Multivariate pattern analysis of fMRI data for imaginary and real colours in grapheme-colour synaesthesia

Mathieu J. Ruiz (1,2), Michel Dojat (2), Jean-Michel Hupé *(1)

(1) Centre de Recherche Cerveau et Cognition, Université de Toulouse Paul Sabatier & CNRS, 31300 Toulouse, France

(2) Grenoble Institut des Neurosciences, Université Grenoble Alpes, INSERM & CHU Grenoble Alpes, 38000 Grenoble, France

* Corresponding author: Jean-Michel Hupé
CNRS CERCO UMR 5549, Pavillon Baudot, CHU Purpan, BP 25202, 31052 Toulouse Cedex 3, France
email: jean-michel.hupe@cnrs.fr

Abstract

Grapheme-colour synaesthesia is a subjective phenomenon related to perception and imagination, in which some people involuntarily but systematically associate specific, idiosyncratic colours to achromatic letters or digits. Its investigation is relevant to unravel the neural correlates of colour perception in isolation from low-level neural processing of spectral components, as well as the neural correlates of imagination by being able to reliably trigger imaginary colour experiences. However, functional MRI studies using univariate analyses failed to provide univocal evidence of the activation of the 'colour network' by synaesthesia. Applying Multivariate (multivoxel) Pattern Analysis (MVPA) on 20 synaesthetes and 20 control participants, we tested whether the neural processing of real colours (concentric rings) and synaesthetic colours (black graphemes) shared patterns of activations. Region of interest analyses in retinotopically and anatomically defined visual regions revealed neither evidence of shared circuits for real and synaesthetic colour processing, nor processing difference between synaesthetes and controls. We also found no correlation with individual differences, characterised by measuring the strength of synaesthetic associations. The whole brain, searchlight, analysis led to similar results. We conclude that identifying the neural correlates of the synaesthetic experience of colours may still be beyond the reach of present technology and data analysis techniques.

Introduction

Synaesthesia is a subjective experience shared by only a fraction of the population [1-5], offering, in principle, an opportunity to study the neural bases of subjective experience, drawing on individual differences just like in neuropsychology, but involving healthy people. Moreover, colour, the typical prototype of a qualia (what it feels like to perceive something) is the most often cited (or at least studied [6]) content of the synaesthetic experience. However, the very subjective nature of the synaesthetic experience represents a major obstacle when trying to set an objective and operational definition, as required in an experimental protocol. Not only subjective descriptions may vary a lot between subjects [7], but also within subjects when asked to fill-up the same questionnaire again [8] or when describing their subjective experience of colour for different letters [9]. Using psychophysical tests, the synaesthetic experience of colour appears more similar to imagined or remembered than perceived colours [10-13]. The experience of synaesthetic colours can be indeed formally described as a form of mental imagery, since it occurs without any corresponding spectral stimulation. The obligatory experience of colour when exposed to letters or digits may therefore justify the label of 'intrusive visual imagery' [14]. Unfortunately, this simplification does not help much with defining the phenomenological content of synaesthesia, since self-reports of mental imagery differ at least as much as those of synaesthesia [15], with mixed evidence about whether the presence of synaesthesia may relate to individual differences in mental imagery [16]. One may, however, study how much synaesthesia requires the neural resources involved in visual perception. This bottom-up approach, which does not address the phenomenological issue, can at least be operationalized. Moreover, grapheme-colour synaesthesia offers a unique opportunity regarding the neural correlates of imagination as it restrains both individual variability and the content specificity of visual imagery. Last but not least, synaesthetic colours are systematically triggered by letters and digits, unlike "regular" mental imagery that depends on both the good will and the (uneven) ability of subjects.

Several brain imaging studies have compared activations in the visual cortex for real and synaesthetic colours, whose majority did not reveal any overlap. There were even questions whether activations triggered by synaesthetic stimuli, when observed, were in fact related to the synaesthetic experience

at all [12]. This surprising 'Null' result may be due to methodological limitations since only massive univariate analysis of brain imaging data were used so far, which may reveal only processes well localized in the brain [9]. Multivariate (multivoxel) Pattern Analysis (MVPA) does not suffer from such a restriction. MVPA provides a way to reveal how information is encoded by the brain [17-20]. It has been applied successfully to the decoding of aspects of mental images [21,22]. Using fMRI, here we simply asked whether classifiers trained on patterns of blood oxygenation dependent signals (BOLD responses) elicited by different coloured stimuli could predict which synaesthetic colours were experienced by synaesthetes when seeing achromatic letters and digits. We studied in particular the early stages of visual processing by identifying cortical areas V1 to V4 in each subject, using retinotopic mapping, thus avoiding the problems related to structural normalization [23,24]. We also explored the whole visual cortex (including parts of the parietal cortex) using regions of interest based on a probabilistic atlas [25], and performed whole brain searchlight analyses [26]. We compared all the measures obtained in synaesthetes with those obtained in a group of non-synaesthetes to take into account any non-specific effect related to the choice of stimuli. We also took into account the individual variability of the synaesthetic experience: without any possibility to characterize objectively the different phenomenological accounts, we measured the strength of the synaesthetic associations [27].

Material and Methods

Participants

Sample size was *a priori* and arbitrary set to 20 synaesthetes and 20 non-synaesthetes. Sample size was small, yet on a par with the literature (the largest studies in the field had 42 synaesthetes *vs.* 19 controls [28] and groups of 20 participants [29]), because of the exploratory nature of the experiment (we know of no other attempt so far of applying MVPA to synaesthesia). Additional power was however expected because we could match each synaesthete with a control subject (paired comparisons). We should stress that our interpretations are not based on p-values, making the requirement of *a priori* power analyses irrelevant. The study was performed in accordance with the Declaration of Helsinki, it received approval by the Institutional Review Board of Grenoble (CPP 12-CHUG-17, approval date 04/04/2012) and written, informed consent was obtained from all subjects. A medical doctor verified that all subjects were without past or current brain disease and had no detected cognitive deficit. All subjects had normal colour perception on the Lanthony D-15 desaturated colour test (Richmond products), and normal or corrected to normal eyesight (then using MRI-compatible glasses).

Synaesthetes (16 women) were between 21 and 42 years old (M = 27.9, SD = 5.5). Recruitment was diverse and opportunistic, based on self-referral following publicity on internet: lab webpage, *Facebook* event, announcements on university networks in Grenoble and Paris. Potential participants, after a first phone interview, were asked by email to fill-up a questionnaire to describe their synaesthetic associations and for grapheme colour associations to send us a list of those. Synaesthetes were included if they had enough letter-colour and digit-colour associations for our experiments. When they came to the lab to perform the experiments, they ran a modified version of the "Synaesthesia Battery Test" [30] to choose precisely the colour of each letter and digit. This procedure was also used as a retest to confirm the validity of the first-person reports [27]. Seven of the included synaesthetes had already participated in psychophysics experiments between 2007 and 2010 [27].

Control participants were recruited after synaesthetes to match their demographic statistics (16 women, age range between 23 and 38 years old, M = 28.5, SD = 4.3), following similar advertisement

strategies as well as soliciting colleagues at the Grenoble Institute of Neuroscience. Interviews were conducted to verify the absence of any type of synaesthesia.

Materials

Stimuli. For each synaesthete, we tried to identify four pairs of graphemes made of one letter and one digit that had similar colour associations. We tried to find pairs of red, green, blue and yellow (R, G, B, Y) graphemes, but we were only partially successful and in some cases we selected a pair from the most saturated colours available. Fig. 1 shows the actual letters and digits with colours used in the experiments. Only 13 subjects named the pairs red, green, blue and yellow; other colours were named orange, violet, fuchsia and brown, as well as light and dark blue or green. Syn08 and syn48 had a pair made of two letters. Since each synaesthete was tested with a different set of stimuli, each control subject was tested with the stimuli of a specific synaesthete (with the exception of syn10 who had no matched control, by mistake; two controls were tested instead with the stimuli of syn11. Paired comparisons were therefore based on 38 subjects).

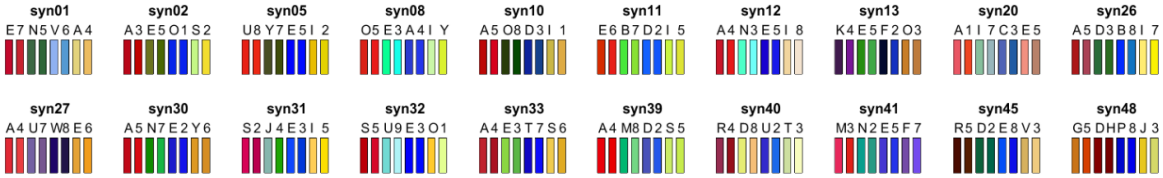

Figure 1. Letters and digits used for each synaesthete, with their corresponding synaesthetic RGB colours (the rendering of the colours using the projector in the scanner was different).

In the MR scanner, we presented these letters and digits in black at the centre of the screen (upper case, Helvetica font, extent up to 2 degrees eccentricity) over a grey uniform background (CIE xyY [0.29 0.3 77.4], half of the maximal luminance of the screen). Stimuli were projected on a translucent screen at the back of the scanner by a video projector Epson EMP 8200. We used a spectrophotometer (PhotoResearch PR 650) for colour and luminance measurements used to compute calibrated images. We also presented dynamic concentric rings (square luminance profile, similar to the stimuli used by Brouwer and Heeger in 2009 [31] except for the absence of anti-aliasing so as to use only the colours selected by each synaesthete), with the exact same (real) colours as those chosen by each synaesthete for each grapheme. The choice of colours matching the individual grapheme 'R, G, B, Y' colour associations was done by each synaesthete in the scanner over the same grey background, using a house-modified MRI compatible, comfortable, 10-button console controller. The same coloured rings were used for each matched control. The rings extent was also up to 2 degree eccentricity and the spatial frequency was 3 cycles/degree (six circles). The phase of the rings changed randomly at 6 Hz to almost nullify visual effects induced by the absence of anti-aliasing.

These stimuli were chosen with the purpose of training and testing classifiers (see below). Briefly, we wanted to use the BOLD responses to the coloured rings to train classifiers on colours, and the BOLD responses to letters to train classifiers on synaesthetic colours. This required choosing pairs of dissimilar stimuli, i.e. a letter and a digit, to try to avoid that the classifier trained on the letters themselves, but rather on their common associated colour. This also implied that decoding should not be feasible that way based on the responses of control subjects. The use of pairs of graphemes also allowed the training on letters and testing on digits (or the reverse), with success in principle

possible only for synaesthetes, based on their synaesthetic colour associations. The careful matching procedure of synaesthetic colours allowed the training of classifiers on real colours and testing on graphemes to identify which brain regions, if any, coded both real and synaesthetic colours. Again, any decoding success would in principle be possible only in synaesthetes.

Classifiers would be trained and tested on four categories, 'R, G, B, and Y', referring either to the real or the synaesthetic prototypical colours that we tried to select. Since two exemplars of each category were presented, it was important that similarity be stronger within pairs than between categories. Fig. 2 represents the actual colours used in the scanner for each synaesthete within the CIE L*a*b* colour space, which is more perceptually uniform than the CIE xyY space. As was already obvious in Fig. 1, differences of luminance were important to distinguish stimuli. Fig. 2 illustrates that the colour and luminance distances were not similar across subjects between categories and within pairs, leading to unequal clusterisation. We could even expect some confusions by the classifiers for some subjects (e.g. "green/yellow" for syn11, "red/blue" for syn13 or "blue/yellow" for syn41). While the maximal theoretical performance achievable by classifiers was therefore below 100%, classifiers could however obtain more than the 25% chance performance in every subject. For all the analyses (described below), we tested if the performance of classifiers across subjects was correlated with the colour distance as measured in the L*a*b* space; we did not find any evidence of that, except for the classification of colours in the searchlight analysis, when testing the group of synaesthetes: we found one significant cluster (106 voxels, 2862 mm^3), in the left fusiform gyrus, peaking at MNI XYZ = [-27 -73 -4], extending from about V4 to FG4, in line with the involvement of these regions in colour processing.

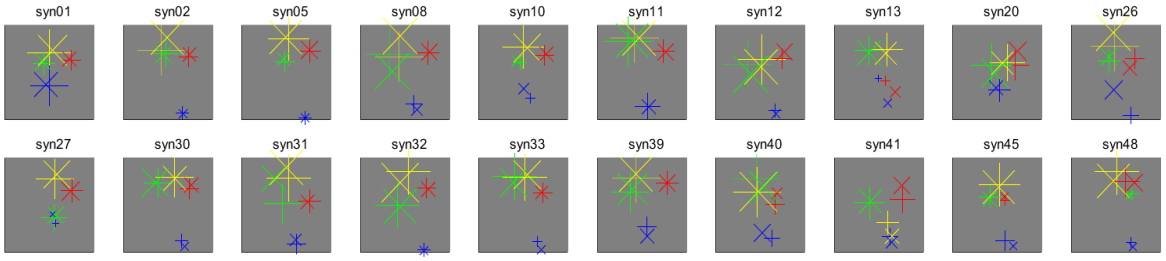

Figure 2. Colour coordinates in the CIE L*a*b* space of the stimuli used for each synaesthete, corresponding to the idiosyncratic synaesthetic colours of letters (+) and digits (x). The colours of the crosses are arbitrary and correspond to the four categories the classifiers had to distinguish. The size of the crosses is proportional to luminance (marker size = 0.4*L, where max(L) = 100; axes limits are +/- 130, possible range being −128 to +127).

Note that luminance variations constitute a major difference, due to the constraint of using synaesthetic colours, with other MVPA studies of the neural correlates of colour processing, which used isoluminant stimuli [31,32]. We do not know how differences along the luminance axis should be perceptually scaled to differences along the green/red opponent colours a* axis and the blue/yellow opponent colours b* axis. This question is probably an ill-posed problem when studying brain correlates of colour perception, since at the cortical level visual circuits rely both (but with different degrees) on the parvo- and magno-cellular pathways [33].

Protocol

Each subject ran three fMRI sessions of about 1 hour. In addition, synaesthetes ran a 1 hour psychophysics experiment (before or interleaved with fMRI experiments, depending on schedule

availability) to measure the strength of their synaesthetic associations using variants of Stroop tasks. All the details of the psychophysics experiment as well as the results of 11 synaesthetes are published [27]. Eight synaesthetes performed the experiments with stimuli presented centrally instead of in the periphery as in our psychophysics study [27], but otherwise the procedure and data analysis were exactly the same. The data of one synaesthete (syn40) could not be analysed because the chosen orange and yellow/green colours revealed too similar (see Fig. 2) and were not named consistently over the course of the experiment.

fMRI experiments

The MR experiments were performed at the IRMaGe MRI facility (Grenoble, France) with a 3T Philips Intera Achieva, using a 32 channels coil. The experiments can be decomposed successively in three "sessions" (about 1 hour each), "runs" (a few minutes), "blocks" (1 minute) and "events" (1 second). One session was dedicated to retinotopic mapping and functional localizer runs using pictures of objects, words and coloured stimuli (Mondrian), in order to define in each subject the Lateral Occipital Complex (LOC [34]), the Visual Word Form Area (VWFA [35]) and "colour centres" [9]. These Regions of Interest (ROIs) functionally defined were however not used in the present study because the results of the other analyses showed that such a refinement was not necessary. Retinotopic mapping was performed strictly as described in a previous study [36], using the Brain Voyager analysis pipeline to define in each subject the ventral and dorsal as well as the left and right parts of areas V1, V2, V3 and V4 (ventral only). The parameters of the EPI functional images were TR/TE: 2000/30 ms, excitation pulse angle: 80°, acquisition matrix: 80x80, bandwidth: 54.3 Hz/pixel, isotropic nominal resolution: 3 mm, 30*2.75 mm thick slices with 0.25 mm interspace covering the whole visual cortex, with four additional dummy scans. To allow the precise alignment of functional scans across sessions, a high-resolution structural image of the brain was also acquired using a T1-weighted MP-RAGE sequence. The sequence parameters were TR/TE: 25/2.3 ms, excitation pulse angle: 9°, 180 sagittal slices of 256*240 (read x phase), bandwidth: 542.5 Hz/pixel isotropic nominal resolution: 1 mm, for a total measurement time of 4 min 31 s.

Another session was dedicated to the "synaesthesia" protocol (a structural image was also acquired with the same parameters as in the first session, in the middle of the functional runs). Twelve functional runs were acquired. The parameters of the EPI functional images were identical to those used for the retinotopic mapping experiment but TR: 2500 ms for an acquisition volume of 45 slices covering the entire brain with a total measurement time of 3 min 47 s. In each functional run, stimuli of one type only were presented: letters, digits, concentric rings with the synaesthetic colours of letters, or concentric rings with the synaesthetic colours of digits. The session contained three successive sequences of four runs, each run with a different stimulus type (with a different random order of stimulus type in each sequence). Each run contained 3*60 s blocks of a rapid-event paradigm, separated by 10 s fixations. Stimuli of different "colours" were presented pseudo-randomly in each block to optimize the estimation of the main effects. For example, in a letter block for syn01 and her matched control, the letters E, N, V and A were presented six times each for 1 s, with 1 s +/- 333 ms fixation only between each letter. This protocol allowed an estimation of the BOLD response to each letter in each block (beta weights, using a General Linear Model, see below) based on six presentations. We obtained three estimations (betas) in each run for each "colour", for a total of thirty-six estimates (9 * 4 "colours") for each type of stimulus to be used by classifiers. The power of classification algorithms depends on both the number and quality (signal to noise ratio) of estimates (called exemplars). The present compromise between quantity and quality was based on Mumford et al. (2012) [37] and on preliminary experiments [38]. Subjects had to fixate the centre of the screen (the fixation point, present between stimuli and at the centre of the coloured rings, or the

centre of the grapheme) and pay attention to the stimuli for the whole duration of each run. To help subjects maintain attention, they performed a one-back task (pressing a button each time the same stimulus was repeated twice in a row).

In the remaining session, a high-resolution, high-contrast structural image of the brain was acquired using a T1-weighted MP-RAGE sequence. The sequence parameters were TR/TE/TI: 25/3.7/800 ms, excitation pulse angle: 15°, acquisition matrix: 180 sagittal slices of 256*240 (read x phase), bandwidth: 191 Hz/pixel, readout in anterio-posterior direction, number of averages: 1, sense factor anterio-posterior: 2.2, right-left: 2, isotropic nominal resolution: 1 mm, with a total measurement time of 9 min 41 s. This image was the structural reference image of each subject. We also acquired diffusion-weighted images, analysed in another study (Dojat, Pizzagalli & Hupé, under review: http://dx.doi.org/10.1101/196865) and a sequence of functional resting state (not analysed yet).

We recorded oculomotor signals during the scans with an ASL EyeTracker 6000. At the beginning of each session, subjects had to fixate each point of a calibration matrix, and were therefore aware that the quality of their fixation was monitored. However, signal quality in some subjects was not good enough or not constant, or even too poor to be of any use for subjects who had to wear non-magnetic glasses in the scanner, so we did not even attempt to analyse these data. We can only speculate that subjects had a better fixation than if they did not know that their gaze was recorded. Whole brain univariate analyses did not reveal any activation along the anterior calcarine and the parieto-occipital sulcus, where activations correspond to the signature of blinks [39], providing indirect evidence that the distributions of blinks were not correlated with our stimuli presented randomly.

Data Analysis

The standard pre-processing procedure of functional images was applied using SPM8: slice-timing correction, then motion correction with realignment, together with correction of spatial distortions of the static magnetic field [40]. The within session structural image was realigned to the mean EPI image, as well as the high resolution high contrast structural image, but no further transformation of the EPI images was performed. No spatial smoothing was applied for MVPA, as maximally differential activation of voxels was shown to maximize the power of classifiers [38]. This was confirmed on these data when testing spatial filters with FWHM = 3, 6 and 9 mm. Transformation matrices were computed between the structural image and the MNI template to allow the transformation and projection of atlas-based masks of specific anatomical structures (Anatomy Toolbox for SPM8 Version 2.2b, 2016) into the subject's space.

Univariate analyses were performed on the groups of subjects to test for differences of magnitude of the BOLD responses to graphemes evoking synaesthetic colours. A 9 mm FWHM spatial smoothing was applied to the subjects' EPI images before testing two contrasts: a T-contrast of all stimuli against the fixation point (we did not have graphemes that did not evoke any synaesthetic colour); an F-contrast of the four pairs of graphemes. Contrast maps were distorted within the study-specific template computed using DARTEL procedure as implemented in VBM8 (see Dojat, Pizzagalli & Hupé, under review, for details) and to the MNI space (resolution 1.5 by 1.5 by 1.5 mm). For second-level analyses, we compared the contrast maps of synaesthetes (N = 20) against controls (N = 20) using t-tests (testing stronger signals either in synaesthetes or controls). We also performed paired t-tests of 19 synaesthetes against their matched control to account for possible differences due to the specific choices of graphemes in each synaesthete. For all comparisons, no individual voxel reached $p < 0.05$, corrected for the family-wise error (FWE, based on the random field theory as implemented in SPM8). We used cluster-based statistics with the cluster-forming threshold set to $p = 0.001$ [41] and $pFWE < 0.05$. We performed the same analyses also for coloured stimuli.

For MVPA, for each subject and each run we ran a General Linear Model (GLM). The six parameters of motion correction were included as factors of non-interest in the design matrix. Thirteen main predictors, four events (grapheme or colour) * three blocks plus one for the fixation point, were obtained by convolving the canonical HRF with Dirac functions corresponding to the time of presentation of each stimulus. The corresponding beta weights estimated by the GLM for each colour and visual feature, divided by the square root of residuals, were used as examples by a Support Vector Classification (SVC) algorithm (Scikit-learn version 0.15.2, implemented in Python version 2.7.9.0 [42]). We used a linear kernel (default value of the C parameter = 1) and a one-versus-one classification heuristic to classify each example in one of the four categories. For all classifications, training and test runs were always fully independent: betas obtained from blocks from the same run were never split between training and test runs. Six runs (eighteen blocks) were used for colour and synaesthesia decoding. The procedure was leave-one-run-out. Six classifiers were therefore trained to classify (5 runs * 3 blocks * 4 colours = 60) colour exemplars in four categories, and tested on (1 run * 3 blocks * 4 colours = 12) independent exemplars. Performance was therefore computed over seventy-two classifications (6 classifiers * 12 tested exemplars), with chance level = 25% and 95% Confidence Interval of chance for each subject = [16 36]% (binomial probability, Agresti-Coull estimation). For grapheme runs, training was performed on pairs made of one letter and one digit. If the decoder learnt only the letters, for example (by being able to filter out the responses to digits), then performance on decoding letters and digits could reach up to 50%, without knowing anything about synaesthetic colours. One could expect, however, that performance of synaesthetes would be higher than for controls because of the additional information provided by synaesthetic colours. A more stringent test of synaesthetic coding was the training of one classifier on letters (3 runs * 12 exemplars) and testing on digits (and the reverse). Learning was achieved using thirty-six exemplars (letters or digits) to be classified in four categories, test was on thirty-six exemplars (digits or letters), for a total performance over seventy-two classifications by combining training on letters and training on digits. To evaluate if brain regions coded both real and synaesthetic colours, training was performed by one classifier on six colour runs (seventy-two exemplars), test on six grapheme runs (seventy-two exemplars). We also performed the reverse classification.

We computed MVPA in regions of interest (ROIs) defined in each native (non-transformed) subject space. We used visual areas defined by individual retinotopic mapping as well as atlas-based ROIs (Fig. 3). We expected synaesthetic colours to involve the ventral visual pathway, anterior to V4, so we tested the four subdivisions of the fusiform gyrus (FG, Fig. 3a). Some studies have also suggested the role of parietal areas, even though no consensus emerged about exactly which part if any may be involved [12], so we defined ROIs in parietal regions (Fig. 3b).

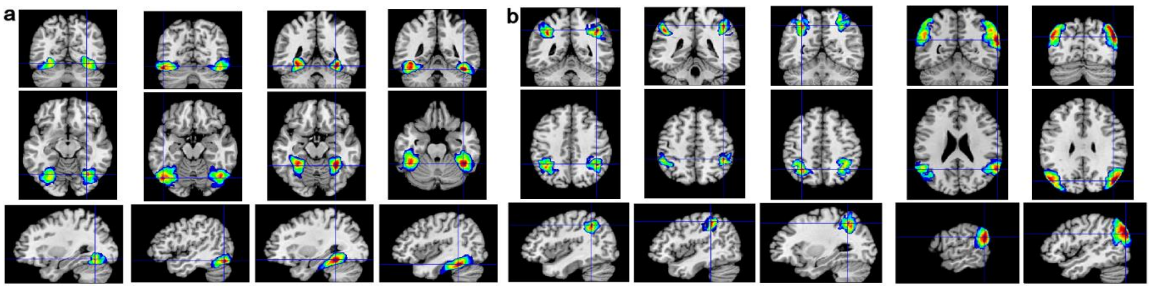

Figure 3. a. Atlas-based regions of interest (ROI) of the fusiform region. From left to right, FG1, FG2, FG3 and FG4. Colour gradients denote the probability of being in the specified ROI, from 0% (dark blue) to 100% (dark red). We considered the largest ROI as the mask of the corresponding region. b. Parietal ROIs. From left to right, AIPS_IP1, AIPS_IP2, AIPS_IP3, IPL_PGa and IPL_PGp. See text for full names and references of these areas.

For each subject, anatomical ROIs were defined as the intersection of the subject's grey matter mask and the mask of the anatomical ROIs (Anatomy Toolbox for SPM8 [25]) projected into the subject's space. Both retinotopic and atlas-based ROIs had different number of voxels within and across subjects. The performance of classifiers may depend on the number of voxels (called "features" for the algorithm), making difficult the comparison of absolute performance in different ROIs. Between-subject differences may also bias group comparisons.

To address this issue, we first tested ROIs of different sizes by regrouping retinotopic areas and subdivisions of the fusiform areas and of the parietal areas. The pattern of results were similar whatever our grouping choice of ROIs. We present the results for ROIs of intermediate size (we indicate the min and max number of voxels across subjects in each ROI), regrouping the right and left parts of retinotopic areas (V1 = [206 441], V2 = [125 420], V3 = [156 340], V4 = [94 268]), the two posterior [43] (left = [98 150], right = [61 125]) and anterior [44] (left = [174 331], right = [127 271]) parts of the fusiform areas, the 3 subdivisions of the Intaparietal Sulcus [45,46] (left = [197 311], right = [191 275]) and the anterior and posterior parts of the Inferior Parietal Lobule [47,48] (left = [131 335], right = [124 279]).

We also defined ROIs using the same number of voxels in each subject and ROI. To do that, for all classifications, we selected 100 voxels with the highest F-scores to colours in each area (we tested different selection sizes and found that 100 was about the optimal number of voxels to reach maximum performance). In order to have enough voxels to choose from in every subject, we selected voxels in only six large areas: the left and right retinotopic areas V1 to V4 (minimum number of voxels across subjects were respectively 352 and 327), the left and right fusiform areas FG1 to FG4 (298 and 188) and the left and right parietal areas (347 and 315). Such a selection provides the best chances for colour classifiers (since we select voxels maximally modulated by colours), but classification is then not independent of selection when measuring colour decoding after selection of F-scores to colours (but classification is independent for grapheme decoding). In order to provide a fair measure of colour decoding performance to compare grapheme decoding with, voxels were selected using F-values computed based only on runs used for training, meaning that each of the six training sets was based on a different set of voxels. For other classifications, the same set of voxels was used based on F-values computed across all colour runs.

In each ROI, we computed 95% CIs of the performance of each group, as well at the 95% CIs of the between group differences (both independent and paired comparisons; results were very similar, we show the CIs for paired comparisons). We also performed paired comparisons by computing the 95% CI of the odds ratio when comparing 19 synaesthetes against their matched controls, using a mixed-effect generalized linear model, with a binomial family and a logit link function, as implemented in the library lme4 [49] in R, version 3.3.3.

Searchlight analysis was also performed over the whole brain [26] using a 15 mm radius and the SVC algorithm. Performance maps were transformed to the common DARTEL space for voxel-wise group comparisons (resolution 3 by 3 by 3 mm). We performed in SPM8 all the analyses equivalent to univariate analyses (both two-sample and paired-sample t-tests between synaesthetes and controls) as well as one-sample t-tests to compare the average performance of each group against chance (= 0.25). For all comparisons, no individual voxel reached pFWE < 0.05. We used cluster-based statistics with the cluster-forming threshold set to p = 0.001 and pFWE < 0.05. These analyses are in principle less powerful than ROI analyses because they constrain to distort each subject's anatomical space within one common space, so the average performance at any given voxel may in reality correspond to different anatomico-functional voxels in different subjects. Moreover, they re-introduce the methodological issues related to spatial smoothness ([50]).

Data Availability. The datasets generated and analysed during the current study are freely available on request (https://shanoir.irisa.fr/Shanoir/login.seam), contact M. Dojat.

Results

Whole brain univariate analyses (normalized anatomical space)

Univariate analyses revealed no difference between controls and synaesthetes at our statistical threshold for T-contrasts of graphemes when performing two-sample t-tests. However, the paired-sample t-tests revealed stronger BOLD signal in a small cluster in synaesthetes, close to the left precentral gyrus, which we treated therefore as a candidate region for the coding of synaesthetic colours (supplementary Table S1). For F-contrasts, we did not observe any stronger modulation in synaesthetes (neither for two-sample nor paired-sample t-tests). Surprisingly, we observed stronger modulation in controls in two clusters (paired comparisons), in the right occipito-parietal cortex (supplementary Fig. S1) and in the left insula. The two-sample t-tests revealed only the occipito-parietal cluster. We did not have any explanation for these differences, which might be false-positives [41]. We note that the analysis by Rouw and Scholte in 2010 [28] revealed a cluster (which they called IPS, cluster extent = 3280 mm3) at equivalent peak coordinates on the left side ([-30 -72 28]), obtained with the contrast synaesthetes>controls for (synaesthetic graphemes)>(non-synaesthetic graphemes). In our case, the weaker modulation by graphemes in synaesthetes would rather argue against the hypothesis of a functional role of this region in synaesthesia. We included these two regions in our post-hoc MVPA analyses for further exploration.

We also tested T- and F-contrasts for the responses to real colours (rings). We observed stronger BOLD signal (T-contrast) in synaesthetes only, in three clusters for paired comparisons (in the left posterior and anterior insula – see supplementary Fig. S2 - and in the left parahippocampal region) and two other clusters for two-sample t-tests (in the right middle temporal gyrus and in the right superior, medial, frontal gyrus - see supplementary Fig. S3). The lack of consistency between paired and two-sample t-tests could again suggest false-positives, but we nonetheless included these five clusters in our post-hoc MVPA analyses, in case those stronger activations be related to the implicit activation of graphemes by the colours associated to them ("bi-directional" synaesthesia [51]). F-contrasts to colours revealed only one cluster of stronger modulation in controls in the frontal region, but in the middle of white matter and thus clearly a false positive. We compared the performances of synaesthetes and controls for our classifiers (see next section) in those eight clusters defined post-hoc, with two-sample and paired-sample t-tests. Only three comparisons came out "significant" at $p < 0.05$, but without correction for multiple comparisons (Table S1). Supplementary Figs. S1 to S3 detail these results.

Multivariate pattern analysis in regions of interest (defined at the individual level)

Fig. 4 shows the performance of all classifiers in all our ROIs, without any voxel selection (ROIs have therefore different number of voxels across regions and subjects).

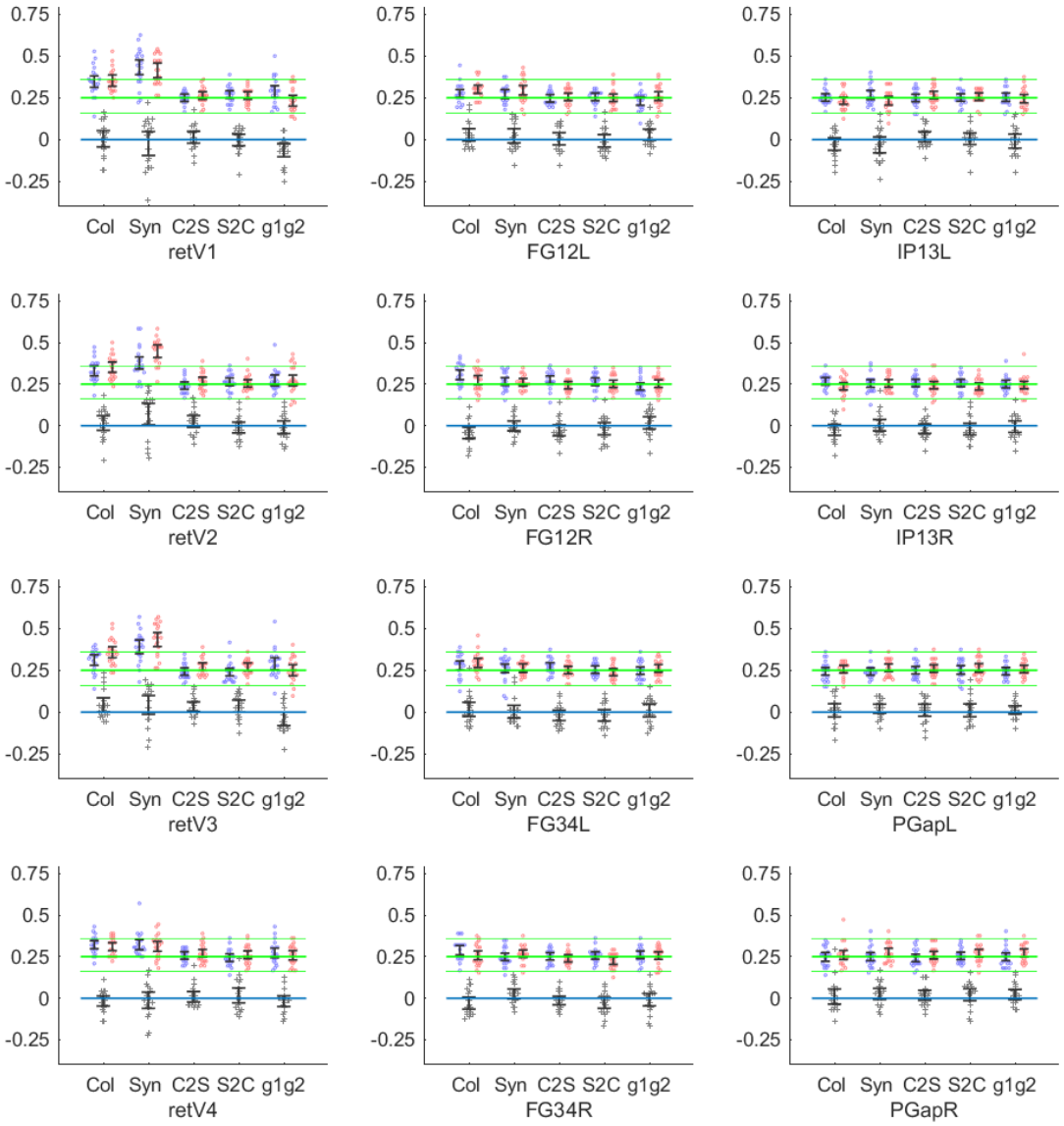

Figure 4. Performance of classifiers in retinotopic areas, the fusiform gyrus and parietal areas. Each ROI regrouped several areas, for example the left and right parts of V1 for 'retV1' in order to provide a large number of voxels in each subject and ROI (at least > 60, and > 100 voxels in most ROIs; see Methods: Data Analysis). 'retV1' to 'retV4' were defined based on retinotopic mapping in each subject; other ROIs were defined as the intersection of the subject's grey matter mask and the mask of atlas-based anatomical ROIs (Anatomy Toolbox for SPM8) projected into the subject's space (see Fig. 3). FG12L = left (FG1 + FG2), IP13L = left (AIPS_IP1 + AIPS_IP2+ AIPS_IP3), PGapL = left (IPL_PGa + IPL_PGp), etc. Each classifier was trained and tested on beta weights computed on voxels in the native subject space with no spatial smoothing. Each panel displays the individual and average performances of five classifiers: 'Col' = training and test on betas for real colours (rings); 'Syn' = training and test for synaesthetic colours (graphemes, letters or digits); 'C2S' = training on real colours (rings) test on synaesthetic colours (graphemes); 'S2C' = training on synaesthetic colours test on real colours; 'g1g2' training on letters test on digits or training on digits test on letters. The y-axis represents both the *performance* of classifiers (between 0 and 1, chance level = 0.25, thick green line; 95%CI of chance for each subject = [0.16 0.36], thin green lines) for individual subjects (blue = controls, red = synaesthetes) and their group average (with 95% Confidence Intervals) and the *difference of performance* (grey crosses) between synaesthetes and their matched controls (0 = no difference between groups, blue line; whiskers denote 95% CI).

As expected, the decoding of real colours ('Col' classification) was above chance (0.25, thick green line) in retinotopic areas as well as in the fusiform gyrus for both controls (blue points) and synaesthetes (red points), with no obvious between group differences (whiskers across the zero blue

line denote the 95% CI for paired comparisons of performance of 19 synaesthetes against their matched control, the difference of performance being denoted by the grey crosses). The whole 95% CI was slightly above 0 in retinotopic V3, 'retV3', and it was slightly below 0 in the subdivisions 1 and 2 of the right fusiform gyrus, 'FG12R' (differences are more visible when estimating the CI by a mixed-effect generalized linear models: supplementary Fig. S4). Without any independent evidence, these small differences could be due to random sampling. Indeed, all the 99.58% CIs included 0 (Bonferroni correction over 12 tests).

Then, we tested if graphemes could be decoded on the basis of synaesthetic colours ('Syn' classification). Since above-chance performance could be achieved based on either a letter or a digit (for example E or 7, both associated to red by syn01: see Fig. 1), we looked for an additional performance due to synaesthetic colours. This was the case in retinotopic V2 (95% CI of the difference of performance, two-sample t-test: [1.5 12.5]%; paired t-test: [0.5 13.4]%; 95% CI of the odds ratio = [1.15 1.56]) and to a lesser extent in retinotopic V3 (but note that performance was lower for synaesthetes in the subdivisions 1 to 3 of the left Intra-Parietal Sulcus, IP13L; such a difference is most likely due to random sampling since none of the group performances in IP13L was above chance). Only the difference in V2 survived Bonferroni correction over 12 tests, for the mixed-effect analysis ($p = 0.0002$). In order to further explore, using independent evidence, the implication of V2 in the coding of synaesthetic colours, we took individual differences into account. We reasoned that synaesthetes with stronger synaesthetic associations might have stronger modulations of the BOLD signal and thus larger decoding values. Fig. 5 shows the performance of each subject as a function of the strength of synaesthetic associations, measured in Stroop-like psychophysics experiments (see Ruiz & Hupé, 2005 [27] for the exact definition of the index of 'Photism Strength') for synaesthetes only (red crosses). Controls (blue circles) were attributed the value of their matched synaesthete.

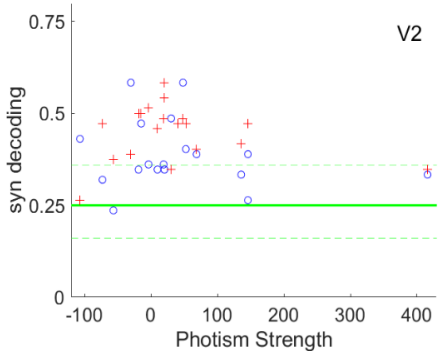

Figure 5. Performance of the classifier trained and tested with synaesthetic colours (pairs of graphemes) in each subject in area V2 defined retinotopically (same data as 'Syn' in the second panel of the first column of Fig. 4) as a function of the strength of synaesthetic associations ('Photism Strength'). This strength was measured for synaesthetes (red crosses); controls (blue circles) were attributed the value of their matched synaesthete.

There was no correlation between both measures, neither for synaesthetes nor controls. We also observed that the difference of score between each synaesthete and her (or his) matched control did not increase with photism strength. Therefore, this analysis did not provide any independent argument in favour of the decoding of synaesthetic colours in V2. We computed similar correlation analyses in every ROI and never found any correlation. We also computed both positive and negative

correlations over the whole brain for the five classifiers, independently for synaesthetes and controls. We never found any significant cluster (cluster forming threshold = 0.001).

We also tested if synaesthetic colours could be decoded based on real colours ('C2S' classification in Fig. 4). This was the case in no ROI and neither in controls, as expected, nor in synaesthetes (the 95% CI of all groups crossed the 0.25 chance baseline). There was no evidence of better classification in synaesthetes, in particular in V2, as would have been suggested if the higher performance for the 'Syn' classification was really due to the coding of synaesthetic colours. We obtained similar results when we tried to decode real colours based on graphemes (and synaesthetic colours in synaesthetes: 'S2C' classification).

Finally, we tested a more stringent classification that should have been possible only on the basis of synaesthetic colours: classifiers were trained on one set of graphemes (digits or letters) and tested on the other set of graphemes ('g1g2' classification). Again, performance was never above chance and we found no difference between groups, in particular in V2. We even observed lower scores for synaesthetes in V1 (where it even survived Bonferroni correction for the mixed-effect analysis: p=0.0002) and V3, where we had yet observed higher performance for 'Syn' decoding. This lack of consistency across different tests addressing the same question confirmed that these small variations, even when statistically "significant", were most likely due to random sampling.

We performed again all these analyses using six larger ROIs (regrouping either the left or the right parts of V1 to V4, FG1 to FG2 and the areas of the inferior parietal lobule and the intraparietal sulcus) in which we selected the 100 voxels with the largest scores to F-tests to real colours (Fig. 6). Performance was never better in synaesthetes.

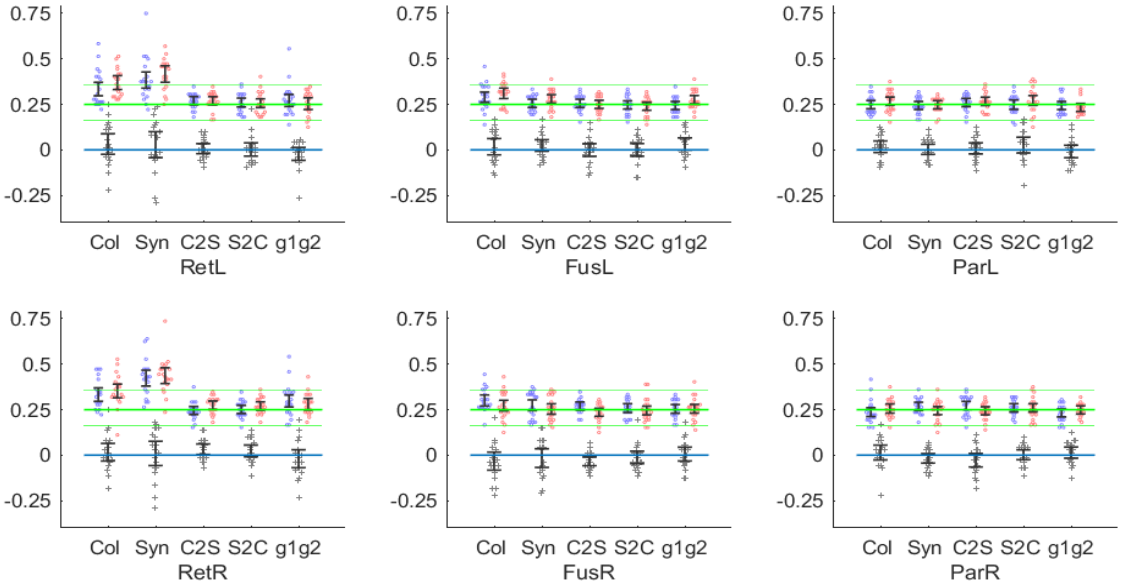

Figure 6. Performance based on the same number of voxels (= 100) in each large ROI (retinotopic areas, fusiform gyrus and parietal regions) and subject. For the classification of real colours ('Col'), the selection of the best F-values to colours was different for each of the six leave-one-out classifications, based each time only on the five runs used for training the classifier, to insure independence of training and test. For the other selections, all colour runs were used to select the voxels with the highest F-scores. The high performance for the 'Syn' classification in retinotopic areas indicates that many voxels respond both to change of colour or luminance and the shape of graphemes, probably thanks to the small receptive fields of lower visual areas. Same conventions as in Fig. 4. The CIs of the odds ratio computed by mixed-effect generalized linear models are shown in supplementary Fig. S5.

Whole brain searchlight multivariate pattern analysis

We complemented our ROI analysis with searchlight analyses over the whole brain (normalized to the MNI space), comparing groups for the five classifications. We found no differences between controls and synaesthetes at our statistical threshold for classifiers trained and tested on colours (rings, 'Col' classifiers). This shows that the clusters with stronger BOLD signal in controls (T-contrast) observed in univariate analyses were not involved in colour coding (see also Table S1). For classifiers trained and tested on synaesthetic colours (graphemes, 'Syn' classifiers), we observed higher performance in synaesthetes in the parietal cortex (Table S1: on the right side with paired t-tests and on the left side with two-sample t-tests; bilateral difference could be observed for both contrasts when using a higher cluster-forming threshold). However, testing synaesthetes against chance revealed no cluster at our threshold around these coordinates of the parietal cortex (performance was above chance in both groups in the occipital cortex, as expected).

We found no difference between controls and synaesthetes at our statistical threshold for the critical test of shared coding of real and synaesthetic colours, when classifiers were trained on coloured rings and tested on graphemes ('C2S' classifiers). Testing synaesthetes against chance also revealed no cluster. The reverse classification (learning on graphemes, 'S2C'), however, revealed two clusters with higher performance in synaesthetes for independent t-tests, in the right occipito-temporal cortex and in the left putamen. Only the first cluster was confirmed by paired-comparisons. When testing performance against chance two clusters emerged for synaesthetes (none for controls), one again in the same part of the right occipito-temporal cortex, and the other in the left parietal cortex, abutting the parietal cluster obtained previously for the higher performance in synaesthetes for the 'Syn' classification (we shall come back to this concordance in the following *post-hoc* analysis).

Finally, for classifiers trained on either letters or digits (and tested respectively on either digits or letters), a critical test of the coding of synaesthetic colours, higher performance was observed, but in controls, in the left inferior frontal gyrus, for both paired and independent t-tests. However, no cluster emerged anywhere in the brain in controls (nor in synaesthetes) when testing performance against chance, so this cluster should be considered as a false positive.

We further explored the performance of classifiers in the two clusters identified by the 'Syn' classifier and the five clusters identified by the 'S2C' classifier, corresponding in fact to two parietal regions (left and right), one right occipito-temporal region and one cluster in the left putamen. In each cluster, we computed the average across voxels of the searchlight scores to compare the performances of our five classifiers for synaesthetes and controls in these seven clusters defined *post-hoc*, with two-sample and paired-sample t-tests. We also compared the performance of each group against chance. Statistically "significant" differences were obtained only for the contrasts used to define the clusters (Table S1). Only one additional comparison was "significant" ($p = 0.012$, not corrected for multiple comparisons) in the left parietal cluster at XYZ = [-33 -28 50], which had been obtained when testing synaesthetes against chance for the 'S2C' classification (training classifiers on graphemes and testing them on colours: Fig. 7): synaesthetes also performed better than controls at decoding graphemes ('Syn' classification), 95%CI = [1 9]% (paired comparisons), and better than chance (95% CI = [26 31]%), but the performance was not correlated with the strength of synaesthetic associations ($p = 0.51$).

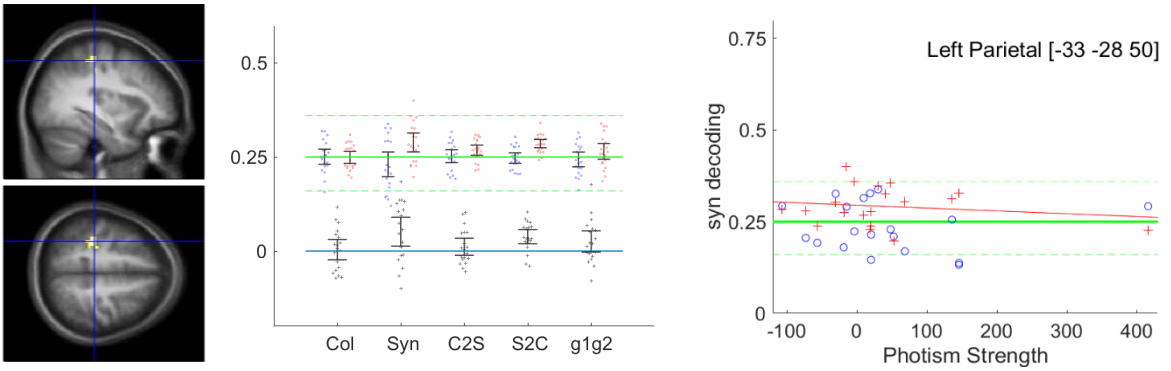

Figure 7. Left: Parietal cluster identified based on whole brain searchlight analysis for 'S2C' decoding, Synaesthetes>chance (27-voxel cluster at XYZ = [-33 -28 50], one-sample t-test). Middle: performance of classifiers in this cluster (same conventions as in Fig. 4). The performance of synaesthetes was logically above chance for the 'S2C' classification, since the cluster was defined based on this contrast. For the independent classifier 'Syn', the performance of synaesthetes was also above chance and above that of controls. Right: Absence of correlation between the strength of synaesthetic associations and 'Syn' decoding (r = -0.16, p = 0.51; for 'S2C' decoding, not shown: r = -0.17, p = 0.49).

Discussion

Across all our analyses, we did not find any convincing evidence of above-chance decoding performance of synaesthetic colours in the visual cortex when training classifiers either on the 3T BOLD responses to pairs of letters or to real colours matching individual synaesthetic colours, neither in retinotopic areas defined at the individual level nor in the fusiform gyrus and parietal regions of interest defined based on a probabilistic atlas. Moreover, exploration outside the visual cortex, across the whole (normalized) brain (searchlight analysis), though it suggested the involvement of regions outside of ROIs initially selected, did not bring any convincing evidence. The mass univariate tests we used for whole brain analysis face the ill-posed problem of correction of multiple comparisons of partly correlated tests, problem not fully solved by the Random Field Theory [41], as well as the production of poorly informative p-value maps [24]. Moreover, since we performed in total at least nine whole brain searchlight analyses and four whole brain univariate comparisons (T- and F-contrasts for responses to graphemes and colours, see Table S1), we could have set a family-wise error level at 0.05/13. We preferred to keep a non-corrected level for easier comparisons with other studies. The whole brain analyses were used only for exploration, and for every "significant" cluster we searched for additional evidence (differential response for other comparisons, or correlation with individual differences). Since we did not find any additional evidence, we conclude that these clusters were likely false positives. We however mention them (see Table S1) in case additional evidence be found in other studies.

For now however, our results further suggest that 3T fMRI studies may not be able yet to identify the neural correlates of the synaesthetic experience of colour [12], probably because those are fine-grained distributed at a resolution lower than our 3-mm^3 voxel resolution, or because the nature of its coding does not translate (well) into BOLD responses. Surprisingly, though, if considering synaesthetic colours simply as a form of mental imagery, we were expecting above-chance decoding performance as observed for other tasks involving mental imagery. Those other tasks, however, typically involved different categories of objects, like food, tools, faces and buildings [22] or objects, scenes, body parts and faces [52], which evoked stronger BOLD signal in specific areas (like the Fusiform Face Area). Other studies involved retinotopic properties [21] where, again, differences of BOLD signal can be easily observed. Here we were trying to decode mental images within only a single category, colour. This

confirms that synaesthetic experiences do not evoke strong BOLD responses, at least when using standard 3T MR scanner, as already suggested by the inconsistency of the published results based on univariate models. Below we consider alternative explanations (e.g., methodological flaws) to the absence in our data of neural traces of the processing of synaesthetic colours.

(1) Our study used a protocol very similar to that used by Bannert and Bartels (2013) [53], who tried to decode the typical colour from eight objects, presented as greyscale photos, with classifiers trained on concentric colour circles designed after Brouwer and Heeger (2009) [31], like in our study. The prototypical colour of the objects was red, green, blue or yellow (like a banana and a tennis ball). Across 18 subjects, decoding accuracy was "significantly" above chance in V1, but reached only 32% on average, which is hardly above the 95% CI ([24 30]%) of the performance observed for our similar classifier ('C2S') in the areas V1 to V4 of synaesthetes. Their experimental procedures, slightly different from ours, may have better optimized the signal to noise ratio and allowed this higher performance (see below). Alternatively, since the colour-diagnostic objects were presented before the coloured concentric rings, subjects may have imagined, when viewing the rings, the very objects that were presented before. Subjects had to do a motion discrimination task to divert their attention (similarly to our one-back task), but such a task (like ours) was not very demanding (though note that Bannert and Bartels argue that their results are due to automatically occurring processes during object vision rather than active imagery). Of course, a similar argument holds even more in our experiment: synaesthetes were very likely to recognize the colour matching exactly their synaesthetic colour of letters and digits, and they might well have imagined the letter or digit when looking at the coloured stimuli. In both cases, decoding would be based on the complex shape of stimuli rather than their colour. In the case of Bannert and Bartels, objects were similar to those used in other successful visual-to-imagery decoding and involved several categories of objects as well as different retinotopic properties (the objects had different orientations but were rotating; however the banana or the coke can, for example, had about 12 deg extent, apparently much more than the Nivea tin or the blue traffic sign), while in our case objects all belonged to the grapheme category, and all spanned the same visual extent. It is therefore possible that in the study by Bannert and Bartels the slightly above chance decoding performance was due to residual category and retinotopic properties, not to colour. With such an interpretation, decoding of imaginary colours would have failed in both their and our study.

(2) Another possible explanation to our chance performance could be linked to our choice of a fast event related paradigm, each stimulus being presented each time for only 1 s, with an ISI = 1 s +/- 333 ms. Bannert and Bartels presented images for 2 s with a 1 s ISI, each repeated four times in a row (miniblocks). One may wonder whether our presentation time was sufficient to trigger synaesthetic associations. However, psychophysical tasks show that the naming of the synaesthetic colours of graphemes takes on average much less than 1 s [27]. Because of our one-back attentional task, though, we cannot be sure that the synaesthetic associations were always conscious. However, synaesthetes did not report any specific difficulty with their synaesthetic experience when viewing, inside the scanner, the proposed paradigm. We designed such a protocol because we did not want synaesthetes to pay too much attention to their synaesthetic colours, then possibly triggering complex attentional and emotional processes. Those components are part of the synaesthetic experience, but they do not tell us anything about the phenomenological experience of colours, our main goal being to try to isolate the possible neural commonalities of the real and synaesthetic experience. The quasi-absence of observed differences of overall activation and modulation between synaesthetes and controls for graphemes indicates that we were successful in synaesthetes having a similar experience to controls for graphemes, in terms of attentional and emotional content. With

different conditions, favouring synaesthetic colours to be experienced intensely, we would expect the overall pattern of brain activity to be different.

(3) Another critical aspect of our fast-event paradigm is related to the slow dynamics of the hemodynamic signal and the signal to noise we could obtain. Here, the critical benchmark was the possibility to decode real colours, since the protocol was identical for synaesthetic and real colours. We were successful in decoding colours above chance in the visual cortex, but not to the extent that we hoped: only 35% on average, chance being 25%. Using 12 s miniblocks, Bannert and Bartels obtained an average performance for colours between 35% and 40% in V1 to V4. Differences other than the timing of the stimuli may explain this only slightly higher performance: their total presentation time of coloured stimuli was about 42 min (20 min in our study; for example, Brouwer and Heeger in 2009 [31] obtained even higher performances with experienced participants tested for much longer durations); their stimuli were much larger (7.19 deg *vs.* 2 deg radius); their stimuli were isoluminant (we do not know whether luminance information in our case helped or hindered decoding). Because they were constrained by the idiosyncratic synaesthetic associations, our stimuli were also not well distributed within the colour space (see Fig. 2). Colour differences between categories (R, G, B and Y) and similarity between colours for pairs (letter-digit) were different between subjects and not always optimal to reach maximal performance by classifiers. Probably, some pairs of supposedly similar colours confused classifiers, as well as short distances between some categories. Retrospectively, we should in fact even consider ourselves lucky to have achieved such a performance for colour decoding. Choosing only three colours would have allowed us to avoid confusions and get more exemplars for each colour (with fewer categories to decode, though, confounding factors are more likely). More repetitions would be welcome, however we wanted to record signals for real and synaesthetic colours within the same scanning session to avoid any spatial smoothing of the voxels (which is often necessary when aligning images obtained in different sessions). Preliminary experiments had showed us indeed that combining the signals from different sessions did not improve performance [38]. Our total session time was about 1 hour, which is about the limit one may ask naïve subjects to lie in a scanner without moving while maintaining fixation and attention over boring stimuli.

(4) Given our moderate performance for colour classification, our absence of above-chance performance for the decoding of synaesthetic colours might be due to a lack of power, since performance across real and imaginary images is typically lower than for real images [22,53]. Indeed, if real differences exist between synaesthetes and controls in the measured BOLD signal, these differences are too small to be detected reliably with sample sizes similar to ours, with no indication about the minimum required sample size. Such a reasoning holds for the average performance, but some subjects did reach performance for colour decoding well above 50%. Yet, the distribution of individual scores were all very similar for controls and synaesthetes (see Figs. 4,6). There was some correlation between the performances of colour ('Col') and synaesthetic ('Syn') classifiers in retinotopic areas (especially V1), but it was similar in synaesthetes and controls (the differences between synaesthetes and controls for the 'Syn' classifier were in fact even weaker when including the 'Col' performance as a covariate).

For the statistical analysis, we adopted the "new statistical approach" proposed by Cumming [54] and focussed on confidence intervals of effect sizes instead of the less informative thresholded p-value maps [24]. In order to facilitate the comparison of our study with previous studies, we indicated when the comparisons could be considered as "significant" (a 95% CI not crossing the chance level corresponds to $p < 0.05$) when correcting the risk level for multiple comparisons. Note however that correction for multiple comparisons corresponds to an ill-posed problem, because there is no unique

and objective way to define the family of tests [24]. Such a problem is pretty obvious in our case, where the number of considered ROIs depends on our choice of regrouping or not ROIs, and by how much. We applied a Bonferroni correction over twelve ROIs, but we could have considered the family across the five types of classifiers (so at least 60 tests). However, by focusing on the extent of the CIs, the conclusions do not change much for different levels of CI (the extent of a 99.58% CIs is just a bit larger than for a 95% CI): for all the cases that may suggest differences between groups, the true differences compatible with our observations may be either close to absent (difference close to or including 0, or odds ratio close to or including 1) or at most up to about 15% (or odds ratio = 1.5), a value that one may consider meaningful. As in most studies currently published in cognitive neuroscience dealing with small effects, the width of our confidence intervals is too wide to reach any definitive conclusion on the sole basis of one test (lack of power). Our choice of CI presentation, however, brings useful information allowing cumulative science [55] and shows that if any real difference exists, it is probably not very large because corresponding to less than a 15% difference of performance.

Our conclusion is that identifying the neural correlates of the synaesthetic experience of colours may still be beyond the reach of present technology, including hardware (3T MR scanner) and advanced data analysis techniques such as MVPA, and that we still do not find any evidence of common neural coding of real and synaesthetic colours [9]. However, across all our analyses, we did find several "significant" differences for several comparisons, which we listed in the Results section and detailed in supplementary Table S1. Other studies also did report so-called "significant" effects, even though the methods to determine the significance levels were questionable in most studies [12]. We applied the latest recommendations for group-level cluster-wise inferences (9-mmm FWHM spatial smoothing, cluster-defining threshold = 0.001, cluster-based pFWE <0.05, groups of 20 participants), yet these criteria do not protect well against false positives [41]. In our study, for each "significant" effect, we had a set of independent measures to further explore any difference that may be real: the performance of other classifiers as well as individual differences (the strength of the synaesthetic associations measured in Stoop-like psychophysics tests). We never found any coherence across different measures. Moreover, the locations of the "significant" effects appeared quite randomly across the brain. As long as no other study replicates any of these "differences", we should conclude that there are false positives indeed. Our study thus further shows that common statistical practices based on Null Hypothesis Significance Tests (NHST) are not adequate for scientific inference [12,56,57]. By stressing that we did not find any evidence of common neural coding of real and synaesthetic colours, based on our data as well as past studies, we do not conclude that such a neural coding does not exist. We bring to light what is required to have any chance to reveal the neural bases of the synaesthetic experience using MRI, like more data by subject, higher signal to noise ratio and spatial resolution (e.g., 7 Tesla scanner [58]) and much larger cohorts. In order to start contributing to this last aim, via the constitution of dedicated data repositories and meta-analyses, our data are freely available on request (https://shanoir.irisa.fr/Shanoir/login.seam, contact M. Dojat). Please refer to the present paper in case of the reuse of these datasets.

References

1. Simner, J. et al. Synaesthesia: The prevalence of atypical cross-modal experiences. *Perception* **35**, 1024-1033 (2006).
2. Chun, C. A. & Hupé, J. M. Mirror-touch and ticker tape experiences in synesthesia. *Front. Psychol.* **4**, 776 (2013).
3. Simner, J. & Carmichael, D. A. Is synaesthesia a dominantly female trait? *Cogn. Neurosci.* **6**, 68-76 (2015).
4. Rouw, R. & Scholte, H. S. Personality and cognitive profiles of a general synesthetic trait. *Neuropsychologia* **88**, 35-48 (2016).
5. Watson, M. R. et al. The prevalence of synaesthesia depends on early language learning. *Conscious Cogn.* **48**, 212-231 (2017).
6. Ward, J. Synesthesia. *Annu. Rev. Psychol.* **64**, 49–75 (2013).
7. Flournoy, T. *Des Phénomènes de Synopsie (Audition Colorée): Photismes, Schèmes Visuels, Personnifications* (Alcan, Paris, 1893).
8. Edquist, J., Rich, A. N., Brinkman, C. & Mattingley, J. B. Do synaesthetic colours act as unique features in visual search? *Cortex* **42**, 222-31 (2006).
9. Hupé, J. M., Bordier, C. & Dojat, M. The neural bases of grapheme-color synesthesia are not localized in real color sensitive areas. *Cereb. Cortex* **22**, 1622:1633 (2012).
10. Witthoft, N. & Winawer, J. Learning, memory, and synesthesia. *Psychol. Sci.* **24**, 258-65 (2013).
11. Chiou, R. & Rich, A. N. The role of conceptual knowledge in understanding synaesthesia: Evaluating contemporary findings from a 'hub-and-spoke' perspective. *Front. Psychol.* **5** (2014).
12. Hupé, J. M. & Dojat, M. A critical review of the neuroimaging literature on synesthesia. *Front. Hum. Neurosci.* **9**, 103 (2015).
13. Janik McErlean, A. B. & Banissy, M. J. Color processing in synesthesia: what synesthesia can and cannot tell us about mechanisms of color processing. *Top. Cogn. Sci.* **9**, 215-227 (2017).
14. Reeder, R. R. Individual differences shape the content of visual representations. *Vision Res.* (2016).
15. Galton, F. Statistics of mental imagery. *Mind* **5**, 301-318 (1880).
16. Chun, C. A. & Hupé, J. M. Are synesthetes exceptional beyond their synesthetic associations? A systematic comparison of creativity, personality, cognition, and mental imagery in synesthetes and controls. *Brit. J. Psychol.* **107**, 397–418 (2016).
17. Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261-70 (2003).
18. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424-30 (2006).
19. Formisano, E. & Kriegeskorte, N. Seeing patterns through the hemodynamic veil--the future of pattern-information fMRI. *Neuroimage* **62**, 1249-56 (2012).
20. Hebart, M. N. & Baker, C. I. Deconstructing multivariate decoding for the study of brain function. *Neuroimage* (2017).
21. Thirion, B. et al. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* **33**, 1104-16 (2006).
22. Reddy, L., Tsuchiya, N. & Serre, T. Reading the mind's eye: decoding category information during mental imagery. *Neuroimage* **50**, 818-25 (2010).
23. Poldrack, R. A. Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* **2**, 67-70 (2007).
24. Hupé, J. M. Statistical inferences under the Null hypothesis: Common mistakes and pitfalls in neuroimaging studies. *Front. Neurosci.* **9**, 18 (2015).
25. Eickhoff, S. B. et al. Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *Neuroimage* **36**, 511-21 (2007).
26. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. USA* **103**, 3863-8 (2006).
27. Ruiz, M. J. & Hupé, J. M. Assessment of the hemispheric lateralization of grapheme-color synesthesia with Stroop-type tests. *PLoS One* **10**, e0119377 (2015).
28. Rouw, R. & Scholte, H. S. Neural basis of individual differences in synesthetic experiences. *J. Neurosci.* **30**, 6205-13 (2010).
29. Gould van Praag, C. D., Garfinkel, S., Ward, J., Bor, D. & Seth, A. K. Automaticity and localisation of concurrents predicts colour area activity in grapheme-colour synaesthesia. *Neuropsychologia* **88**, 5-14 (2016).

30. Eagleman, D. M., Kagan, A. D., Nelson, S. S., Sagaram, D. & Sarma, A. K. A standardized test battery for the study of synesthesia. *J. Neurosci. Methods* **159**, 139-45 (2007).

31. Brouwer, G. J. & Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992-4003 (2009).

32. Parkes, L. M., Marsman, J. B., Oxley, D. C., Goulermas, J. Y. & Wuerger, S. M. Multivoxel fMRI analysis of color tuning in human primary visual cortex. *J. Vis.* **9**, 1 1-13 (2009).

33. Tootell, R. B. H. & Nasr, S. Columnar segregation of magnocellular and parvocellular streams in human extrastriate cortex. *J. Neurosci.* **37**, 8014-8032 (2017).

34. Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. The lateral occipital complex and its role in object recognition. *Vision Res.* **41**, 1409-22 (2001).

35. Dehaene, S. & Cohen, L. The unique role of the visual word form area in reading. *Trends Cogn. Sci.* **15**, 254-62 (2011).

36. Bordier, C., Hupé, J. M. & Dojat, M. Quantitative evaluation of fMRI retinotopic maps, from V1 to V4, for cognitive experiments. *Front. Hum. Neurosci.* **9**, 277 (2015).

37. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**, 2636-43 (2012).

38. Ruiz, M. J., Hupé, J. M. & Dojat, M. Use of pattern-information analysis in vision science: a pragmatic examination. In: Wang F., Shen D., Yan P., Suzuki K. (eds) *Machine Learning in Medical Imaging* (MLMI 2012). *Lecture Notes in Computer Science* **7588**, 103-110, Springer: Berlin, Heidelberg (2012).

39. Hupé, J. M., Bordier, C. & Dojat, M. A BOLD signature of eyeblinks in the visual cortex. *Neuroimage* **61**, 149–161 (2012).

40. Vasseur, F. et al. fMRI retinotopic mapping at 3 T: Benefits gained from correcting the spatial distortions due to static field inhomogeneity. *J. Vis.* **10 (12)**, 30 (2010).

41. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **113**, 7900-5 (2016).

42. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

43. Caspers, J. et al. Cytoarchitectonical analysis and probabilistic mapping of two extrastriate areas of the human posterior fusiform gyrus. *Brain Struct. Funct.* **218**, 511-26 (2013).

44. Lorenz, S. et al. Two new cytoarchitectonic areas on the human mid-fusiform gyrus. *Cereb. Cortex* **27**, 373-385 (2017).

45. Choi, H. J. et al. Cytoarchitectonic identification and probabilistic mapping of two distinct areas within the anterior ventral bank of the human intraparietal sulcus. *J. Comp. Neurol.* **495**, 53-69 (2006).

46. Scheperjans, F. et al. Observer-independent cytoarchitectonic mapping of the human superior parietal cortex. *Cereb. Cortex* **18**, 846-67 (2008).

47. Caspers, S. et al. The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. *Neuroimage* **33**, 430-48 (2006).

48. Caspers, S. et al. The human inferior parietal lobule in stereotaxic space. *Brain Struct. Funct.* **212**, 481-95 (2008).

49. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* **67**, 48 (2015).

50. Stelzer, J., Lohmann, G., Mueller, K., Buschmann, T. & Turner, R. Deficient approaches to human neuroimaging. *Front. Hum. Neurosci.* **8** (2014).

51. Gebuis, T., Nijboer, T. C. & Van der Smagt, M. J. Multiple dimensions in bi-directional synesthesia. *Eur. J. Neurosci.* **29**, 1703-10 (2009).

52. Cichy, R. M., Heinzle, J. & Haynes, J. D. Imagery and perception share cortical representations of content and location. *Cereb. Cortex* **22**, 372-80 (2012).

53. Bannert, M. M. & Bartels, A. Decoding the yellow of a gray banana. *Curr. Biol.* **23**, 2268-72 (2013).

54. Cumming, G. *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis* (Routledge, New York, 2012).

55. Yarkoni, T., Poldrack, R. A., Van Essen, D. C. & Wager, T. D. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* **14**, 489-96 (2010).

56. Wasserstein, R. L. & Lazar, N. A. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* **70**, 129-133 (2016).

57. Hupé, J. M. Comment on “Ducklings imprint on the relational concept of ‘same or different’”. *Science* **355**, 806-806 (2017).

58. Turner, R. Uses, misuses, new uses and fundamental limitations of magnetic resonance imaging in cognitive science. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **371** (2016).

Acknowledgements

Research funded by *Agence Nationale de la Recherche* ANR-11-BSH2-010. Mathieu J. Ruiz was supported by a Ph.D. allowance from the Université Grenoble Alpes. Grenoble MRI facility IRMaGe was partly funded by the French program *Investissement d'avenir* run by the Agence Nationale de la Recherche: grant *Infrastructure d'avenir en Biologie Santé* ANR-11-INBS-0006.

Author Contributions

M.J.R., M.D, and J.-M.H. designed research; M.J.R. and M.D. performed research; M.J.R., M.D, and J.-M.H. analysed data; J.-M.H. wrote the paper; M.J.R., M.D, and J.-M.H. revised the paper.

Competing Financial Interests

The authors declare no competing financial interests.

Supplementary Information

Table S1. Clusters identified based on whole brain analyses and tested *post-hoc* with MVPA.

Figure S1. Right occipito-parietal cortex cluster identified based on whole brain univariate analysis

Figure S2. Left anterior insula cluster identified based on whole brain univariate analysis

Figure S3. Right frontal cortex cluster identified based on whole brain univariate analysis

Figure S4. Alternative version of Fig. 4, based on mixed-effect generalized linear models

Figure S5. Alternative version of Fig. 6, based on mixed-effect generalized linear models