Behavioral/Systems/Cognitive

# Associative Learning Increases Trial-by-Trial Similarity of BOLD-MRI Patterns

**Renée M. Visser,**[1,2] **H. Steven Scholte,**[2,3] **and Merel Kindt**[1,2]

[1]Department of Clinical Psychology, [2]Priority Program Brain and Cognition, and [3]Department of Brain and Cognition, University of Amsterdam, 1018 WB Amsterdam, The Netherlands

**Associative learning is a dynamic process that allows us to incorporate new knowledge within existing semantic networks. Even after years, a seemingly stable association can be altered by a single significant experience. Here, we investigate whether the acquisition of new associations affects the neural representation of stimuli and how the brain categorizes stimuli according to preexisting and emerging associations. Functional MRI data were collected during a differential fear conditioning procedure and at test (4–5 weeks later). Two pictures of faces and two pictures of houses served as stimuli. One of each pair coterminated with a shock in half of the trials (partial reinforcement). Applying Multivoxel Pattern Analysis (MVPA) in a trial-by-trial manner, we quantified changes in the similarity of neural representations of stimuli over the course of conditioning. Our findings show an increase in similarity of neural patterns throughout the cortex on consecutive trials of the reinforced stimuli. Furthermore, neural pattern similarity reveals a shift from original categories (faces/houses) toward new categories (reinforced/unreinforced) over the course of conditioning. This effect was differentially represented in the cortex, with visual areas primarily reflecting similarity of low-level stimulus properties (original categories) and frontal areas reflecting similarity of stimulus significance (new categories). Effects were not dependent on overall response amplitude and were still present during follow-up. We conclude that trial-by-trial MVPA is a useful tool for examining how the human brain encodes relevant associations and forms new associative networks.**

## Introduction

From the very first moment we interact with our environment, semantic networks are formed and continuously updated through integrating new information within the context of previous experience. Central to the formation of these semantic networks is the learning of associations between novel and meaningful events. Even after years, a seemingly stable semantic network can be altered by a single experience, especially when this experience carries an affective load, sometimes resulting in long-lasting fear memory. Given the power of associative fear learning, an intriguing question is to what extent an aversive event overshadows the original semantic network.

Associative fear learning is typically studied in a classical fear conditioning paradigm by the pairing of an initially neutral or ambiguous stimulus [conditioned stimulus (CS)] and an intrinsically aversive consequence [unconditioned stimulus (US)] (Pavlov, 1927). The neural circuits that mediate human fear conditioning are well delineated (for review, see Sehlmeyer et al., 2009; Mechias et al., 2010), yet little is known about how fear conditioning alters the neural representation and categorization of a stimulus (e.g., once bitten by a dog, the association "a dog is a pet" may be overshadowed by "a dog is primarily a dangerous animal"). Multivoxel pattern analysis (MVPA) is a technique that, by decoding distributed patterns of BOLD-MRI data, offers the opportunity to examine the neural representation of stimuli. Unlike analysis of mean activation, MVPA yields a distinctive stimulus signature that can be used to assess semantic similarity, providing a tool for examining how the brain categorizes information (Haxby et al., 2001; Kriegeskorte et al., 2008; for review, see Norman et al., 2006). To date, it remains unknown what happens to the neural representation of dominant, well defined categories when new associations are acquired. In this light, the application of MVPA in a trial-by-trial manner seems particularly valuable, as it enables the assessment of gradual changes in the neural representation and categorization of a stimulus.

Previous studies have shown that the neural representation of a (neutral) stimulus refines as a function of repeated stimulus presentation (Li et al., 2009; Xue et al., 2010; Zhang et al., 2010). A similar refinement may be observed for associative fear learning, which may not only be visible in perceptual areas, but also in areas that are more directly involved in the processing of stimulus significance. Although changes in the spatial pattern of activity have been shown as a result of fear conditioning (Li et al., 2008), gradual changes in similarity between stimulus-evoked activation patterns have not been directly assessed.

Combining differential fear conditioning and MVPA in a trial-by-trial manner, this study tests two hypotheses. First, fear conditioning will refine the neural representation of affectively

Received April 29, 2011; revised June 15, 2011; accepted June 29, 2011.

Author contributions: R.M.V., H.S.S., and M.K. designed research; R.M.V. performed research; H.S.S. contributed unpublished reagents/analytic tools; R.M.V. and H.S.S. analyzed data; R.M.V. and M.K. wrote the paper.

This work was supported by a Vici grant (M.K.) from the Netherlands Organization for Scientific Research. We thank B. Molenkamp and M. Spaan for technical assistance.

The authors declare no competing financial interests.

This article is freely available online through the *J Neurosci* Open Choice option.

Correspondence should be addressed to Dr. Merel Kindt, Department of Clinical Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: M.Kindt@uva.nl.

DOI:10.1523/JNEUROSCI.2178-11.2011

 0270-6474/11/3112021-08$15.00/0

significant stimuli, which will be expressed in a differential increase of within-stimulus correlations on consecutive trials. Second, the formation of new categories based on affective significance (threat/no threat) will alter the preponderance of preexisting categories (faces/houses) over the course of conditioning.

## Materials and Methods

*Participants.* Twenty-two undergraduate psychology students of the University of Amsterdam participated in the first part of the experiment (mean age, 22.4 ± 3.8 years; five male; 18 right-handed). All participants gave their written informed consent before participation, had normal or corrected-to-normal vision, and were naive to the purpose of the experiment. Procedures were executed in compliance with relevant laws and institutional guidelines and were approved by the local ethics committee. Due to excessive movement, data from three participants had to be discarded from analyses. Twenty-one subjects returned for a follow-up experiment 4–5 weeks after they participated in the first part. Again, data from three participants had to be discarded from analyses due to excessive movement.

*Experimental design.* The experiment consisted of two sessions: an acquisition phase and a test phase 4–5 weeks later. For the acquisition phase, a differential fear conditioning paradigm was used (Fig. 1*A*). Two faces and two houses were repeatedly presented for 6.5 s and served as the to-be-conditioned stimuli. One face and one house coterminated with a mild electric stimulus in half of the trials (CS+, partial reinforcement), while the other face and house were never followed by an electric stimulus (CS−, unreinforced). Faces were selected from the NimStim set based on how neutral they were (Tottenham et al., 2009) and were converted to grayscale. Houses (Haxby et al., 2001) were already in grayscale. All stimuli were presented on a gray background to minimize afterimages. The electrical stimulation served as US and was delivered for 2 ms to the right shinbone. Before the experiment, the intensity of the electric stimulus was individually adapted to be aversive but not painful (intensity range, 11–39 mA). Participants were explicitly told that two of four stimuli could be followed by the shock, the other two would never be followed by the shock, and that they had to learn these contingencies. Participants had no problem identifying afterward which house and which face was paired with a shock.

The test phase, 4–5 weeks later, was exploratory and was combined with a pilot study (the results of the pilot study are not discussed here). This pilot preceded the test phase and consisted of 15 min of functional scanning while participants viewed numerous pictures of faces and houses. The pictures included the four stimuli that were used during the acquisition phase (each was presented four times). However, participants knew that none of the stimuli would be followed by a shock (half of the participants did not have electrodes attached to their leg and the other half was explicitly told that no shocks would be delivered during that particular run). During the subsequent test phase, only the four stimuli that were used during acquisition were presented and participants were asked to memorize what they had learned about the stimuli 4 weeks earlier. All participants had electrodes attached to their legs and the shock intensity was set at the individual level as determined during the first session. Given that this follow-up served as a memory test, no actual shocks were delivered. After scanning, participants had to indicate which pictures were followed by a shock during acquisition to assess retention of the acquired contingencies. During the acquisition phase and for at least 3 weeks following acquisition, participants were not aware that they would be asked to participate in this follow-up.

In both sessions, participants were repeatedly instructed not to move their eyes, but instead fixate on the center of the screen for as long as a stimulus was presented. This was done to prevent variation of image representation in retinotopically organized areas in the visual cortex. During the stimulus intervals, a white fixation cross appeared on the screen and turned pink 500 ms before a stimulus was presented. This gave participants the chance to focus in time.

Interstimulus intervals were fixed and long enough (21.5 s) to reduce intrinsic noise correlations. The onset of each trial was triggered by the start of the acquisition of a BOLD-MRI volume. The order of stimulus presentation was fixed (counterbalanced across participants) and consisted of a repeating sequence of four target trials, with filler trials of the same stimuli in between. The interval between two consecutive target trials (e.g., trial 5 and 6) of a stimulus (e.g., CS+ face) was exactly as long as the interval between the same consecutive trials (5 and 6) of the other three stimuli (CS− face, CS+ house, CS− house). In total, the acquisition phase consisted of 52 trials: 28 target trials (seven per stimulus type) and 24 filler trials (six per stimulus type), including all CS+ trials that coterminated with a shock. The test phase consisted of 24 trials: 16 target trials (four per stimulus type) and eight filler trials (two per stimulus type). For both phases, we constrained our analyses to target trials. Using target and filler trials was necessary for three reasons. First, the interval between two consecutive presentations of a stimulus had to be equal for all four stimuli to ascertain that differences between conditions in the strength of trial-to-trial correlations were not affected by factors other than the experimental manipulation, such as temporal proximity. Second, filler trials were necessary to make the order of stimuli to be perceived as random by the subject. Third, in our design, activity related to the US may confound activity related to the CS during reinforced trials. For the acquisition phase, reinforced trials were therefore treated as filler trials and not analyzed. Although no shocks were administered during

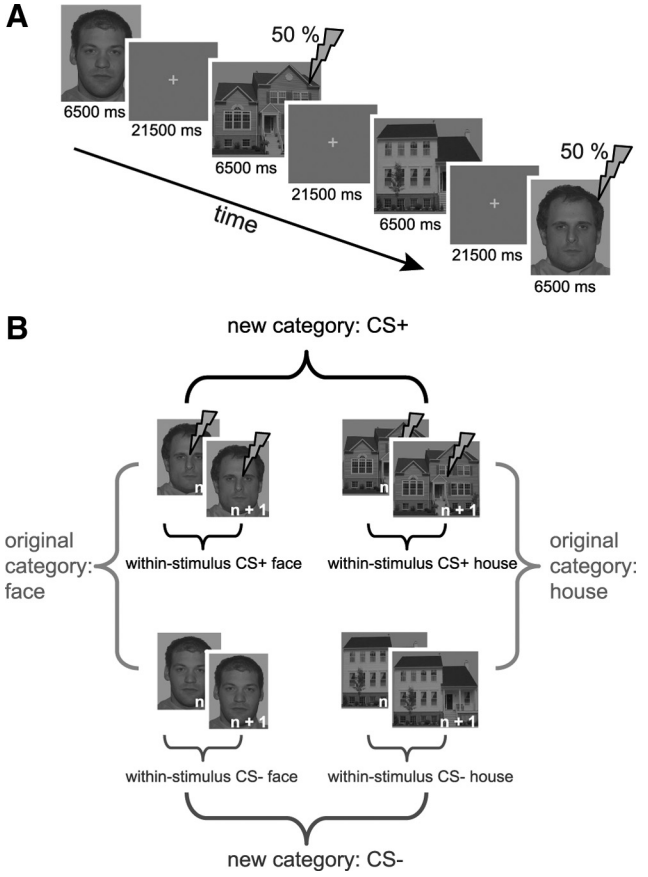


**Figure 1.** Conditioning paradigm. ***A***, A partial reinforcement paradigm was used during functional scanning. Four stimuli were repeatedly presented for 6.5 s, two of which coterminated with a shock on half of the trials. Intertrial intervals were fixed. ***B***, Correlations were calculated between consecutive trials of the same stimulus (within-stimulus), between trials of stimuli that belong to the same category (original category: faces and houses), and between trials of stimuli that share (non)reinforcement (new category: shock vs no shock). Our analyses were restricted to target trials (i.e., trials in which no shock was given). Note that the number of category correlations is equal to the number of target trials, whereas the number of within-stimulus correlations is equal to the number of target trials minus one. Results are reported with original category correlations and within-stimulus correlations being averaged over faces and houses (see Materials and Methods).

the test phase, filler trials were still necessary to make the order of stimuli unpredictable.

*Image acquisition and data analysis.* Scanning was performed on a 3 T Philips Achieva MRI scanner using an eight-channel head-coil. Functional data were acquired using a gradient-echo, echo-planar pulse sequence (TR, 2000 ms; TE, 27.63 ms; FA, 90°; 38 sagittal slices with interleaved acquisition; voxel size, 2.04 × 2.04 × 3.1 mm; 96 × 96 matrix; FOV, 192 × 192 × 129; SENSE factor, 1) covering the whole brain. For the acquisition phase, 730 volumes were acquired and 366 volumes were acquired for the test phase. Foam pads minimized head motion. A high-resolution T1-weighted image (TR, 8.141 ms; TE, 3.74 ms; FOV, 256 × 256 × 160) was collected for anatomical visualization. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil.

FEAT (fMRI Expert Analysis Tool) version 4.1, part of FSL [Oxford Centre for Functional MRI of the Brain (FMRIB) Software Library; www.fmrib.ox.ac.uk/fsl] was used to analyze the fMRI data. Preprocessing steps included slice-time correction, motion correction, high-pass filtering in the temporal domain ($\sigma = 100$ s), and prewhitening (Woolrich et al., 2001). Structural images were coregistered to the functional images and transformed to MNI standard space (Montreal Neurological Institute) using FLIRT (FMRIB's Linear Image Registration Tool; FSL). The resulting normalization parameters were applied to the functional images. No spatial smoothing was applied.

All trials were modeled as separate events. The resulting single trial data were further analyzed in Matlab (version 7.4; MathWorks) as described below. Regions of interest (ROI) were obtained, where possible, from the Juelich Histological atlas (Eickhoff et al., 2007) and otherwise from the Harvard-Oxford cortical and subcortical structural atlases (Harvard Center for Morphometric Analysis). Selected ROIs included the insula, amygdala, hippocampus, anterior cingulate cortex (ACC), superior frontal gyrus (SFG), middle frontal gyrus (MFG), medial frontal cortex (MFC), inferior frontal gyrus (IFG), orbitofrontal cortex (OFC), occipital cortex (OC), inferior temporal gyrus (ITG), medial temporal gyrus (MTG), and superior parietal lobule (SPL).

For each participant, a vector was created containing the spatial pattern of BOLD-MRI signal related to a particular event (*Z*-values per voxel) in a certain ROI. Pairwise Pearson correlations were calculated between all vectors of all single trials, resulting in a similarity matrix containing correlations among all trials for each participant for each ROI. Correlations were then Fisher-transformed and averaged across participants.

From the average correlation matrix, three different types of correlations were selected (Fig. 1*B*). First, we examined within-stimulus correlations on consecutive target trials. Second, we examined correlations between adjacent target trials of original-related stimuli (original categories: CS+ face with CS− face and CS+ house with CS− house) and correlations between adjacent target trials that share (non)reinforcement (new categories: CS+ face with CS+ house and CS− face with CS− house). The strength of these correlations indicates to what degree the neural response to two stimuli is similar and was used as a metric of similarity (Xue et al., 2010). Fisher-transformed within-stimulus and original category correlations were then averaged over face and house stimuli (Fig. 1*B*). This was done to reduce the number of comparisons and because we were not interested in whether a face–shock association was learned differently from a house–shock association, or whether the original category "house" changed differently from the original category "face" over the course of conditioning.

We performed repeated-measures ANOVA on Fisher-transformed correlation values using Statistical Package for the Social Sciences (SPSS, version 17; SPSS). For the acquisition phase, we assumed differential fear learning to be expressed by a significant interaction of trials × stimulus type or category. For the test phase, we assumed memory to be expressed by a significant main effect of stimulus type or category. Predictions were tested while correcting for multiple comparisons (13 ROIs) by limiting the false discovery rate (FDR) (Benjamini and Hochberg, 1995). This method corrects the $p$ value at which significance is evaluated (in this case, $p = 0.05$) for the number of tests being performed. In contrast to Bonferroni correction, which controls the chance of any false positive among all tests, FDR correction fixes the expected proportion of false positives (Benjamini and Hochberg, 1995) and is therefore more suited for tests that are not independent. However, this method is not suitable when predictions differ for the to-be-tested effects, as high $p$ values on some tests reduce the chance of finding any meaningful differences on other tests. As we did not formulate specific predictions about how exactly effects would be represented across different cortical regions, uncorrected effects are additionally reported. All $p$ values are reported two-sided.

Finally, traditional activation-based analyses were performed. This was done for two reasons. First, we tested whether trial-to-trial fluctuations of BOLD activity for the CS+ stimulus relative to the CS− stimulus could account for an increase in trial-to-trial correlations. This analysis was necessary to determine whether trial-by-trial correlation curves contain unique information, or whether instead they are paralleled by deflecting curves of overall response–amplitude. We used the same ROIs for this analysis as for the correlation analysis. Second, to integrate our findings with the existing body of literature regarding human fear conditioning, we examined activation in the left and right amygdala, dorsal ACC (dACC), and left and right anterior insula (Sehlmeyer et al., 2009; Mechias et al., 2010). The amygdala was atlas-based; for the anterior insula (40/−40, 16, −6) and the dACC (−2, 14, 40), 10 mm spheres were created around previously reported coordinates (coordinates are in MNI space).

## Results

All 22 participants were aware of the CS–US contingencies (i.e., which pictures were/were not followed by the shock) immediately after they underwent differential fear conditioning. After scanning, participants rated the US as being aversive (ranging from moderately to highly aversive). After the test phase 4–5 weeks later, participants were again asked about the CS–US contingencies; five of 21 misidentified one or both reinforced stimuli. All participants indicated they were not expecting shocks during the first pilot run, but were expecting shocks to follow CS+ stimuli during the test phase. For this test phase, only data were analyzed from participants that correctly remembered the CS–US associations and that were not confounded by movement ($N = 13$). For the acquisition phase, all data that were not confounded by movement were analyzed ($N = 19$).

### Within-stimulus correlations

In line with our expectations, results show that associative learning coincides with a differential increase in similarity of BOLD-MRI patterns on consecutive trials (Fig. 2). This increase was visible in multiple cortical regions, with effect sizes ranging from medium ($\eta^2 = 0.11$) to large ($\eta^2 = 0.24$) (Cohen, 1988; Tabachnick and Fidell, 2007). The largest effects [trial (6) × stimulus type (2), FDR corrected] were found in the SFG ($F_{(5,90)} = 5.55$, $p < 0.0005$), MFG ($F_{(5,90)} = 3.76$, $p = 0.004$), ACC ($F_{(5,90)} = 3.70$, $p = 0.004$), OC ($F_{(5,90)} = 3.18$, $p = 0.011$), and the OFC ($F_{(5,90)} = 3.31$, $p = 0.009$). Smaller effects were found in the IFG ($F_{(5,90)} = 2.70$, $p = 0.026$), the SPL ($F_{(5,90)} = 2.50$, $p = 0.034$), insula ($F_{(5,90)} = 2.62$, $p = 0.029$), and MTG ($F_{(5,90)} = 2.52$, $p = 0.035$); however, these effects did not survive FDR correction. Although a similar pattern was observed in the ITG and MFC, interaction effects of trial × stimulus type did not reach significance. In the amygdala and hippocampus, no clear differential learning curves were observed.

During the test phase 4–5 weeks later, correlations of BOLD-MRI patterns in two ROIs still showed a significant difference between the CS+ and CS− [main effect of stimulus type (2)] in the IFG ($F_{(1,12)} = 7.14$, $p = 0.020$, $\eta^2 = 0.37$) and the OFC ($F_{(1,12)} = 6.14$, $p = 0.029$, $\eta^2 = 0.34$). However, these significant effects did not survive FDR correction. Given that the effect sizes

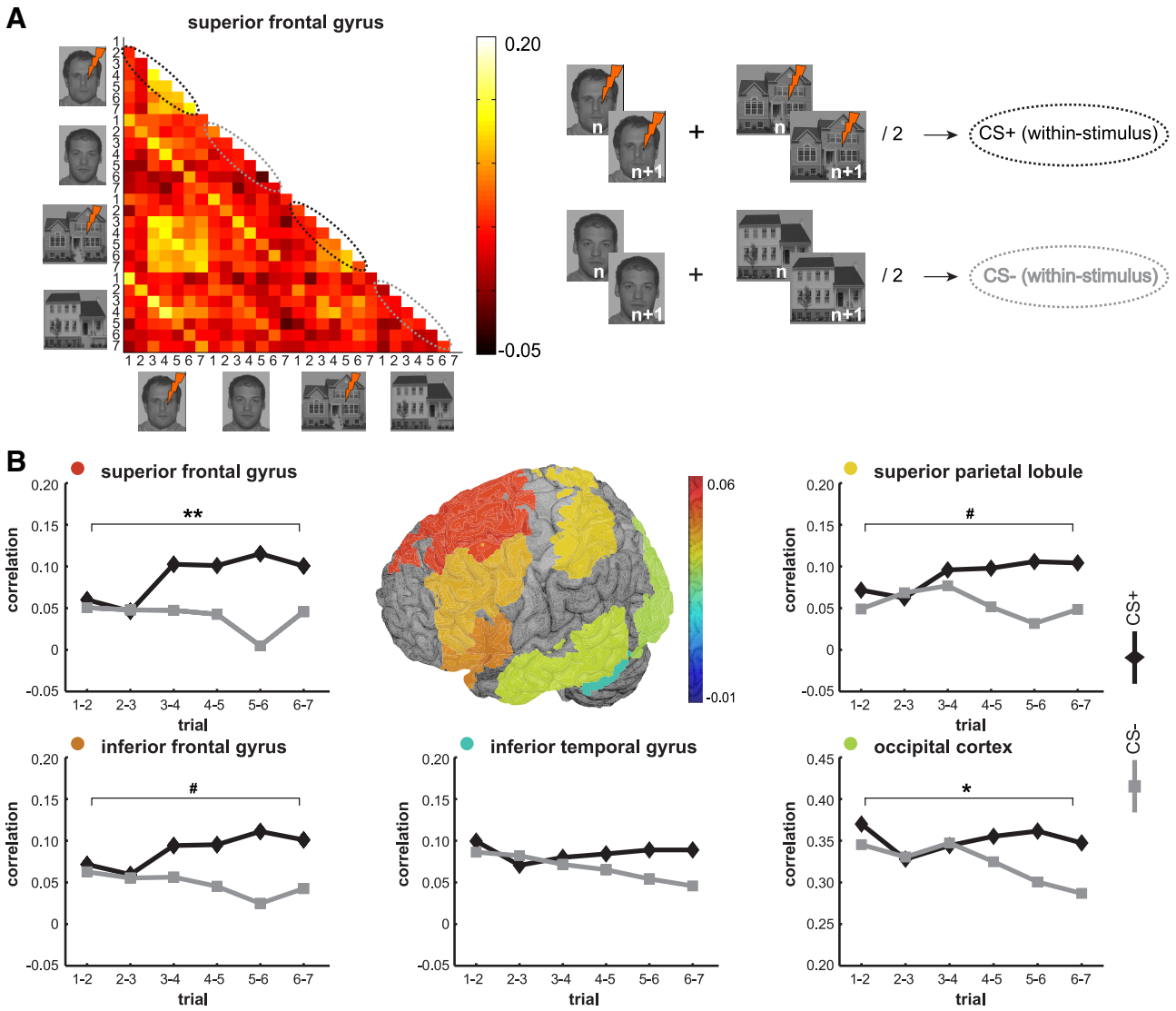


**Figure 2.** Within-stimulus correlations of BOLD-MRI patterns during acquisition. ***A***, A 28 × 28 correlation matrix was created for each participant ($N = 19$) and each ROI (shown is half of the correlation matrix for the superior frontal gyrus). The off-diagonal represents the correlations between two consecutive target trials. The first ellipse shows correlations between trials of the reinforced (CS+) face stimulus, the second ellipse shows correlations between trials of the unreinforced (CS−) face stimulus, the third ellipse shows correlations between trials of the reinforced (CS+) house stimulus, and the fourth ellipse shows correlations between trials of the unreinforced (CS−) house stimulus. For reinforced stimuli, an increase of within-stimulus correlations over the course of conditioning is visible. ***B***, Graphs refer to the off-diagonal of the correlation matrix and represent the trial-to-trial within-stimulus correlations in several cortical ROIs. The first CS+ trial that is paired with a shock (filler trial, see Materials and Methods) occurs between the second and third target trials. Colors indicate the average difference between CS+ and CS− stimuli over seven trials and illustrate a selective increase of within-stimulus correlations for CS+ stimuli. Statistics refer to the interaction of trials (6) × stimulus type (2) performed on Fisher-transformed correlations. $^{*}p < 0.05$; $^{**}p < 0.001$, FDR corrected; $^{\#}p < 0.05$, uncorrected.

are large, this may be due to the smaller sample size at test compared with acquisition. Also in other areas (ACC, SFG, MFG, MTG, insula), correlations for the CS+ were higher than for the CS− (effect sizes ranging from 0.11 to 0.25), but main effects of stimulus type did not reach statistical significance.

### Between-stimulus correlations

Aside from within-stimulus correlations, between-stimulus correlations were calculated. The strength of correlations for the original categories and new categories revealed a shift from old categories (faces/houses) toward new categories (CS+/CS−) over the course of conditioning (Fig. 3). Although strong main effects of category were observed in the ACC, SPL, IFG, SFG, OFC, MTG, ITG, mPFC, and OC (effect sizes ranging from 0.20 to 0.60), in none of the brain areas did interaction effects (category [3] × trial [7]) survive FDR-corrected significance testing (effect sizes ranging from 0.04 to 0.10). Surprisingly, without FDR correction, a small interaction effect was observed in the OC ($F_{(12,216)} = 1.89$, $p = 0.037$, uncorrected), while a graphical presentation of the data shows that the emergence of an affectively significant (CS+) category was more pronounced in frontal areas (Fig. 3*B*).

During the test-phase 4–5 weeks later, the affectively significant category was still prominent in frontal areas (Fig. 4). Main effects of category were found in the ACC ($F_{(2,24)} = 3.95$, $p = 0.033$), IFG ($F_{(2,24)} = 4.55$, $p = 0.021$), insula ($F_{(2,24)} = 8.86$, $p = 0.001$), MFG ($F_{(2,24)} = 3.49$, $p = 0.047$), OFC ($F_{(2,24)} = 7.19$, $p = 0.004$), and SFG ($F_{(2,24)} = 3.63$, $p = 0.042$), with effect sizes ranging from 0.23 to 0.43, but only effects in the insula and OFC survived FDR correction. In the OC, there was also a main effect

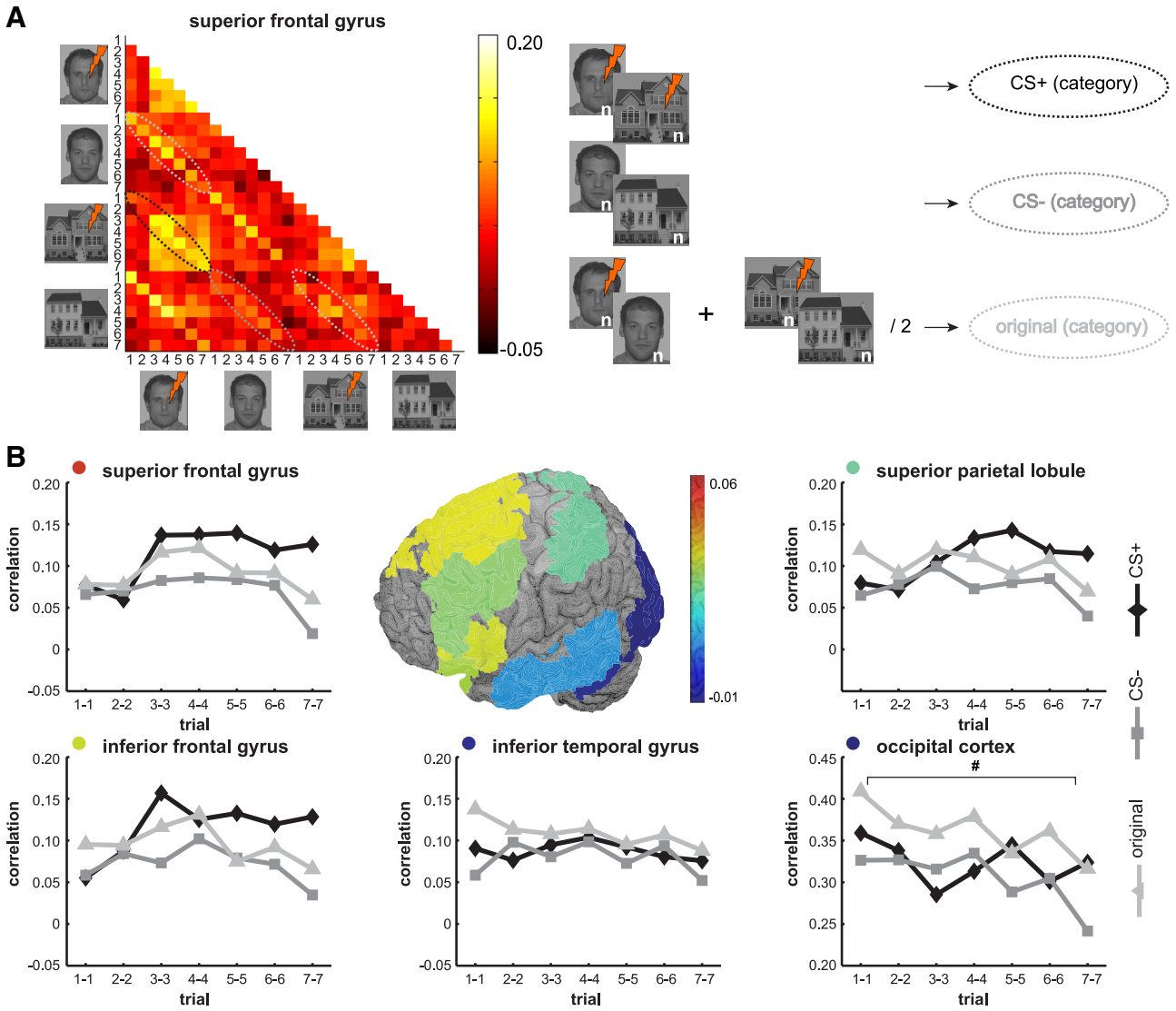


**Figure 3.** Between-stimulus correlations of BOLD-MRI patterns (formation of categories) during acquisition. ***A***, A 28 × 28 correlation matrix was created for each participant ($N = 19$) and for each ROI (shown is half of the correlation matrix for the superior frontal gyrus). Highlighted are trial-by-trial correlations between seven target trials of different stimuli. The first ellipse shows correlations between adjacent trials of the reinforced face and house stimuli (new category: CS+), the second ellipse shows correlations between trials of the unreinforced face and house stimuli (new category: CS−), the third ellipse shows correlations between trials of the reinforced and unreinforced face stimuli (original category: faces), and the fourth ellipse shows correlations between trials of the reinforced and unreinforced house stimuli (original category: houses). For the CS+ category, an increase of correlations over the course of conditioning is visible. ***B***, Graphs refer to the between-stimulus correlations in several cortical ROIs. The first CS+ trial that is paired with a shock (a filler trial, see Materials and Methods) occurs between the second and third target trials. Colors indicate the sum of the average difference between CS+ and CS− and the difference between CS+ and the original categories over seven trials. This illustrates how the new category, which is based on affective significance, becomes dominant in frontal and parietal areas, while the original categories remain dominant in temporal and occipital areas. Statistics refer to the interaction of trials (7) × category (3) performed on Fisher-transformed correlations. $^{\#}p < 0.05$, uncorrected.

of category, but, in contrast with frontal and parietal areas, this was caused by higher correlations for the original categories than for the reinforced category ($F_{(2,24)} = 4.10$, $p = 0.029$) (Fig. 4).

Interestingly, when data from the acquisition phase were reanalyzed for selected participants who correctly remembered the CS–US associations 4–5 weeks later ($N = 14$), category-formation patterns were enhanced and interaction effects did reach statistical significance (uncorrected) in multiple areas including the IFG ($F_{(12,156)} = 1.89$, $p = 0.041$), SFG ($F_{(12,156)} = 1.88$, $p = 0.041$), MTG ($F_{(12,156)} = 2.13$, $p = 0.018$), OFC ($F_{(12,156)} = 1.87$, $p = 0.042$), and OC ($F_{(12,156)} = 1.89$, $p = 0.041$), effect sizes ranging from 0.13 to 0.15. The current sample sizes do not allow for classification analyses, hence we cannot make inferences about the implications of these findings for successful memory encoding.

### Activation-based analysis

To explore whether increased BOLD activity for the CS+ stimulus relative to the CS− stimulus could account for an increase in trial-to-trial correlations, we performed traditional activation-based analyses on the same ROIs that we used for the correlation analyses. Repeated-measures ANOVA revealed a significant main effect of stimulus type in the SFG ($F_{(1,18)} = 7.54$, $p = 0.026$) (Fig. 5*A*), as well as in the MFG, ITG, MTG, hippocampus, and OC (all $Fs > 5.38$). However, these main effects were explained by more activation in response to the CS− compared with the CS+, ruling out the possibility that higher activation could account for higher CS+ correlations that were found in some of these areas. No significant effects of mean activation were found in OFC, IFG, or SPL (all $Fs < 2.40$).

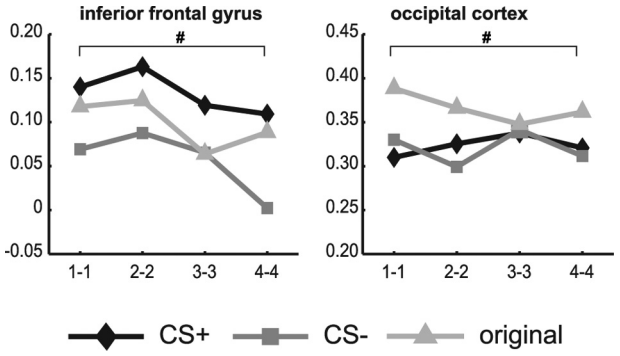


**Figure 4.** Between-stimulus correlations of BOLD-MRI patterns during the test phase. Correlations ($N = 13$) show long-term effects of aversive conditioning in several ROIs 4–5 weeks after a fear-association was learned. Frontal ROIs still show higher correlations for the reinforced (CS+) category, whereas the visual cortex shows higher correlations for the original categories (i.e., faces and houses). Statistics refer to main effects of category performed on Fisher-transformed correlations. $^{\#}p < 0.05$, uncorrected.

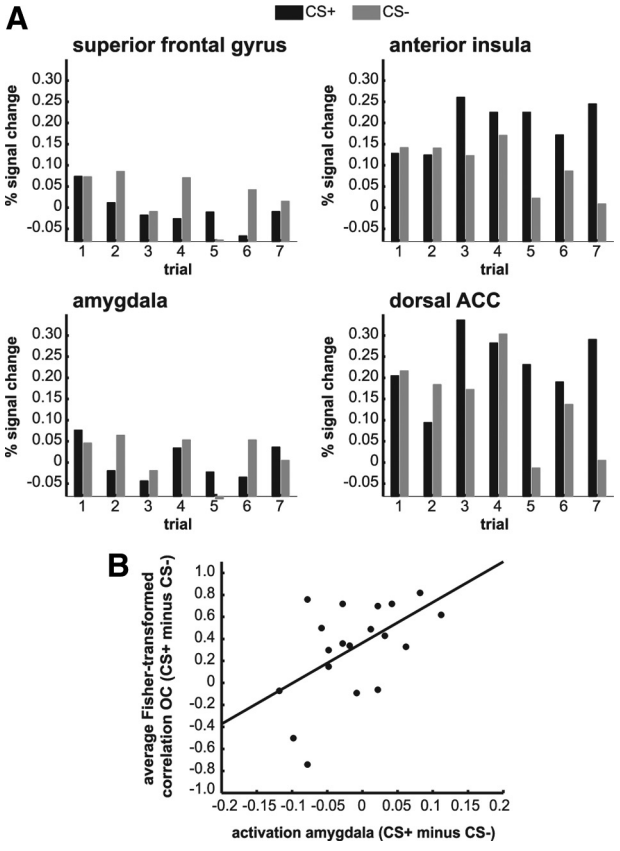


**Figure 5.** Single trial activation-based analyses. ***A***, Data ($N = 19$) show higher activation for CS+ stimuli compared with CS− stimuli in the dorsal ACC and the anterior insula, but not in the amygdala or superior frontal gyrus. In the superior frontal gyrus, differential mean activation could not account for differential correlations. ***B***, A correlation between differential amygdala activation (CS+ − CS−, averaged over seven trials) and differential within-stimulus correlations (CS+ − CS−, averaged over six trial-to-trial Fisher-transformed correlations) in the visual cortex are in line with theories about the role of the amygdala in refinement of visual processing.

We additionally examined activation in areas associated with fear conditioning. Repeated-measures ANOVA revealed an interaction of stimulus type (2) × trial (7) in the anterior insula ($F_{(6,108)} = 2.47$, $p = 0.028$), dACC ($F_{(6,108)} = 2.57$, $p = 0.023$), and amygdala ($F_{(6,108)} = 2.26$, $p = 0.043$) (Fig. 5*A*). In the anterior insula and in the dACC, this interaction originated from an increase in activation in response to the CS+ over the course of conditioning (main effect of stimulus type: $F_{(1,18)} = 13.35$, $p = 0.002$ and $F_{(1,18)} = 8.92$, $p = 0.008$, respectively). In the amygdala, the interaction was not explained by a main effect of stimulus type ($F_{(1,18)} = 0.84$, $p = 0.37$). Nevertheless, its role in refinement of visual processing (Vuilleumier and Driver, 2007; Lim et al., 2009; Sabatinelli et al., 2009; Pessoa and Adolphs, 2010) was indirectly supported by a correlation between differential amygdala activation (CS+ − CS−, averaged over seven trials) and differential within-stimulus correlations (CS+ − CS−, averaged over six trial-to-trial correlations) in the visual cortex ($r = 0.53$, $p = 0.019$) (Fig. 5*B*).

Together, our design replicates and extends previous findings from neuroimaging studies on human fear conditioning. Importantly, only differential activation in the anterior insula and dACC resembled the learning curve as quantified with BOLD-MRI correlations, suggesting that in most areas, the stronger CS+ correlations were not purely driven by an increase in overall response amplitude.

## Discussion

Here we show that associative learning increases similarity of BOLD-MRI patterns throughout the cortex on consecutive trials of the reinforced stimulus, but not on trials of the unreinforced stimulus. Additionally, we examined how the brain categorizes stimuli according to preexisting and emerging associations. Our findings show dissociable roles of different brain areas in the formation of new categories, on a trial-by-trial basis, with visual areas primarily reflecting similarity of low-level stimulus properties (old categories) and frontal areas reflecting similarity of stimulus significance (new categories). Differential pattern similarity was not explained by overall response amplitude and was still present during follow-up.

As one may have noticed, correlations in this study are rather low or even close to zero. This may not be surprising considering that the technique here uses single trials. Consequently, the signal-to-noise ratio is limited compared with analyses where BOLD-MRI patterns from multiple trials are used to train classifiers. On top of that, stimulus intervals are quite long and ROIs are atlas-based instead of based on masks of responsive voxels. The visual cortex, which is retinotopically organized, showed substantially higher correlations. Nonetheless, despite the low correlations, learning-dependent changes were clearly observed in multiple cortical areas, with effect sizes varying from moderate to large, suggesting that the underlying neural mechanism must be fairly robust.

Our study touches on four lines of neuroimaging research: (1) the investigation of human fear conditioning using traditional activation-based approaches (Büchel et al., 1998; LaBar et al., 1998); (2) research showing a link between the acquired emotional significance of a stimulus and more refined responses in perceptual cortices (Keil et al., 2007; Li et al., 2008; Padmala and Pessoa, 2008; Damaraju et al., 2009); (3) the application of MVPA to examine how the brain categorizes (Haxby et al., 2001; Polyn et al., 2005; Ethofer et al., 2009); and 4) the use of MVPA to assess learning-dependent changes (Li et al., 2009; Xue et al., 2010; Zhang et al., 2010). To begin with the last item, it has been fairly well established that learning how to discriminate nonemotional stimuli, such as Glass patterns, results in a more refined neural representation (i.e., a tuned response) for those particular stimuli (Li et al., 2009; Zhang et al., 2010). Additionally, refinement of neural processing has been found for repeatedly presented, non-

emotional faces, and this refinement was associated with better memory (Xue et al., 2010). This suggests that cortical refinement is not restricted to stimuli that are difficult to distinguish. The neural representation of a stimulus may not only be altered by multiple repetitions of the same stimulus, it may be altered by associative learning as well (Li et al., 2008). Comparing spatial activation maps before and after aversive conditioning, Li and colleagues (2008) found changes for a pair of initially indiscriminable stimuli (odor cues) of which one was reinforced, but not for a second pair of indiscriminable stimuli, which were both unreinforced. In light of the current findings, these changes putatively reflect a refined response for reinforced stimuli. Increased similarity of activation patterns for these threat-associated stimuli fits well within an evolutionary perspective on how enhanced processing of threatening stimuli allows for the rapid selection of appropriate defensive behaviors.

For some time, a predominant view has been that the processing of threatening stimuli is subserved by specialized neural pathways (e.g., a fast subcortical pathway). Here we show that learning-dependent tuning is at least to some degree represented across the entire cortex, which is in line with a more cortical view of how affectively significant stimuli are processed (van den Hout et al., 2000; Pessoa and Adolphs, 2010). It should be noted that some overlap existed between different atlas-based cortical ROIs. Our goal, however, was not to attribute certain functions to certain regions. Instead, we show a gradient of cortical tuning to threat-associated stimuli, with local variation with regard to the size of the effects.

A second observation in this study—the formation of new categories—seemed differentially represented in the cortex. In frontal areas, the strength of trial-by-trial correlations for the original categories and threat-associated categories revealed a shift from old categories (faces/houses) toward new categories (reinforced/unreinforced) over the course of conditioning. In contrast, visual areas primarily reflected similarity of low-level stimulus properties (faces/houses). Although we did not test a priori predictions regarding the spatial distribution of these effects, our observations are in line with previous findings (for review, see Seger and Miller, 2010). Exploratory analyses revealed that the formation of a relevance-based superordinate response pattern during acquisition was restricted to individuals that recalled the (un)reinforced contingencies during follow-up. Although the current sample sizes did not allow for classification analyses, our findings give rise to the intriguing question of whether long-term memory for relevant associations could be predicted from how individuals form superordinate categories based on these associations.

Four weeks later, neural correlates of the new associations were still present. This was to some degree expressed by higher neural pattern similarity on repeated presentations of the same reinforced stimulus, but was more prominently expressed by similarity across reinforced stimuli. Under the same circumstances, the dominance of the affectively significant category appears to be reinstated immediately. Note, however, that it would be very unfortunate if, under all circumstances, new categories would surpass strong semantic categories after learning a new association. A more plausible assumption is that thousands of categories coexist and become alternately dominant depending on which one is most relevant in a particular context. A better understanding of the circumstances that activate a certain category would provide new insights into the flexibility of memory traces and, more specifically, individual differences in acquired fear. As a starting point, it would be interesting to examine to what degree the learned associations generalize to other stimuli (different faces/different houses) and to other contexts (without the threat of a shock). The present study illustrates the utility of MVPA for studying how the brain categorizes; a process that cannot be studied using traditional activation-based approaches.

Similar classical conditioning paradigms have been used for decades as experimental models of how fear is acquired (LeDoux, 2000). The effectiveness of this paradigm as a model for fear acquisition has been demonstrated in studies that used this paradigm during functional MRI scanning, using additional measures of arousal such as skin conductance and eye-blink responses (Sehlmeyer et al., 2009 and Mechias et al., 2010). Given that all participants in the current study judged the US to be aversive, and given that all were aware of the CS–US contingencies, there is no reason to assume that we induced something else than associative fear. However, it is unclear whether we measured associative fear with our trial-by-trial correlations, as our study—like most neuroimaging studies on fear conditioning—did not partial out the effects of heightened attention or mere anticipation of a US. In accordance with these studies (Sehlmeyer et al., 2009; Mechias et al., 2010), activation-based ROI analyses did reveal more activation for reinforced stimuli compared with unreinforced stimuli in the dACC and anterior insula. We did not show increased activation in response to the reinforced stimuli in the amygdala, which contrasts with some human studies (Morris and Dolan, 2004; but see Sehlmeyer et al., 2009; Mechias et al., 2010). In our study, the correlation between differential amygdala activity and the within-stimulus trial-to-trial correlations in the visual cortex corroborates theories about the mediating role of the amygdala in the enhancement of perception of emotional-significant stimuli (Vuilleumier and Driver, 2007; Pessoa and Adolphs, 2010). It also suggests a close link between activation of the traditional fear circuit and changes in the spatial representation of threat-associated stimuli. It is nevertheless possible that these within-stimulus correlations reflect the arousing aspect of the fear association and not the negative tone. They may also mirror different components of the fear-association depending on brain area; some areas may reflect the negative tone of an association, while other areas mainly reflect arousal or attention. As attention and emotion are inextricably linked (Pessoa, 2008; Bradley, 2009), they will often indirectly reflect the presence of an affectively significant association. Only varying the valence and arousing properties of a US could disentangle these different components.

Together, the results suggest that trial-by-trial MVPA provides a promising tool for examining how the human brain encodes relevant associations and forms semantic networks. This type of learning is essential for the survival of a species, but can become dysfunctional in the case of anxiety disorders. Experimental research on mechanisms of fear learning and modulation of fear memory may benefit from advanced neural pattern analyses. They offer the possibility to closely monitor how semantic associations evolve over time.

## References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a new and powerful approach to multiple testing. J Roy Stat Soc B 57:1289–1300.

Bradley MM (2009) Natural selective attention: orienting and emotion. Psychophysiology 46:1–11.

Büchel C, Morris J, Dolan RJ, Friston KJ (1998) Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron 20:947–957.

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd ed. Mahwah: Lawrence Erlbaum Associates.

Damaraju E, Huang YM, Barrett LF, Pessoa L (2009) Affective learning enhances activity and functional connectivity in early visual cortex. Neuropsychologia 47:2480–2487.

Eickhoff SB, Paus T, Caspers S, Grosbras MH, Evans AC, Zilles K, Amunts K (2007) Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. Neuroimage 36:511–521.

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P (2009) Decoding of emotional information in voice-sensitive cortices. Curr Biol 19:1028–1033.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.

Keil A, Stolarova M, Moratti S, Ray WJ (2007) Adaptation in human visual cortex as a mechanism for rapid discrimination of aversive stimuli. Neuroimage 36:472–479.

Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Sys Neurosci 2:1–28.

LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA (1998) Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron 20:937–945.

LeDoux JE (2000) Emotion circuits in the brain. Annu Rev Neurosci 23:155–184.

Li S, Mayhew SD, Kourtzi Z (2009) Learning shapes the representation of behavioral choice in the human brain. Neuron 62:441–452.

Li W, Howard JD, Parrish TB, Gottfried JA (2008) Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. Science 319:1842–1845.

Lim SL, Padmala S, Pessoa L (2009) Segregating the significant from the mundane on a moment-to-moment basis via direct and indirect amygdala contributions. Proc Natl Acad Sci U S A 106:16841–16846.

Mechias ML, Etkin A, Kalisch R (2010) A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. Neuroimage 49:1760–1768.

Morris JS, Dolan RJ (2004) Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. Neuroimage 22:372–380.

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430.

Padmala S, Pessoa L (2008) Affective learning enhances visual detection and responses in primary visual cortex. J Neurosci 28:6202–6210.

Pavlov IP, ed (1927) Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford: Oxford UP.

Pessoa L (2008) On the relationship between emotion and cognition. Nat Rev Neurosci 9:148–158.

Pessoa L, Adolphs R (2010) Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. Nat Rev Neurosci 11:773–783.

Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. Science 310:1963–1966.

Sabatinelli D, Lang PJ, Bradley MM, Costa VD, Keil A (2009) The timing of emotional discrimination in human amygdala and ventral visual cortex. J Neurosci 29:14864–14868.

Seger CA, Miller EK (2010) Category learning in the brain. Annu Rev Neurosci 33:203–219.

Sehlmeyer C, Schöning S, Zwitserlood P, Pfleiderer B, Kircher T, Arolt V, Konrad C (2009) Human fear conditioning and extinction in neuroimaging: a systematic review. PLoS One 4:e5865.

Tabachnick B, Fidell L (2007) Using multivariate statistics. Boston: Pearson Education.

Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C (2009) The NimStim set of facial expressions: judgements from untrained research participants. Psychiatry Res 168:242–249.

van den Hout MA, De Jong P, Kindt M (2000) Masked fear words produce increased SCRs: an anomaly for Öhman's theory of pre-attentive processing in anxiety. Psychophysiology 37:283–288.

Vuilleumier P, Driver J (2007) Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. Philos Trans R Soc Lond B Biol Sci 362:837–855.

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370–1386.

Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA (2010) Greater neural pattern similarity across repetitions is associated with better memory. Science 330:97–101.

Zhang J, Meeson A, Welchman AE, Kourtzi Z (2010) Learning alters the tuning of functional magnetic resonance imaging patterns for visual forms. J Neurosci 30:14127–14133.