



UvA-DARE (Digital Academic Repository)

The neural dynamics of fear memory

Visser, R.M.

Publication date

2016

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Visser, R. M. (2016). *The neural dynamics of fear memory*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Chapter 5

Optimizing study designs for single-trial multi-voxel pattern analysis: a systematic comparison

Renée M. Visser

Michelle I. C. de Haan

Tinka Beemsterboer

Pia Haver

Merel Kindt

H. Steven Scholte

Abstract

Single-trial analysis is particularly useful for assessing cognitive processes that are intrinsically dynamic, such as learning. Studying these processes with functional magnetic resonance imaging (fMRI) is problematic, as the low signal-to-noise ratio of fMRI requires the averaging over multiple trials, obscuring trial-by-trial changes in neural activation. The superior sensitivity of multi-voxel pattern analysis (MVPA) over univariate analyses has opened up new possibilities for single-trial analysis, but until now few studies have focused on optimizing study designs for this type of analysis. Here, we present two experiments in which we systematically compare study designs to determine empirically which of these designs are optimal for single-trial pattern analysis. In Experiment 1 we employed slow event-related fMRI combined with classical fear conditioning to assess associative learning in a trial-by-trial manner. This experiment consisted of four between-subject conditions ($n = 50$), which varied in order and spacing of trials (8.1-18.5 s). In Experiment 2 we examined the discriminability of stimulus categories, using rapid event-related fMRI in six within-subject conditions ($n = 18$), which varied in number and spacing of trials (2-6 s, with and without null-events, equal scan durations). Representational similarity analysis (RSA) on data from Experiment 1 revealed clear learning curves in all conditions, but showed the strongest effects when trial order was counterbalanced, such that temporal autocorrelations affected the comparisons of interest to a similar degree. Furthermore, support vector machine classification on data from Experiment 1 and 2 showed that classification of stimulus category was above chance in all conditions and comparable for different pattern estimation techniques (Least Squares Single [LSS] and Least Squares All [LSA]). Yet, our data indicate that - given a fixed amount of time - longer intervals are preferred over more stimulus repetitions for single-trial pattern analysis, while at the same time confirming that these designs are inefficient for univariate analyses. In sum, the current findings emphasize the importance of deciding on the type of data analysis before carrying out an experiment.

Introduction

Over the last two decades much effort has been devoted to optimizing study designs for functional magnetic resonance imaging (fMRI). The signal-to-noise ratio (SNR) of fMRI is low, given that events often do not evoke more than a 1% change in the blood-oxygenation-dependent (BOLD) signal (Huettel, Song, & McCarthy, 2004). A common method for improving SNR in event-related fMRI studies is to collect multiple observations for each experimental condition and to combine these as a single regressor in a general linear model (GLM). This method reduces noise and allows for a better estimation of the amplitude of the hemodynamic response. Rapid event-related designs (intervals of less than 10 s) generally produce the strongest effects, as more trials can be presented in the same amount of time, increasing the total variance in the BOLD signal and thereby the experimental power (Dale & Buckner, 1997; Huettel & McCarthy, 2001).

However, many psychological constructs are intrinsically dynamic. The first time a picture is presented is not equivalent to the second time it is presented, as the picture may have become familiar. As a result, at least part of the brain will respond differently. This *change* may in some cases be of specific interest, as for example in learning paradigms. In other cases, one is interested in the subtle differences and similarities between many different stimuli (stimulus rich design). In both scenarios, averaging across trials or stimulus categories would obscure the type of information that is supposed to be extracted from the data and would therefore defeat the purpose of the study. The information of interest is simply not available in the average responses and can only be obtained with single-trial analyses (Chadwick et al., 2012; Rey, Ahmadi, & Quiroga, 2015). However, single-trial fMRI analyses are quite challenging due to the low SNR and the sluggishness of the hemodynamic response.

The advent of multi-voxel pattern analysis (MVPA) opened up new avenues for single-trial analysis of BOLD-MRI data. Instead of average signal change, MVPA assesses distributed (multi-voxel) patterns of BOLD-signal to characterize the distinctive neural representation of a stimulus or condition. Although single-trial analysis suffers from a reduction in SNR due to a decrease in variance of the explanatory variable (temporal domain), single-trial MVPA compensates for this by evaluating multiple voxels at the same time (spatial domain). Over the last decade, numerous studies have underscored the superior sensitivity of MVPA compared to analysis of average activation for reading cognitive states from BOLD-MRI data (Haxby et al., 2001; Haynes & Rees, 2005; Kamitani & Tong, 2005) and quantifying the relationships between patterns induced by different states or stimuli (Kriegeskorte et al., 2008). For most applications of MVPA, trials are modeled as single regressors in a general linear model (GLM), instead of combined into one regressor per condition. The response

patterns related to the different events are then used either for (binary) classification analysis, or (continuous) similarity analysis. In classification analyses some of the response patterns are used to train a classifier (e.g., a support vector machine, SVM) and other patterns are used to test the classifier. In representational similarity analysis (Kriegeskorte et al., 2008) similarity values (e.g., Pearson's r) are calculated between different response patterns, resulting in matrices that display the representational (dis)similarity between stimuli or trials. In a way, these analytical techniques are not strictly single-trial: with classification analysis you need multiple trials within a condition to train and test a classifier, and with RSA you always need *pairs* of trials to calculate similarity indices. Yet, both types of MVPA may require the estimation of single-trial response patterns.

Despite the growing popularity of MVPA, relatively little is known about how to optimize study designs for this particular type of analysis. Rapid event-related designs with many trial repetitions and jittered periods of prolonged stimulus intervals (null events) are clearly the most efficient designs for estimating univariate signal changes. However, these designs pose problems for the estimation of single-trial activation patterns, since overlapping BOLD signals cannot be decorrelated unless multiple trials are combined into a single regressor (Mumford, Davis, & Poldrack, 2014; Mumford, Turner, Ashby, & Poldrack, 2012). Furthermore, temporally autocorrelated noise introduces false positive correlations between the activation patterns of trials that are close in time. When multiple trials of the same stimulus are presented in a row (e.g., mini blocks, or other types of structured stimulus presentation), which is often done in univariate designs to optimize detection power (Liu, Frank, Wong, & Buxton, 2001), both collinearity in the model and temporal autocorrelations in the data can lead to inflated classification accuracies or, in the case of RSA techniques, they may differentially affect the correlations of interest (Mumford et al., 2012). As mentioned above, the response patterns used in MVPA are usually obtained by modeling trials as separate regressors using a single GLM, that is, a Least Squares All (LSA) approach. An alternative approach is to use a separate GLM for each trial, in which the trial is modeled as the regressor of interest and all other trials are combined into a single nuisance regressor per condition (Least Squares Single, LSS; (Mumford et al., 2012). Although this latter technique has recently been shown to provide somewhat better parameter estimations in rapid-event related designs (Mumford et al., 2012), it does not completely solve the problems caused by collinearity (Mumford et al., 2014). Alternatively, between-run analysis alleviates this problem as the BOLD-signal that is used for training is never mixed-up with the BOLD-signal that is used for testing, reducing biases in classification analysis (Mumford et al., 2014) and promoting a classifiers' ability to generalize across exemplars (Coutanche & Thompson-Schill, 2012). However, while between-run analyses are clearly

beneficial in most cases, we can only assess trial-by-trial changes in activation patterns - the measure of interest in learning paradigms - by employing within-run analyses (Visser, Kunze, Westhoff, Scholte, & Kindt, 2015; Visser et al., 2013, 2011).

In this paper we address several questions. First of all, is reliable classification and RSA possible with single-trial data obtained from a rapid event-related fMRI design? Does it matter how the data are modeled (LSA versus LSS)? Second, given a fixed amount of time, is it better to increase trial spacing (e.g., more independence of hemodynamic responses) or is it better to have more trials (e.g., more observations)? Related to this: what is the effect of jittering on the estimation of single-trial response patterns? Third, what are optimal design choices for assessing trial-by-trial changes in activation patterns (i.e., learning-paradigms)? Is it important to keep temporal noise constant, by presenting the trials in a counterbalanced order?

Here, we present two experiments in which we systematically compare study designs and statistical approaches (LSA versus LSS). The first experiment focuses on learning-paradigms, while the second experiment focuses on the tradeoff between trial spacing and number of stimuli that can be presented in a fixed amount of time. The aim of these experiments is to empirically test the effects of different design choices on the estimation of single-trial response patterns.

Experiment I

Experiment I consisted of one session of fMRI scanning during which we used discriminant conditioning to assess associative learning in a trial-by-trial manner. In four (between-subject) conditions we compared the effects of trial spacing and trial order on neural pattern similarity. We did not use the generally more powerful within-subject comparisons (i.e., where individuals participate in all conditions) because the learning experience with electrical stimulation might influence subsequent learning within the same experimental context.

Experiment I: Methods

Participants

Fifty-five participants were recruited by advertisements in the social media and the university website. For the analyses of BOLD-MRI data, participants were excluded because of sleep ($n = 1$), excessive head motion ($n = 3$) or substantial signal drop-out ($n = 1$). In the remaining sample of 50 participants, 2 participants lacked eye-tracker data. Hence, BOLD-MRI data are reported for 50 participants (11 male, 3 left-handed, mean = 22.4, \pm 2.8 s.d. yrs. of age), and pupil data are reported for 48 participants. Participants earned a small amount of money or partial course credit for their

participation. All participants gave their written informed consent before participation and had normal or corrected-to-normal vision. Procedures were executed in compliance with relevant laws and institutional guidelines, and were approved by the University of Amsterdam's ethics committee (2012-CP-2638).

Apparatus and materials

Stimuli and conditioning procedure. A classical fear-conditioning paradigm was used, with delay conditioning and partial reinforcement (Figure 1a). Two pictures of neutral faces, derived from the Todorov database (Oosterhof & Todorov, 2008), which was generated with FaceGen Modeller 3.1, and two pictures of houses (Visser et al., 2013) were converted to grey scale and presented on a grey background. Each picture was presented 11 times for 4.5 seconds and served as a to-be conditioned stimulus (CS). One face and one house were followed by a mild electrical stimulus in 5 out of 11 presentations (CSs+). The electrical stimulation served as an unconditioned stimulus (UCS) and was delivered at CS+ offset for 2 ms, by a Digitimer DS7A through MRI-compatible carbon electrodes attached to the right shinbone. The intensity of the electric stimulus was individually adapted at a level that was aversive but not painful (mean = 30.89 mA \pm 14.97 s.d.). The other two stimuli were never reinforced (CSs-). Participants were told that two out of four stimuli might be followed by electrical stimulation while the other two would never be reinforced, and were instructed to learn the specific contingencies. Stimuli were backward-projected onto a screen that was viewed through a mirror attached to the head-coil.

pupil diameter was the average value during the 500 ms prior to each CS onset. The pupil response to the CS was calculated as the peak change from baseline in a window from 0 to 4 seconds after picture onset. Trials that suffered substantial signal loss, affecting more than 50% of either the baseline samples or the 4 seconds after stimulus onset were eliminated and replaced entirely by estimating the linear trend at point over trials for each condition separately. Participants that ended up missing more than 25% of the trials were excluded ($n = 2$), leaving 48 participants for the analysis of pupil responses (0 - 20.5 % replaced trials per participant, median = 0%). Next, data were Z-transformed, to reduce between-subjects variability.

Image acquisition. Scanning was performed on a 3T Philips Achieva TX MRI scanner using a 32-channel head-coil. Functional data were acquired using a gradient-echo, echo-planar pulse sequence (TR = 2000 ms; TE = 27.63 ms; FA = 76.1°; 39 sagittal slices with interleaved acquisition; 3 x 3 x 3.3 mm voxel size; 64 x 64 matrix; 192 x 192 x 141.24 FoV) covering the whole brain. In condition I and II 500 volumes were recorded; in condition III and IV 304 volumes were recorded. Foam pads minimized head motion. A high-resolution 3D T1-weighted image (TR = 8.11 ms, TE = 3.72 ms, FA = 8°; 1 x 1 x 1 mm voxel size; 240 x 220 x 188 FoV) was additionally collected for anatomical visualization.

Pre-processing. fMRI data processing was carried out using FEAT (fMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). Pre-processing included motion correction using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002); slice-timing correction using Fourier-space time-series phase-shifting; non-brain removal using BET (Smith, 2002); high-pass temporal filtering ($\sigma = 50$ s), and pre-whitening (Woolrich et al., 2001). No spatial smoothing was applied. Registration to high resolution structural images was carried out using FLIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001). Registration from high resolution structural to standard space was then further refined using FNIRT nonlinear registration (Andersson, Jenkinson, Smith, & others, 2007).

Region of interest selection. Regions of interest (ROI) were selected based on their role in fear learning and (extinction) memory and included the anterior cingulate cortex (ACC, 9213 voxels), the insula (6591 voxels), amygdala (2967 voxels), hippocampus (5837 voxels) and ventromedial prefrontal cortex (vmPFC, 4160 voxels). We additionally included the superior frontal gyrus (SFG, 18946 voxels), a region outside the salience network, for its large learning-effect as revealed by

previous similarity analysis, in the absence of differences in average activation (Visser et al., 2015, 2013, 2011). ROIs were obtained from the Harvard-Oxford cortical and subcortical structural atlases (Harvard Center for Morphometric Analysis).

Trial spacing and order of presentation

Participants were randomly assigned to one of four conditions (Figure 1a). In condition I, inter-stimulus intervals were fixed (17.5 s) and the onset of each trial was triggered by the start of the acquisition of a BOLD-MRI volume. In condition II, inter-stimulus intervals varied randomly between 16.5 and 18.5 seconds. In condition III and condition IV, inter-stimulus intervals were drawn from a truncated exponential distribution (8.1-12.5 seconds, mean 9.2 seconds). In condition I, II and III, 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 (Figure 1b). In total, the experiment consisted of 44 trials: 24 target trials (6 per stimulus type) and 20 filler trials (5 per stimulus type), including all CS+ -trials that co-terminated with electrical stimulation. In condition IV, the order of stimulus presentation was semi-random, with the restriction that stimuli were roughly equally distributed across the experiment and that (like the other three conditions) the experiment started with two unreinforced presentations of each stimulus, to estimate a pre-conditioning baseline response to the pictures. For each run, we constrained our analyses to target trials, to be certain that UCS-related activity would not confound CS-related activity (all four conditions) and b), that the time between two consecutive target trials was equal over the four CS types (condition I, II and III), while filler trials ensured that the stimulus presentation remained unpredictable for the participant (Visser et al., 2015, 2013, 2011). The relatively high temporal resolution of the pupil dilation response allowed for the analysis of reinforced trials, so for the analysis of pupil data both filler and target trials are included.

Trial-by-trial similarity analysis

For the trial-by-trial representational similarity analysis we employed a Least Squares – All (LSA) approach, modeling each trial as a separate regressor in a voxelwise whole-brain analysis using a single general linear model (GLM) and including six motion parameters as regressors of no interest. The resulting single-trial parameter estimates were transformed into *t*-values to down-weight noisy voxels (Misaki, Kim, Bandettini, & Kriegeskorte, 2010). To this end, each voxel's parameter estimate was divided by the standard error of that voxel's residual error term after fitting the first-level GLM. In Matlab (version 8.0; MathWorks) we created for each participant, for each ROI, a vector

containing t-values per voxel, for a particular trial (i.e., the 'spatial representation' of that trial). Next we calculated pair-wise Pearson correlations (i.e., 'representational similarity') between all vectors of all single trials, resulting in a similarity matrix containing correlations among trials, for each participant, for each ROI (Figure 2, left panels). From this matrix, two different types of correlations were selected (Figure 1c, Figure 2, right panels), discarding filler trials. The strength of these correlations was used as a metric of similarity. First, we examined *within-stimulus* correlations on consecutive target trials. Second, we examined *between-stimulus* correlations between adjacent target trials that shared (non)reinforcement (learned associations: CS+ face with CS+ house and CS- face with CS- house). Note that the number of between-stimulus 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. Next, data were Z-transformed, to reduce between-subjects variability.

Trial-by-trial univariate analysis

To visualize trial-by-trial changes in average activation, we analyzed data as described in the previous section, except that when we analyzed the normalized single-trial parameter estimates, we averaged across voxels in a ROI. Thus, instead of preserving the spatial information by creating a vector of voxels per ROI, we obtained one value per ROI (average response amplitude), which we then Z-transformed across trials to reduce between-subject variability.

Statistical analyses

Z-transformed pupil dilation responses, average activation, and within-stimulus correlations were averaged over face- and house stimuli. This was done to reduce the number of comparisons and because we were not interested in the difference between faces and houses with regard to the experimental manipulation.

Statistical comparisons of the learned associations were performed by within-subjects Analysis of Variance (ANOVA), using Statistical Package for the Social Sciences (SPSS, version 21; SPSS Inc.). Statistical tests are equivalent for pupil dilation and neural measures, the only difference being the number of trials that is included in the analysis (i.e., 11 for pupil dilation, 5 for within-stimulus similarity, 6 for between-stimulus similarity and average activation). Differential fear learning was assessed by the interaction of trials x stimulus type (2 levels [CS+ and CS-, averaged over faces and houses]), but was only tested when there was also a significant main effect of stimulus type. Likewise, the reported average effect sizes represent the average over the tested main effects, and if significant also over the tested interaction effects (see also Supplementary Table 1-4). Note that the

sample sizes are too small to directly compare the learning effects (which include numerous parameters, i.e., different trials and stimuli) between groups, but the estimated effect sizes within each scenario may guide choices for study designs.

In case that the assumption of sphericity was violated a Greenhouse-Geisser correction was applied. All p -values are reported two-sided, with the significance level set at $\alpha = 0.05$.

Experiment I: Results

Conditioned pupil dilation response

Consistent with previous work (Visser et al., 2015, 2013), successful fear conditioning was evident from a trial-by-trial change in pupil dilation in response to the CS+, relative to the CS- in all four conditions (Figure 3a). Interaction effects of trial (11) and stimulus type (2) reached significance in condition II ($F_{3.95, 39.52} = 4.59$, $p = 0.004$, $\eta_p^2 = 0.31$), condition III ($F_{10, 110} = 2.61$, $p = 0.007$, $\eta_p^2 = 0.19$) and condition IV ($F_{10, 130} = 2.77$, $p = 0.004$, $\eta_p^2 = 0.18$), but not in condition I ($F_{4.16, 41.58} = 1.45$, $p = 0.232$, $\eta_p^2 = 0.13$). Follow-up tests revealed strong main effects of stimulus type in condition I ($F_{1, 10} = 15.04$, $p = 0.003$, $\eta_p^2 = 0.60$) in condition II ($F_{1, 10} = 28.06$, $p < 0.0005$, $\eta_p^2 = 0.74$), condition III ($F_{1, 11} = 13.84$, $p = 0.003$, $\eta_p^2 = 0.56$) and condition IV ($F_{1, 13} = 85.15$, $p < 0.0005$, $\eta_p^2 = 0.87$), indicating that pupil size increased when a CS+ was presented.

Neural pattern similarity

The left panels in Figure 2 present similarity matrices in the superior frontal gyrus, showing all trials (filler and target trials) in order of presentation. In condition III and IV, which have shorter inter-trial intervals, high correlations were observed between adjacent trials. Still, when controlling stimulus presentation such that the time between consecutive target trials was equal over conditions (condition I, II, and III), higher pattern similarity was observed within and between CS+ stimuli compared to CS- stimuli (Figure 2, right panels). This differential pattern similarity was weaker when the trial presentation was random (condition IV).

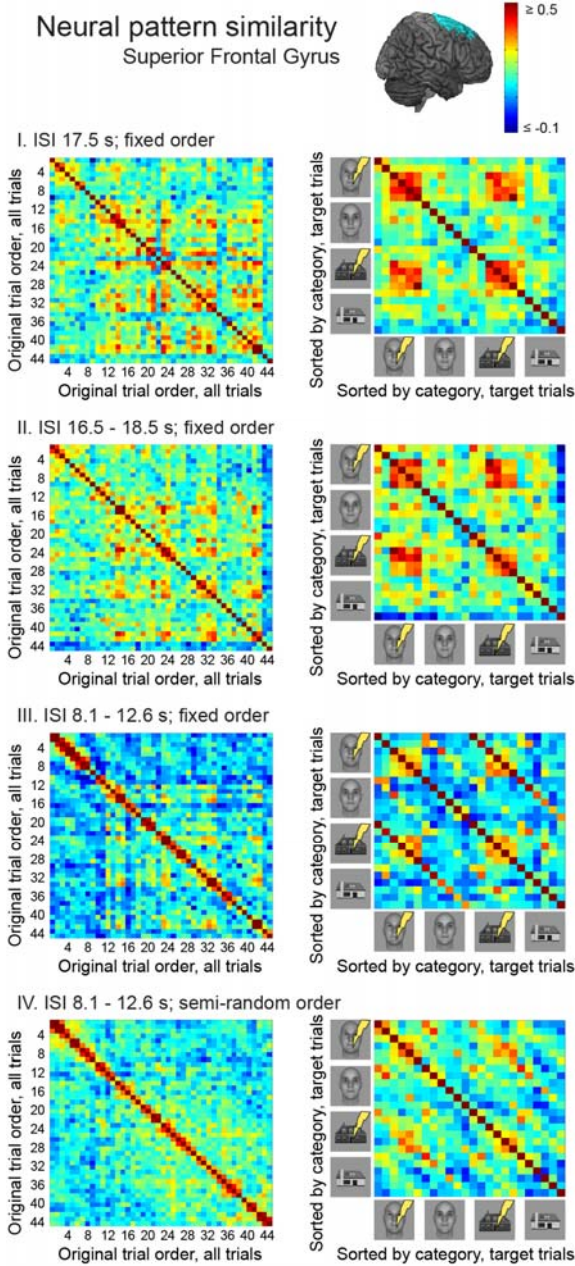


Figure 2 Neural pattern similarity in the superior frontal gyrus, showing all trials (filler and target trials) in order of presentation (left panels), averaged over participants. In condition III and IV, which have shorter inter-trial intervals, high correlations are observed between adjacent trials, suggesting stronger temporal autocorrelations. Still, when controlling stimulus presentation such that the time between consecutive target trials is equal over conditions (condition I, II, and III), we clearly observe higher pattern similarity within and between CS+ stimuli compared to CS- stimuli (right panels). This differential pattern similarity is weaker when the trial presentation is random (condition IV). ISI = inter-stimulus interval.

A trial-by-trial assessment of within-stimulus pattern similarity (Figure 3b) revealed successful learning as evidence by an increase in similarity for CS+ stimuli, relative to CS- stimuli. This is in line with our previous work (Visser et al., 2015, 2013, 2011), although effects did not reach statistical

significance in all areas and all conditions (Supplementary Table I-4). For between-stimulus pattern similarity, learning curves were observed when trial presentation was controlled (condition I, II, and III), but not when presentation was semi-random (condition IV). In general, main and interaction effects were stronger in ACC (mean $\eta_p^2 = 0.34$), insula (mean $\eta_p^2 = 0.28$) and SFG (mean $\eta_p^2 = 0.35$) than in amygdala (mean $\eta_p^2 = 0.16$), hippocampus (mean $\eta_p^2 = 0.17$) and vmPFC (mean $\eta_p^2 = 0.20$). Furthermore, effects were stronger in condition I (mean $\eta_p^2 = 0.29$) and II (mean $\eta_p^2 = 0.30$) than in condition III (mean $\eta_p^2 = 0.25$) and IV (mean $\eta_p^2 = 0.19$).

Trial-by-trial mean activation

Results obtained with single-trial univariate analysis dissociated from results obtained with similarity analyses in some, but not all of the regions. Typical learning curves were observed in areas in the ‘salience-network’ (ACC and insula), but were absent in hippocampus, amygdala, vmPFC and SFG (Figure 3c). For an overview of the statistics per ROI see Supplementary Table I-4. The fact that effect sizes were substantially smaller for average activation (mean $\eta_p^2 = 0.18$) than for pattern similarity is consistent with previous results (Visser et al., 2015, 2013, 2011). This again shows the high sensitivity of pattern analysis compared to analysis of average activation for quantifying the changes in (fear) associations over time.

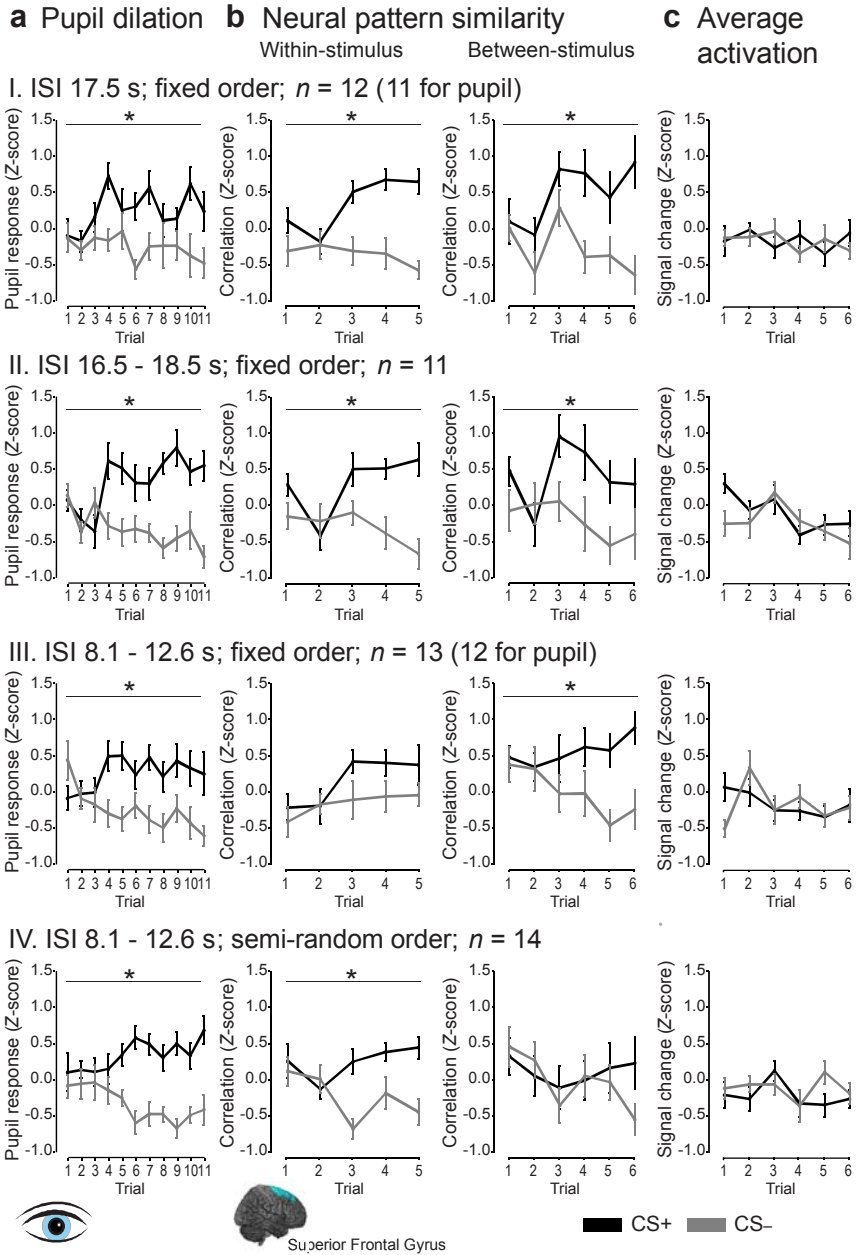


Figure 3 a) Pupil dilation responses and b) Neural pattern similarity in the superior frontal gyrus - both within-stimulus and between-stimulus - show clear acquisition of fear over the course of conditioning in each of the four conditions, with the exception of between-stimulus pattern similarity in condition IV. c) For average activation in the superior frontal gyrus no learning-dependent changes are observed. ISI = inter-stimulus interval. Error bars represent SEM. * $p < 0.05$.

Experiment 1: Discussion

Consistent with previous work (Visser et al., 2015, 2013, 2011), Experiment 1 shows that the application of trial-by-trial similarity analysis produces clear learning curves that index the formation of aversive associations, even in the absence of differences in mean activation. Furthermore, although longer inter-trial intervals yielded stronger effects, we were able to detect learning-dependent changes in designs with shorter intervals as well, provided that the order in which stimuli were presented was fixed. Given that longer scans are costly, boring, and increase the risk of head motion, inter-trial intervals of medium length (8-12 seconds) may be preferred, at the cost of signal strength, as long as the experiment is designed in such a way that pattern similarity will not be biased by temporal proximity of the trials of interest. However, while in learning paradigms the number of trials is usually based on theoretical considerations, in other paradigms the number of trials may solely depend on how much information is needed to obtain a reliable signal (see Experiment 2).

A limitation might be that we have not tested whether learning effects would be observed in a true rapid event-related design (an interval of 8.1-12.6 is still quite long). Shorter intervals are problematic in fear-conditioning paradigms (and therefore rarely used; Fullana et al., 2015) for several other reasons. For example, the sensation of the electrical stimulus always requires a few seconds to fade. This poses problems if many trials were to be presented in a short amount of time, causing unwanted effects such as backward conditioning (Moscovitch & LoLordo, 1968) and possibly additional discomfort as the uncomfortable stimulation summates. Furthermore, the number of conditioning trials is usually limited, given that learning in these paradigms often reaches an asymptote within a few trials, and/ or peripheral indices of sympathetic activity tend to habituate.

Experiment 2: Methods

Experiment 2 consisted of one session of fMRI scanning, in which we compared six designs to measure BOLD-MRI in response to face and house stimuli. As the experiment did not induce aversive learning, it was possible to compare the effects of interest in a within-subject design.

Participants

Twenty-one participants were recruited by means of advertisements in the social media and the university website. Three participants were excluded because of excessive head motion ($n = 1$),

because of equipment failure ($n = 1$), and because of excessive sleepiness ($n = 1$). The remaining sample included 18 participants (3 male, 3 left-handed, mean = 25.0, \pm 2.2 s.d. yrs. of age). Participants earned a small amount of money or partial course credit for their participation. All participants gave their written informed consent before participation and had normal or corrected-to-normal vision. Procedures were executed in compliance with relevant laws and institutional guidelines, and were approved by the University of Amsterdam’s ethics committee (2012-CP-2638).

Apparatus and materials

Stimuli. The experiment consisted of six blocks of functional scanning. In each block we used a new stimulus set (Figure 4), consisting of two pictures of neutral faces, derived from the Karolinska Directed Emotional Faces dataset (Lundqvist et al., 1998), and two pictures of houses, derived from the Web. Stimuli were separated from their background and converted to greyscale. Each picture was presented for 300 ms, with stimulus intervals varying per block (see section on trial spacing). Participants were instructed to pay close attention to the stimuli. During stimulus intervals a fixation cross turned either green or blue: the participant pressed a button with their right middle finger for blue and with their left middle finger for green.







Block	A	B	C	D	E	F
Stimuli						
Trial duration	2 s	2 s	4 s	4 s	6 s	6 s
# trials per stimulus	48	34	24	17	16	11
# null events	0	56	0	28	0	12
Total scan duration	390 s	390 s	390 s	390 s	390 s	390 s

Figure 4 Within-subject design consisting of six conditions (blocks), varying in trial spacing and use of null events. In each block we used a new stimulus set, consisting of two pictures of neutral faces and two pictures of houses. The order of blocks and the order of stimulus-sets were independently counterbalanced across participants. Each picture was presented for 300 ms, with trial durations varying per block. In each block, the number of trials and null events together fill 390 seconds of functional scanning. Images are not to scale.

Image acquisition. Scanning was performed on a 3T Philips Achieva TX MRI scanner using a 32-channel head-coil. Functional data were acquired using a gradient-echo, echo-planar pulse sequence (TR = 2000 ms; TE = 27.63 ms; FA = 76.1°; 37 axial slices with ascending acquisition; 3 x 3 x 3 mm voxel size; 80 x 80 matrix; 240 x 121.8 x 240 FoV) covering the whole brain. Each of the six conditions consisted of 195 volumes. Foam pads minimized head motion. A high-resolution 3D T1-weighted image (TR = 8.16 ms, TE = 3.73 ms, FA = 8°; 1 x 1 x 1 mm voxel size; 240 x 220 x 188 FoV) was additionally collected for anatomical visualization.

Pre-processing. Preprocessing was performed in the same way as in Experiment 1.

Region of interest selection. Since our aim was to assess how well we could classify faces and houses in different study designs, we selected the temporal occipital fusiform cortex (TOFC; 7485 voxels), a region that is known to be involved in the processing of these stimuli (e.g., Epstein & Kanwisher, 1998; Haxby et al., 2001). The ROI was obtained from the Harvard-Oxford cortical atlases (Harvard Center for Morphometric Analysis). No feature selection was performed. To identify temporal correlations in the data that were not related to hemodynamic responses a control region was examined (ventricles, 556 voxels).

Trial spacing and order of presentation

Each participant underwent six 6.5 min blocks of fMRI scanning, across which the number of trials and the length of the intervals were varied (Figure 4). The order of blocks and the order of stimulus-sets were independently counterbalanced across participants. Trial duration was 2 seconds in block A and B, 4 seconds in block C and D and 6 seconds in block E and F. The onset of each trial was triggered by the start of the acquisition of a BOLD-MRI volume. While in block A, C, and E no null events were used, block B and D contained 29.2% null events and block contained 18.8 % null events. The inclusion of null events is a convenient means of achieving a stochastic distribution of stimulus intervals, typically used in fMRI to decorrelate the different regressors. The question here is whether such a stochastic distribution is required and/ or beneficial for the estimation of single-trial parameters.

The total number of stimulus presentations was 192 in block A (48 per stimulus), 136 in block B (34 per stimulus), 96 in block C (24 per stimulus), 68 in block D (17 per stimulus), 64 in block E (16 per stimulus) and 44 in block F (11 per stimulus). To the extent that this was possible (given our demands), we used a genetic algorithm (Kao et al., 2009) to create for each block three

optimized sequences of stimulus presentation, and we counterbalanced these sequences across participants. Note that this algorithm is not designed for single-trial analysis. We are currently not aware of any algorithm that is designed for optimization of single-trial analysis.

Representational similarity analysis

To get an impression of the effect of study design on the correlation structure of the BOLD-MRI data, we examined pattern similarity in the TOFC. To this end, each trial was modelled separately in a GLM. For the LSA analysis, *t*-values per voxel were obtained in the way that is described in Experiment I (section on trial-by-trial similarity), i.e., using a single GLM containing all trials as separate regressors. For the LSS (Least Squares — Single) analysis, a separate GLM was run for each trial where the trial was modeled as the regressor of interest and all other trials were combined into a single nuisance regressor. Note that the LSS analysis is explicitly developed to cope with problems occurring when combining short intervals with classification analysis, not necessarily similarity analysis (Mumford et al., 2014, 2012), but that similarity analysis is a useful way to visualize the difference between the two estimation techniques. The resulting parameter estimates were transformed into *t*-values in the same way as for the LSA analysis. In Matlab (version 8.0; MathWorks) we created for each participant, a vector containing *t*-values per voxel, for a particular trial (i.e., the ‘spatial representation’ of that trial). Next, we calculated pair-wise Pearson correlations (i.e., ‘representational similarity’) between all vectors of all single trials, resulting in a similarity matrix containing correlations among trials, for each participant. Finally, data were averaged over stimulus category (see Figure 5 & 6), excluding the diagonal (i.e., autocorrelations).

Data from Experiment I were reanalyzed using pattern similarity in the TOFC, to assess whether we could reliably distinguish faces from houses, independent from any learning-dependent effects. To enable a fair comparison between the different study designs we ‘matched’ scanning time, by limiting the classification analysis to a number of target trials that could have been presented in a 6.5 minute scan. For condition III and IV (with stimulus intervals varying between 8.1 and 12.6 s), this did not change the number of target trials per condition (i.e., 6), but for condition I and II (intervals 16.5-18.5 s) we restricted our analysis to 4 out of 6 target trials per condition. Note that we only performed a LSA analysis on these data, as LSS analysis is specifically recommended for rapid event-related fMRI (Mumford et al., 2012).

Classification analyses

The aim of the classification analysis was to assess how well each design allowed for the decoding of stimulus categories from the neural response patterns in the TOFC.

Single trial response patterns were obtained as described in the previous sections. Data were further analyzed in a leave-two-out classification analysis, with 2000 iterations using a one-class support vector machine with a linear kernel function (LIBSVM, Chang & Lin, 2011). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>). Within each iteration, one face trial and one house trial were separated as test dataset and the rest of the data were used as training data. Iterations were semi-random, such that the selection of test trials was equally distributed over trials and each trial was sampled at least 10, 14, 20, 29, 31 or 45 times for block A-F respectively. For the data from Experiment I, each trial was sampled at least 125 times in condition I and II and 83 times in condition III and IV.

Univariate analyses

To examine if guidelines for design optimization differed for MVPA and traditional fMRI analyses, we ran a standard voxel-wise univariate analysis, modeling all trials within a category (faces and houses) as one regressor and including filler trials (data from Experiment I), motion parameters and temporal derivatives as regressors of no interest. Next, we counted within the TOFC those voxels that differentiated between faces and houses, thresholded at $Z > 2.3$ and divided this by the total number of voxels within this region. This yielded a percentage of activated voxels per participant, per condition (i.e., extent of activation). Furthermore, we averaged across these voxels to obtain the average signal strength of the activated voxels (i.e., average response amplitude).

Statistical analyses

Statistical comparisons of neural pattern similarity were performed by within-subjects ANOVA (SPSS, version 21; SPSS Inc.), with spacing (6 blocks) as within-subject level, and the difference in pattern similarity (within-category minus between-category) as dependent variable, for the two approaches (LSA versus LSS) separately. Within-category similarity was averaged over faces and houses. For data from Experiment I a between-subjects ANOVA was conducted, with condition (I, II, III, IV) as between-subject factor and differences in pattern similarity (within-category minus between-category) as dependent variable. Follow-up paired t-tests were performed per condition to test whether within-category pattern similarity was higher than between-category similarity.

Statistical comparisons of classification performance were performed by within-subjects ANOVA, with spacing (6 blocks) as within-subject level and proportion of correctly classified trials as dependent variable, for the two approaches (LSA versus LSS) separately. For data from Experiment 1 a between-subjects ANOVA was conducted, with condition (I, II, III, IV) as between subject factor and proportion of correctly classified trials as dependent variable. To examine classification performance in each condition separately, one-sample *t*-tests were performed, comparing the proportion correctly classified trials in each condition to chance level (0.5). A within-subjects ANOVA with block and approach (LSA and LSS) as within-subjects level, as well as paired *t*-tests, were used to directly compare differential pattern similarity and classification performance obtained with the LSA and the LSS approach (data from Experiment 2 only). For the univariate analyses, the extent of activation, as well as the average response amplitude were compared between conditions (Experiment 1 and 2). All *p*-values are reported two-sided, with the significance level set at $\alpha = 0.05$.

Experiment2: Results

Pattern similarity

Figure 5 (LSA) and Figure 6 (LSS) present neural pattern similarity in the TOFC in the different blocks, calculated on data from Experiment 2. Left panels show the trials in order of presentation, middle panels show the trials sorted per stimulus category and right panels show pattern similarity averaged over trials within stimulus categories.

Not surprisingly, a regular LSA analysis was not possible in block A and in one of the three (for univariate analyses optimized) presentation orders in block B ($n = 6$), because the design matrix was rank deficient (i.e., the extremely short intervals increased the correlations between the predictors to an unacceptable height). In contrast, the LSS analysis did not have any problems with block A and B. Yet, ANOVAs that directly compare the two approaches are conducted on 5 blocks (B-F) and on data from 12 participants (listwise inclusion of cases). Especially the blocks without null events (A, C and E) suffered from temporal autocorrelations (i.e., higher correlations for trials close in time), although this problem appeared less severe in the LSS approach (Figure 6) than in the LSA approach (Figure 5). The problem was negligible in data from Experiment 1 (Figure 7), in which relatively long stimulus intervals were used.

Neural pattern similarity

Temporal Occipital Fusiform Cortex

Least Squares All (LSA)

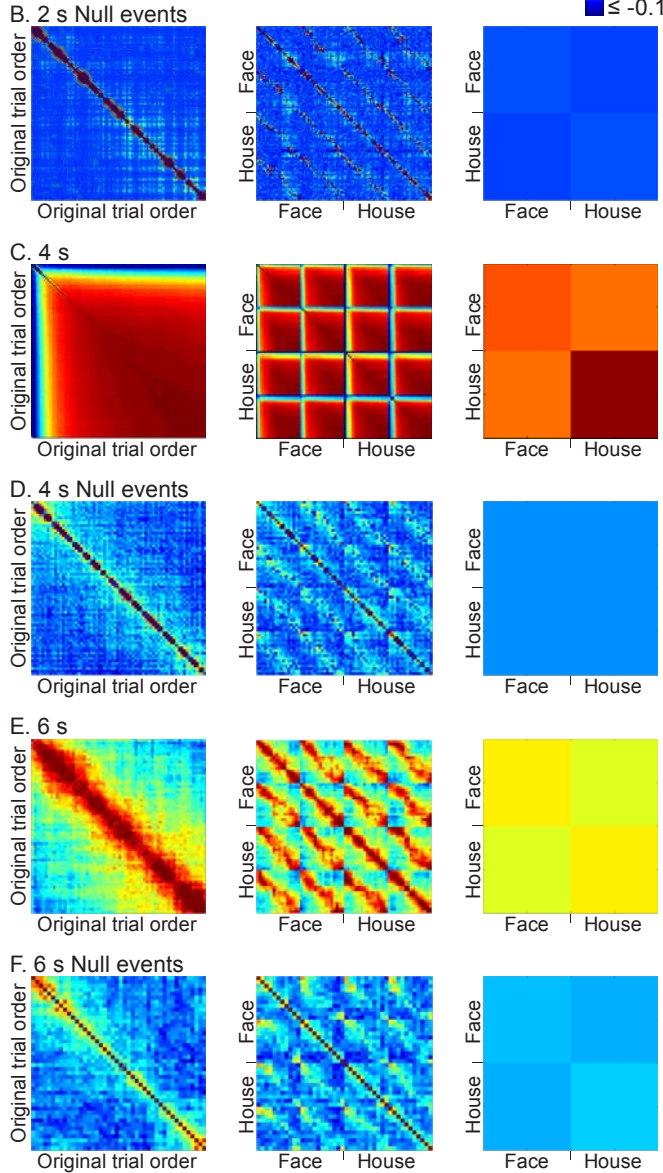


Figure 5 Average neural pattern similarity in the temporal occipital fusiform cortex, showing all trials (Experiment 2) in order of presentation (left), sorted by stimulus category (middle), and averaged over trials within each stimulus category (right). Values on which correlations are calculated are obtained using a Least Squares All approach. As the correlations in conditions C are all above 0.95 these matrices are plotted on a scale from 0.4-1.0. The difference between within-category similarity and between-category similarity is significant in all conditions and largest in block C and D, which are designs without null events.

Neural pattern similarity
Temporal Occipital Fusiform Cortex

Least Squares Single (LSS)

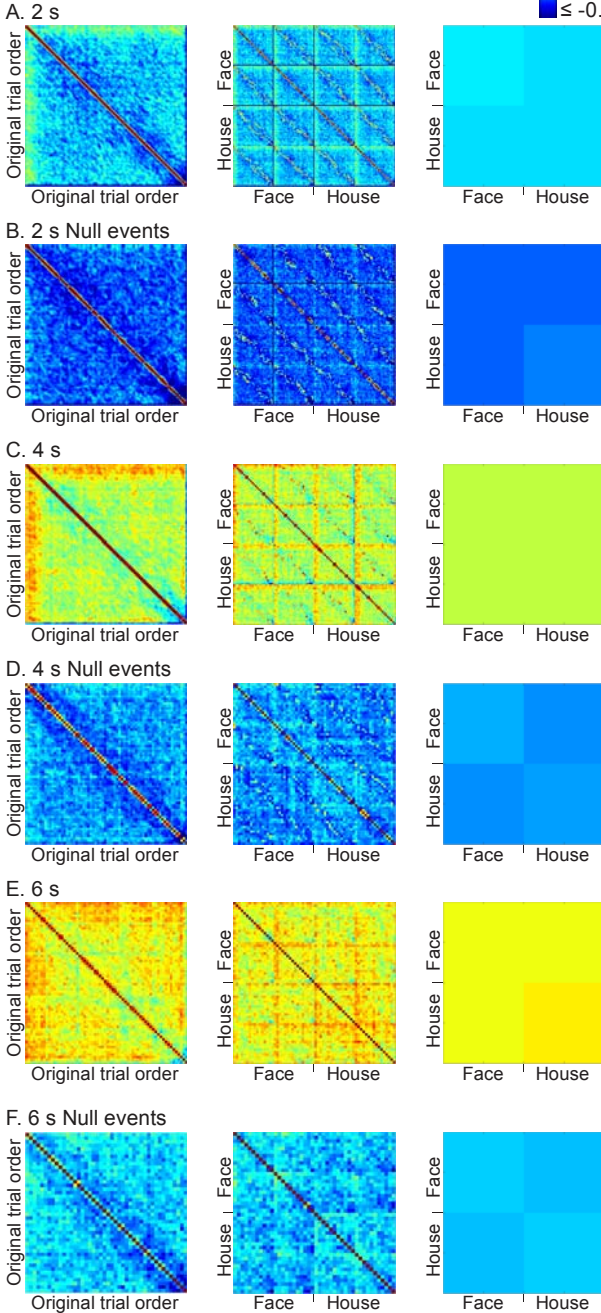
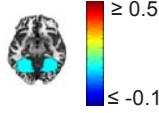
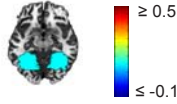


Figure 6 Average neural pattern similarity in the temporal occipital fusiform cortex, showing all trials (Experiment 2) in order of presentation (left), sorted by stimulus category (middle) and averaged over trials within each stimulus category (right). Values on which correlations are calculated are obtained using a Least Squares Single approach. The difference between within-category similarity and between-category similarity is significant in conditions A, B, D and F, and greatest in block B.

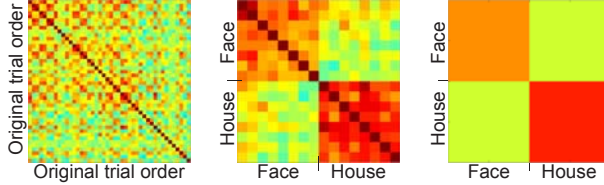
Neural pattern similarity

Temporal Occipital Fusiform Cortex

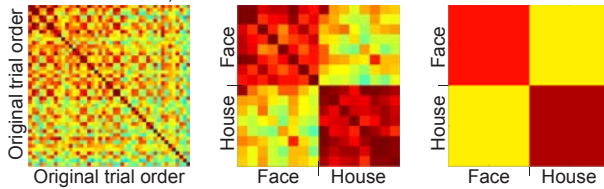
Least Squares All (LSA)



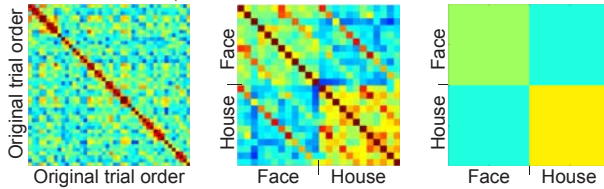
I. ISI 17.5 s; fixed order



II. ISI 16.5-18.5 s; fixed order



III. ISI 8.1 - 12.6 s; fixed order



IV. ISI 8.1 - 12.6 s; semi-random order

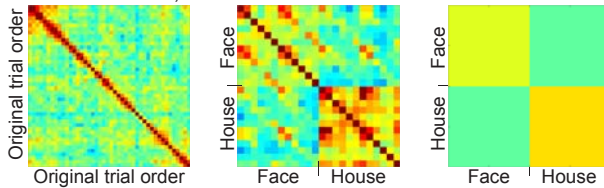


Figure 7 Average neural pattern similarity in the temporal occipital fusiform cortex, showing all trials (Experiment I) in order of presentation (left), sorted by stimulus category (middle), and averaged over trials within each stimulus category (right). Values on which correlations are calculated are obtained using a Least Squares All approach. The difference between within-category and between-category similarity is largest in condition II, and this difference is significantly larger than in condition III and IV. ISI = inter-stimulus interval.

What is also evident from looking at the matrices is that the difference between within-category similarity and between-category was most pronounced in designs with longer intervals. A crucial question is to what extent temporal autocorrelations impede the ability to distinguish between stimulus categories. To this end, we first compared average within-category similarity with average between-category similarity. There was an effect of block (5) in the LSA approach ($F_{4, 44} = 9.19$, $p = 0.001$, $\eta_p^2 = 0.46$), and an effect of block (6) in the LSS approach ($F_{5, 85} = 4.44$, $p = 0.001$, $\eta_p^2 = 0.21$), with the difference between within- and between-category similarity being largest in the designs without null-events (an overview of the statistics from Experiment 2 is presented in Supplementary

Table 5). Results differed for the two approaches, as indicated by a significant interaction between block and approach ($F_{4, 44} = 19.44$, $p < 0.0005$, $\eta_p^2 = 0.64$) and a significant main effect of approach ($F_{1, 11} = 44.50$, $p < 0.0005$, $\eta_p^2 = 0.80$). Follow-up t -tests showed that in designs without null events (block C and E) the LSA approach showed more differential pattern similarity (within- minus between-category) than the LSS approach ($ps < 0.0005$). In block B and D (designs with null events) the LSS approach yielded slightly better results than the LSA approach ($p = 0.080$ and $p = 0.052$ respectively).

In Experiment 1 (Supplementary Table 6), the difference between within- and between-category pattern similarity was significant in all conditions ($p < 0.0005$). Furthermore, there was a trend significant effect of condition ($F_{3, 46} = 2.78$, $p = 0.051$, $\eta_p^2 = 0.15$). Independent t -tests revealed more differential pattern similarity in condition II than condition III ($t_{22} = 2.10$, $p = 0.048$, $d = 0.85$) and condition IV ($t_{23} = 3.05$, $p = 0.006$, $d = 1.20$). None of the other post-hoc comparisons were significant ($ps > 0.205$).

Classification

Figure 8a presents the classification performance for the LSS and LSA approach for data from Experiment 2 (left panels) and for the LSA approach for data from Experiment 1 (right panels), sorted on ascending intervals. Average classification performance for the data from Experiment 2 was above chance in each of the six conditions (all $ps < 0.0005$, Cohen's d ranging from 1.10-2.65), regardless of the approach, with the exception of the blocks for which a LSA analysis was not possible (block A and part of block B). The interaction between block (5) and approach (2) was marginally significant ($F_{4, 44} = 2.42$, $p = 0.063$, $\eta_p^2 = 0.18$). For both the LSA and the LSS approach, there was an effect of block (5 blocks for LSA: $F_{4, 44} = 4.04$, $p = 0.007$, $\eta_p^2 = 0.27$, and 6 blocks for LSS: $F_{5, 85} = 11.35$, $p < 0.0005$, $\eta_p^2 = 0.40$, respectively), indicating that average performance differed as a function of trial-duration. Follow-up tests indicated that there was a significant increase from block B to C for the LSA approach ($t_{11} = 4.13$, $p = 0.002$, $d = 1.19$) and the LSS approach ($t_{17} = 2.21$, $p = 0.041$, $d = 0.52$), and for the LSA approach there was also a significant increase from block D to E ($t_{17} = 2.65$, $p = 0.017$, $d = 0.62$). Furthermore, follow-up tests revealed that the LSA approach outperformed the LSS approach in block C ($t_{17} = 2.27$, $p = 0.037$, $d = 0.53$) and E ($t_{17} = 2.67$, $p = 0.016$, $d = 0.63$), which is in line with the results from the similarity analyses. A complete overview of the statistics per block is presented in Supplementary Table 7.

Classification performance for the data from Experiment 1 was at ceiling in all four conditions ($ps < 0.0005$, Cohen's d ranging from 10.03-19.16). As a consequence, no effect of

condition was observed ($F_{3, 46} = 0.57$, $p = 0.635$, $\eta_p^2 = 0.04$). A complete overview of the statistics per condition can be found in Supplementary Table 8.

To verify that the significant effects were not explained by nonspecific autocorrelations in the data, or Type I errors caused by the specific trial orderings (Mumford et al., 2014), we examined performance in a control region (ventricles, Supplementary Figure 1a & b). This performance remained at chance level for the LSA approach in Experiment 1 ($ps > 0.239$) and Experiment 2 ($ps > 0.273$). However, the LSS approach yielded false positives in block A ($p = 0.010$), block B ($p = 0.062$) and block D ($p = 0.074$).

Univariate analyses

In contrast to MVPA analyses, traditional activation analyses did not benefit from very long inter-stimulus intervals at the cost of information (i.e., number of trials that can be presented in a same amount of time). As expected, a design with intermediate trial intervals and null events produced the strongest differential activation in the TOFC (Figure 8b), both in terms of extent of activation and average signal amplitude.

In Experiment 2 (Figure 8b, left panels), a within-subject ANOVA (5 blocks) was not significant for extent and strength of activation ($ps > 0.280$). Paired t -test revealed a trend significant difference between extent of activation in block A and block D ($t_{17} = 1.96$, $p = 0.066$, $d = 0.46$), and higher average amplitude in block D compared to block C ($t_{17} = 2.72$, $p = 0.015$, $d = 0.64$) and block E ($t_{17} = 2.31$, $p = 0.034$, $d = 0.54$).

In Experiment 1 (Figure 8b, right panels), there was a significant effect of condition on the extent of activation ($F_{3, 46} = 3.48$, $p = 0.023$, $\eta_p^2 = 0.18$), but not on average amplitude ($p = 0.182$). Independent t -tests revealed a larger extent of activation in condition III compared to condition I ($t_{23} = 2.67$, $p = 0.014$, $d = 1.07$) and condition II ($t_{18.34} = 2.05$, $p = 0.055$, $d = 0.82$). A larger extent of activation was also observed in condition IV compared to condition I ($t_{24} = 2.12$, $p = 0.045$, $d = 0.83$). Higher average activation was observed in condition III compared to condition I ($t_{23} = 1.95$, $p = 0.063$, $d = 0.78$). None of the other post-hoc comparisons were significant ($ps > 0.136$).

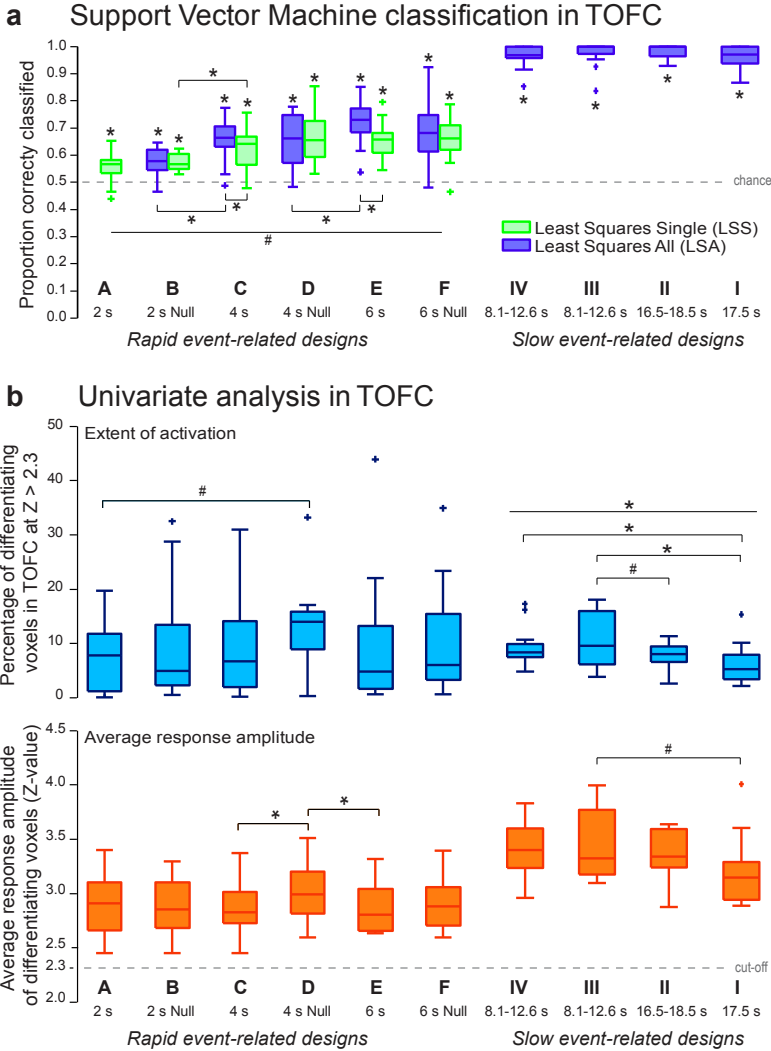


Figure 8 a) Results from the support vector machine classification on rapid event-related fMRI data (Experiment 2, left) and slow event-related fMRI data (Experiment 1, matched for scan duration, right) in the temporal occipital fusiform cortex (TOFC), depicting classification performance for each individual (box plots). Classification scores are obtained using a Least Squares All (LSA) approach (purple) and a Least Squares Single (LSS) approach (green). In each of the six within-subject conditions (left) average performance is above chance. Both approaches yield similar results, except for block A, where a LSA approach could not be calculated, and block C and E, where a LSA approach outperformed the LSS approach. Data from the slow event-related designs (right) show that in each of the four between-subject conditions average performance was above 0.8. **b)** Average activation derived from rapid event-related fMRI designs (Experiment 2, left), and slow event-related designs (Experiment 1, matched for scan duration, right). Top panels depict the percentage of voxels that was activated ($Z > 2.3$) in the TOFC; bottom panels depict the average response amplitude of these activated voxels. In contrast to MVPA analyses, slow-event related designs are inefficient for traditional univariate analyses. Note that data are presented from short to long intervals; hence the conditions in Experiment 1 (right panels) are presented in reversed order. * $p < 0.05$; # $p < 0.08$.

Experiment 2: Discussion

The results from this experiment demonstrate that design optimization for univariate analyses cannot be automatically applied to single-trial MVPA. This is exemplified by the fact that a regular LSA analysis was not possible for trial durations of 2 seconds, in all trial orders in which null events were omitted (block A), and in a third of the trial orders in which null events were present (block B), because the design matrix turned out to be rank deficient. Interestingly, the LSS approach did not have problems with these short trial durations. Although this seems to suggest that LSS is preferred in rapid event-related designs, the story seems to be more complicated. First of all, aside from the failed LSA analyses in block A and part of block B, results from Experiment 2, and the reanalysis of data from Experiment 1, show that faces and houses can be reliably distinguished using a SVM, independent of study design. However, when comparing within- and between-category pattern similarity the distinction can be made in all designs using a LSA approach, but using a LSS approach this distinction can only be made if designs include null events. Still, except for block A, the LSA approach outperforms the LSS approach, not only in its performance on designs *without* null events, but also in its performance in designs *with* null events. This is seen for both classification analysis and similarity analysis. In sum, LSA benefits most and LSS suffers most from the omission of null events, while the omission of null events seems the best design choice for MVPA. Second, although a LSS approach may be the solution for very short trial durations, classification analysis on a control region (ventricles) was above chance level, indicating that there may be the risk of Type I error inflation using this approach. Third, results from Experiment 2, but especially the reanalysis of data from Experiment 1, revealed that longer inter-stimulus intervals produced larger effects, despite the smaller number of trials included in the model, in which case a LSS approach is not required.

Note that while the outcome of our experiments favor long trial durations, we have not systematically varied the effect of stimulus duration. In Experiment 1, stimulus durations were 4.5 seconds, which is substantially longer than the 300 ms used in Experiment 2. Perhaps the observed benefit of longer intervals is in fact explained by longer stimulus durations. Given that within each experiment conditions with longer intervals (and equal stimulus durations) yielded better results than conditions with shorter intervals, it seems unlikely that stimulus duration alone can account for the observed advantage of slow event-related designs.

In sum, given a fixed amount of scanning time, designs with longer intervals yield better results in MVPA. However, the univariate results showed a different pattern than the classification and similarity analysis. Here, the optimal tradeoff between trial duration and number of trials included in the model yielded a design with trial durations of 4 seconds and null events. The

discrepancy between single trial MVPA and univariate analysis in the degree to which each profits from null events seems to be explained by the fact that in the parameter estimation of single trials null events induce a rather arbitrary and uneven distribution of statistical power across trials, causing some response patterns to be estimated much more reliable than others. Indeed, this can be observed in more variability in classification performance (Figure 8a).

General discussion

The first question that we aimed to answer in this study was whether it is possible to apply single-trial MVPA to data obtained in a rapid event-related fMRI design. Our results show differential pattern similarity and reliable classification using both very short intervals and longer intervals. Next, we asked whether, given a fixed amount of scanning time, one should opt for more trials, or longer intervals. The results clearly advocate for the latter, that is, long inter-stimulus intervals should be used (average of 10 seconds or longer) if single-trial MVPA is the analysis of choice, at the cost of reducing the number of trials. Additionally, it is important to keep the durations of stimulus intervals fixed, as periods of rest (null events) are detrimental for a stable estimation of single-trial parameters. Finally, our third question was specifically relevant for learning paradigms and regarded the intervals and order in which stimuli should be presented if one is interested in assessing changes in representational similarity over the course of a task. Although longer inter-trial intervals yielded the strongest effects, we were able to detect learning-dependent changes in designs with shorter (8.1-12.6 s) intervals as well, but only if the order of stimulus presentation was controlled for temporal proximity.

The present findings show that it is important to decide which type of data analysis has priority before carrying out an experiment, as they show that rapid event-related designs are sub-optimal for MVPA and at the same time confirm that slow event-related designs are inefficient for univariate analyses. For univariate analyses, multiple trials usually contribute to a single regressor, facilitating the deconvolution of the different regressors by using a jitter to obtain a stimulus onset asynchrony. In contrast, the use of null events seems to be detrimental for single-trial MVPA analysis, leading to more variability in the estimation of single-trial response patterns. An explanation for this is that the variance that is introduced by a jitter cannot be used to decorrelate all regressors, as is done in standard fMRI analysis, hence only the trials surrounding a longer interval profit from an increase in statistical power. Estimating a parameter weight with only the information of a single observation is unreliable unless the intervals are very long. With longer intervals, the trial-

specific activation patterns have more power and are more independent from each other, as both the impact of temporal autocorrelation in the data and collinearity in the model is reduced. The discrepancies between univariate analysis and MVPA can thus be explained by the fact that for MVPA each trial is modeled as a separate regressor. The problems that occur for single-trial MVPA in rapid event-related designs would probably also apply to single-trial univariate analysis.

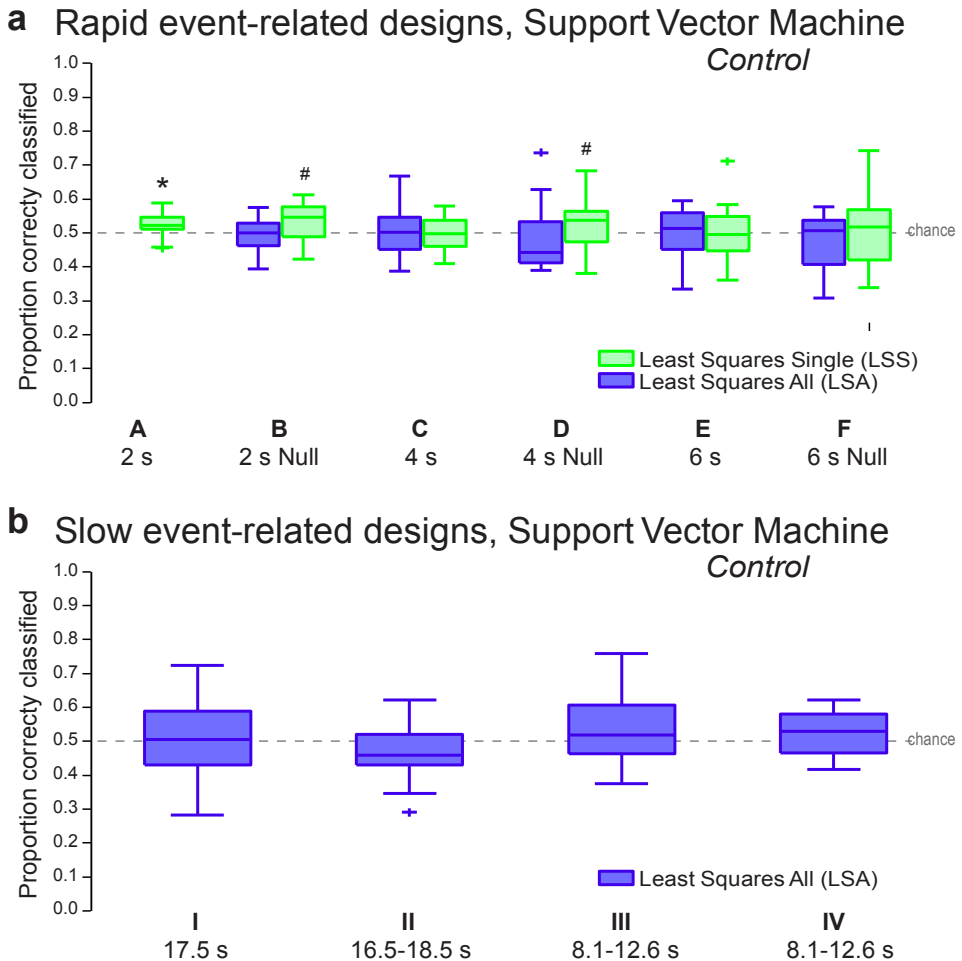
A potential weakness of the study could be that the trials were not randomly ordered for each subject. Simulations and analyses on resting-state data (Mumford et al., 2014) have shown that single-trial pattern analysis on a design with a blocked trial presentation and short intervals (mean of 3 s or 7 s) results in higher within-category correlations than between-category correlations, which is not surprising given the collinearity in the model. Although we did not present trials in a blocked fashion, the risk of having sequences that are optimized for univariate analyses is that such sequences often contain mini blocks where there are several trials in a row of the same type (which in univariate analysis increases detection power, (Kao et al., 2009; Liu et al., 2001). Such mini blocks may bias pattern analysis and classification performance (Mumford et al., 2014), especially when intervals are short (e.g., Experiment 2). However, we do not think that the results from the two experiments reflect an inflated Type I error. First, the order of stimulus presentation in Experiment 1 consisted of a repeated sequence of trials that was carefully counterbalanced across participants. This trial ordering did not contain mini blocks and was designed in such a way that all pair-wise correlations of interest were unbiased with regard to temporal proximity. Still, learning-dependent effects were evident from trial-by-trial changes in pattern similarity, and stimulus categories could be reliably distinguished using similarity and classification analysis. Second, if the 18 optimized sequences in Experiment 2 (3 per block) contained mini blocks, this should have led to an inflated Type I error in the control region (ventricles) as well, as temporal correlations in the hemodynamic response are not the only source of temporal autocorrelations in fMRI data. However, at least with an ordinary LSA approach, classification accuracy was at chance in this control region. Third, mini-blocks would be expected to have their greatest effect for designs with the shortest intervals (i.e., they contain more trials and therefore the chance of clustering is higher). So any effect of mini blocks would only further enhance our argument that short intervals need to be avoided in single-trial analysis. Of note, many stimulus-rich designs (e.g., Kriegeskorte et al., 2008) are aimed at assessing the similarity between patterns related to single *items*, but not necessarily single *trials*. In that case, regressors can include multiple presentations of the same stimulus and/ or between-run classification can be performed. Our conclusions primarily apply to experiments that completely rely on single-trial

parameter estimation; the degree to which they apply to stimulus-rich designs needs to be elucidated.

In conclusion, our data show that the optimal trade-off between trial duration and number of trials is different for single-trial MVPA and classical univariate approaches. While the latter benefits from rapid-event related designs and a jittered stimulus presentation, single-trial analysis benefits from slow event-related designs and fixed inter-trial intervals. We therefore recommend the use of slow event-related designs if single-trial pattern analysis is the main analysis of interest.

Supplementary Material – chapter 5

Supplementary Figure



Supplementary Figure 1 a) Control analyses using a support vector machine classification on rapid event-related fMRI data (Experiment 2) in the ventricles. For the Least Squares All (LSA) approach (purple), performance was at chance in all blocks. In contrast, false positive findings were found for the Least Squares Single (LSS) approach (green) in block A, B and D. b) Results from the support vector machine classification on slow-event related fMRI data (Experiment 1), depicting classification performance for each individual (box plots). In each of the four between-subject conditions performance is at chance. * $p < 0.05$; # $p < 0.08$.

Supplementary Tables

Supplementary Table 1 Summary of statistics of the fMRI data in condition 1 ($n = 12$), within-subjects ANOVA, in six anatomical ROIs

	Within-stimulus				Between-stimulus				Mean activation			
	Main effect of stimulus (2)		Interaction of stim(2) x trial(5)		Main effect of stimulus (2)		Interaction of (2 x 6)		Main effect of stimulus (2)		Interaction of stim(2) x trial(6)	
	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2
	<i>p-val</i>			<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		
ACC	15.00	0.58	5.39	0.33	10.80	0.50	1.23	0.10	6.72	0.38	1.74	0.14
	0.003		0.001		0.007		0.313		0.025		0.142	
Amygdala	0.98	0.08	NT	NT	0.56	0.05	NT	NT	0.72	0.06	NT	NT
	0.344		NT		0.469		NT		0.416		NT	
Hippocampus	1.91	0.15	NT	NT	2.32	0.17	NT	NT	3.20	0.23	NT	NT
	0.194		NT		0.156		NT		0.101		NT	
Insula	12.99	0.54	0.63	0.05	6.65	0.38	2.22	0.17	4.90	0.31	3.90	0.26
	0.004		0.641		0.026		0.065		0.049		0.004	
SFG	23.84	0.68	4.07	0.27	16.78	0.60	1.69	0.13	0.04	0.00	NT	NT
	<0.0005		0.007		0.002		0.152		0.856		NT	
vmPFC	1.36	0.11	NT	NT	3.62	0.25	NT	NT	2.33	0.18	NT	NT
	0.268		NT		0.084		NT		0.155		NT	

All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for interaction effects. Main effects are calculated over all acquisition trials, 5 for within-stimulus pattern similarity, 6 for between-stimulus pattern similarity and 6 for average activation. stim = stimulus. ^a effect caused by significantly higher values for CS- stimuli. ACC = Anterior Cingulate Cortex; SFG = Superior Frontal Gyrus; vmPFC = ventromedial Prefrontal Cortex.

Supplementary Table 2 Summary of statistics of the fMRI data in condition II ($n = 11$), within-subjects ANOVA, in six anatomical ROIs

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect of stimulus (2)		Interaction of stim(2) x trial(5)		Main effect of stimulus (2)		Interaction of (2 x 6)		Main effect of stimulus (2)		Interaction of stim(2) x trial(6)	
	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2
	<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>	
ACC	16.50	0.62	2.17	0.18	15.62	0.61	2.44	0.20	8.22	0.45	1.06	0.10
	0.002		0.090		0.003		0.094		<i>0.017</i>		0.393	
Amygdala	3.28	0.25	NT	NT	0.96	0.09	NT	NT	1.91	0.16	NT	NT
	0.100		NT		0.351		NT		0.197		NT	
Hippocampus	9.21	0.48	2.26	0.02	6.30	0.39	0.61	0.06	1.32	0.12	NT	NT
	0.013		0.904		0.031		0.691		0.277		NT	
Insula	7.07	0.41	1.63	0.14	16.45	0.62	1.85	0.16	17.88	0.64	4.02	0.29
	0.024		0.186		0.002		0.121		0.002		0.004	
SFG	11.79	0.54	5.25	0.34	14.10	0.59	1.14	0.10	1.10	0.10	NT	NT
	0.006		0.002		0.004		0.344		0.320		NT	
vmPFC	1.29	0.11	NT	NT	1.01	0.09	NT	NT	0.64	0.06	NT	NT
	0.283		NT		0.339		NT		0.443		NT	

All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for interaction effects. Main effects are calculated over all acquisition trials, 5 for within-stimulus pattern similarity, 6 for between-stimulus pattern similarity and 6 for average activation. stim = stimulus. ^a effect caused by significantly higher values for CS-stimuli. ACC = Anterior Cingulate Cortex; SFG = Superior Frontal Gyrus; vmPFC = ventromedial Prefrontal Cortex.

SUPPLEMENTARY MATERIAL - CHAPTER 5

Supplementary Table 3 Summary of statistics of the fMRI data in condition III ($n = 13$), within-subjects ANOVA, in six anatomical ROIs

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect of stimulus (2)		Interaction of stim(2) x trial(5)		Main effect of stimulus (2)		Interaction of (2 x 6)		Main effect of stimulus (2)		Interaction of stim(2) x trial(6)	
	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2
	<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>	
ACC	5.32	0.31	2.79	0.19	12.81	0.52	1.31	0.10	0.90	0.07	NT	NT
	<i>0.040</i>		<i>0.036</i>		0.004		0.273		0.361		NT	
Amygdala	1.79	0.13	NT	NT	2.40	0.17	NT	NT	0.15	0.01	NT	NT
	0.205		NT		0.147		NT		0.706		NT	
Hippocampus	2.65	0.18	NT	NT	0.00	0.00	NT	NT	0.43	0.03	NT	NT
	0.129		NT		0.949		NT		0.522		NT	
Insula	15.62	0.57	2.42	0.17	10.11	0.46	0.41	0.03	5.96	0.33	1.98	0.14
	0.002		0.061		0.008		0.843		<i>0.031</i>		0.095	
SFG	2.62	0.18	NT	NT	12.00	0.50	1.61	0.12	0.01	0.00	NT	NT
	0.132		NT		0.005		0.172		0.921		NT	
vmPFC	9.52	0.44	1.36	0.10	8.68	0.42	2.05	0.15	31.26	0.72	2.61	0.18
	0.009		0.263		0.012		0.084		<0.0005^a		<i>0.031</i>	

All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for interaction effects. Main effects are calculated over all acquisition trials, 5 for within-stimulus pattern similarity, 6 for between-stimulus pattern similarity and 6 for average activation. stim = stimulus. ^a effect caused by significantly higher values for CS-stimuli. ACC = Anterior Cingulate Cortex; SFG = Superior Frontal Gyrus; vmPFC = ventromedial Prefrontal Cortex.

Supplementary Table 4 Summary of statistics of the fMRI data in condition IV ($n = 14$), within-subjects ANOVA, in six anatomical ROIs

Region	Within-stimulus				Between-stimulus				Mean activation			
	Main effect of stimulus (2)		Interaction of stim(2) x trial(5)		Main effect of stimulus (2)		Interaction of (2 x 6)		Main effect of stimulus (2)		Interaction of stim(2) x trial(6)	
	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2	<i>F</i>	η_p^2
<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		<i>p-val</i>		
ACC	19.05	0.59	1.74	0.12	2.09	0.14	NT	NT	1.05	0.08	NT	NT
	0.001		0.155		0.172		NT		0.324		NT	
Amygdala	10.75	0.45	0.58	0.04	2.90	0.18	NT	NT	0.66	0.05	NT	NT
	0.006		0.681		0.112		NT		0.432		NT	
Hippocampus	6.09	0.32	0.51	0.04	1.42	0.10	NT	NT	0.00	0.00	NT	NT
	0.028		0.728		0.255		NT		0.994		NT	
Insula	6.83	0.34	2.53	0.16	0.72	0.05	NT	NT	11.06	0.46	2.19	0.14
	0.021		0.051		0.413		NT		0.005		0.066	
SFG	17.76	0.58	3.00	0.19	1.18	0.08	NT	NT	0.91	0.07	NT	NT
	0.001		0.027		0.297		NT		0.358		NT	
vmPFC	2.92	0.18	NT	NT	1.70	0.12	NT	NT	2.24	0.15	NT	NT
	0.111		NT		0.215		NT		0.158		NT	

All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. NT = not tested: Areas without significant main effect of stimulus type are not tested for interaction effects. Main effects are calculated over all acquisition trials, 5 for within-stimulus pattern similarity, 6 for between-stimulus pattern similarity and 6 for average activation. stim = stimulus. ^a effect caused by significantly higher values for CS- stimuli. ACC = Anterior Cingulate Cortex; SFG = Superior Frontal Gyrus; vmPFC = ventromedial Prefrontal Cortex.

Supplementary Table 5 Summary of statistics of the similarity analysis (within-category correlation – between-category correlation) in the temporal occipital fusiform cortex on data from Experiment 2

	A		B		C		D		E		F	
	LSA	LSS	LSA	LSS	LSA	LSS	LSA	LSS	LSA	LSS	LSA	LSS
Mean diff.	-	0.01	0.01	0.02	0.03	0.00	0.01	0.01	0.03	0.00	0.02	0.01
Std. Dev.	-	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01
<i>t</i>	-	4.39	6.39	7.18	13.61	1.13	2.08	4.74	9.57	1.40	3.56	2.81
df	-	17	11	17	17	17	17	17	17	17	17	17
<i>p</i> -value	-	<0.0005	<0.0005	<0.0005	<0.0005	0.274	0.053	<0.0005	<0.0005	0.181	0.002	0.012
Cohen's <i>d</i>	-	1.04	1.84	1.69	3.21	0.27	0.49	1.12	2.26	0.33	0.84	0.66

P-values are obtained with paired-sample *t*-tests, comparing within-category and between-category similarity. All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. LSA = Least Squares All; LSS = Least Squares Single.

Supplementary Table 6 Summary of statistics of the similarity analysis (within-category correlation – between-category correlation) in the temporal occipital fusiform cortex on data from Experiment I

	I	II	III	IV
	<i>LSA</i>	<i>LSA</i>	<i>LSA</i>	<i>LSA</i>
Mean diff	0.12	0.16	0.11	0.09
Std. Dev.	0.07	0.06	0.05	0.04
<i>t</i>	6.09	8.51	8.01	8.45
df	11	10	12	13
<i>p</i> -value	<0.0005	<0.0005	<0.0005	<0.0005
Cohen's <i>d</i>	1.76	2.57	2.22	2.26

P-values are obtained with paired-sample *t*-tests, comparing within-category and between-category similarity. All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. LSA = Least Squares All.

Supplementary Table 7 Summary of statistics of the classification performance in the temporal occipital fusiform cortex on data from Experiment 2

	A		B		C		D		E		F	
	<i>LSA</i>	<i>LSS</i>	<i>LSA</i>	<i>LSS</i>	<i>LSA</i>	<i>LSS</i>	<i>LSA</i>	<i>LSS</i>	<i>LSA</i>	<i>LSS</i>	<i>LSA</i>	<i>LSS</i>
Mean	-	0.56	0.57	0.58	0.66	0.62	0.65	0.66	0.71	0.66	0.68	0.66
Std. Dev.	-	0.05	0.05	0.03	0.07	0.07	0.09	0.09	0.09	0.06	0.11	0.08
<i>t</i>	-	4.67	4.99	10.72	9.52	6.96	6.81	7.96	10.02	11.23	7.29	8.93
df	-	17	11	17	17	17	17	17	17	17	17	17
<i>p</i> -value	-	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005	<0.0005
Cohen's <i>d</i>	-	1.10	1.44	2.53	2.24	1.64	1.60	1.87	2.36	2.65	1.72	2.10

P-values are obtained with one-sample *t*-tests (compared to the 0.5 chance level). All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. LSA = Least Squares All; LSS = Least Squares Single.

Supplementary Table 8 Summary of statistics of the classification performance in the temporal occipital fusiform cortex on data from Experiment I

	I	II	III	IV
	<i>LSA</i>	<i>LSA</i>	<i>LSA</i>	<i>LSA</i>
Mean	0.96	0.98	0.98	0.96
Std. Dev.	0.04	0.03	0.05	0.04
<i>t</i>	39.95	63.56	36.15	40.08
df	11	10	12	13
<i>p</i> -value	<0.0005	<0.0005	<0.0005	<0.0005
Cohen's <i>d</i>	11.53	19.16	10.03	10.71

P-values are obtained with one-sample *t*-tests (compared to the 0.5 chance level). All significant values ($p < 0.05$) are in italics, and those that reach FDR-corrected significance are in bold. LSA = Least Squares All.