Computations of confidence are modulated by mentalizing ability

E van der Plas[1,2,3], D Mason[3], LA Livingston[3,4], J Craigie[5], F Happé[3] and SM Fleming[1,2,6]

[1]Wellcome Centre for Human Neuroimaging, University College London, WC1N 3BG, London UK
[2]Department of Experimental Psychology, University College London, WCH1H 0AP London, UK
[3]Social, Genetic and Developmental Psychiatry Centre, King's College London, SE5 8AF London, UK
[4]School of Psychology, Cardiff University, CF10 3AT Cardiff, UK
[5]Centre of Medical Law and Ethics, Dickson Poon School of Law, King's College London, WC2R 2LS London, UK
[6]Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, WC1B 5EH London, UK

Email: elisa.plas.18@ucl.ac.uk

Code and data availability: All anonymized behavioural data and analysis scripts supporting the findings of this work are available on GitHub (github.com/metacoglab/MetaMenta-project).

Preregistration availability: https://osf.io/u6ecx/

Author Contributions: All authors conceptualized the project. EvdP and DM collected the data, EvdP analysed the data, EvdP and SMF wrote the paper with revisions from DM, JC, LAL, and FH. All authors approved the manuscript.

Competing Interest Statement: The authors declare no competing interest.

Keywords: Metacognition, mentalizing, autism spectrum condition, perceptual decision-making

This PDF file includes:

Main Text (8,180 words)
Figures 1 to 4

Supplementary Materials (1,295 words)
Supplementary Figures 1 to 6

ABSTRACT

Do people have privileged and direct access to their own minds, or do we infer our own thoughts and feelings indirectly, as we would infer the mental states of others? In this study we shed light on this question by examining how *mentalizing ability*—the set of processes involved in understanding other people's thoughts and feelings—relates to *metacognitive efficiency*—the ability to reflect on one's own performance. In a general population sample (*N* = 477) we showed that mentalizing ability and self-reported socio-communicative skills are positively correlated with perceptual metacognitive efficiency, even after controlling for choice accuracy. By modelling the trial-by-trial formation of confidence we showed that mentalizing ability predicted the association between response times and confidence, suggesting those with better mentalizing ability were more sensitive to inferential cues to self-performance. In a second study we showed that both mentalizing and metacognitive efficiency were lower in autistic participants (*N* = 40) when compared with age, gender, IQ, and education-matched non-autistic participants. Together, our results suggest that the ability to understand other people's minds predicts self-directed metacognition.

"The sorts of things that I can find out about myself are the same as the sorts of things that I can find out about other people, and the methods of finding them out are much the same."
– G. Ryle in The Concept of Mind (1949)

INTRODUCTION

In 1949, Ryle proposed that the cognitive mechanisms employed to understand ourselves are similar to those involved in understanding the feelings and experiences of other people (Ryle, 1949). Since then, various proposals have echoed Ryle in suggesting that *explicit metacognition*—the capacity for conscious evaluation of one's own mental states (Fleming et al., 2010; Fleming & Lau, 2014; Frith, 2012; Yeung & Summerfield, 2012) and *mentalizing*—the capacity to evaluate and understand other people's mental states (Abell et al., 2000; David et al., 2008; Rosenblau et al., 2015; White et al., 2009; White et al., 2011) have a common neurocomputational basis (Carruthers, 2009; Dimaggio et al., 2008; Fleming & Daw, 2017; Frith, 2012; Vaccaro & Fleming, 2018).

According to recent perspectives on the developmental trajectory of metacognition, while "core" or implicit mechanisms for self-monitoring and tracking uncertainty may be in place early in infancy (Goupil & Kouider, 2016), explicit metacognition emerges around the ages of 2-3 (e.g. Hembacher & Ghetti, 2014; see Goupil & Kouider, 2019 for a review), and continues to be shaped in childhood and adolescence (Fandakova et al., 2017; Weil et al., 2013). One potential driver of this continued development of explicit metacognition is that a growing understanding of other people's mental states may be used to refine awareness of ourselves (Carruthers, 2009). For example, repeatedly perceiving a parent expressing uncertainty together with their hesitation may allow a child to recognize and express uncertainty when they themselves are hesitating. This hypothesis predicts that introspection is not a distinct natural kind, but is instead grounded in the same processes used to understand the mental states of others (Carruthers, 2009; Gazzaniga, 1995, 2000; Gopnik, 1993; Wegner, 2002; Wilson, 2002). This view makes several testable predictions, for example, that people with a good mentalizing ability should also have good metacognitive ability; and that if children have problems with inferring the mental states of others (e.g., because of a neurodevelopmental condition such as autism), they may also develop difficulties with understanding their own minds.

The second prediction can be directly studied in the context of Autism Spectrum Condition (ASD)—a neurodevelopmental condition that is, in part, characterised by nonverbal and verbal communicative problems, deficits in socio-emotional reciprocity (American Psychiatric Association, 2013) and mentalizing difficulties (Happé, 2015; Livingston & Happé, in press). If our view is correct, difficulties with understanding other people's thoughts and social communication (as is typical in autism) should also affect the development of metacognition in this condition.

Metacognition is often quantified in laboratory tasks as the ability to provide accurate confidence ratings about self-performance in a range of cognitive domains. "Good" metacognitive ability is indicated by reporting lower confidence when wrong, and higher confidence when right (Fleming et al., 2010; Fleming & Lau, 2014; Frith, 2012; Yeung & Summerfield, 2012). This is known as metacognitive "sensitivity" and is distinct from metacognitive "bias", the tendency to be more or less confident overall (Fleming & Lau, 2014). Mentalizing, on the other hand, is often assessed as participants' ability to understand what agents are thinking or intending from observations of their actions and expressions (Abell et al., 2000; Baron-Cohen et al., 2001; White et al., 2011). "Good" mentalizing ability is indicated by correct assessment of others' mental states. To date, six studies have examined associations between metacognition and mentalizing in children or adults with autism

(Carpenter et al., 2019; Grainger et al., 2016; Nicholson et al., 2019; 2020; Wojcik et al., 2013; Williams et al., 2018). Three of the six papers suggest, in line with the idea that mentalizing and metacognition have a similar neuro-computational mechanism, that autistic individuals have metacognitive difficulties that are commensurate with their mentalizing capacity (Grainger et al., 2016; Nicholson et al., 2020; Williams et al., 2018). However, the remaining three studies did not find deficits in metacognition in autistic compared with non-autistic participants despite finding deficits in mentalizing ability (Wojcik et al., 2013; Carpenter et al., 2019). Taken together, the existing data indicate a link between metacognition and mentalizing, but not unequivocally so.

One difficulty with interpreting findings on metacognition is that its measurement is often confounded by other aspects of task performance, which itself may vary across individuals and clinical groups. For example, many of the studies reviewed above computed people's metacognitive sensitivity as the Goodman-Kruskall gamma correlation between trial-by-trial accuracy and confidence (Nelson, 1984), a measure known to be confounded by *type 1 sensitivity* (task performance) and *metacognitive bias* (people's average confidence scores) (Fleming & Lau, 2014; Maniscalco & Lau, 2012, 2014; Masson & Rotello, 2009; Rahnev & Fleming, 2019; **Figure 1a)**. The impact of this confound may be particularly pertinent in studies comparing autistic and non-autistic people, as sensory (hyper-) sensitivity (Ewbank et al., 2016; Lieder et al., 2019; Pirrone et al., 2017) and over-confidence (McMahon et al., 2016; Milne et al., 2002; Zalla et al., 2015) are sometimes found to be higher in autistic compared to non-autistic groups. In other words, previously reported measures of *metacognitive* sensitivity may have been confounded by higher *sensory* sensitivity in autistic participants.

A powerful approach to control for task performance confounds in studies of metacognition is to use model-based metrics derived from signal detection theory, that allow metacognitive sensitivity to be expressed in the same units as task performance while also controlling for metacognitive bias (meta-*d'*; Maniscalco & Lau, 2012, 2014). Notably, a recent study identifying a positive correlation between metacognitive and mentalizing ability when using this meta-*d'* metric to quantify metacognitive sensitivity (Nicholson et al., 2020). Nicholson and colleagues (2020) measured both implicit (behavioural) and explicit (verbal) metrics of choice uncertainty (defined as 'opting-out' from choosing or verbally reporting lower confidence, respectively) and measured mentalizing ability from participants' descriptions of short animations of abstract figures that vary in their level of intentionality (Abell et al., 2000). The authors found that explicit, but not implicit, metacognitive sensitivity was positively correlated with mentalizing ability, and significantly lower among autistic children. In a second study on neurotypical adults, the authors leveraged a dual-task condition in which participants completed a mentalizing or non-mentalizing-related cognitive task alongside a metacognition task and found that the dual mentalizing task significantly lowered metacognitive sensitivity compared to conditions in which the dual task did not require mentalizing (Nicholson et al., 2020). Together these findings suggested that mentalizing and metacognitive ability share a common neurocognitive basis which is commensurately impaired in autistic individuals.

However, despite this promising result, further limitations in the measurement of both mentalizing and metacognition in Nicholson et al (2020) are worth considering. First, mentalizing ability was scored from participants' written descriptions of the triangles' mental states. It has been proposed that this type of question is more prone to confounds of verbal fluency than, for example, multiple-choice assessments of mentalizing (White et al., 2011). This may be particularly problematic in studies of autism given that differences in verbal fluency are commonly observed in this condition (Livingston, Carr, et al., 2019; Livingston et al., in press; Spek et al., 2009). Second, in the metacognition task, decisions were of varying choice difficulty, with some perceptual discriminations (of colour, or dot density) being easier than others. When task difficulty is varying between trials and subjects, it may affect measures of metacognitive ability, even when d’ is controlled for (Rahnev & Fleming, 2019). Finally, participants received trial-by-trial feedback on their confidence ratings, where they were rewarded for reporting higher confidence on correct trials and lower confidence on error trials (i.e., better metacognition was incentivized). This may have created a disadvantage for autistic participants who may have difficulties with interpreting and learning from ambiguous or implicit feedback (Broadbent & Stokes, 2013; Greene et al., 2019; Reed, 2019; Robic et al., 2015; Sapey-Triomphe et al., 2018; Zwart et al., 2018). In other words, it could be that the lower metacognitive ability in the autistic group was a consequence of failing to maximize rewards on the basis of the ambiguous feedback.

Across two studies, we set out to control for some of the factors that might have influenced the results of these previous studies by adopting experimental and computational methods that are considered optimal for the assessment of metacognitive sensitivity (Rahnev & Fleming, 2019; Fleming, 2017). Specifically, we measured metacognition using a psychophysical task on which participants make repeated perceptual judgements and rated their confidence in being correct. In order to match sensory sensitivity across participants and over the course of the experiment within the same participant, we employed a staircase procedure that continually adjusted sensory evidence strength on the basis of people’s responses. In addition, we measured the same participants’ *mentalizing ability* on a separate task in which they watched short animations of abstract figures that moved across the screen according to distinct types of interaction (Abell et al., 2000), similar to that used by Nicholson et al. (2020). Instead of providing a verbal description of each interaction, participants indicated their answer using multiple choice selection (White et al., 2011; Livingston et al., in press). We controlled for type 1 performance in the measurement of metacognition by computing *metacognitive efficiency* (meta-*d'/d'*), which controls for type 1 sensitivity and metacognitive bias using the meta-*d'* model (Maniscalco & Lau, 2012, 2014). Moreover, we estimated metacognitive efficiency within a Bayesian hierarchical model that allows optimal estimation of the relationship between metacognitive efficiency and individual differences in mentalizing ability, while also taking into account uncertainty surrounding each individual subject's parameter estimates (Fleming, 2017; Harrison et al., 2020).

Having confirmed a link between metacognition and mentalizing, in a second set of analyses we investigated *how* the computation of confidence is modulated by mentalizing ability by building hierarchical regression models of trial-by-trial confidence ratings. We reasoned that, if metacognition and mentalizing rely on similar inferential processes and cues, mentalizing ability should facilitate the use of behavioural cues that are similarly predictive of the mental states of others. Work in cognitive psychology has often shown that people have poor access to the reasons for their actions but instead infer these from contextual cues (even if these cues are experimentally decoupled from the true underlying intention; Gazzaniga, 1995, 2000; Nisbett & Wilson, 1977; Wegner, 2002; Wilson, 2002). For example, when asked to rate their confidence in a previous decision, people's confidence reports may be affected by various (behavioural) cues that are more or less related to the decision, such as response times (Kiani et al., 2014; Patel et al., 2012), social context (Bang et al., 2017, 2020; Van der Plas et al., 2021), as well as the quantity and reliability of evidence (Campbell-Meiklejohn et al., 2010; De Martino et al., 2017; Kiani & Shadlen, 2009; Pleskac & Busemeyer, 2010). Intriguingly, response times have been shown to have a causal impact on the confidence levels people ascribe not only to themselves, but also to others (Palser et al., 2018; Patel et al., 2012). In a series of exploratory analyses, we therefore asked whether confidence was more tightly coupled to response times among participants with better mentalizing ability.

In two independent behavioural experiments, we tested three pre-registered premises of the hypothesis that metacognition and mentalizing are inter-related, namely that: (1) metacognition and mentalizing ability are positively correlated, even after controlling for first-order performance; (2) metacognitive efficiency is lower in people with autism, and in participants with greater autistic traits; and (3) especially in those with greater difficulties with social communication and understanding but not non-social autistic traits. We also assessed the extent to which response times predict confidence on a trail-by-trial level by conducting exploratory hierarchical regression models, asking whether the predictions of confidence interacted with mentalizing ability.

METHODS

Experiment 1.

Participants. We recruited N = 501 proficient English speaking participants via Prolific (https://www.prolific.com), a recruitment platform more representative of real populations than standard student samples (Palan & Schitter, 2018). All participants accessed the experiment with a desktop computer or laptop (no tablets or smartphones). Exclusion criteria were responding incorrectly to a "catch" question (e.g., "If you are still paying attention, please select x as your answer"); performing below or above pre-defined accuracy cut-offs (60% and 90% respectively) on the metacognition task; or rating the same confidence on more than 90% of the trials on the metacognition task. This resulted in the exclusion of N = 23 participants (5% of the total sample), leaving N = 477 participants for further analysis (168 female, mean age: 28.73, SEM = 0.52 years). All participants gave informed consent before the experiment, which was approved by the University College London Ethics Committee (1260/003).

Metacognition task. Stimuli were programmed in JavaScript using JSPsych (version 5.0.3) and hosted on the online research platform Gorilla (https://gorilla.sc//). Participants made 168 decisions across four blocks concerning which box was filled with a higher density of dots (left or right, indicated by pressing the "W" or "E" key, respectively without a time limit). The boxes were two black squares (each 250 x 250 pixels) which were each subdivided into grids of 625 cells that were filled with 313 dots. Choice difficulty was manipulated by adjusting the dot difference between boxes according to a "2-down-1-up" staircase procedure: dot difference increased after every error and decreased after two consecutive correct answers. Dots seemed to flicker, an effect created by replotting five different configurations of the same dot difference level for 150 ms each, for a full stimulus duration of 750 ms (Rollwage et al., 2018). On 26 practice trials participants received immediate feedback. During the remaining trials, participants did not receive feedback but had to rate their confidence that their decision was correct (on a scale from 1 “*Guessing*” to 6 “*Certainly correct*”, without a time limit; Rouault et al., 2018).

Mentalizing task. We used a validated online version of the Frith-Happé Triangle Task (Abell et al., 2000; Livingston et al., in press). Participants were shown twelve short (34-35 sec.) animations of one large red and one small blue triangle. The way in which the triangles moved was manipulated across three conditions: in random animations they moved purposelessly around; in Goal-Directed animations they interacted behaviourally; and in four Theory of Mind (ToM) animations they interacted in a way that involves responding to the other’s mental states. Participants were scored on their accuracy in classifying which category the interaction pertained to (mentalizing classification) giving a score ranging between 0-12 (i.e., participants could score one point after each animation). In addition, we computed participants’ accuracy in categorizing the feelings of the triangles (mentalizing ability; White et al., 2011). Mentalizing ability was scored as the number of correctly identified mental states of each of the two triangles, after each ToM animation that had been correctly identified in the mentalizing classification question. This type of mental state attribution requires tracking the triangle’s intentions throughout the animation and cannot simply be deduced from the general kinematics of the triangle, therefore making it less susceptible to compensatory strategies. Participants had to watch the complete animation before the questions appeared, after which they were allowed to decide without a time limit. All animations were presented in pseudo-randomized order and after three practice animations on which participants received immediate feedback.

Additional measures. After the two computer tasks, which were presented in counterbalanced order, the following questionnaires were administered: (1) the Autism

Quotient-10 (AQ-10) a brief assessment of autistic traits (a higher score indicates more autistic traits; Allison et al., 2012); (2) the RAADS-14, a screening tool for autistic traits in adult populations which asks whether each trait was present either in childhood, currently, both or neither (with a higher score indicating more autistic traits; Eriksson et al., 2013); (3) the Beck Cognitive Insight Scale (BCIS) an assessment of people's ability to distinguish between objective reality and subjective experience (Beck et al., 2004); and (4) the International Cognitive Ability Resource (ICAR) a brief assessment of fluid intelligence (Condon & Revelle, 2016). More details on these questionnaires are provided in **Supplementary Materials**.

Statistics. The hypotheses and analyses for this study were pre-registered (https://osf.io/vgy7a/). Validation checks are reported in the **Supplementary Material** and consisted of Spearman's rho correlations (which are recommended for ordinal data) to assess relationships between main composite survey scores. Equal variances were assumed if not otherwise specified. We report *P* values at a 0.05 alpha level and the 95% confidence interval (95% CI) of the test statistic. Type-1 cognitive and type-2 metacognitive parameters were estimated using the open source HMeta-d toolbox (https://github.com/metacoglab/Hmeta-d) implemented in MATLAB (version 9.7.0). Type-2 $meta - d'$, the ability to determine one's accuracy with confidence ratings, was inferred using Markov chain Monte Carlo (MCMC) Bayesian sampling procedures using JAGS (http://mcmc-jags.sourceforge.net) across 30,000 samples after a burn-in of 1,000 samples distributed across three chains. Our parameter of interest was Mratio (meta-*d'*/*d'*), or metacognitive efficiency, which expresses metacognitive sensitivity (*meta-d'*) relative to task performance (*d'*; in other words, an Mratio of 1 implies participants have optimal metacognitive efficiency; Fleming, 2017).

We assessed model convergence for each HMeta-d model by ensuring that the consistency of the posteriors within and between chains, the Gelman-Rubin (G-R) $\hat{R}$ statistic, was below 1.1 (Gelman & Rubin, 1992) and by visually inspecting the chains (**Supplementary Materials**). In addition, each reported model was checked for reliability by conducting posterior predictive checks which are summarized in the **Supplementary Materials**.

To test the first pre-registered hypothesis of a positive association between metacognitive efficiency and mentalizing ability, we incorporated a simultaneous hierarchical estimation of the beta coefficient (β) of the impact of our standardized mentalizing ability score, $menta,$ on the log of metacognitive efficiency, $log(Mratio)$:

$$log(Mratio)_s \sim log(Mratio)_0 + \beta\, menta_s + \varepsilon_s \qquad (1.1)$$

$log(Mratio)_0$ denotes baseline group-level metacognitive efficiency; $menta_s$ is the mentalizing score for subject s; and ε_s refers to noise that is drawn from a T-distribution with variance σ_δ and 5 degrees of freedom, multiplied by a noise parameter ζ. We used priors found to provide the most efficient regression parameter recovery (Harrison et al., 2020), which were drawn from Gaussians $N(\mu, \sigma)$, half-Gaussians $HN(\mu, \sigma)$ and T-distributions $T(\mu, \sigma, df)$:

$$meta_0 \sim N(0,1)$$
$$\beta \sim N(0,1)$$
$$\sigma_\delta \sim HN(1)$$
$$\zeta \sim Beta(1,1)$$
$$\delta_s = T(0, \sigma_\delta, 5)$$
$$\varepsilon_s = \zeta * \sigma_\delta$$

The highest density interval (HDI) represents the ‘credible’ posterior range within which 95% of the estimated regression coefficient falls. We plotted the HDI for the regression coefficient and assessed significance by computing the probability that it differed from zero: $P_\theta(HDI < 0 \mid HDI > 0)$, where a higher probability suggests a stronger effect (Kruschke, 2010).

We also calculated $log(Mratio)_s$ at the individual level for use in post-fit frequentist analyses. We used a linear model with $log(Mratio)_s$ as the dependent variable and $menta_s$ and covariates (standardized age, IQ, gender [-1: female, 1: male] and education (edu) [1: no education, 2: high school or equivalent, 3: some college, 4: BSc, 5: MSc, 6: doctoral]) as independent variables:

$$\log(\text{Mratio})_s \sim \log(\text{Mratio})_0 + \beta_1 \text{menta}_s + \beta_2 \text{age}_s + \beta_3 \text{IQ}_s + \beta_4 \text{gender}_s + \beta_5 \text{edu}_s + \varepsilon_s \qquad (1.2)$$

To test the effect of autistic traits on $log(Mratio)_s$ we ran the same models specified in Equations 1.1 and 1.2 but now replacing menta_s with the RAADS-14 main composite autistic trait scores (Eriksson et al., 2013). In preliminary analyses we failed to replicate previous findings of a negative correlation between mentalizing ability and AQ-10 scores (Allison et al., 2012), and therefore (deviating from our pre-registration plan) we decided to conduct all further analysis of questionnaire data using RAADS-14 scores alone (Bertrams, 2021).

To assess the effects of trial-by-trial standardized (log) response times $logRT$ and accuracy on confidence, we conducted hierarchical mixed-effect regression models using the 'lme4' package in R (version 3.3.3) and plotted the standardized fixed-effect beta coefficients of the model fits. We obtained the P-values of the regression coefficients using the *car* package. All models include a random effect at the participant level and all statistics are computed at the group level. We report type III Wald chi-square tests (χ^2), degrees of freedom (*df*) for fixed effects, and estimated beta-coefficients (β) together with their standard errors of the mean (± SEM) and *P*-values of the associated contrasts.

To investigate if $logRT$ informs confidence differently as a function of individual differences in autistic traits, we tested whether a hierarchical mixed-effect regression model better predicts trial-by-trial confidence (conf) when the predictor variables accuracy (acc) [-1: error, 1: correct], z-score of the log response time (RT) and their interactions (Equation 2.1) were allowed to vary as a function of individual differences in standardized autistic trait scores (ASD; Equation 2.2.):

$$conf \sim acc + logRT + acc{:}logRT \; + \; (1 + acc + logRT + acc{:}logRT | subj) \qquad (2.1)$$

$$conf \sim \mathrm{ASD}{:}(acc + logRT + acc{:}logRT) + (1 + acc + logRT + acc{:}logRT | subj) \qquad (2.2)$$

The results of the Likelihood Ratio Test are expressed in terms of the *Akaike Information Criterion* (*AIC*): $\Delta\mathrm{AIC} = \mathrm{AIC}_{\text{Equation 2.1}} - \mathrm{AIC}_{\text{Equation 2.2}}$, and the *Log Likelihood* (*LL*): $\Delta\mathrm{LL} = \mathrm{LL}_{\text{Equation 2.1}} - \mathrm{LL}_{\text{Equation 2.2}}$ with associated *P* values extracted from a type III Wald chi-square tests (χ^2).

Experiment 2

Participants. We recruited a sample of N = 43 autistic participants via the research charity Autistica (www.autistica.org.uk). Interested participants first completed an online pre-screening questionnaire that included questions about mental health and demographics. Participants that met the inclusion criteria (i.e., aged between 18 and 50 years old and a self-reported diagnosis of autism spectrum condition by a health professional) were sent a link to the online experiment that could be accessed with a desktop computer or laptop (no tablets or smartphones). Exclusion criteria were the same as in Experiment 1. Three participants were excluded: one participant performed below the *a priori* accuracy cut-off and two participants performed above the *a priori* accuracy cut-off. This resulted in the exclusion of N = 3 participants (7.5% of the total sample, which is consistent with Experiment 1), leaving data from N=40 participants for analysis (37 female, mean age: 37.90, SEM = 1.59 years). All participants gave informed consent before experiment onset which was approved by the Research Ethics Office at King’s College London (HR-19/20-17704).

To obtain an equal number of comparison participants we re-analysed a subset of the dataset from Experiment 1 which used the same experimental paradigms and questionnaire battery. The dataset from Experiment 1 consisted of N = 477 English speaking participants from the general population (198 female, mean age: 28.73, SEM = 0.52). Data on mental health conditions was not collected. To ensure that the participants from this dataset provided a comparison group with low autistic traits, we first reduced the number to N = 97 participants scoring in the lowest 50% quantile of RAADS-14 and AQ-10 responses (a score lower than 16 and 5, respectively, which is more stringent than the clinical cut-off score; Ashwood et al., 2016; Eriksson et al., 2013). Next, to ensure the groups were well-matched on other characteristics, for each included autistic participant we manually selected a comparison participant of similar gender (a high proportion of females in the autism group meant that it was not possible to find a 1:1 gender match for three participants); who was within ±5 years from the target age; ±2 levels from the target education; and ±5 ICAR points from the target fluid intelligence level. These criteria were identified after initial exploration indicated they provided sufficient flexibility to provide a reasonable match between the two groups on all relevant dimensions. Importantly, participant selection was carried out prior to hypothesis testing.

Experimental paradigm. The experimental procedure was the same as in Experiment 1.

Statistics. Statistical inference was conducted similarly to analysis of Experiment 1. Validation checks are reported in the **Supplementary Material**. To investigate if metacognitive efficiency was different between the autism and comparison group, we fitted a linear model with $Mratio$ from a single-subject fit as dependent variable, clinical group [autism: -0.5, comparison: 0.5] and covariates (standardized age, IQ, gender [-1: female, 1: male] and education (edu) [1: no education, 2: high school or equivalent, 3: some college, 4: BSc, 5: MSc, 6: doctoral]) as independent variables:

$$log(Mratio)_s \sim \beta_0 + \beta_1 \mathrm{group}_s + \beta_2 \mathrm{age}_s + \beta_3 \mathrm{IQ}_s + \beta_4 \mathrm{gender}_s + \beta_5 \mathrm{edu}_s + \varepsilon_s \qquad (3.1)$$

We also conducted hierarchical regressions using the HMeta-d toolbox in which $Mratio$ in the autism and comparison groups were estimated in separate models that controlled for the following covariates:

$$meta_s \sim \beta_0 + \beta_1 \mathrm{age}_s + \beta_2 \mathrm{IQ}_s + \beta_3 \mathrm{gender}_s + \beta_4 \mathrm{edu}_s + \varepsilon_s \qquad (3.2)$$

To assess significance, we computed the probability P_θ of overlap between the HDI posterior distribution of $Mratio$ in the autism and comparison group:

$$P_\theta(HDI_{autism} < HDI_{comparison})$$

To assess whether the effect of $logRT$ on confidence was different for autistic and comparison participants, we conducted hierarchical mixed-effect regression models using the “lme4” package in R (version 3.3.3), similar to the method used in Experiment 1, but now using a dummy variable denoting clinical group (group [autism: -0.5, comparison: 0.5]) instead of continuous autistic trait scores. To visualize the direction of significant effects we obtained the beta-coefficients of $logRT$ on confidence for each clinical group and on error and correct trials, separately:

$$conf_{acc/group} \sim \beta_0 + \beta_1 \mathrm{logRT}_s + \beta_2 \mathrm{gender}_s + \beta_3 \mathrm{edu}_s + \varepsilon_s \qquad (3.3)$$

RESULTS

Experiment 1

The staircase converged to a stable performance level within and between participants (choice accuracy: M = 75%, SEM = 0.23). Given that staircase variability can affect estimates of metacognitive sensitivity (Rahnev & Fleming, 2019), we also computed each individual's experienced stimulus variability (the ratio between the standard deviation of stimulus difficulty and average stimulus difficulty) and established that stimulus variability was not correlated with metacognitive efficiency (rs_{475} = -0.068, P = 0.137; **Supplementary Figure 1.2b**).

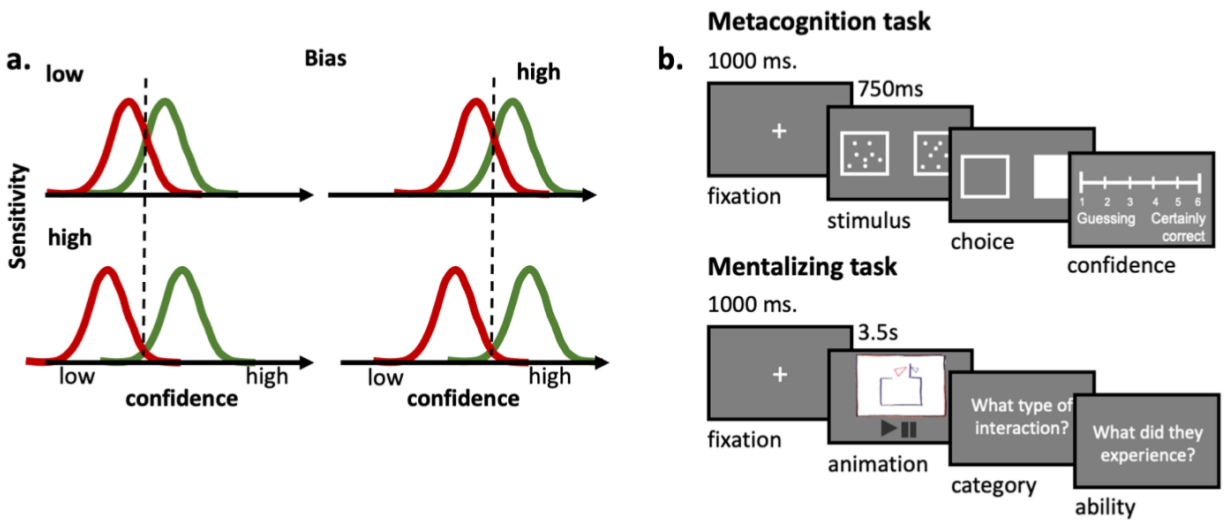


Figure 1. Task design and dissociation between metacognitive sensitivity and bias. a. *Hypothetical Gaussian distributions of confidence for correct (green) and incorrect (red) decisions. The left panel represents a decider with low confidence; the right panel represents a decider with high confidence. Metacognitive sensitivity is defined as the separation in confidence between correct and incorrect decisions; metacognitive bias is the overall confidence expressed.* ***b.*** *On the metacognition task, participants made judgments about which patch with dots had a higher density (left or right). After this, they were asked to rate their confidence on a scale from 1 "Guessing" to 6 "Certainly correct". On the mentalizing task, participants watched animations of moving triangles and were asked to categorize and interpret the interaction of the triangles.*

We next investigated the hypothesis of a positive association between metacognitive efficiency and mentalizing ability within the hierarchical meta-*d'* model. When we examined the beta coefficient representing the impact of mentalizing ability on metacognitive efficiency, the HDI was positive and did not encompass zero (hierarchical estimation: 95% HDI [0.01, 0.09]), with 99% of the sampled beta values being higher than zero ($P_{\theta\ (\text{HDI mentalizing ability} > 0)}$ = 0.99; **Figure 2a**) indicating a significant positive relationship. To confirm this effect while controlling for covariates of age, gender, IQ and education, we used a linear regression model with the standardized log metacognitive efficiency from a single-subject model as a dependent variable and mentalizing ability and these covariates as predictor variables. This approach again revealed a positive relationship between mentalizing and metacognition (linear regression model: $\beta_{mentalizing\ efficiency}$ = 0.11, SE = 0.05, t_{476} = 2.26, *P* = 0.02) and no effects of the covariates (*P* > 0.05), suggesting that participants who were better at inferring the mental states and interactions on the mentalizing task were also better at tracking their performance on the metacognition task.

To investigate how mentalizing was related to metacognition, we next tested the hypothesis that mentalizing is associated with a greater impact of response times on confidence. Specifically, we estimated a hierarchical mixed-effects model predicting trial-by-trial explicit confidence levels on the metacognition task from differences in standardized log response times (logRT) and accuracy [error: -0.5, correct: 0.5] (**Equation 2.1**), and asked whether this model provided a better fit when these predictors were allowed to vary as a function of the participants' mentalizing ability (**Equation 2.2**). A Likelihood Ratio Test indicated that this was the case ($\chi^2(4)$ = 27.59, *P* = 1.51e-05) which was also confirmed by several goodness-of-fit indices (log likelihood (LL): ΔLL = 13, Akaike Information Criterion (AIC): ΔAIC = -20, Bayesian Information Criterion (BIC): ΔBIC = 17 and ΔDeviance: -28), suggesting a significant relationship between mentalizing and the computations underpinning confidence formation.

We next asked how mentalizing modulated the construction of confidence by investigating which predictor variables interacted with mentalizing ability. We found that participants with better mentalizing ability reported lower overall confidence in their own responses than participants with lower mentalizing ability (hierarchical linear regression, main effect of mentalizing ability: $\chi^2(1) = 6.08$, $P = 0.01$, β = -0.04, SE = 0.02). In addition, participants with higher mentalizing ability scores modulated their confidence ratings more on the basis of their response times than participants with lower scores of mentalizing ability (interaction effect of logRT x mentalizing ability: $\chi^2(1) = 21.92$, $P = 2.84\text{e-}06$, β = -0.03, SE = 0.006; **Figure 2b**), consistent with the idea that mentalizing facilitates metacognition by facilitating self-inference on the basis of externally visible behavioural cues.

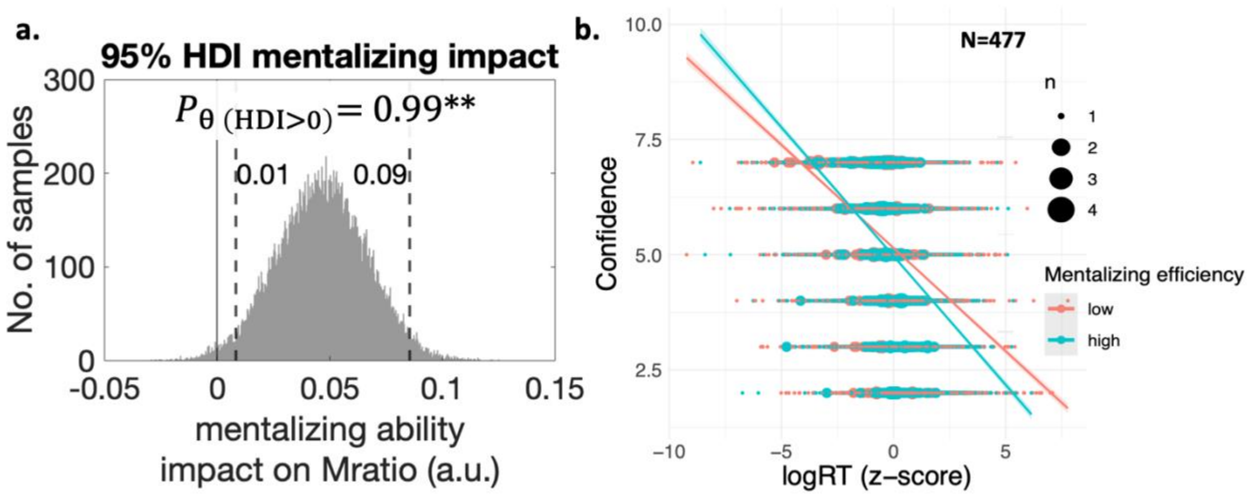


Figure 2. Mentalizing modulates computation of confidence. a. *Posterior distribution over the regression coefficient relating mentalizing ability to metacognitive ability. The dashed lines represent the 95% highest density interval (HDI), P_θ indicates the probability that the posterior samples are greater than zero, ** P < 0.01 in the frequentist linear model.* ***b.*** *Confidence was negatively related to response times (logRT). Trial-by-trial response times have a higher impact on the estimated confidence of participants scoring above the median of mentalizing ability scores (in turquoise) than participants scoring below the median (in pink). Shaded area represents the Standard Deviation from the Mean (±SDM).*

Next, we addressed the second hypothesis of a negative association between metacognitive efficiency and autistic traits in the general population, as assessed with the

AQ-10 (Allison et al., 2012) and the RAADS-14 questionnaires (Eriksson et al., 2013). First, we evaluated whether participants with higher scores of autistic traits had lower mentalizing ability, by conducting a linear regression model with mentalizing ability as the dependent variable and autistic trait scores and the covariates (age, gender, education, IQ) as predictor variables. We found the expected negative relationship between mentalizing ability and RAADS-14 scores (linear regression model: $\beta_{RAADS-14}$ = -0.002, SE = 0.0009, t_{476} = -2.21, P = 0.03) but not AQ-10 scores (linear regression model: β_{AQ10} = 0.006, SE = 0.004, t_{476} = 1.33, P = 0.19). This unexpected finding, together with recent re-evaluations of the reliability of the AQ-10 scale (Bertrams, 2021), and the greater developmental information captured by the RAADS-14, led us to focus on RAADS-14 scores in the remainder of the analyses.

Next, we asked whether compromised mentalizing ability in participants with higher scores of autistic traits was associated with lower metacognitive efficiency. To test this, we estimated the correlation between metacognitive efficiency and RAADS-14 scores within a hierarchical regression model. The 95% HDI for the coefficient of RAADS-14 scores was negative on average, ranging from [-0.057, 0.019], but encompassed zero (hierarchical estimation: $P_{\theta\ (HDI\ RAADS < 0)}$ = 0.82). A frequentist linear model that controlled for the covariates also confirmed that participants with higher scores of autistic traits do not necessarily also have compromised metacognitive efficiency (linear regression model: $\beta_{RAADS14}$ = -0.05, SE = 0.05, t_{476} = -1.09, P = 0.28).

An alternative explanation hypothesis is that autistic traits as measured by the RAADS-14 do not have a direct impact on the metacognitive efficiency score, but rather affect the construction of confidence. To examine this, we tested if our mixed-effect hierarchical regression model better predicts trial-by-trial confidence levels on the metacognition task when the predictors (accuracy, logRT and their interactions) were allowed to vary as a function of differences in autistic traits. A Likelihood Ratio Test indeed suggests that an interaction term on autistic traits improved the fit of the model ($\chi^2(4) = 14.52$, P = 0.006) which was further confirmed by several goodness-of-fit metrics (ΔLL: 7, ΔBIC: -31, ΔAIC: 7 and ΔDeviance: -15), indicating that the computation of confidence differs as a function of individual differences in autistic traits.

We next asked in what way people with higher scores for autistic traits constructed their confidence differently, by testing which predictor variables interacted with RAADS-14 scores. We found that participants with higher scores for autistic traits reported lower confidence overall (hierarchical linear regression, main effect of RAADS-14: $\chi^2(1) = 4.86$, P = 0.027, β = -0.008, SE = 0.004). In addition, explicit confidence was more informed by logRT among participants with lower scores for autistic traits than among participants with higher scores for autistic traits (interaction effect of logRT x RAADS-14: $\chi^2(1) = 6.46$, P = 0.011, β = 0.004, SE = 0.001). In **Figure 3a** we plot the extracted beta coefficients of the impact of response times on confidence for participants scoring above and below the median cut-off on autistic traits on error and correct trials separately, which shows that this effect was driven by participants with higher autistic trait scores having a lower impact of response times on error-trials than participants with lower autistic traits (three-way interaction of logRT x RAADS-14 x accuracy: $\chi^2(1) = 4.63$, P = 0.031, β = -0.003, SE = 0.001). Together these results suggest that participants with higher autistic traits use response times less to infer they have committed an error than participants with lower autistic trait scores.

These results suggest that compromised mentalizing ability may specifically affect the relationship between response times and confidence. We next asked whether specifically social aspects of the autistic phenotype, rather than non-social aspects, negatively impact metacognition. In an exploratory analysis we estimated the correlation between metacognitive efficiency and self-reported social skills with hierarchical regression models. This analysis revealed that participants with self-reported difficulties in everyday types of social interaction, measured by the 'mentalizing' sub-scale of the RAADS-14, had lower metacognitive efficiency than participants with better self-reported social skills (hierarchical estimation: HDI: [-0.07, 0.00], $P_{\theta\ (HDI\ social\ skills < 0)}$ = 0.97; frequentist linear regression: β= -0.09, SE = 0.05, t_{476} = -1.84, P = 0.067; **Figure 3b**). In contrast, the non-social sub-scale of the RAADS-14 was not associated with metacognitive efficiency (hierarchical estimation: HDI: [-0.04, 0.04], $P_{\theta\ (HDI\ nonsocial\ skills < 0)}$ = 0.43; frequentist linear regression: β= -0.007, SE = 0.05, t_{476} = -1.14, P = 0.89; **Figure 3c**). Together, these results suggest that self-reported social, but not non-social, autistic traits are negatively associated with metacognitive efficiency.

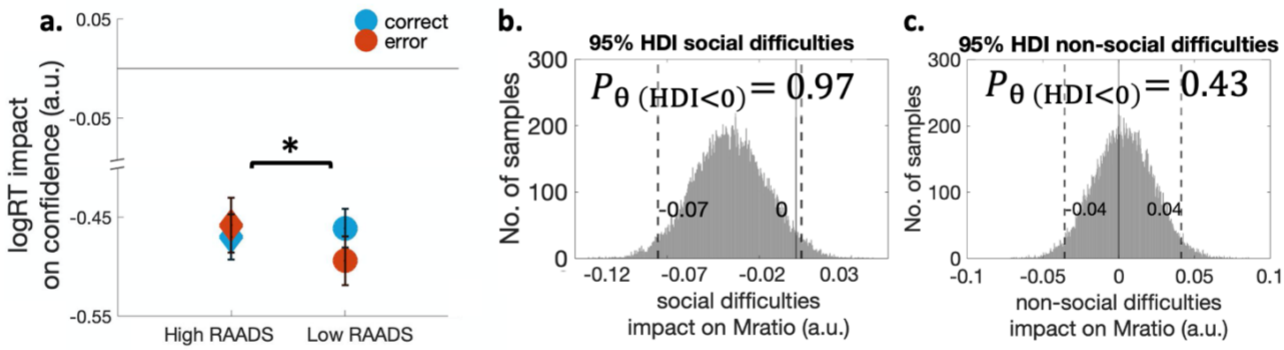


Figure 3. Autistic trait differences modulate metacognitive efficiency. a. *Standardized beta coefficients of the impact of logRT on confidence from a hierarchical mixed-effect regression model on error trials (red) and correct trials (blue) for participants with high and low RAADS scores (above and below the median cut-off, respectively).* ***b.*** *Posterior estimates of the hierarchically estimated beta coefficient relating the social subscale of RAADS-14 to metacognitive efficiency.* ***c.*** *Posterior estimates of the hierarchically estimated beta coefficient relating the non-social subscale of RAADS-14 to metacognitive efficiency. The dashed lines represent the 95% highest density intervals (HDI),* P_{θ} *indicates the probability that the posterior samples are different from zero. Error bars represent group means ± SEM, * P < 0.05 of the interaction effect between RAADS and logRT on confidence.*

In summary, in Experiment 1 we found a metacognitive benefit for participants with better mentalizing ability. We further disentangled the mechanism of this effect by showing that mentalizing ability is associated with a tighter coupling between response times and confidence in errors. Metacognition was less efficient in participants with higher scores for autistic traits, in particular, among participants who report greater difficulties with self-reported social difficulties. Together these results provide initial evidence that metacognitive processes are related to mentalizing capacity.

Experiment 2

In Experiment 1 we found that metacognitive and mentalizing abilities are related, potentially by affecting the extent to which response times modulate confidence. Against our expectation, we did not find a statistically significant negative correlation between autistic traits and metacognitive efficiency. One explanation of this null result is that the variation in autistic traits was not pronounced enough in our general population sample to allow estimation of this relationship. In Experiment 2 we sought to compare data from N = 40 autistic participants recruited via the charity organization Autistica to a matched comparison group of N = 40 participants subsampled from the dataset of Experiment 1. As a result of the selection procedure described in **Methods**, both groups had similar age (independent samples t-test, t_{78} = 0.90, P = 0.37), gender (independent samples t-test, t_{78} = 1.07, P = 0.29), education (M_{autism} = 4.00, SE = 0.06; $M_{comparison}$ = 3.92, SE = 0.19; independent samples t-test, t_{78} = 0.25, P = 0.80) and IQ scores (M_{autism} = 9, SE = 0.54; $M_{comparison}$ = 7.90, SE = 0.52; independent samples t-test, t_{76} = 1.45, P = 0.15). In addition, as a result of the calibration procedure, first-order performance on the metacognition task was not statistically different between groups (M_{autism} = 0.75, SE = 0.01; $M_{comparison}$ = 0.74, SE = 0.008; independent samples t-test, t_{78} = 0.52, P = 0.60; see **Supplementary Material** for other reliability checks).

Having shown that the two groups were matched in terms of demographics and general cognitive ability, we next asked if autistic participants had lower mentalizing ability than comparison participants by testing a linear regression model with mentalizing ability as independent variable and clinical group [autism: -0.5, comparison: 0.5] and the covariates (age, gender, IQ, and education) as predictor variables. When we do this, we find that mentalizing ability was indeed lower for autistic participants than comparison participants, but not significantly so (linear regression: β_{group} = -0.43 (0.25), t_{68} = -1.72, P = 0.089).

Next, we use a similar linear regression model to test if the autism group had lower metacognitive efficiency than the comparison group. In line with our pre-registered hypotheses, this indeed revealed significantly lower metacognitive efficiency in autistic participants than in comparison participants (linear regression model: β_{group} = -0.60 (0.25), t_{63} = -2.46, P = 0.016; **Figure 4a**) with no effects of the covariates. We next estimated metacognitive efficiency within a hierarchical model fitted to each group separately, while accounting for the effects of IQ, age, gender and education. The HDI of metacognitive efficiency in the autism group (HDI [0.92, 0.55]) was quantitatively lower than that of the comparison group (HDI [0.84, 0.52]) in 78% of the samples $P_{\theta\,(HDI\,ASD\,<\,HDI\,comparison)} = 0.78$ (**Figure 4b**), although did not reach significance at the classical 95% threshold. Taken together these analyses provide some evidence in support of our pre-registered hypothesis of lower metacognitive efficiency in autism.

Finally, building upon a hierarchical mixed-effect regression model of trial-by-trial predictions of confidence on the metacognition task, we next tested whether the model could better predict confidence levels when the predictors (**Equation 2.1**), were allowed to vary as a function of whether the subject was autistic or not (**Equation 2.2**). A likelihood ratio test indicated that this was the case ($\chi^2(4) = 966.46,\ P < 2.20\text{e}^{-16}$) which was further strongly confirmed by goodness-of-fit indices (ΔLL: -484, ΔAIC: 958, ΔBIC: 929 and ΔDeviance:

966), supporting the prediction that confidence formation in autistic participants is qualitatively distinct to comparison participants.

Consistent with the results of Experiment 1 we found that autistic participants report lower confidence than comparison participants in general (hierarchical regression model, main effect of group: $\chi^2(1) = 768.50,\ P < 2.0\text{e}^{-16}$, $\beta = 0.82, SE = 0.03$). Autistic participants show a marginally lower impact of response times in error trials than comparison participants (three-way interaction logRT x group x accuracy: $\chi^2(1) = 3.086,\ P = 0.060$, $\beta = 0.10, SE = 0.06$). In **Figure 4c** we plot the impact of response times on confidence on error and correct trials separately, which shows that the negative impact of RT on confidence was less negative in autistic participants than in comparison participants, suggesting a weaker influence on response times on confidence in error trials.

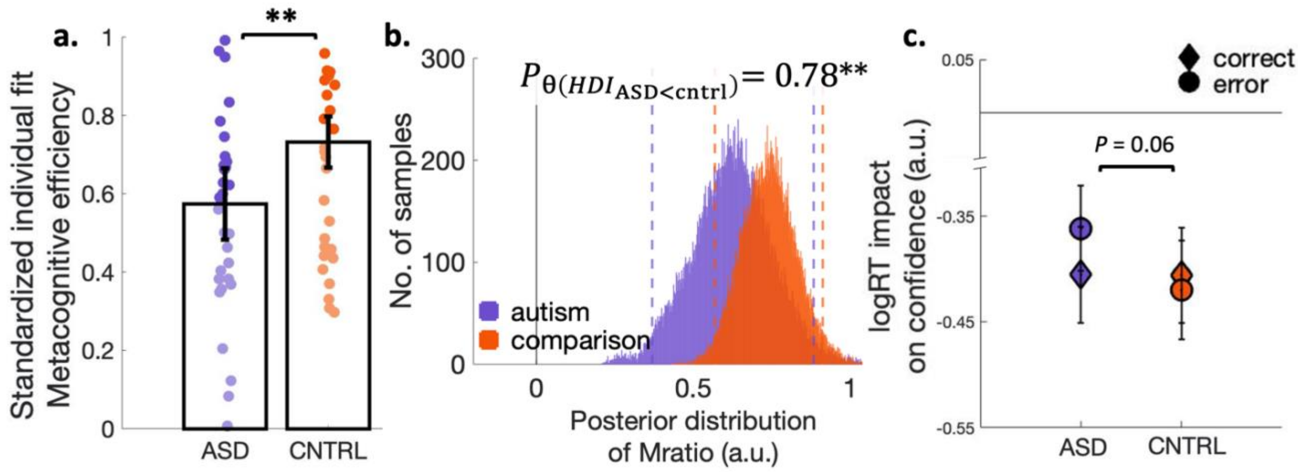


Figure 4. Differences in metacognitive efficiency and confidence formation in autism. a. *Metacognitive efficiency estimated from a single-subject Bayesian model fit is significantly lower in the autism group (N=40) than in the comparison group (N=40). Error bars represent group mean ± SEM.* ***b.*** *Posterior estimates of metacognitive efficiency from independent group model fits (autism in purple, controls in orange) where the dashed lines represent the highest density intervals (HDI) and P_θ represents the probability that the HDI of the autism group is lower than the HDI of the comparison group.* ***c.*** *Impact of logRT on confidence on error and correct trials for autism and comparison participants. Error bars represent group means ± SEM.*

In summary, in Experiment 2 we show that metacognitive efficiency is compromised in autism and reveal a weaker association between response times and confidence in autistic participants in contrast to matched comparison participants.

DISCUSSION

Across two behavioural experiments we show that mentalizing ability is positively related to metacognition. In a general population sample of N = 477 participants we found that individuals who were better at self-reported social skills and mentalizing could also more reliably track their own accuracy on a perceptual discrimination task. By investigating the trial-by-trial computations of confidence, we were able to investigate precisely how mentalizing relates to metacognition. Notably, mentalizing ability was associated with a tighter coupling between response times and confidence, suggesting that mentalizing ability may facilitate inference on cues to self-performance. In a second dataset with autistic participants, we show that the mentalizing difficulties that characterize this condition are associated both with compromised metacognitive ability and replicate the findings of Experiment 1 that autistic traits are associated with a weaker link between response times and confidence. Together, these findings suggest that processes involved in inferring other people's mental states may also facilitate self-directed metacognition, and vice versa.

We quantified metacognition as the ability to reliably separate correct from incorrect decisions with confidence ratings (Flavell, 1979; Fleming et al., 2010; Rollwage et al., 2018; Rouault et al., 2018). Several studies have suggested confidence is 'read out' from how much reliable evidence has been seen, either during the course of the decision itself (Kiani & Shadlen, 2009; Pleskac & Busemeyer, 2010) or after an initial decision has been made (post-decisional evidence processing; Fleming et al., 2018; Resulaj et al., 2009; Talluri et al., 2018; van den Berg et al., 2016). Other studies suggest that response times also provide a behavioural cue to confidence (Kiani et al., 2014; Patel et al., 2012). How, then, might mentalizing play a role in confidence construction? Recent theoretical models suggest that confidence estimates reflect an inference about the state of the decider, informed by behavioural and cognitive cues—suggesting a computational parallel between self- and other-evaluation (Fleming & Daw, 2017). Indeed, evidence strength (Campbell-Meiklejohn et al., 2017) and response times (Patel et al., 2012) appear to be used similarly to infer both one's own and others' confidence. However, isolating such metacognitive capacity requires tight control over the evidence going into a decision, to avoid first-order performance and stimulus factors confounding estimates of the confidence-accuracy correlation (Masson & Rotello, 2009; Rahnev & Fleming, 2019). Here we used a staircase procedure to control perceptual performance within a narrow range and used a metric of metacognition that is unconfounded by both metacognitive bias and first-order performance. In addition, we used a Bayesian inference approach to estimate the impact of mentalizing ability on metacognitive ability within the same hierarchical model, which ensured that both within- and between-subject variability are appropriately taken into account. These methodological advances may explain why here we found a more robust between-subjects relationship between metacognition and mentalizing than reported previously (Carpenter et al., 2019; Nicholson et al., 2020).

Our results are also in line with previous work on autism, suggesting that metacognitive ability may be compromised in autistic individuals to a similar extent to the ability to evaluate other people's mental states. Autism was characterised as a general "mind-blindness" in 1985 (Simon Baron-Cohen et al., 1985) but, since then, only a handful of studies have extended the study of mentalizing in autism to that of metacognitive ability about one's own behaviour and mental states (Carpenter et al., 2019; Grainger et al., 2016; Nicholson et al., 2019, 2020; Williams et al., 2018; Wojcik et al., 2013). Some of these studies (Grainger et al., 2016; Nicholson et al., 2020; Williams et al., 2018) but not others (Carpenter et al., 2019; Wojcik et al., 2013), found, in line with our pre-registered hypotheses and findings, that mentalizing and metacognitive ability were commensurately compromised in autism. A notable exception to this general picture is that we unexpectedly found that self-reported autistic traits on the RAADS-14 were not negatively associated with metacognitive efficiency in our general population dataset of Experiment 1. One candidate explanation for this inconsistency is that variation in autistic traits in the general population may not have been pronounced enough to find statistically significant differences in metacognitive efficiency. Another explanation is that metacognitive ability in autism may not be worse on average but rather more extreme (both extremely strong *and* weak; Pariser, 1981; Shields-Wolfe & Gallagher, 1992)—as hinted at by the greater variance in the autistic group estimates (see overlayed dots in **Figure 4a).** Future studies should investigate whether this is the case in larger samples and, if so, whether it can be attributed to autistic people engaging in alternative, perhaps more cognitively demanding, processes to compensate for metacognitive difficulties (Livingston, Colvert, et al., 2019; Livingston, Shah, et al., 2019). Given the range of cues people may use to inform confidence, it will be important for future studies to focus on how the construction of confidence or other mentalizing processes varies across participants. It could be that, in real life, the metacognitive ability of some autistic people is above average but achieved via different routes than those studied in this experiment.

Our work goes beyond estimating correlations between metacognition and mentalizing by revealing a potential mechanism through which mentalizing may affect metacognitive processes. Specifically, we show that better mentalizing ability is associated with a tighter coupling between response times and confidence. Previous work has experimentally manipulated response times and found this to have a causal effect on the construction of confidence: when response times are manipulated to be faster, people are subsequently more likely to report being confident (Kiani et al., 2014; Palser et al., 2018). The mentalizing-is-prior theory suggests metacognition consists of a re-application of inferential processes used to understand other people to understand our own mental states

(Carruthers, 2009). Our findings are consistent with this view, showing that people with greater proficiency in self-reported social skills and objectively measured mentalizing also had better metacognitive efficiency. In addition, we found that mentalizing ability not only correlated with overall metacognitive efficiency, but specifically with the ability to infer confidence from behavioural cues that would also be visible markers of other people's decision confidence in everyday situations. An important limitation of the current study is that we cannot draw causal conclusions about how mentalizing affects metacognition or vice versa. Future longitudinal work is needed to ask whether exposure to situations requiring mental state inference from behaviour causally affects the development of explicit metacognition. Another limitation of this study is that of domain-generality. There is reason to believe that metacognitive efficiency measured from perceptual decision-making is similar to metacognitive efficiency measured in other domains, such as from mnemonic or numerical decision-making tasks (Bronfman et al., 2015; Rouault, McWilliams, et al., 2018; Talluri et al., 2018; van der Plas et al., 2021). However, other studies found selective differences in perceptual metacognition between groups, in the absence of differences in memory metacognition (Fleming et al., 2014). The possibility of dissociations between domains suggests an unlikely, albeit possible, chance that mentalizing ability is only related to metacognitive efficiency when the latter is measured in the context of a perceptual task. Future studies should test the interplay between metacognition and mentalizing across a wider range of cognitive domains.

In summary, across two behavioural experiments we demonstrate that mentalizing ability is associated with both greater metacognitive efficiency, and tighter links between response times and confidence. In a general population sample, participants with better social skills were also better at reflecting upon their own performance. In a second dataset we show that autistic participants with generally lower mentalizing ability also had weaker metacognitive ability, in the absence of differences in first-order performance. Together, these results suggest that inferring other people's mental states is related to the ability to evaluate our own decisions.

REFERENCES

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16. https://doi.org/10.1016/S0885-2014(00)00014-9

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., Findon, J., Eklund, H., Spain, D., Wilson, C. E., Cadman, T., Young, S., Stoencheva, V., Murphy, C. M., Robertson, D., Charman, T., Bolton, P., Glaser, K., Asherson, P., … Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. *Psychological Medicine*, *46*(12), 2595–2604. PubMed. https://doi.org/10.1017/S0033291716001082

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117. https://doi.org/10.1038/s41562-017-0117

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private-public mappings in human prefrontal cortex. *BioRxiv*, 2020.02.21.954305. https://doi.org/10.1101/2020.02.21.954305

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251.

Baron-Cohen, Simon, Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Beck, A. T., Baruch, E., Balter, J. M., Steer, R. A., & Warman, D. M. (2004). A new instrument for measuring insight: The Beck Cognitive Insight Scale. *Schizophrenia Research*, *68*(2), 319–329. https://doi.org/10.1016/S0920-9964(03)00189-0

Bertrams, A. (2021). Internal reliability, homogeneity, and factor structure of the ten-item Autism-Spectrum Quotient (AQ-10) with two additional response categories. *Experimental Results*, *2*, e3. https://doi.org/10.1017/exp.2020.70

Broadbent, J., & Stokes, M. A. (2013). Removal of negative feedback enhances WCST performance for individuals with ASD. *Research in Autism Spectrum Disorders*, *7*(6), 785–792. https://doi.org/10.1016/j.rasd.2013.03.002

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1810), 20150228. https://doi.org/10.1098/rspb.2015.0228

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the Opinion of Others Affects Our Valuation of Objects. *Current Biology*, *20*(13), 1165–1170. https://doi.org/10.1016/j.cub.2010.04.055

Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D., & Daw, N. D. (2017). Independent Neural Computation of Value from Other People's Confidence. *The Journal of Neuroscience*, *37*(3), 673. https://doi.org/10.1523/JNEUROSCI.4490-15.2016

Carpenter, K. L., Williams, D. M., & Nicholson, T. (2019). Putting Your Money Where Your Mouth is: Examining Metacognition in ASD Using Post-decision Wagering. *Journal of Autism and Developmental Disorders*, *49*(10), 4268–4279. https://doi.org/10.1007/s10803-019-04118-6

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *The Behavioral and Brain Sciences*, *32*(2), 121–138; discussion 138-182. https://doi.org/10.1017/S0140525X09000545

Condon, D. M., & Revelle, W. (2016). Selected ICAR Data from the SAPA-Project: Development and Initial Validation of a Public-Domain Measure. *Journal of Open Psychology Data*, *4*(1), 1. https://doi.org/10.5334/jopd.25

David, N., Gawronski, A., Santos, N. S., Huff, W., Lehnhardt, F.-G., Newen, A., & Vogeley, K. (2008). Dissociation Between Key Processes of Social Cognition in Autism: Impaired Mentalizing But Intact Sense of Agency. *Journal of Autism and Developmental Disorders*, *38*(4), 593–605. https://doi.org/10.1007/s10803-007-0425-x

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social Information Is Integrated into Value and Confidence Judgments According to Its Reliability. *The Journal of Neuroscience*, *37*(25), 6066–6074. https://doi.org/10.1523/JNEUROSCI.3880-16.2017

Dimaggio, G., Lysaker, P. H., Carcione, A., Nicolò, G., & Semerari, A. (2008). Know yourself and you shall know the other… to a certain extent: Multiple paths of influence of self-reflection on mindreading. *Consciousness and Cognition*, *17*(3), 778–789. https://doi.org/10.1016/j.concog.2008.02.005

Eriksson, J. M., Andersen, L. M., & Bejerot, S. (2013). RAADS-14 Screen: Validity of a screening tool for autism spectrum disorder in an adult psychiatric population. *Molecular Autism*, *4*(1), 49. https://doi.org/10.1186/2040-2392-4-49

Ewbank, M. P., von dem Hagen, E. A. H., Powell, T. E., Henson, R. N., & Calder, A. J. (2016). The effect of perceptual expectation on repetition suppression to faces is not modulated by variation in autistic traits. *Cortex*, *80*, 51–60. https://doi.org/10.1016/j.cortex.2015.10.011

Fandakova, Y., Selmeczy, D., Leckey, S., Grimm, K. J., Wendelken, C., Bunge, S. A., & Ghetti, S. (2017). Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proceedings of the National Academy of Sciences*, *114*(29), 7582. https://doi.org/10.1073/pnas.1703079114

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., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

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

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

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

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

Fleming, S. M, Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain : A Journal of Neurology*, *137*(Pt 10), 2811–2822. PubMed. https://doi.org/10.1093/brain/awu221

Fleming, S. M., van der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, *21*(4), 617–624. https://doi.org/10.1038/s41593-018-0104-6

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223. https://doi.org/10.1098/rstb.2012.0123

Gazzaniga, M. S. (1995). Principles of human brain organization derived from split-brain studies. *Neuron*, *14*(2), 217–228. https://doi.org/10.1016/0896-6273(95)90280-5

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, *123*(7), 1293–1326. https://doi.org/10.1093/brain/123.7.1293

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14, 29–113. https://doi.org/10.1017/S0140525X00028636

Goupil, L., & Kouider, S. (2016). Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology : CB*, *26*(22), 3038–3045. https://doi.org/10.1016/j.cub.2016.09.004

Goupil, L., & Kouider, S. (2019). Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Current Directions in Psychological Science*, *28*(4), 403–408. https://doi.org/10.1177/0963721419848672

Grainger, C., Williams, D. M., & Lind, S. E. (2016). Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy. *Consciousness and Cognition*, *42*, 65–74. https://doi.org/10.1016/j.concog.2016.03.003

Greene, R. K., Zheng, S., Kinard, J. L., Mosner, M. G., Wiesen, C. A., Kennedy, D. P., & Dichter, G. S. (2019). Social and nonsocial visual prediction errors in autism spectrum disorder. *Autism Research*, *12*(6), 878–883. https://doi.org/10.1002/aur.2090

Happe, F. (2015). Autism as a neurodevelopmental disorder of mind-reading. *Journal of the British Academy*, *3*. https://doi.org/10.5871/jba/003.197

Harrison, O. K., Garfinkel, S. N., Marlow, L., Finnegan, S., Marino, S., Nanz, L., Allen, M., Finnemann, J., Keur-Huizinga, L., Harrison, S. J., Stephan, K. E., Pattinson, K., & Fleming, S. M. (2020). The Filter Detection Task for measurement of breathing-related interoception and metacognition. *BioRxiv*, 2020.06.29.176941. https://doi.org/10.1101/2020.06.29.176941

Hembacher, E., & Ghetti, S. (2014). Don’t look at my answer: Subjective uncertainty underlies preschoolers’ exclusion of their least accurate memories. *Psychological Science*, *25*(9), 1768–1776. https://doi.org/10.1177/0956797614542273

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

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

Kruschke, J. K. (2010). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS* (1st ed.). Academic Press, Inc.

Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M., & Ahissar, M. (2019). Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience*, *22*(2), 256–264. https://doi.org/10.1038/s41593-018-0308-9

Livingston, L. A., & Happé, F. (in press). Understanding atypical development through social cognitive theory: Lessons from autism. In *The cognitive basis of social interaction across the lifespan.* H. J. Ferguson, V. E. A. Brunsdon & E.E.F. Bradford (Eds.). Oxford University Press.

Livingston, L. A., Carr, B., & Shah, P. (2019). Recent Advances and New Directions in Measuring Theory of Mind in Autistic Adults. *Journal of Autism and Developmental Disorders*, *49*(4), 1738–1744. https://doi.org/10.1007/s10803-018-3823-3

Livingston, L. A., Colvert, E., the Social Relationships Study Team, Bolton, P., & Happé, F. (2019). Good social skills despite poor theory of mind: Exploring compensation in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *60*(1), 102–110. https://doi.org/10.1111/jcpp.12886

Livingston, L.A., Shah, P., & Happé, F. (2019). Compensatory strategies below the behavioural surface in autism: A qualitative study. *The Lancet Psychiatry*, *6*(9), 766–777. https://doi.org/10.1016/S2215-0366(19)30224-X

Livingston, L. A., Shah, P., White, S., & Happé, F. (in press). Further developing the Frith-Happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. *Autism Research*.

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

Maniscalco, B., & 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. *The Cognitive Neuroscience of Metacognition.*, 25–66. https://doi.org/10.1007/978-3-642-45190-4_3

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

McMahon, C. M., Henderson, H. A., Newell, L., Jaime, M., & Mundy, P. (2016). Metacognitive Awareness of Facial Affect in Higher-Functioning Children and Adolescents with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(3), 882–898. https://doi.org/10.1007/s10803-015-2630-3

Milne, E., Swettenham, J., Hansen, P., Campbell, R., Jeffries, H., & Plaisted, K. (2002). High motion coherence thresholds in children with autism. *Journal of Child Psychology and Psychiatry*, *43*(2), 255–263. https://doi.org/10.1111/1469-7610.00018

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nicholson, T., Williams, D., Lind, S., Grainger, C., & Carruthers, P. (2020). Linking metacognition and mindreading: Evidence from autism and dual-task investigations. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0000878

Nicholson, T., Williams, D. M., Grainger, C., Lind, S. E., & Carruthers, P. (2019). Relationships between implicit and explicit uncertainty monitoring and mindreading: Evidence from autism spectrum disorder. *Consciousness and Cognition*, *70*, 11–24. https://doi.org/10.1016/j.concog.2019.01.013

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Palser, E. R., Fotopoulou, A., & Kilner, J. M. (2018). Altering movement parameters disrupts metacognitive accuracy. *Consciousness and Cognition: An International Journal*, *57*, 33–40. https://doi.org/10.1016/j.concog.2017.11.005

Pariser, D. (1981). Nadia's Drawings: Theorizing about an Autistic Child's Phenomenal Ability. *Studies in Art Education*, *22*(2), 20–31. https://doi.org/10.1080/00393541.1981.11650279

Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4853–4860. https://doi.org/10.1098/rspb.2012.1847

Pirrone, A., Dickinson, A., Gomez, R., Stafford, T., & Milne, E. (2017). Understanding perceptual judgment in autism spectrum disorder using the drift diffusion model. *Neuropsychology*, *31*(2), 173–180. https://doi.org/10.1037/neu0000320

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., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(niz009). https://doi.org/10.1093/nc/niz009

Reed, P. (2019). Unpredictability reduces over-selective responding of individuals with ASD who have language impairments. *Research in Autism Spectrum Disorders*, *57*, 35–45. https://doi.org/10.1016/j.rasd.2018.10.006

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

Robic, S., Sonié, S., Fonlupt, P., Henaff, M.-A., Touil, N., Coricelli, G., Mattout, J., & Schmitz, C. (2015). Decision-Making in a Changing World: A Study in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *45*(6), 1603–1613. https://doi.org/10.1007/s10803-014-2311-7

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, *28*(24), 4014-4021.e8. https://doi.org/10.1016/j.cub.2018.10.053

Rosenblau, G., Kliemann, D., Heekeren, H. R., & Dziobek, I. (2015). Approximating Implicit and Explicit Mentalizing with Two Naturalistic Video-Based Tasks in Typical Development and Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(4), 953–965. https://doi.org/10.1007/s10803-014-2249-9

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

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Ryle, G. (1949). Meaning and Necessity. *Philosophy*, *24*(88), 69–76. JSTOR.

Sapey-Triomphe, L.-A., Sonié, S., Hénaff, M.-A., Mattout, J., & Schmitz, C. (2018). Adults with Autism Tend to Undermine the Hidden Environmental Structure: Evidence from a Visual Associative Learning Task. *Journal of Autism and Developmental Disorders*, *48*(9), 3061–3074. https://doi.org/10.1007/s10803-018-3574-1

Shields-Wolfe, J., & Gallagher, P. A. (1992). Functional utilization of splinter skills for the employment of a young adult with autism. *Focus on Autistic Behavior*, *7*(4), 1–16.

Spek, A., Schatorjé, T., Scholte, E., & van Berckelaer-Onnes, I. (2009). Verbal fluency in adults with high functioning autism or Asperger syndrome. *Neuropsychologia*, *47*(3), 652–656. https://doi.org/10.1016/j.neuropsychologia.2008.11.015

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology*, *28*(19), 3128-3135.e8. https://doi.org/10.1016/j.cub.2018.07.052

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and Neuroscience Advances*, *2*, 239821281881059. https://doi.org/10.1177/2398212818810591

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence Is the Bridge between Multi-stage Decisions. *Current Biology*, *26*(23), 3157–3168. https://doi.org/10.1016/j.cub.2016.10.021

van der Plas, E. A. A., Shiqi, Z., Keer, D., Bang, D., Nicholas, W., Jian, L., & Stephen, F. (2021). *Isolating cultural contributors to confidence*. *(2021, March 8).* https://doi.org/10.31234/osf.io/sjh7d

Wegner, D. M. (2002). *The illusion of conscious will.* (pp. xi, 405). MIT Press.

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

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the Strange Stories: Revealing Mentalizing Impairments in Autism. *Child Development*, *80*(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research*, *4*(2), 149–154. https://doi.org/10.1002/aur.174

Williams, D. M., Bergström, Z., & Grainger, C. (2018a). Metacognitive monitoring and the hypercorrection effect in autism and the general population: Relation to autism(-like) traits and mindreading. *Autism*, *22*(3), 259–270. https://doi.org/10.1177/1362361316680178

Williams, D. M., Bergström, Z., & Grainger, C. (2018b). Metacognitive monitoring and the hypercorrection effect in autism and the general population: Relation to autism(-like) traits and mindreading. *Autism*, *22*(3), 259–270. https://doi.org/10.1177/1362361316680178

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wojcik, D. Z., Moulin, C. J. A., & Souchay, C. (2013). Metamemory in children with autism: Exploring “feeling-of-knowing” in episodic and semantic memory. *Neuropsychology*, *27*(1), 19–27. https://doi.org/10.1037/a0030526

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zalla, T., Miele, D., Leboyer, M., & Metcalfe, J. (2015). Metacognition of agency and theory of mind in adults with high functioning autism. *Consciousness and Cognition*, *31*, 126–138. https://doi.org/10.1016/j.concog.2014.11.001

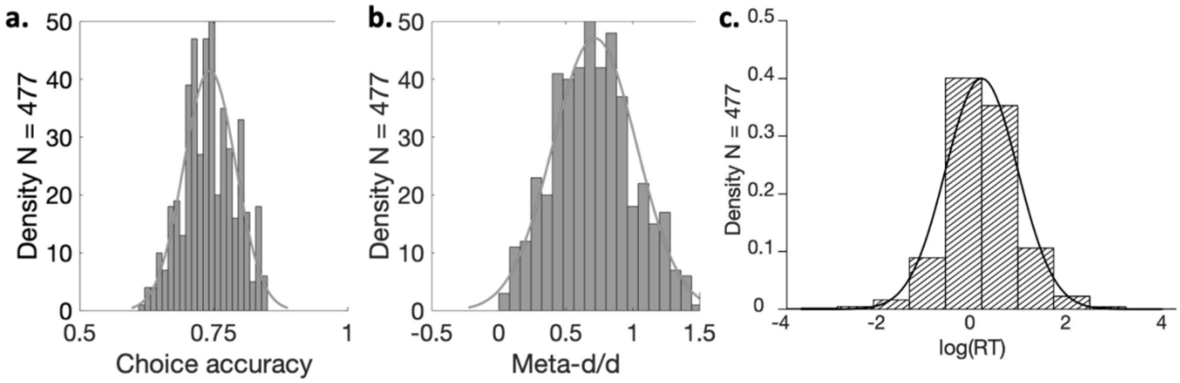
Zwart, F. S., Vissers, C. Th. W. M., Kessels, R. P. C., & Maes, J. H. R. (2018). Implicit learning seems to come naturally for children with autism, but not for children with specific language impairment: Evidence from behavioral and ERP data: Implicit learning intact in ASD but altered in SLI. *Autism Research*, *11*(7), 1050–1061. https://doi.org/10.1002/aur.1954

SUPPLEMENTARY MATERIAL

Experiment 1

1.1 Performance and validation checks

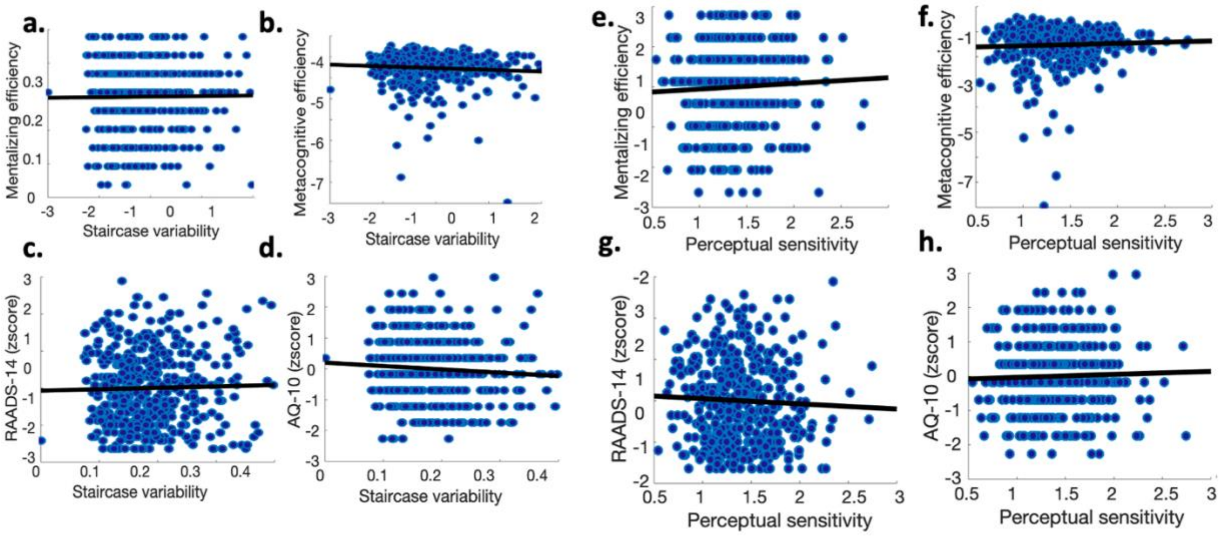
Average choice accuracy on the metacognition task (M = 74.16%, SEM = 0.002; **Supplementary Figure 1.1a**), metacognitive efficiency (M = 0.693, SEM = 0.016; **Supplementary Figure 1.1b**) and the log of response times (logRT; M = -1.405e^{-07}, SEM = 0.046; **Supplementary Figure 1.1c**) were similar to those of previous studies (Rouault et al., 2018; Rollwage et al., 2018).



Supplementary Figure 1.1. Choice accuracy on the metacognition task. a. *Histogram distribution of choice accuracy.* ***b.*** *Histogram distribution of metacognitive efficiency (meta-d'/d').* ***c.*** *Histogram distribution of the log of standardized response times (logRT). All variables are derived from the metacognition task and plotted for the group as a whole (N=477).*

As an indication of the reliability of mentalizing task variables, we asked whether the two mentalizing measures from the Happé-Frith Triangle Task were measuring a similar mentalizing construct. This was the case, with a positive correlation between the mentalizing feelings and mentalizing category scores: Spearman's r = 0.37, P = 2.73e-16. In addition, to establish whether the autistic trait surveys and Frith-Happé triangle task were measuring a similar mentalizing construct, we tested whether people with more autistic traits on the mentalizing subscale of the RAADS-14 also had lower mentalizing ability on the Frith-Happé Triangle Task, which was also the case (Spearman's r = -0.11, P = 0.017).

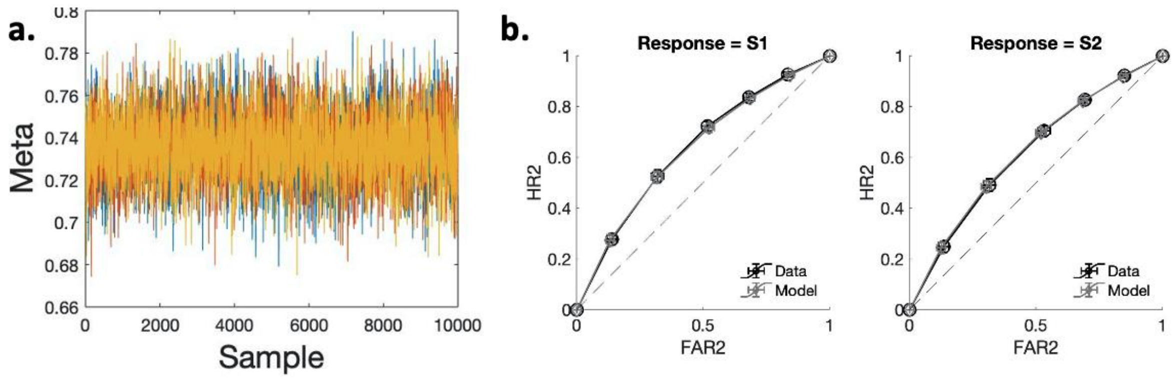
We next sought to ensure key variables related to metacognition and mentalizing were independent of first-order perceptual task performance. We first calculated each individual's experienced stimulus variability (the ratio between the standard deviation of stimulus difficulty and average stimulus difficulty) and correlated this with the main variables of interest. Staircase variability was not correlated with mentalizing ability (rs_{475} = 0.005, P = 0.91; **Supplementary Figure 1.2a**) metacognitive efficiency (rs_{475} = -0.068, P = 0.137; **Supplementary Figure 1.2b**), RAADS-14 scores (rs_{475} = 0.0015, P = 0.974; **Supplementary Figure 1.2c**) or AQ-10 scores (rs_{475} = -0.066, P = 0.149; **Supplementary Figure 1.2d**). The same validation checks were conducted for perceptual sensitivity, which was not correlated with mentalizing ability (rs_{475} = 0.0655, P = 0.1524; **Supplementary Figure 1.2e**) metacognitive efficiency (rs_{475} = -0.0513, P = 0.264; **Supplementary Figure 1.2f**), RAADS-14 scores (rs_{475} = -0.0536, P = 0.2437; **Supplementary Figure 1.2g**) or AQ-10 scores (rs_{475} = 0.0359, P = 0.435; **Supplementary Figure 1.2h**).



Supplementary Figure 1.2. Correlations between the main variables of interest. a-d: *Staircase variability, the ratio of the standard deviation and the mean dot difference, was not correlated with* ***a.*** *mentalizing ability,* ***b.*** *metacognitive efficiency (meta-d'/d'),* ***c.*** *autistic traits as measured by the RAADS-14* ***d.*** *autistic traits as measured with the AQ-10.* ***e-h:*** *Perceptual sensitivity (d') was not correlated with* ***e.*** *mentalizing ability,* ***f.*** *metacognitive efficiency (meta-d'/d'),* ***g.*** *autistic traits as measured by the RAADS-14,* ***h.*** *autistic traits as measured with the AQ-10.*

1.2. Posterior predictive checks

Next, we test whether the HMeta-d models used in estimating metacognitive efficiency were reliable by means of convergence checks and posterior predictive checks. The hierarchical regression model predicting metacognition from mentalizing ability scores converged well, indicated by the Gelman-Rubic statistics ($\hat{R}$ = 0.99997 and see plotted chains in **Supplementary Figure 1.3a)**. In addition, posterior predictive plots captured key patterns of the participants' confidence responses, with model and predicted type ROCs closely overlapping (**Supplementary Figure 1.3b**). The same was true for the hierarchical regression models with RAADS-14 scores ($\hat{R}$ = 1.0003 **Supplementary Figure 1.3c, d**) and AQ-10 scores ($\hat{R}$ =1.0006 **Supplementary Figure 1.3e, f**).



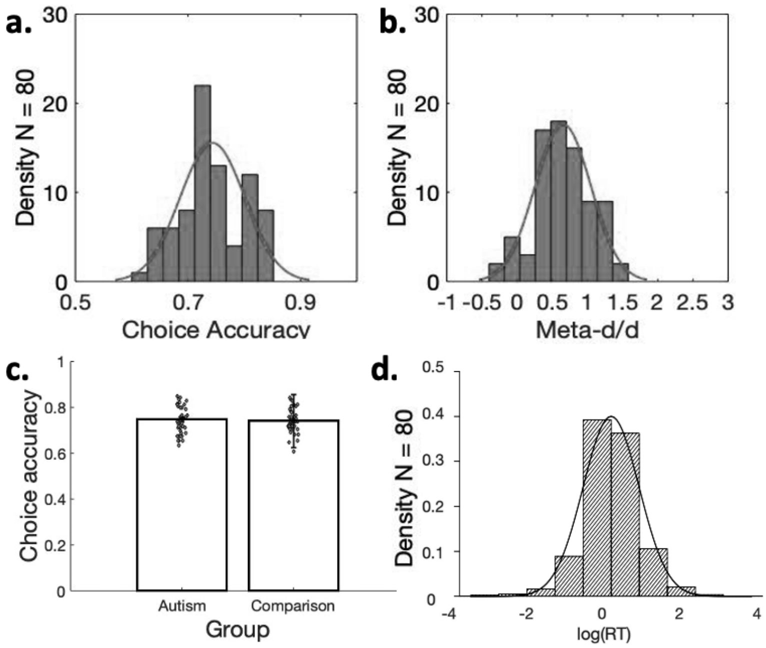
Supplementary Figure 1.3. Posterior predictive checks on HMeta-d fits in Experiment 1. a. *MCMC chains for parameter meta-d'/d' (metacognitive efficiency) from the hierarchical regression model.* ***b.*** *Observed and model estimates for the Type 2 ROC curves for leftward (S1) and rightward (S2) responses from the regression meta-d model fits. Error bars represent the mean ± standard error of the mean.*

Experiment 2

2.1. Performance and validation checks

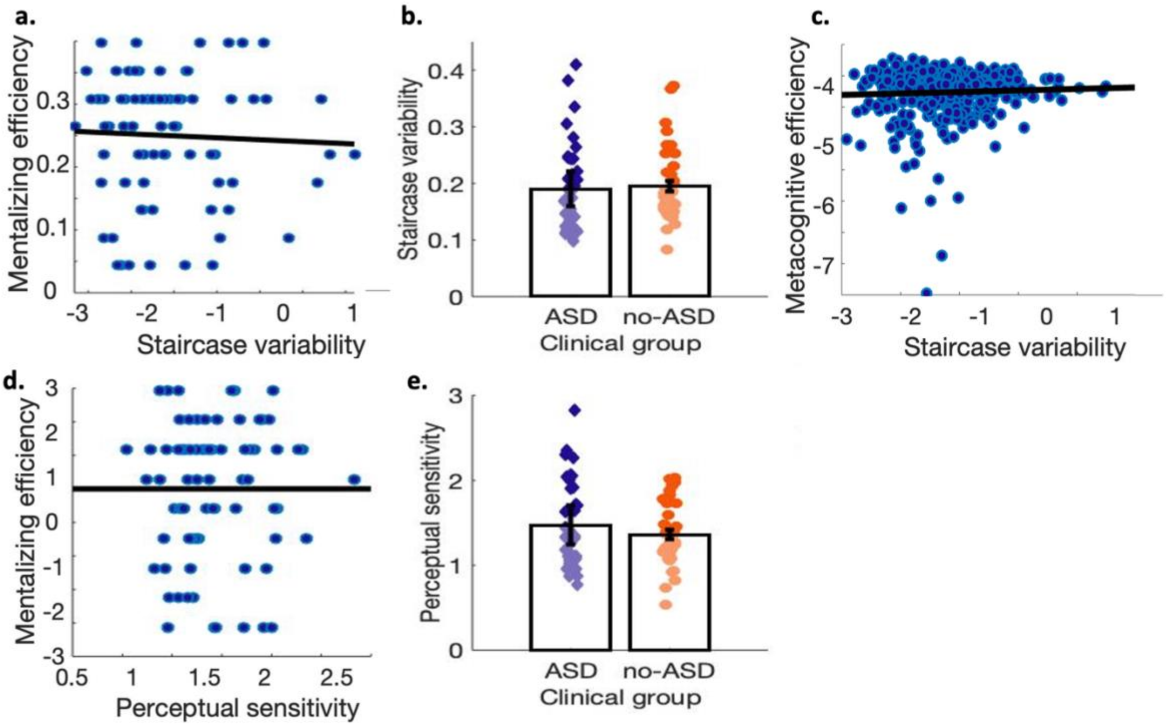
Average choice accuracy on the metacognition task (M=74.34% ± 0.006) was normally distributed (W = 0.98, *P* = 0.12; **Supplementary Figure 2.1a**) and is visually similar to those of the larger dataset (**Supplementary Figure 1.1**). Metacognitive efficiency or $meta-d'/d'$

(M=0.653 ± 0.045) was also normally distributed (W = 0.987, P = 0.60; **Supplementary Figure 2.1b**) and similar to that in Experiment 1(**Supplementary Figure 1.1**). As a result of the calibration procedure, first-order performance was not statistically different between the autism (M = 0.75 ± 0.01) and comparison groups (M = 0.74 ± 0.008; equal variances: P = 0.73, K = 0.15; independent samples t-test, t_{78} = 0.519, 95% CI = [-0.019, 0.032], P = 0.61; **Supplementary Figure 2.1c**). Finally, we averaged the log of response times (logRT) across trials of the metacognition task for each subject and plotted the distribution in **Supplementary Figure 2.1d.** Average logRT in the autism group (M= -7.39e^{-17} ± 6.05e^{-17}) and in the comparison group (M= -2.59e^{-17} ± 6.17e^{-17}) were not statistically different (t_{71} = 0.49, 95% CI = [-2.43, 1.47], P = 0.63).



Supplementary Figure 2.1. Choice accuracy on the metacognition task. a. *Histogram distribution of choice accuracy on the metacognition task in the group as a whole (N=80).* ***b.*** *Histogram distribution of metacognitive efficiency (meta-d'/d') on the metacognition task in the group as a whole (N=80).* ***c.*** *Average choice accuracy was matched for autism (N=40) and comparison participants (N=40) on the metacognition task. Error bars represent group mean ± SEM.* ***d.*** *Histogram distribution of the log of standardized response times (logRT) on the metacognition task in the group as a whole (N=80).*

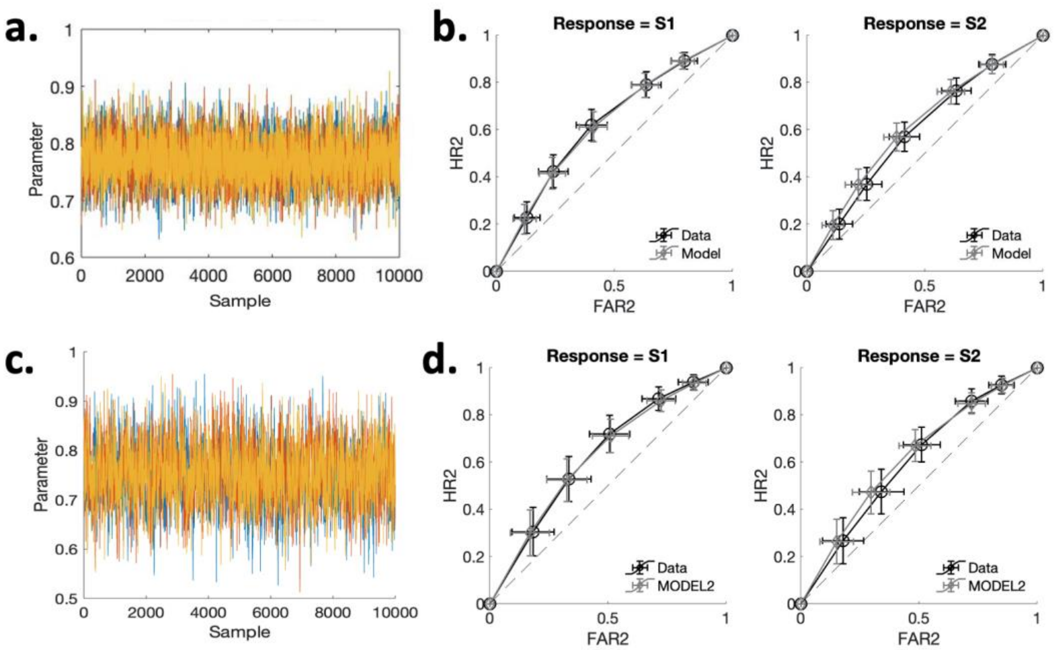
We again sought to ensure key variables related to metacognition and mentalizing were independent of first-order perceptual task performance. Staircase variability was not correlated with mentalizing ability (rs_{78} = -0.044, P = 0.71; **Supplementary Figure 2.2a**) and was not statistically different between groups (95% CI = [-0.036, 0.026], t_{78} = -0.31, P = 0.756; **Supplementary Figure 2.2b**). In addition, staircase variability was not correlated with metacognitive efficiency (rs_{78} = 0.031, P = 0.782; **Supplementary Figure 2.2c**). Perceptual sensitivity (*d'*) was not correlated with mentalizing ability (rs_{78} = 0.011, P = 0.924; **Figure 3.2d}**) and was not statistically different between groups (95% CI = [-0.083, 0.309], t_{78} = 1.15, P = 0.253; **Supplementary Figure 2.2e**).



Supplementary Figure 2.2. Correlations between the main variables of interest. a. *Mentalizing ability and staircase variability in the sample as a whole (N=80) were not correlated.* ***b.*** *Staircase variability was not different between the autism (N=40) and comparison groups (N=40).* ***c.*** *Metacognitive efficiency (meta-d'/d') and staircase variability in the sample as a whole (N=80) were not correlated.* ***d.*** *Mentalizing ability and perceptual ability (d') in the sample as a whole were not correlated.* ***e.*** *Perceptual ability (d') was not statistically different between autism (N=40) and comparison participants (N=40). Error bars represent the group means ± SEM.*

2.2 Posterior predictive checks

Finally, we asked whether the two HMeta-d models fitted to Experiment 2 data were reliable by means of convergence checks and posterior predictive checks. The hierarchical regression model converged well, indicated by the Gelman-Rubic statistics ($\hat{R}_{Mratio}$=1.0001 and plotted chains in **Supplementary Figure 2.3a**). In addition, posterior predictive plots recaptured key patterns of the participants' confidence responses correctly (**Supplementary Figure 2.3b**). The same was true for separate model fits to the comparison group ($\hat{R}_{Mratio}$=1.0014, **Supplementary Figure 2.3c, d**) and autism group ($\hat{R}_{Mratio}$=1.002, **Supplementary Figure 2.3e, f**).



Supplementary Figure 2.3. Posterior predictive checks on HMeta-d fits in Experiment 2. a. *MCMC chains for parameter meta-d'/d' (metacognitive efficiency) from the hierarchical regression model on autistic participants' data (N = 40) and* ***c.*** *on comparison participants' data (N = 40).* ***b.*** *Observed and model estimates for the Type 2 ROC curves for leftward (S1) and rightward (S2) responses from the hierarchical regression model are plotted for autistic participants' data (N=40).) and* ***d.*** *on comparison participants' data (N = 40). Error bars represent the mean ± standard error of the mean.*

SUPPLEMENTARY REFERENCES

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, *2019*(niz009). https://doi.org/10.1093/nc/niz009

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, *28*(24), 4014-4021.e8. https://doi.org/10.1016/j.cub.2018.10.053

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017