**Supplementary Materials**

*Participants*

28 adults with autism spectrum disorder (ASD) and 24 neurotypical adults (NTs) came to the UCL Institute of Cognitive Neuroscience to take part in a testing day. ASD participants had previously been diagnosed by an independent clinician, according to the DSM-IV(American Psychiatric Association, 2000) or ICD-10 criteria(Organization, 2012) [21 Asperger Syndrome, 3 Autism Spectrum Disorder, 4 Autism]. The Wechsler Adult Intelligence Scale (WAIS 3rd edition UK) had previously been administered to assess IQ (Wechsler and Hsiao-pin, 2011). Following exclusions, see below, 24 ASD participants (21 males) and 21 NT’s (16 males) were included in the final behavioural analysis, whereas 24 ASD participants (21 males) and 22 NT’s (17 males) were included in the pupillometry analysis. Demographic information pertaining to these samples are included in Table S1. Participants were matched for age and IQ. All participants provided written informed consent and were compensated financially for their time and travel expenses. The study was approved by the UCL Graduate School Ethics Committee (4357/001).

*Stimuli and Procedure*

The experiment was programmed in MATLAB (Mathworks R2013b.) using Psychtoolbox (Brainard, 1997). Participants were sat in front of a gamma corrected SONY CPD-G520P 21" CRT computer screen at a viewing distance of 600mm to complete the task. Participants were asked to make a two-alternative-forced-choice (2AFC) judgement about what they had seen, followed by a confidence rating in their decision (Figure 1A). Two circles (diameter 5.1°), each with a small cross hair in the centre, appeared simultaneously on the left and the right of the screen for 1000ms. When the crosshairs disappeared a variable number of dots (0.4°) appeared inside the circles for 700ms. The circles and dots were black, displayed on a grey background. Participants were then required to make a 2AFC judgement as to which circle contained more dots (e.g. the one on the left or the one on the right) via button press (left or right arrow key). The number of dots in each circle was bounded between 1 and 100. One circle, selected at random, always contained 50 dots. The number difference in number of dots between the two circles (Δ dots) was titrated to hold each participants performance at a constant level (~75% correct) using a one-up-two-down staircase procedure as used previously (Fleming *et al.*, 2010, 2012, 2014a). After two consecutive correct responses Δ dots was decreased by one dot and after one incorrect response Δ dots was increased by one dot. The staircase allowed us to equate the difficulty of the task across individuals. After making the 2AFC response participants were presented with a sliding scale to rate their confidence in the perceptual decision they had just made. The scale ranged from 1 (low confidence) to 6 (high confidence) and participants were encouraged to make use of the whole scale. The left and right arrow keys slid the cursor on the scale, allowing participants to indicate their confidence level. The cursor position initialised at a random location between points 3 and 4 on each trial and, after 3000ms timed out and turned red (for 500ms). During the main task participants received no feedback about the accuracy of their responses. Each participant completed 200 trials (eight blocks x 25 trials per block), taking ~25 minutes with breaks as needed. Following previous convention (Allen *et al.*, 2016), two ASD participants and two NTs were excluded on the basis of sustained extreme confidence rating behaviour (>65% responses rated as 6 or 1).

Before the main task participants received training and practice. First, example stimuli were shown with text below each circle indicating the number of dots in each circle (e.g. 40 vs. 60). Next participants practiced a series of 2AFC dot number judgements without confidence ratings. This not only familiarised participants with the task but also began to titrate the Δ dots by initialising the staircase. Finally, participants completed 10 practice trials of the complete task to gain familiarity with giving the confidence judgement.

*Data Analysis*

To estimate metacognitive efficiency we computed meta-d’ (Maniscalco and Lau, 2014). Meta-d’ was fit to each participant’s data using MATLAB code freely available at: <http://www.columbia.edu/~bsm2105/type2sdt/>. A full mathematical treatment of this model-based measure is available in (Maniscalco and Lau, 2012, 2014). As previously (Fleming *et al.*, 2014b), we binned continuous confidence ratings into four quartiles before analysis. In the decision theoretic framework of signal detection theory meta-d’ is a measure of type-2 sensitivity, in other words, the extent to which a subject’s confidence ratings can discriminate correct from incorrect judgements. Meta-d’ is expressed in the same units as d’, the extent to which subjects can distinguish which circle contains more dots (type-1 sensitivity). Accordingly, d’ is the sensory evidence available for perceptual decisions expressed in units of signal-to-noise ratio and meta-d’ is the sensory evidence available for metacognition, also expressed in units of signal-to-noise ratio (Fleming and Lau, 2014). Since meta-d’ (metacognitive sensitivity) is expressed in the same units as d’ (perceptual sensitivity), their relative comparison, meta-d’/d’ is a theoretically meaningful way to examine metacognitive ability relative to performance. A meta-d’/d’ value of 1 indicates a theoretically ideal value of metacognitive efficiency, whereas a value of 0.65, for example, would indicate 65% metacognitive efficiency. Using this ratio as a measure of metacognition eliminates response bias and performance confounds affecting other metacognitive measures (Galvin *et al.*, 2003; Maniscalco and Lau, 2012, 2014; Barrett *et al.*, 2013), making it an ideal to compare metacognitive ability in different participant groups (Fleming and Lau, 2014). To ensure that there were no differences in the model fits to the data when estimating meta-d’ we compared the sums of squared errors (SSE) for the fits for each individual between the groups. There was no significant group difference in the goodness of model fit (t (43) = 0.24, P = 0.81) and the correlation between meta-d’/d’ and model fit across individuals did not reach significance (r = 0.15, P = 0.33).

Parametric statistical tests were conducted on meta-d’/d’ measures and also log(meta-d’/d), to correct for the fact that ratio measures often violate assumptions of normality (Howell, 2012). Meta-d’ is bounded below by zero in theory, but in practice when fit using inbounded methods can produce negative values. Following previous practice (Fleming *et al.*, 2014a), data from two ASD participants and one NT were excluded from this analysis due to negative meta-d’ estimates, precluding the calculation of log-transformed metacognitive efficiency scores. Parametric analyses were supplemented with bootstrapped 95% confidence intervals (CI) and bootstrapped t-tests (1000 samples with replacement). Prior to analysis, Shapiro-Wilk tests confirmed that the meta-d’/d’ metric was normally distributed in each group (ASD: W(24)=0.95, P=0.247; NT: W(21)=0.98, P=0.917) and Grubbs test for outliers confirmed no influential datapoints in either group (ASD: most extreme value=0.03, Z=1.83, Grubbs critical Z=2.80; NT: most extreme value=0.28, Z=1.98, Grubbs critical Z=2.73).

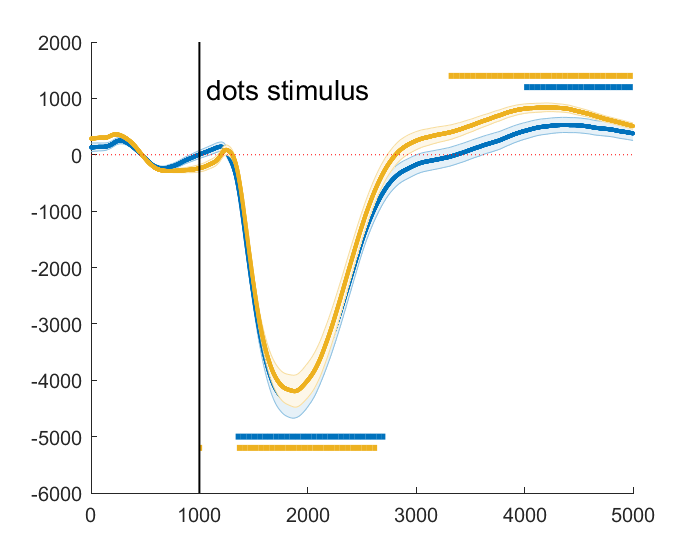
In a complementary analysis we estimated meta-d’/d’ using a hierarchical Bayesian version of the standard metacognitive efficiency model (HMeta-d toolbox (Fleming, 2017), <https://github.com/smfleming/HMM>) which allows estimation and comparison of *group-level* parameters. The parameters were estimated using Markov-Chain Monte-Carlo methods (MCMC, here: 3 chains of 10’000 samples each, burn-in of 1000 samples) as implemented in JAGS (<http://mcmc-jags.sourceforge.net>). We followed the standard analysis approach by comparing the posterior group distributions in metacognitive efficiency (Fleming, 2017). Significance was assessed as whether the difference of the group posteriors significantly overlapped with 0 (similar to a classical or frequentist statistical test, it assesses the probability of the difference between the groups to be 0), as well as the 95% high density intervals of the difference distribution (analogous to confidence intervals).

All statistical analyses of eye-tracking data were performed in MATLAB (Mathworks). For pupil data, blinks were treated with linear interpolation, and the resulting pupil traces were baseline corrected, epoched and smoothed. Only trials in which 80% or more of the samples were successfully tracked were included in the analysis. The same two ASD and two NT participants excluded from the behavioural analysis (for exhibiting extreme confidence behaviour) were excluded from the confidence-linked pupillometry analysis. In addition, eye tracking was not successful in two ASD participants leaving n=24 ASD and N=22 NT participants in this analysis (see Table S1 for participant demographics).

Pupil size in rodents, non-human primates and humans (Murphy *et al.*, 2014; Joshi *et al.*, 2016; Larsen and Waters, 2018) is a peripheral proxy for phasic activity of the noradrenergic locus coeruleus. For the pupillometry analyses we applied a general linear modelling approach to estimate the encoding of confidence in pupil size across time (while controlling for confounding variables). Two 2000ms time windows within each trial were examined (stimulus presentation and confidence judgement). Regression analyses were conducted for each individual time point, with confidence rating as the predictor of interest. RT to make the confidence judgement, RT to make the initial perceptual decision, difficulty (∆ dots), and stimulus side (left, right) were all entered as control regressors. The resultant timeseries of β weights (multiple regression conducted at every time point) provided estimates of when and how subjective confidence is encoded in pupil size. Positive β weights indicate that when confidence was rated high, pupil size increased. Negative β weights indicate that when confidence was rated high, pupil size decreased.

At the group level, we then conducted t tests for the positive or negative effects of the regressors of interest, and the independent-samples difference between groups, corrected for multiple comparisons with a cluster-based permutation approach at 2,000 permutations (cluster α = 0.05) (Groppe *et al.*, 2011).

**Supplemental Figure 1**



Arbitrary Units

**Figure S1:** Shows the basic pupil response to the presentation of the dots stimulus in the ASD participants (blue) and NT participants (yellow). Trials were self-paced and varied in duration, so this figure shows phasic changes in pupil size in the 4000ms after the stimulus first appeared. In the main analysis (see Figure 4), the effect of confidence on pupil size is time-locked to the onset of the dots and the onset of the confidence rating respectively. Blue solid horizontal lines indicate significant time clusters when the pupil response differed from zero in ASD participants. The yellow solid horizontal line indicates the same for NT participants. There were no significant clusters in which the pupil response different in the ASD and NT groups (2,000 permutations; cluster *α* = 0.05).

**Supplemental Table 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group** |  |  |  |  |  |  |
|  | **ASD (n=24)** |  | **NT(n=21)** |  |  |  |  |
| Behavioural analysis | **mean(sd)** | **range** | **mean(sd)** | **range** | ***t*** | ***df*** | ***P*** |
| **Age(years)** | 28.6(6.3) | 18-38 | 28.1(5.9) | 21-40 | 0.289 | 43 | 0.77 |
| **Full Scale IQ** | 119.6(13.7) | 89-152 | 123.4(14.6) | 98-151 | -0.908 | 43 | 0.37 |
| **Verbal IQ** | 122.0(15.1) | 85-155 | 124.0(15.27) | 96-150 | -0.450 | 43 | 0.66 |
| **Performance IQ** | 113.2(13.0) | 89-132 | 116.8(15.1) | 80-148 | -0.867 | 43 | 0.39 |
| **ADOS-total** | 8.9(2.9) | 4-17 |  |  |  |  |  |
| **ADOS-C** | 2.7(1.4) | 1-6 |  |  |  |  |  |
| **ADOS-SI** | 6.1(1.8) | 2-11 |  |  |  |  |  |
|  | **Group** | |  | |  | | |
|  | **ASD (n=24)** | | **NT(n=22)** | |  | | |
| Pupillometry analysis | **mean(sd)** | **range** | **mean(sd)** | **range** | ***t*** | ***df*** | ***P*** |
| **Age(years)** | 28.7(6.1) | 18-38 | 28.0(5.8) | 21-40 | 0.380 | 44 | 0.71 |
| **Full Scale IQ** | 118.6(12.8) | 89-152 | 123.5(14.3) | 98-151 | -1.210 | 44 | 0.23 |
| **Verbal IQ** | 121.0(14.6) | 85-155 | 124.7(12.2) | 96-150 | -0.826 | 44 | 0.41 |
| **Performance IQ** | 112.8(12.2) | 95-132 | 116.2(15.1) | 80-148 | -0.860 | 44 | 0.39 |
| **ADOS-total** | 8.7(2.8) | 4-17 |  |  |  |  |  |
| **ADOS-C** | 2.7(1.4) | 1-6 |  |  |  |  |  |
| **ADOS-SI** | 5.9(1.8) | 2-11 |  |  |  |  |  |

Table S1 – Participant demographics. The participants were matched on both age and IQ ADOS, autism diagnostic observation schedule for both the behavioural and pupillometry analyses; C, communication; SI, social interaction; IQ, intelligence quotient.

**Supplementary References**

Allen M, Frank D, Schwarzkopf DS, Fardo F, Winston JS, Hauser TU, et al. Unexpected arousal modulates the influence of sensory noise on confidence. Elife 2016; 5: e18103.

American Psychiatric Association. Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000). American Psychiatric Association; 2000.

Barrett AB, Dienes Z, Seth AK. Measures of metacognition on signal-detection theoretic models. Psychol. Methods 2013; 18: 535.

Brainard DH. The psychophysics toolbox. Spat. Vis. 1997; 10: 433–436.

Fleming SM. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. Neurosci. Conscious. 2017; 2017: nix007.

Fleming SM, Huijgen J, Dolan RJ. Prefrontal contributions to metacognition in perceptual decision making. J. Neurosci. 2012; 32: 6117–6125.

Fleming SM, Lau HC. How to measure metacognition. 2014

Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain 2014a; 137: 2811–2822.

Fleming SM, Ryu J, Golfinos JG, Blackmon KE. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain 2014b; 137: 2811–2822.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science 2010; 329: 1541–1543.

Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychon. Bull. Rev. 2003; 10: 843–876.

Groppe DM, Urbach TP, Kutas M. Mass univariate analysis of event‐related brain potentials/fields II: Simulation studies. Psychophysiology 2011; 48: 1726–1737.

Howell D. Statistical methods for psychology. Cengage Learning; 2012.

Joshi S, Li Y, Kalwani RM, Gold JI. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. Neuron 2016; 89: 221–234.

Larsen RS, Waters J. Neuromodulatory Correlates of Pupil Dilation. Front. Neural Circuits 2018; 12: 21.

Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. 2012; 21: 422–430.

Maniscalco B, Lau H. Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d′, Response-Specific Meta-d′, and the Unequal Variance SDT Model. In: The Cognitive Neuroscience of Metacognition. Springer; 2014. p. 25–66.

Murphy PR, O’connell RG, O’sullivan M, Robertson IH, Balsters JH. Pupil diameter covaries with BOLD activity in human locus coeruleus. Hum. Brain Mapp. 2014; 35: 4140–4154.

Organization WH. International classification of diseases (ICD). 2012

Wechsler D, Hsiao-pin C. WASI-II: Wechsler abbreviated scale of intelligence. Pearson; 2011.