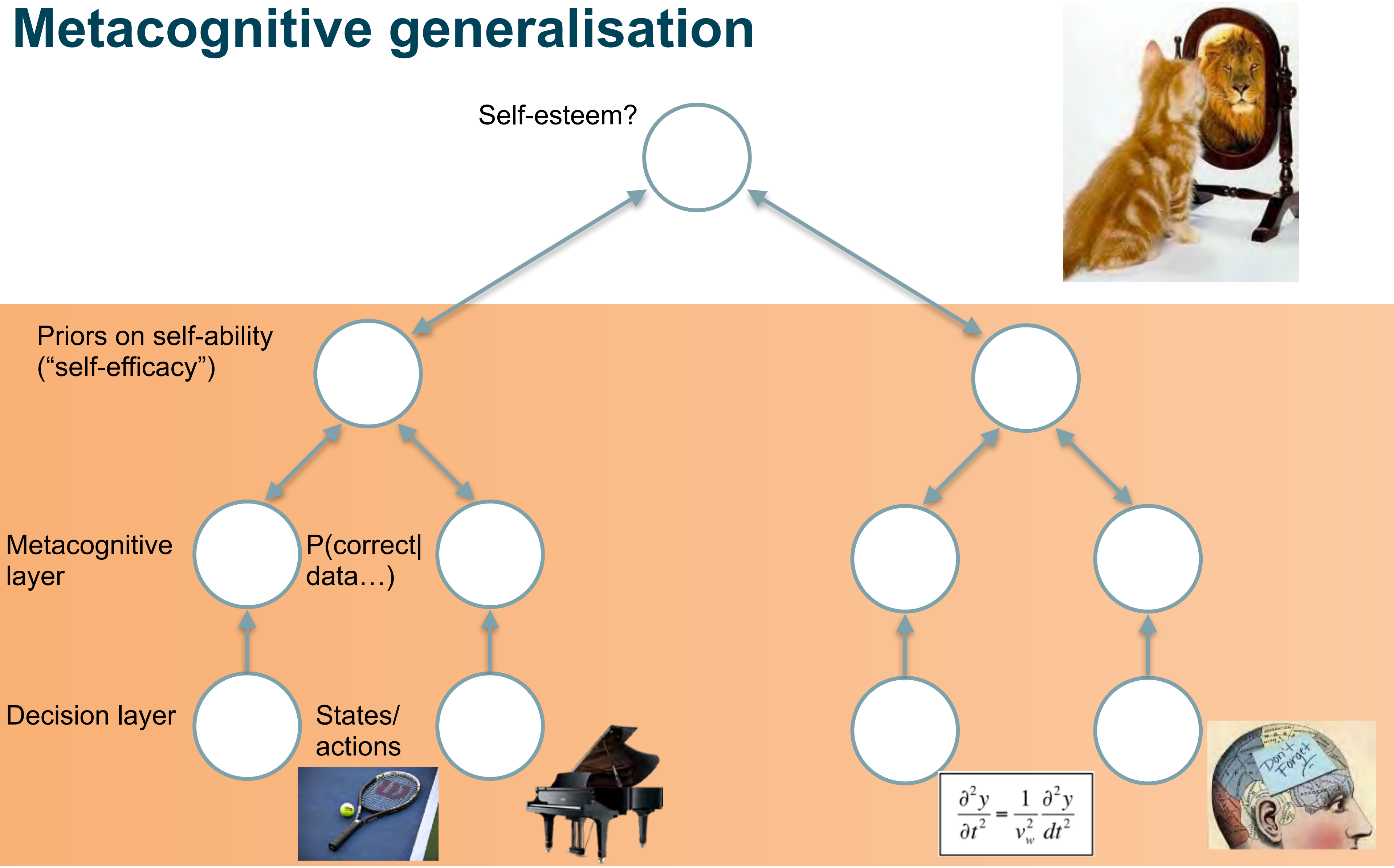


Hauser et al. (2018) *Sci Rep*

Summary

- We can measure metacognition across different tasks as the statistical association between behaviour and self-evaluation (confidence)
- Adopting a signal detection theory framework allows simultaneous estimation of both first-order (d') and metacognitive (meta-d') sensitivity
- Psychiatric symptom dimensions are associated with changes in metacognitive beliefs over and above differences in behavioural performance
- These confidence estimates are encoded in (domain-general?) mPFC activation patterns
- Modelling generalisation of metacognitive beliefs holds promise for understanding distorted self-beliefs / self-esteem

Metacognitive generalisation



Rouault et al. (2018) *Personality Neuroscience*

UCL

metacoglab.org

Dan Bang
Marion Rouault
Max Rollwage
Matan Mazor
Elisa van der Plas
Andrew McWilliams
Xiao Hu
Sara Ershadmanesh
Alisa Loosen
Rylan Schaeffer

NIH National Institutes of Health

THE ROYAL SOCIETY

MHJ mental health & justice

wellcome centre human neuroimaging

wellcome

MAX PLANCK UCL CENTRE for Computational Psychiatry and Ageing Research