Task-specific Frontopolar and temporoparietal contributions to decision confidence

Matan Mazor & Stephen M. Fleming

1 Objective

In a previous study we compared the parametric effect of subjective decision confidence on brain activation in two perceptual decision-making tasks: a discrimination task (*was the grating tilted clockwise or counterclockwise?*) and a detection task (*was there any grating present at all?*)[5]. As expected, we found a linear effect of confidence in a set of pre-defined regions of interest, with high confidence levels associated with a stronger (ventromedial prefrontal cortex, vmPFC; precuneus; ventral striatum) or weaker (posterior medial frontal cortex, pMFC) signal, across both tasks and responses.

Exploratory analysis additionally revealed a widespread positive quadratic effect of confidence, with stronger signal associated with using the extreme ends of the confidence scale. In the right frontopolar cortex, right superior temporal sulcus (STS) and right pre-SMA this quadratic effect was stronger for the detection task, where participants made judgments about the presence or absence of a grating. Additionally, in the right temporoparietal junction (TPJ), the linear effect of confidence was stronger following judgments about target absence compared with judgments about target presence.

The difference in confidence-related activations between detection and discrimination can originate from the differences in the shape and spread of the distributions of incoming perceptual input. Alternatively, this difference can reflect the involvement of distinct metacognitive components in decisions and confidence-formation about presence and absence. The design of our previous study did not allow us to decide between these alternative accounts. Here, we introduce a third condition to our experimental design: a discrimination task with the distributional properties of a detection task (*tilt recognition*; similar to [2]). The two explanations make different predictions for this third condition. The first account predicts confidence effects similar to those observed in detection, as the two tasks share a similar distributional structure (unequal variance). Conversely, the second account predicts confidence effects similar to those observed in discrimination, as this task does not involve inference abour true absences or presences.

The objectives of this study are to:

1. Replicate our finding of an interaction between task (discrimination/detection) and the quadratic effect of confidence, in medial and lateral frontopolar cortex, as well as in the STS and pre-SMA.

2. Replicate our finding of an interaction between detection response (PRESENT/ABSENT) and the linear effect of confidence in the right TPJ.

3. Compare the quadratic effect of confidence in the tilt-recognition task with the quadratic effect of confidence in the detection and discrimination tasks in the frontopolar cortex, the STS and the pre-SMA.

4. Compare the response-specific linear effects of confidence in the tilt-recognition task with the response-specific linear effects of confidence in the detection and discrimination tasks in the right TPJ.

All design and analysis details will be pre-registered before data acquisition and time-locked using pre-RNG randomization [6].

2 Design

We will test healthy subjects in a 3 Tesla MRI scanner at the Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London. Participants will be acquainted with the task in a preceding behavioural session. During this session, task difficulty will be adjusted independently for the three tasks, targeting approximately 70 % accuracy. We will aim to acquire usable data from 35 participants after applying our pre-specified inclusion criteria.

2.1 Statistical power

Assuming that the effect size estimates observed in [5] are not inflated and that reducing the number of trials per condition will not affect group-level sensitivity, with a sample size of 35 we will have:

1. statistical power of 85% to replicate the interaction between task and the quadratic effect of confidence in the FPm cluster (sensitivity for detecting an effect of 0.46 standard deviations from zero with a one-sided t test, $\alpha = 0.05$).

2. statistical power of 66% to replicate the interaction between task and the quadratic effect of confidence in the FPl cluster (sensitivity for detecting an effect of 0.36 standard deviations from zero with a one-sided t test, $\alpha = 0.05$).

3. statistical power of 66% to replicate the interaction between task and the quadratic effect of confidence in the rTPJ cluster (sensitivity for detecting an effect of 0.36 standard deviations from zero with a one-sided t test, $\alpha = 0.05$).

4. statistical power of 65% to replicate the interaction between detection response and the linear effect of confidence in the rTPJ cluster (sensitivity for detecting an effect of 0.35 standard deviations from zero with a one-sided t test, $\alpha = 0.05$).

2.2 Procedure

Participants will be acquainted with the task in a preceding behavioural session. During this session, difficulty will be adjusted independently for the three tasks, targeting around 70 % accuracy. This will be achieved by using a 1 up 2 down procedure on stimulus visibility (discrimination and detection task) and on the standard deviation of the orientation distribution (tilt recognition). The scanning session will start with a structural MP-RAGE scan, followed by a fieldmap scan. During this time, participants will perform a staircase procedure similar to that performed in the preceding behavioural session. This calibration phase is used to further calibrate participants' performance on the three tasks inside the MRI scanner. After completing the calibration phase, participants will undergo 5 ten-minute functional scanner runs, each comprising one block of 26 trials from each experimental condition, presented in a random order.

After a temporally jittered rest period of 500-4000 milliseconds, each trial will start with a fixation cross (500 milliseconds), followed by a presentation of a target for 500 milliseconds. In all three conditions, stimuli will consist of 10 grayscale frames presented at 20 frames per second within a circle of diameter 3°. Stimuli will be generated in the following way:

1. Generate 10 grayscale frames ($F_1, ...F_{10}$), each an array of 142 by 142 random luminance values.

2. Create a 142 by 142 sinusudial grating (G; 24 pixels per period, random phase). The orientation of the grating is determined according to the trial type.

3. The grating visibility for frame i is $p_i = v \times exp(-|i-5|/2)$ with v being the visibility level in this trial (0 for target-absent trials).

4. for each pixel in the frame F_i, j, k, replace the luminance value for this pixel with the luminance value of this pixel in the grating (G_j, k) with a probability of p_i.

Example stimuli can be found in the instructions slides. For the detection and discrimination tasks, stimulus visibility will be calibrated by controlling the value v on a logarithmic scale. Participants will perform the following three tasks:

- **Discrimination** Decide whether the grating was tilted clockwise (50% of trials; 45° relative to a vertical baseline) or anticlockwise (-45° relative to a vertical baseline).

- **Tilt Recognition** Decide whether the grating was vertical (50% of trials; 0°) or tilted (sampled from a normal distribution with mean 0° and standard deviation that will be calibrated to reach 70% accuracy; $\sigma_{orientation}$). Stimuli will be presented with a fixed v value of 0.2 at which stimuli are clearly visible.

- **Detection** Decide whether the grating was present (50% of trials) or absent. Gratings in the 'present' trials will be sampled from a normal distribution with mean 0 and standard deviation $\sigma_{orientation}$ (yoked to the tilt recognition task).

After stimulus offset, participants will use their right-hand index and middle fingers to make a perceptual decision about the orientation of the grating (discrimination and tilt-recognition blocks), or about the presence or absence of a grating (detection blocks). The response mapping will be counterbalanced between blocks, such that an index finger press will be used to indicate a clockwise tilt on half of the trials, and an anticlockwise tilt on the other half. Similarly, in half of the tilt-recognition trials the index finger will be mapped to a VERTICAL response, and on the other half to a TILTED response. Lastly, in half of the detection trials the index finger will be mapped to a YES ('target present') response, and on the other half to a NO ('target absent') response.

Immediately after making a decision, participants will rate their confidence on a 6-point scale by using two keys to increase and decrease their reported confidence level with their left-hand thumb. Confidence levels will be indicated by the size and color of a circle presented at the center of the screen. The initial size and color of the circle will be determined randomly at the beginning of the confidence rating phase, to decorrelate the number of button presses and the final confidence rating. The mapping between color and size to confidence will be counterbalanced between participants: for half of the participants high confidence will be mapped to small, red circles, and for the other half high confidence will be mapped to large, blue circles. This counterbalancing is employed to isolate confidence-related activations from activations that originate from the perceptual properties of the confidence scale or from differences in the motor requirement to press the upper and lower buttons. The perceptual decision and the confidence rating phases will be restricted to 1000 and 2500 milliseconds, respectively. No feedback will be delivered to subjects about their performance.

To avoid stimulus-driven fluctuations in confidence, grating visibility (v) and $\sigma_{orientation}$ will be kept fixed within each experimental block. Nevertheless, following experimental blocks with markedly bad ($\leq$52.5%) or good ($\geq$85%) accuracy, v or $\sigma_{orientation}$ will be adjusted for the next block of the same task (divided or multiplied by a factor of 0.95 for bad and good performance, respectively).

Compensation

Participants will be payed £10 for the behavioural session and £20 for the scanning session. To incentivize participants to do their best at the task and rate their confidence accurately, we will offer a bonus payment according to the following payment schedule: bonus = £$\frac{\overrightarrow{accuracy}\cdot\overrightarrow{confidence}}{200}$ Where $\overrightarrow{accuracy}$ is a vector of 1 and -1 for correct and incorrect responses, and $\overrightarrow{confidence}$ is a vector of integers in the range of 1 to 6, representing confidence reports for all trials. We will explain the payment structure to participants in the preceding behavioural session. Specifically, we will advise participants that to maximize their bonus they should do their best at the main task, rate the confidence higher when they believe they are correct, and rate their confidence lower when they believe they might be wrong.

2.3 Scanning parameters

Scanning will take place at the Wellcome Centre for Human Neuroimaging, London, using a 3 Tesla Siemens Prisma MRI scanner with a 64-channel head coil. We will acquire structural images using an MPRAGE sequence (1x1x1mm voxels, 176 slices, in plane FoV = 256x256 mm^2), followed by a double-echo FLASH (gradient echo) sequence with TE1=10ms and TE2=12.46ms (64 slices, slice thickness = 2mm, gap = 1mm, in plane FoV= $192 \times 192 mm^2$, resolution = 3×3 mm ^{2}) that will later be used for field inhomogeneity correction. Functional scans will be acquired using a 2D EPI sequence, optimized for regions near the orbitofrontal cortex (3.0x3.0x3.0mm voxels, TR=3.36 seconds, TE = 30 ms, 48 slices tilted by -30 degrees with respect to the T¿C axis, matrix size = 64x72, Z-shim=-1.4).

3 Analysis

3.1 Behavioural analysis

1. We will perform an analysis of variance (ANOVA) to compare accuracy levels between the three tasks across subjects.

2. Using a t-test on the subject-wise response probabilities ((p(response='yes'), p(response='vertical') and p(response='CW')) we will test for consistent response bias effects across participants.

3. We will compare metacognitive sensitivity (meta-d') and bias (mean confidence) between the three tasks.

4. Based on previous work and pilot data, we predict lower metacognitive sensitivity for NO than for YES responses in the detection task, and for VERTICAL than for TILTED in the tilt recognition task. To quantify this effect, we will plot the response-conditional type 2 Receiver Operator Characteristic (ROC) curves within the detection task, and compare the areas under the curves for the different responses [7].

3.2 fMRI data preprocessing

Data preprocessing will follow the procedure described in [8]:

> Imaging analysis will be performed using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run will be discarded to allow for T1 stabilization. Functional images will be realigned and unwarped using local field maps (Andersson et al., 2001) and then slice-time corrected (Sladky et al., 2011). Each participant's structural image will be segmented into gray matter, white matter, CSF, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner and Friston, 2005). This mapping will be applied to both structural and functional images to create normalized images to Montreal Neurological Institute (MNI) space. Normalized images will be spatially smoothed using a Gaussian kernel (6 mm FWHM). We set a within-run 4 mm affine motion cutoff criterion. Preprocessing and construction of first- and second-level models will use standardized pipelines and scripts available at https://github.com/metacoglab/MetaLabCore/

3.3 Exclusion criteria

Block exclusion

Individual experimental blocks will not be analysed in the following cases:

1. More than 20% of the trials in the block were missed.
2. Mean accuracy was lower than 60%.
3. The participant used the same response in more than 80% of the trials.
4. For a particular response, the same confidence level was reported for more than 90% of the trials.

The first trial of each block will be excluded from all analyses, leaving 25 usable trials per block.

Subject exclusion

Subjects will be excluded from all analyses in case more then two blocks of one task were excluded based on the above block exclusion criteria.

4 Regions of Interest

In addition to an exploratory whole-brain analysis (corrected for multiple comparisons at the cluster level), our analysis will focus on the following a priori regions of interest, following the ROIs used in [3] and the results from [5]:

1. *Medial frontopolar cortex (FPm).* We will use a connectivity-based parcellation [9] to define this region. The right hemisphere mask will be mirrored to create a bilateral mask.

2. *Lateral frontopolar cortex (FPl).* We will use a connectivity-based parcellation [9] to define this region. The right hemisphere mask will be mirrored to create a bilateral mask.

3. *Brodman area 46 (BA46).* We will use a connectivity-based parcellation [9] to define this region. The right hemisphere mask will be mirrored to create a bilateral mask.

4. *Ventromedial prefrontal cortex (vmPFC).* The vmPFC ROI will be defined as a 8-mm sphere around MNI coordinates [0,46,-7], obtained from a meta-analysis of subjective-value related activations [1] and aligned to the cortical midline.

5. *Right temporoparietal junction (rTPJ).* Defined using the contrast $confidence_{No} - confidence_{Yes}$ from [5] (peak voxel [54,-46, 26], see mask attached to the protocol folder at 'ROIs/rTPJ.nii').

6. *Right superior temporal sulcus (rSTS).* Defined using the rSTS cluster from the contrast $confidence^2_{Detection} - confidence^2_{Discrimnation}$ from [5] (peak voxel [60,-43,2], see mask attached to the protocol folder at 'ROIs/rSTS.nii').

7. *Pre-supplementary moror area (preSMA).* Defined using the preSMA cluster from the contrast $confidence^2_{Detection} - confidence^2_{Discrimnation}$ from [5] (peak voxel [0,35,47], see mask attached to the protocol folder at 'ROIs/preSMA.nii').

4.1 Univariate fMRI analysis

Univariate analysis will follow a similar procedure to that described in [5]. After preprocessing, runs will be temporally concatenated and a design matrix will be fitted to the entire timecourse. Here we chose not to exclude entire runs, but specific blocks of trials. This will be achieved by modeling excluded blocks with a separate nuisance regressor. We will estimate two design matrices:

Quadratic-Confidence Design Matrix (QC-DM)

The quadratic-confidence design matrix for the univariate GLM analysis will consist of 18 regressors of interest. There will be a regressor for each of the six responses: YES, NO, TILTED, VERTICAL, CLOCKWISE and ANTICLOCKWISE. Similar to the main design matrix, the relevant trials will be modeled by a boxcar regressor with nonzero entries at the 4000 millisecond interval starting at the onset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). Each of these primary regressors will be accompanied by two parametric modulators, representing the linear and quadratic effects of confidence. Together, the design matrix will include 18 regressors (6 responses + 6 linear confidence regressors + 6 quadratic confidence regressors). The QC-DM will include the same set of nuisance regressors as the main design matrix.

Table 1: List of regressors in the quadratic confidence design matrix (QC-DM).

		Task	Condition
1	CW		
2	CW_conf	Discrimination	Clockwise
3	CW_conf ^{2}		
4	ACW		
5	ACW_conf	Discrimination	Anticlockwise
6	ACW_conf ^{2}		
7	P		
8	P_conf	Detection	Present
9	P_conf ^{2}		
10	A		
11	A_conf	Detection	Absent
12	A_conf ^{2}		
13	V		
14	V_conf	Tilt recognition	Vertical
15	V_conf ^{2}		
16	T		
17	T_conf	Tilt recognition	Tilted
18	T_conf ^{2}		

Trials in which the participant did not respond within the 1000 millisecond time frame will be modeled by a separate regressor. Similarly, the first trial in each block will be modeled by a separate regressor. The design matrix will also include a run-wise constant term regressor, an instruction-screen regressor for the beginning of each block, motion regressors (the 6 motion parameters and their first derivatives as extracted by SPM in the head motion correction preprocessing phase) and regressors for physiological measures. Button presses will be modeled as stick functions, convolved with the canonical HRF, and separated into three regressors: two regressors for the right and left right-hand buttons, and one regressor for both up and down left-hand presses.

We will perform the following contrasts on the mean beta estimates from DM-1 for our 7 ROIs, as well as in an exploratory whole-brain contrast (corrected for multiple comparisons at the cluster level):

1. Main linear effect of confidence, across tasks and responses (T contrast).
2. Main effect of task (F contrast).
3. $task \times confidence$ interaction (F contrast).

4. Within detection, $response \times confidence$ interaction (T contrasts).
5. Within tilt recognition, $response \times confidence$ interaction (T contrasts).
6. Main effect of $confidence^2$, across tasks and responses (F contrast).
7. $task \times confidence^2$ interaction (F contrast).
8. $confidence^2_{Detection} - confidence^2_{Discrimination}$ (T contrast)
9. $confidence^2_{TiltRecongnition} - confidence^2_{Discrimination}$ (T contrast)

Categorical-Confidence Design Matrices CC-DM

A set of three design matrices - one for each task - in which confidence level is modeled as a categorical variable will be created. This design matrix will consist of only one regressor of interest for all included trials, modeled by a boxcar with nonzero entries at the 4000 millisecond interval starting at the onset of the stimulus and ending immediately after the confidence rating phase, convolved with the canonical hemodynamic response function (HRF). This regressor will in turn be modulated by a series of 12 dummy (0/1) parametric modulators - one for every response (YES and NO for detection, VERTICAL and TILTED for tilt recognition and CLOCKWISE and ANTICLOCKWISE for discrimination) and confidence rating (1-6 for both tasks). Using three design matrices instead of one will allow us to set trials from the remaining two tasks to serve as baseline for the task of interest. These design matrices will include the same set of nuisance regressors as the main design matrix.

For each participant, beta-estimates from the categorical-confidence design matrices will be given as input to six response-specific multiple linear regression models, with linear confidence and quadratic confidence as predictors, in addition to an intercept term. The subject-specific coefficients will then be subjected to ordinary least squares group-level inference, to compare linear and quadratic effects of confidence between responses. The rationale for employing this two-step approach is its ambivalence to differences in the confidence distributions for the six responses, that may bias the estimation of quadratic and linear terms.

We will perform the following comparisons within our regions of interest and within clusters discovered in contrasts on the beta-estimates from DM-1 and QC-DM:

1. Linear effect of confidence (main effect)
2. Qudratic effect of confidence (main effect)
3. Interaction of the linear effect of confidence with task.
4. Interaction of the quadratic effect of confidence with task.
5. Within detection, interaction of the linear effect of confidence and response.

6. Within detection, interaction of the quadratic effect of confidence and response.
7. Within tilt recognition, interaction of the linear effect of confidence and response.
8. Within tilt recognition, interaction of the quadratic effect of confidence and response.

4.2 Multivariate analysis

Representation Similarity Analysis (RSA)

Representational Similarity Analysis [4] will be used to detect consistent spatiotemporal structures in the representation of choice and confidence across tasks and responses, within our 7 pre-specified ROIs. High and low confidence trials will be defined using a median split within each response category. The empirical Representational Dissimilarity Matrix (RDM) will be compared against the following set of a-priori RDMs:

1. **Task** (figure 1, panel a). Trials of the same task are similar, trials of different tasks are different.
2. **Variance structure** (figure 1, panel b). Discrimination trials are similar to each other. Detection PRESENT trials and tilt recognition TILTED trials are similar (high variance), and detection ABSENT trials are similar to tilt recognition VERTICAL trial (low variance).
3. **Decision: detection only** (figure 1, panel c). Only detection decision trials are reliably represented.
4. **Decision: unequal variance only** (figure 1, panel d). Only detection and tilt recognition trials are reliably represented.
5. **Confidence** (figure 1, panel e). High and low confidence trials are represented differently, without an effect of task or response.
6. **Confidence and variance structure interaction** (figure 1, panel f). High and low confidence trials are represented differently. This effect is modulated by the variance structure of the trial category.
7. **Confidence in detection only** (figure 1, panel g). High and low confidence trials are represented differently in detection only.
8. **Confidence in unequal variance only** (figure 1, panel h). High and low confidence trials are represented differently in detection and tilt recognition only.

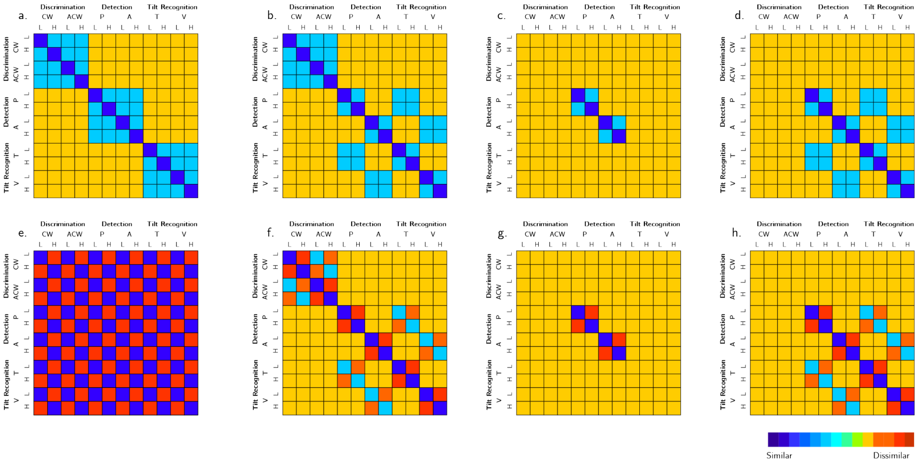

Figure 1: Our eight a priori RDMs: task, variance structure, detection, unequal variance, confidence, confidence and variance structure interaction, confidence in detection, and confidence in unequal variance.

4.3 Group level inference

For exploratory whole-brain analysis, group level inference will follow an ordinary least squares (OLS) procedure on the subject-specific contrast maps. Correction for multiple comparisons will be performed at the cluster level, using a significance threshold of P=0.05 and a cluster defining threshold of P=0.001. No correction for multiple comparisons will be applied to our prespecified ROIs.

References

[1] Oscar Bartra, Joseph T McGuire, and Joseph W Kable. The valuation system: a coordinate-based meta-analysis of bold fmri experiments examining neural correlates of subjective value. *Neuroimage*, 76:412–427, 2013.

[2] Rachel N Denison, William T Adler, Marisa Carrasco, and Wei Ji Ma. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115(43):11090–11095, 2018.

[3] Stephen M Fleming, Elisabeth J Van Der Putten, and Nathaniel D Daw. Neural mediators of changes of mind about perceptual decisions. *Nature neuroscience*, 21(4):617, 2018.

[4] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[5] Matan Mazor, Karl Friston, and Stephen M Fleming. Distinct neural contributions to metacognition for detecting (but not discriminating) visual stimuli. *BioRxiv*, page 853366, 2019.

[6] Matan Mazor, Noam Mazor, and Roy Mukamel. A novel tool for time-locking study plans to results. *European Journal of Neuroscience*, 49(9):1149–1156, 2019.

[7] Julia DI Meuwese, Anouk M van Loon, Victor AF Lamme, and Johannes J Fahrenfort. The subjective experience of object recognition: comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, 76(4):1057–1068, 2014.

[8] Jorge Morales, Hakwan Lau, and Stephen M Fleming. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, 38(14):3534–3546, 2018.

[9] Franz-Xaver Neubert, Rogier B Mars, Adam G Thomas, Jerome Sallet, and Matthew FS Rushworth. Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron*, 81(3):700–713, 2014.