Pattern analyses reveal separate experience-based fear memories in the human right amygdala

Senne Braem[a,b], Jan De Houwer[b], Jelle Demanet[a], Kenneth S. L. Yuen[c], Raffael Kalisch[c,d,1], & Marcel Brass[a,1]

[a]Department of Experimental Psychology, Ghent University, Henri-Dunantlaan 2, Ghent, 9000, Belgium

[b]Department of Experimental Clinical and Health Psychology, Ghent University, Henri-Dunantlaan 2, Ghent, 9000, Belgium

[c]Neuroimaging Center Mainz (NIC), Focus Program Translational Neuroscience, Johannes Gutenberg University Medical Center, Langenbeckstr. 1, Geb. 701, EG, Raum 0.36, Mainz, 55131, Germany

[d]Deutsches Resilienz-Zentrum (DRZ), Johannes Gutenberg University Medical Center, Langenbeckstr. 1, Geb. 701, EG, Raum 0.36, Mainz, 55131, Germany

[1]contributed equally to the present work.

In press. Journal of Neuroscience.

Corresponding author:

Senne.Braem@Ugent.be

Henri-Dunantlaan 2, 9000 Gent, Belgium

Figures (8), Tables (0), Words: Abstract (191), Introduction (615), & Discussion (804)

Acknowledgements: S.B. is supported by FWO - Research Foundation Flanders (12K6316N). The research was funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (IUAPVII/33), the Research Foundation Flanders (G.0231.13N), and Ghent University (BOF09/01M00209). The authors declare no competing financial interests. We would like to thank Dominik R. Bach, Timothy Behrens, Lauri Nummenmaa, and two anonymous reviewers, for their helpful comments on earlier versions of this manuscript.

Abstract

Learning fear via the experience of contingencies between a conditioned stimulus (CS) and an aversive unconditioned stimulus (US) is often assumed to be fundamentally different from learning fear via instructions. An open question is whether fear-related brain areas respond differently to experienced CS-US contingencies than to merely instructed CS-US contingencies. Here, we contrasted two experimental conditions where subjects were instructed to expect the same CS-US contingencies while only one condition was characterized by prior experience with the CS-US contingency. Using multi-voxel pattern analysis of fMRI data, we found CS-related neural activation patterns in the right amygdala (but not in other fear-related regions) that dissociated between whether a CS-US contingency had been instructed and experienced versus merely instructed. A second experiment further corroborated this finding by showing a category-independent neural response to instructed and experienced, but not merely instructed, CS presentations in the human right amygdala. Together, these findings are in line with previous studies showing that verbal fear instructions have a strong impact on both brain and behaviour. However, even in the face of fear instructions, the human right amygdala still shows a separable neural pattern response to experience-based fear contingencies.

Significance statement

In our study we addressed a fundamental problem of the science of human fear learning and memory, namely whether fear learning via experience in humans relies on a neural pathway that can be separated from fear learning via verbal information. Using two new procedures and recent advances in the analysis of brain imaging data, we localize purely experience-based fear processing and memory in the right amygdala, thereby making a direct link between human and animal research.

Introduction

As a product of evolution, animals are equipped with the ability to learn relations that impact their survival. For example, by recognizing contingencies between certain stimuli in the environment (CSs) and harmful events (USs), animals can learn to anticipate these events in the future, a process which is thought to underlie Pavlovian fear conditioning (Maren, 2001; Pavlov, 1927). Pavlovian learning is often distinguished from other forms of fear acquisition (e.g., via instructions or observation; Olsson & Phelps, 2007) as it necessitates first-hand experiences of paired events rather than information transfer from an instructor or model. The experience-based nature and strong evolutionary conservation of Pavlovian conditioning led many theorists to think that conditioning in humans happens relatively automatically and independently from verbal processing or even awareness (Dolan & Vuilleumier, 2003; Grillon, 2009; LeDoux, 2014; Mineka & Öhman, 2002; Olsson & Phelps, 2004; Schultz & Helmstetter, 2010). Accordingly, it has been proposed that there is an evolutionary old fear module in the human brain, centered around the amygdala, that contributes to the acquisition and expression of Pavlovian fear by specifically mediating its putative non-verbal, experience-based element (Öhman & Mineka, 2001). In its strongest form, this theory postulates the amygdala-centered fear module to be "encapsulated", i.e., impenetrable to conscious cognitive control (Öhman & Mineka, 2001).

If there is a brain module responsible for purely experience-based Pavlovian fear learning, it must operate relatively independently from verbally-mediated or instructed fear learning. Therefore, to test whether Pavlovian fear learning can operate independently from verbally-mediated fear learning, previous studies tried to isolate neural correlates of Pavlovian learning by evidencing conditioning in the absence of CS awareness, that is, with backward-masked or subliminally presented CSs. These studies often pointed towards the amygdala (Critchley et al., 2002; Knight et al., 2009; Morris et al., 1998; Tabbert et al., 2011) as the neural substrate of Pavlovian fear learning, but also received substantial criticism based on methodological (potential residual CS awareness, Mitchell et al., 2009) and statistical grounds (Vadillo et al.,

2016). Even evidence for fear conditioning in non-verbal human children (Watson & Rayner, 1920) still leaves open the main question that motivated our research: Once a human becomes verbal, can these verbal processes override learning pathways via experience, or do we keep separable pathways for experience-based fear conditioning instead? In other words, it remains unclear whether the human brain reserves space for the unique impact of actually experiencing CS-US pairings, in the face of explicit fear instructions.

If we want to demonstrate a truly independent, separable neural response to experience-based Pavlovian conditioning, we must contrast it to verbally-mediated instruction-based learning. Therefore, rather than trying to exclude conscious or language-based processing, as in previous studies, we here developed an experiment where conditions were optimal for verbally-mediated language-based processing to override the hypothesized separate experience-based component to fear learning. To this end, we compared neural pattern responses to two CSs that were both part of explicitly instructed contingencies but of which one was (CS^+P, CS^+ Paired) and the other was not (CS^+U, CS^+ Unpaired) previously paired with the US (Mertens et al., 2016; Raes et al., 2014). Hence, whereas both stimuli were expected to activate the same instruction-based fear memory during the memory testing phase, only the previously paired CS (CS^+P) should additionally activate experience-based memory elements, which we here call the Pavlovian trace. This way, we studied a unique, experience-based component of Pavlovian fear conditioning in humans. In a second experiment, we aimed to replicate and further extend this finding by testing whether a similar neural signature could be observed when comparing entirely novel (i.e., merely instructed) to old (i.e., instructed and experienced) CS presentations.

Material and Methods

General Method

In our first and main experiment, participants were first instructed that two visual stimuli (CS^+P, CS^+U) could be followed by a painful electrotactile stimulus (US). A third stimulus (CS^-) was introduced as a control stimulus that would never be paired with the US. Participants were

then told that, in a preparatory training phase, only one of the two CS^+s (CS^+P) would be occasionally paired with an electrical stimulus, whereas the other CS^+ (CS^+U) would be occasionally followed by a placeholder (drawing of a lightning bolt) – under the false pretense that this limitation of the absolute number of electrical stimulation given would allow the subjects to gradually adjust to the aversive task conditions. During this training phase, we randomly presented each CS nine times; three out of nine CS^+P presentations were followed by a US, making subjects experience the CS^+P-US contingency, while three out of nine CS^+U presentations were followed by the placeholder (Figure 1). Before the subsequent test phase, participants were re-instructed that, from then on, both CS^+s could be followed by a US. In fact, no more USs were delivered, keeping this critical phase of the experiment free of any Pavlovian fear learning. Because instructions in the test phase were identical for both CS^+s, neural activity patterns associated exclusively with the CS^+P during the test phase would reflect a Pavlovian trace of actual CS-US pairings. This procedure has been tested before by Raes and colleagues (2016) and Mertens and colleagues (2016) but without simultaneous MRI recordings. Therefore, the behavioral data will be discussed in close comparison to those studies.

A second experiment was performed to examine whether our main finding could also be observed in a different, but conceptually similar procedure. As an additional motivation, this second experiment also allowed us to examine whether a higher visual category-independent neural response to instructed and experienced CS presentations could be observed, as compared to merely instructed CS presentations. This experiment was different in design from Experiment 1, but nonetheless allowed for a similar comparison between instructed and experienced versus merely instructed fear. Specifically, we used a large range of different CSs (house or face pictures), and participants were instructed on the relevant CS-US contingencies before each CS presentation (Figure 8a). That is, each trial of this experiment started with the presentation of an instruction screen that defined the CS^+ and the CS^-, followed by the presentation of the CS^+ or CS^-. Each CS^+ was in its turn followed by a US presentation. This way, participants saw many different CS instructions. Importantly, however, whereas some contingency instructions (and

the experience thereof) recurred multiple times throughout the experiment (i.e., old CS stimuli), others were always new (i.e., new CS stimuli). Hence, some contingencies had been "instructed and experienced" whereas, for new CSs, this was not the case. The new CSs could therefore be considered "merely instructed".

In contrast to Experiment 1, Experiment 2 also allowed for a comparison within CS conditions but across visual stimuli or visual categories. Specifically, the different CS instructions could be further subdivided into those that used pictures of faces as CSs, and those that used pictures of houses as CSs. This way, we could compute the similarity in neural activation patterns at the time of CS presentation between faces versus houses separately (see Figure 8b) for each of the four different CS presentation conditions: CS^{+}old, CS^{-}old, CS^{+}new, and CS^{-}new. By computing similarities between responses to faces and houses, this study allowed us to investigate to which extent fear-related regions showed a neural response that was independent from CS object category, and thus specific to whether a CS carried a representation of threat (CS^{+} vs. CS^{-}) or whether it was old or novel (CSold vs. CSnew). Following up on the findings of the first experiment, we zoomed in on the pattern similarities in the left and right amygdala. In particular, we were interested in whether the right amygdala would show a higher similarity between face and house evoked patterns when a CS^{+} was old, which would indicate a representation of threat sensitive to whether a contingency had been experienced before.

Participants

Twenty participants took part in Experiment 1 (twelve women and eight men, mean age = 24, SD = 2.5, range = 19 - 28), and another twenty in Experiment 2. One participant from Experiment 2 was excluded from analyses due to self-reported nausea and inattention to the task, so the final sample of Experiment 2 contained 19 participants (ten women and nine men, mean age = 24, SD = 3.7, range = 18 - 34). All participants had normal or corrected to normal vision, and were right-handed as assessed by the Edinburgh Handedness Inventory. They gave their informed written consent and reported no current or history of neurological, psychiatric or

major medical disorder. Every participant was paid 35€ for participating. The work has been completed with the approval of the Ghent University Hospital Ethical Committee.

Stimuli and Procedure: Experiment 1

Stimuli. The conditioned stimuli consisted of three dissociable blue fractal figures (snowflakes) that were presented on a white background. Counterbalanced across participants, one of the fractals served as a conditioned stimulus (CS) that would never be followed by an electrotactile pain stimulus (CS^-), another served as a CS that could occasionally be followed by a pain stimulus (CS^+P), and a third as a CS that subjects were told could be followed by a pain stimulus, but was actually only occasionally followed by a placeholder (CS^+U). The electrotactile pain stimulus that served as the aversive stimulus (unconditioned stimulus, US) consisted of a train of twelve square-wave pulses of 2 ms duration each (interval 18 ms). The US was delivered through a surface electrode with platinum pin (Specialty Developments, Bexley, UK) onto the right leg over the retromalleolar course of the sural nerve using a DS5 electrical stimulator (Digitimer, Welwyn Garden City, UK). The intensity was determined through a standard work-up procedure prior to the experiment (more info below). The placeholder consisted of a centrally presented yellow drawing of a lightning bolt.

Ratings. Self-reported CS fear and US expectancy were assessed for all CSs in separate rating blocks interspersed between conditioning trials. These ratings were performed on screen. On a typical rating trial, the CS was presented centrally, while the question on fear or US expectancy was situated on top of the CS and a rating scale was presented below. Before each rating phase, participants were instructed to respond to the questions that would appear at the top of the screen through selecting the response possibility that felt most appropriate to them. Furthermore, it was stressed that these questions pertained to their most recent encounter with the CSs during the foregoing (conditioning) phase. In addition, participants were instructed that "Whenever you are asked about your expectancy of an electrical stimulus, we refer to the actual stimulation, not to the picture of the lightning bolt". The questions that appeared were "How much fear did you experience when looking at this figure?" (self-reported CS fear) and "To

what extent did you expect an electrical stimulation while looking at this figure?" (US expectancy). Participants responded verbally on a 9-point Likert scale. Numbers 1, 3, 5, 7, and 9 of this scale carried a response label that was presented right above the number (with 1 = none at all/certainly not; 3 = very little/rather not; 5 = uncertain; 7 = quite some/to some extent; 9 = very much/most certainly).

Procedure. The experimenter attached the electrodes to the participant, who was positioned on the scanner table, right before the table was inserted into the scanner. Next, the tolerance level of the pain stimulus was determined for each participant individually, by means of an adapted interleaved staircase procedure. This procedure consisted of 20 trials where USs were presented and participants rated the subjectively experienced pain intensity on a scale from zero (not painful at all) to ten (extremely painful). In order to increase reliability of the threshold procedure, the 20 trials were divided into two separate sequences of ten trials differing in current amplitude on their first trial, which was randomly drawn from either .5 to 1.0 mA, or 1.0 to 1.5 mA, respectively. After the first trial of each sequence, the current amplitude of the following trial depended on the participant's rating in the previous trial of the respective sequence. If the rating was below five, current amplitude would increase by 0.1 mA in the following trial of that sequence, whereas it would decrease by 0.1 mA in case of a rating above five, or stay the same if the rating equaled five. Thus, ratings from both sequences would approach five on the rating scale for each participant. The two sequences were presented intermixed and the final electrical current amplitude was then calculated as the mean of the final values from both sequences. Participants were instructed to rate an intensity five when it was experienced as unpleasant, but not intolerable, and informed that the pain stimuli used throughout the experiment would not surpass this value.

After this preparation phase, the participants were instructed to wait for five minutes, during which an anatomical scan would be administered. Upon completion of the anatomical scan, the participants were presented with the instructions. Participants were informed that three fractal figures (see Figure 1a) would appear successively for 8 seconds each. Following the

presentation of all three fractals on a white background, the participants were instructed that two of the fractals would sometimes be followed by an electrical stimulation, whereas the third fractal would never (in capital letters) be followed by an electrical stimulation. Subsequently, participants were informed that these fractal-pain contingencies would be clearly displayed and were encouraged to closely attend to these contingencies. After this instruction, a slide containing both CS^+ fractals and the text "+ electrical stimulation!" was presented during 8 seconds. This was followed by another 8 second presentation of a slide containing the CS^- fractal and the text "This figure will never be followed by a stimulation".

Following these general instructions, the participants were informed that they would first be allowed to familiarize themselves with the stimuli and the procedure in an initial training phase. The training phase was said to be very similar to the test phase that would follow, except that some of the electrical stimulations would be replaced by a picture of a lightning bolt (i.e., the placeholder stimulus, see Figure 1a). Participants were told that this was to prevent them from getting too many pain stimuli before the real experiment actually started and asked to keep in mind that whenever a lightning bold would be presented, this meant that in the actual test phase, a real electrical stimulation would occur. This was followed by a slide presenting the CS^+P (with electrical stimulation) and CS^+U (with lightning bolt) contingencies during 8 seconds. The final page of instructions informed participants that they would be asked to perform fear and US expectancy ratings at regular intervals during the upcoming phase. They were told that no stimuli would be administered during the ratings and asked to remember the most recent encounter with the fractals while answering the questions.

The actual training phase consisted of 27 conditioning trials (9 for each CS) interspersed with blocked ratings. Each conditioning trial started with a 4 second presentation of a fixation cross followed by a CS presentation for 8 seconds, with an inter-trial interval of 13, 15, or 17 seconds (see Figure 1a). On reinforced trials, the US or placeholder was presented at CS^+ offset. The US was presented for 300 ms. The placeholder remained on screen for a duration of 500 ms. The CSs were presented in "triplets" of three CS presentations so that each CS-type had

been presented once before the next triplet started. Trial order was randomized within triplets. Blocked ratings of fear and US expectancy were presented after 9, 18 and 27 conditioning trials (3, 6 and 9 triplets) respectively. As such, three mini-blocks were created within the training phase, each containing 3 trials of $CS^{+}P$, $CS^{+}U$ and CS^{-} presentation. Three out of nine $CS^{+}P$ and $CS^{+}U$ presentations were reinforced during the training phase. For half of the participants, the first, third, and penultimate presentation of the $CS^{+}P$ was followed by the US and the first, second, and last presentation of the $CS^{+}U$ was followed by the placeholder. The other half of the participants had counterbalanced reinforcement schedules (e.g., the first, second and last presentation of the $CS^{+}P$ would be followed by the US).

Each block of ratings contained six ratings (two for each CS). The order of rating trials within each rating block was fully randomized. However, due to technical difficulties, the CS presentation and question type were independently randomized within each rating block for the first five participants, resulting in repeated measurements or empty cells for some of the questions. For this reason, the rating analyses are reported for the last 15 subjects only. Brain-behavior correlation analyses, however, which focus on averaged CS ratings for an entire phase (across the three rating blocks), were possible to perform on the entire set of participants. Before the start of each rating block, it was stressed that by electrical stimulation (expectancy), we referred solely to real electrical stimulations (i.e., not placeholders).

After the training phase, participants again received on-screen instructions. They were informed that now the test phase would start, meaning that all electrical stimulations would be presented for real and no placeholders would be used. Participants were instructed that the test phase would evolve similarly to the training phase in all other respects. The course of the test phase was very similar to that of the test phase, with 27 trials and 3 rating blocks in between. However, no USs or placeholders were presented during this phase.

Stimuli and Procedure: Experiment 2

Stimuli. The stimuli for Experiment 2 were gray-scaled images selected from a database of 252 images of faces and 252 images of houses as previously used by Muhle-Karbe and

colleagues (2016). For each subject separately, 128 pictures (64 from each category) were randomly selected and assigned to sixty-four pairs of pictures with the only restriction that both pictures should be from the same category. When the category was faces, we further assured that both pictures were from the same gender (and an equal amount of pairs were formed per gender), to avoid that participants could dissociate the pictures based on gender. Per pair, one picture served as a conditioned stimulus (CS) that would never be followed by a electrical stimulation (CS^-), while another served as a CS that would always be followed by a electrical stimulation (CS^+). While eight pairs of CS^- and CS^+-pictures would re-occur throughout the experiment (four pairs of houses, two pairs of male faces, and two pairs of female faces), the 56 remaining pairs only appeared once, hereafter referred to as "old" and "new" stimuli, respectively. The electrotactile stimulus (US) consisted of the same sequence of pulses, and was applied to the same location, as in Experiment 1. Different from Experiment 1, however, the intensity could either be low or high (see below). The instructions would indicate the intensity of the potentially following electrotactile stimulus by showing a grey intensity meter which either pointed to the left or the right indicating a low or high intensity, respectively (see Figure 8a).

Procedure. Before entering the scanner, the participants were shown an example trial of the experiment (without electrotactile stimulation) and were instructed about the general procedure of the experiment. Namely, participants were informed that they would encounter several instruction screens where two pictures were presented above one another on the left side of the screen and an intensity meter on the right side of the screen, next to one of the two pictures (see Figure 8a). It was further explained to the participants that after these instruction screens, one of the two pictures would be presented (i.e., CS presentation) in the center of the screen. If this picture was presented next to the intensity meter during the instructions, they would receive an electrotactile stimulation shortly afterwards. The intensity of this stimulation was dependent on which direction the intensity meter pointed to. To ensure that participants paid attention to the task, one out of eight CS presentations (or one out of four in the practice block) were replaced

by a catch question where participants were shown either one of the two CSs, or a third picture that they had never seen before (of the same category). On these trials, their task was to indicate whether this picture was instructed to be followed by a stimulation, not followed by a stimulation, or never presented before (all participants performed above chance level on these catch questions; mean = 87.1%, SD = 10.6 %, minimum = 67%, maximum = 100%). Last, on some trials participants would also see a centrally presented intensity meter, which would indicate that an electrotactile stimulus could follow with a 50% probability. These trials are hereafter referred to as control trials. Briefly, this last condition was designed to control for general US expectancy effects during instruction presentation, but is not important for the current focus of analysis.

Next, participants were placed on the scanner table and after attaching the electrode, the tolerance level was determined for each subject separately. The two intensities were determined through a work-up procedure where gradually increasing current amplitudes were presented to the participant. Participants were asked to determine which amplitude was the first noticeable, and when we should stop increasing the amplitude, upon which the work-up procedure automatically ended. The first noticeable current level was used as the low intensity, the last tolerable as the high intensity. We assured subjects that only those two intensities could be used in the remainder of the experiment.

After the anatomical scan, we presented the instructions to the participant once more. They were further informed that they would receive a practice block half the length of the following three experimental blocks. This practice block was to ensure that participants were familiarized with all the "old" pairs, and was not included in the analyses.

During each block, the participant was presented with an equal amount of all three possible trial types: old instructed trials, new instructed trials, and control trials. The general structure of each of those trial types was that they started with 2500 ms instruction presentation, followed by a 2000 to 4200 ms instruction-CS interval, a 1000 ms CS presentation, 2000 to 4200 ms CS-US interval, a 300 ms US presentation (if the CS was a CS^{+}), and, finally, a 2500 to 4700 ms inter-

trial interval. For instructed trials, instruction presentation consisted of a presentation of the CS^- and CS^+ picture above one another on the left side of the screen (location of CS-type was randomly determined per trial), and an intensity meter next to the CS^+ picture. On control trials, instruction presentation consisted of a centrally presented intensity meter. CS presentation on instructed trials consisted of one of the two CSs (CS^- or CS^+) centrally presented on screen. During control trials, this was replaced by the presentation of a central fixation cross.

Each test block consisted of 16 new instructed trials, 16 old instructed trials (each specific pair was presented twice per block), and 16 control trials. All trial types were further subdivided in eight high intensity and eight low intensity trials. The instructed trial types also showed an equal amount of faces and houses (and an equal amount of male and female pictures among the faces). The different trials were presented in a random order.

fMRI data acquisition

In both experiments, participants were positioned headfirst and supine and instructed not to move their heads to avoid motion artefacts. Images were collected using a 3T Magnetom Trio MRI scanner system (Siemens Medical Systems, Erlangen, Germany) with a standard thirty-two-channel radio-frequency head coil. First, a 3D high-resolution anatomical image of the whole brain was acquired for co-registration with the functional images using a T1-weighted 3D MPRAGE sequence (TR = 2530 ms, TE = 2.58 ms, TI = 1100 ms, acquisition matrix = 256 × 256 x 176, sagittal FOV = 220 mm, flip angle = 7°, voxel size = 0.9 × 0.86 × 0.86 mm). Next, whole brain functional images were collected using a T2*-weighted EPI sequence, sensitive to BOLD contrast (TR = 2000 ms, TE = 28 ms, image matrix = 64 × 64, FOV = 224 mm, flip angle = 80°, slice thickness = 3 mm, distance factor = 17%, voxels resized to 3.0 × 3.0 × 3.0 mm, 34 axial slices). The number of images per run varied depending on response speed during the rating blocks (Experiment 1) or catch questions (Experiment 2).

Experimental Design and Statistical Analysis

Behavioral data analyses for Experiment 1. We carried out two ANOVAs with the three within-subjects factors phase (training versus test), CS type (CS^+P, CS^+U and CS^-), and block

(first, second, or third rating block) for each rating scale separately. Further ANOVAs focused on specific CS contrasts (e.g., $CS^{+}P$ vs. $CS^{+}U$) to allow for a more detailed picture of the significant interactions observed in the main ANOVA. Due to incomplete data collection for the first five subjects, we performed all behavioral ANOVAs on fifteen subjects. Analyses excluding the factor block allowed us to study nineteen subjects and all twenty subjects for the fear and US expectancy data (one subject did not have any fear ratings for the CS-), respectively. Including these subjects did not change the significance of our findings, but, naturally, are limited to analyses excluding the factor block. Therefore, we will only report the overall ANOVA. We did, however, use these general ratings (across blocks) to investigate brain-behavior correlations.

Introduction to Representational Similarity Analyses. A traditional mass univariate voxel-wise comparison of CS-related activations in canonical fMRI analyses can inform about gross differences in the recruitment of brain regions at the macroscopic scale but is blind to differences in the recruitment of separable neural ensembles *within* a brain region. More recently, multivariate multi-voxel pattern analysis (MVPA) techniques have permitted to investigate intra-regional spatial patterns of neuronal activation (Haxby et al., 2001; Kriegeskorte et al., 2008; Norman et al., 2006), including in fear conditioning (Bach et al., 2011; Dunsmoor et al., 2013; Hauner et al., 2013; Li et al., 2008; Visser et al., 2011; 2013), with high sensitivity, often going beyond conclusions that can be derived from univariate analyses (notably, the latter did not show any significant differences in $CS^{+}P$ versus $CS^{+}U$ activity during the test phase, neither in a whole-brain corrected contrast, nor when restricting the analysis to any of the below-defined regions). We therefore investigated (dis)similarities in neural processing of different CSs or of the same CS at different time points of the experiment by extracting and comparing multi-voxel activation pattern data from previously selected regions of interest (ROIs) event-locked to the presentation of the different CSs (Kriegeskorte et al., 2008).

fMRI data analyses. Data processing and analyses were performed using the SPM8 Matlab-package software (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). The first four volumes of each run in which no stimulation occurred were discarded before estimating statistical models. Anatomical images were spatially normalized to the SPM T1-template image and resliced to a voxel size of 1 x 1 x 1 mm. Functional data were slice-time corrected and spatially realigned to the first volume of the task. Next, EPIs were spatially normalized based on the T1-derived normalization parameters and a temporal high-pass filter of 128 s was applied to remove low-frequency drifts. No spatial smoothing was used.

In Experiment 1, BOLD responses were modelled with boxcar functions at CS onset till CS offset (eight seconds) or with delta functions for all other regressors, which were then convolved with a standard hemodynamic response function (HRF). Event-related regressors were created corresponding to the onset of the CS and defined by CS-type (CS^-, CS^+P, or CS^+U), to the onset of US and defined by US-type (electrotactile stimulation or placeholder), or to the onset of a ratings question. Specifically, the model included three (or more, see below) regressors denoting the CS-type, one for the ratings, two for the US-type (in the training phase only), and six movement parameters derived from the realignment procedure, for the two runs separately (which corresponded with the training and test phase, respectively). The statistical parameter estimates were computed separately for each voxel for all columns in the design matrix. Three different first-level models were fitted. The CS-type regressors per phase were either further split up in a single regressor for each trial separately (i.e., the trial-based model), a regressor for each mini-block separately (i.e., the mini-block model), or a regressor per phase (i.e., the phase model). We analyzed the resulting data by computing pair-wise Pearson correlations between all CS event-related spatial patterns of activation. For a visual depiction of the resulting similarity matrices from these three types of models in the anterior cingulate cortex (ACC), see Figure 2. The strength of these correlations was used as a metric of similarity. The correlations were Fisher-transformed for statistical analyses. Different types of correlations were selected for the analyses of interest and analyzed in ANOVAs using Statistical Package for

the Social Sciences (version 22, SPSS). Each analysis started with an overall ANOVA including the factor region, whose six levels corresponded to the six regions of interest (ROIs) identified below. Only when effects interacted with the factor region, the effects were further studied for each region separately. All reported *p*-values are two-sided and a Greenhouse-Geisser correction was applied whenever the assumption of sphericity was violated, but uncorrected degrees of freedom are reported for ease of reading.

Task events in Experiment 2 were similarly modelled as in Experiment 1, only now we also added separate regressors for instruction presentations (also modelled with boxcar functions from instruction onset till instruction offset, 2500 milliseconds). In total, 31 event-related regressors were created. Eight regressors corresponded to the onset of instruction presentation and were defined by instruction-type (coding for either old versus new instructions, house versus face stimuli, and low versus high US intensity), sixteen to the onset of CS presentation defined by CS-type (similarly determined by the same conditions as instruction-type and further coding for being a CS^- or CS^+ presentation), four to the onset of control instructions and the subsequent fixation cross presentation (further determined by low or high intensity), two to the onset of US presentations and defined by being either of low or high intensity, and, finally, one to the onset of catch questions. Although instruction presentations, US presentations, and catch trials were modelled separately, they were not included in our contrasts and considered regressors of non-interest. For the analysis of this dataset, we only focused on high intensity trials, where the CS^+ was instructed and expected to be followed by a high intensity electrical stimulation. Specifically, we focused on voxel patterns evoked by CS presentation in any of the six below-defined ROIs. We again analyzed the resulting data by computing Fisher-transformed pair-wise Pearson correlations. In particular, we were interested in the correlations between house and face presentations of the same trial type (see Figure 8b). This way, we created four possible correlations of interest: the correlations between multi-voxel patterns of activity to face presentation versus house presentation, for CS^-old trials, CS^+old trials, CS^-new trials, and CS^+new trials, separately. These values were interpreted as the degree to which a certain region

showed a response independent of visual category. Next, these values were analyzed using two by two ANCOVAs with the factors fear relevance (CS^+ versus CS^-) and novelty (old versus new), and, to control for differences in preferred US intensity, the standardized covariate US intensity (baseline-corrected by dividing the high intensity value by the low intensity value), for each ROI separately.

ROI analyses. We extracted β estimates from the separate voxels of anatomically defined ROIs, known to be relevant in fear conditioning (the same fear related regions as were used in Visser et al., 2013). Specifically, in addition to the amygdala, we focused on the anterior cingulate cortex (ACC), superior frontal gyrus (SFG), insula, and ventromedial prefrontal cortex (vmPFC), areas previously implicated in fear conditioning (Visser et al., 2013; Fullana et al., 2015; Mechias et al., 2010). We extracted the mean voxel activation from those regions, as obtained from the Harvard-Oxford cortical and subcortical structural atlases (Harvard Center for Morphometric Analysis), thresholded at 25%, and included a factor "region" in each of our analyses to identify potential between-region differences.

Results

Experiment 1: Rating data

In the rating data, we tested whether we could replicate the behavioral results by Raes and colleagues (2014), and Mertens and colleagues (2016), by analyzing the mean fear and US expectancy ratings per phase and block. There was a clear main effect of CS type for both ratings, both *p*s < .001, which interacted with phase, both ps < .001 (see Figure 1b and 1c). Marginal significant three-way interactions between phase, CS type, and block for both the fear, $F(4, 11) = 3.4$, $p = .050$, and US expectancy ratings, $F(4, 11) = 2.7$, $p = .088$, hinted at a differential evolution of CS ratings over time depending on the phase.

Next, we investigated these interactions for each CS type comparison and phase separately, by running separate block × CS type ANOVAs. In the training phase, both the CS^+P and CS^+U elicited a higher US expectancy and fear rating, relative to the CS^-, all *p*s < .001. Moreover, the

CS^+P elicited higher ratings on both scales relative to the CS^+U, both *p*s < .005. In the testing phase, again, both the CS^+P and CS^+U elicited higher ratings relative to the CS^-, all *p*s < .001. The CS^+P, however, was no longer significantly different from the CS^+U, on both the fear, $F(1, 14) = 1.3$, $p = .272$, as well as the US expectancy scale, $F(1, 14) = 1.2$, $p = .277$, suggesting similar fear responses.

For all CS^+ to CS^- comparisons, on both phases, there was a significant interaction between CS type and block, all *p*s < .001, suggesting that CS^+ ratings did, while CS^- ratings did not, decay over time, irrespective of the phase (see Figures 1b and 1c). Interestingly, however, similar interactions between CS type and block between CS^+P and CS^+U were absent in the training phase, *F*s < 1.7, *p*s > .21, but present in the testing phase, with a significant interaction for the fear ratings, $F(2, 13) = 12.0$, $p = .001$, and a marginally significant interaction in the US expectancy data, $F(2, 13) = 1.2$, $p = .052$. These interactions demonstrated that while there was no overall difference between both CS^+ ratings in the test phase, there were some initial differences in the first mini-blocks of the test phase (see Figures 1b and 1c). Specifically, the fear rating for CS^+P relative to CS^+U was significantly higher in the first block, $t(14) = 2.22$, $p = .044$, marginally significant in the second block, $t(14) = 1.87$, $p = .082$, and absent in the third block, $t(14) = .76$, $p = .458$. The US expectancy rating for CS^+P relative to CS^+U was marginal significantly higher in the first block, $t(14) = 1.79$, $p = .095$, but not significantly different in the second or third block, $t(14) = 1.08$, $p = .301$, $t(14) = 0.52$, $p = .610$, respectively.

In sum, these results clearly replicate the results by Raes and colleagues (2014) and Mertens and colleagues (2016). Most importantly, the CS^+P elicited slightly but reliably higher fear ratings than the CS^+U, especially at the beginning of the test phase (Figure 1; for similar findings in fear potentiated startle responses, but not skin conductance responses, see Mertens et al., 2016; Raes et al., 2014). This indicates a dissociable contribution of prior actual CS-US pairings to the fear reaction to the CS^+P in the testing phase.

Similarity analyses: Experiment 1

Different pattern similarity analyses are reported to allow for a comprehensive picture of the data. Importantly, each of those analyses were motivated by specific hypotheses, which are detailed below when discussing each analysis separately. We will first discuss an analysis that tested whether the different CSs were responded to similarly within the training and test phases. That is, did the ROIs respond more similarly to the $CS^{+}P$ and the $CS^{+}U$, than, for example, the $CS^{+}P$ and the CS^{-}? Thereafter, we report an analysis that looked a the internal consistency of the separate patterns to each CS within a phase. Namely, did certain regions respond in a more consistent manner to one CS as opposed to another CS? Third, and most importantly, we tested whether the pattern response to $CS^{+}P$ during the training phase was a better predictor of itself during the test phase, than it was to $CS^{+}U$ or CS^{-}. As a post-hoc analysis following up on this hypothesized result, we also tested whether regions that show this differential processing of the $CS^{+}P$ also show a relation with the difference in fear ratings. That is, we wanted to explore whether the difference in subjective fear experience as observed in the present study (replicating previous findings by Mertens et al., 2016; Raes et al., 2014) can be linked back to a neural trace of experience-based fear. Last, we will report a targeted pattern-informed connectivity analyses by investigating parallels in trial-to-trial pattern similarities between the ACC, and left and right amygdala, depending on CS type, as well as a broader connectivity analysis involving all six regions. Note that every analysis started with an omnibus ANOVA that included the factor region, to detect between-region differences.

Inter-CS similarities per phase. We first confirmed that the selected ROIs process learned stimulus qualities (rather than merely processing the perceptual properties of the CSs). Specifically, we observed that the trial-averaged multi-voxel activation patterns evoked by the $CS^{+}P$ and the $CS^{+}U$ were more similar to each other ($CS^{+}PCS^{+}U$ inter-CS similarity, green bars in Figure 3) than each of them was to the trial-averaged patterns evoked by the CS^{-} (comparison with $CS^{-}CS^{+}P$ inter-CS similarity: $F(1,19) = 11.08$, $p = .004$, blue bars in Figure 3; with $CS^{-}CS^{+}U$ inter-CS similarity: $F(1,19) = 18.59$, $p < .001$, red bars; for statistical procedures, see Method). This was the case during both training and test (effects of phase: both Fs < 1), but,

intriguingly, the effect differed significantly between regions (both *p*s < .004). Specifically in the left and right amygdalae, the CS^+P, which is the only CS whose acquired qualities result (in part) from experience, evoked activation patterns that were not more similar to CS^+U patterns than to CS^- patterns (both *F*s < 1). This analysis suggests that while participants were instructed to treat the CS^+P and CS^+U similarly (and most fear-related regions also seem to reflect this), the amygdala could be sensitive to whether the fear value results from experience, and therefore did respond to them differently. This difference in responding was even comparable to the difference in responding to CS^+P versus CS^-. Whereas we cannot prove that this effect is not a floor effect, we would like to note that the other analyses below do show reliable differences in the amygdala (sometimes even exclusively in the amygdala), speaking against the idea that activity in the amygdala regions was simply too variable or too noisy to detect reliable differences.

Intra-CS similarities per phase. Areas responsible for processing acquired as opposed to mere perceptual stimulus qualities are thought to exhibit more consistent activation patterns to each of the CS^+s than to the CS^- from one trial to the next of the experiment (Visser et al., 2011; 2013). However, note that one could also expect the opposite. Namely, regions responsive to dynamic trial-to-trial changes in a certain condition might show a smaller internal consistency. Therefore, these analyses only allow us to conclude that regions which show differences in these internal consistency measures between conditions must respond to these two conditions differently. Such a result would be another confirmation that such regions encode acquired stimulus qualities. For simplicity and robustness, we grouped trials into three mini-blocks per phase and computed intra-CS similarities from one mini-block to the next. As expected, intra-CS^+P similarity was significantly higher than intra-CS^- similarity (main effect of CS type: $F(1,19) = 18.11$, $p < .001$; Figure 4), and this difference decreased in the test phase relative to the training phase, consistent with the observed extinction in fear ratings in this phase ($F(1,19) = 4.49$, $p = .047$; no interactions with region, *F*s(5,95) < 1.95, *p*s > .135). In stark contrast, the intra-CS^+U versus intra-CS^- similarity analysis (main effect of CS type: $F(1,19) = 7.31$, $p =$

.014) exhibited an interaction effect with region ($F(5,95) = 2.82$, $p = .020$; but no interactions with phase, Fs < 1). The amygdalae were the only ROIs not showing significantly higher intra-CS^+U than intra-CS^- similarities (both $Fs(1,19) < 1.25$, $ps > .278$). That is, an analysis based on the temporal consistence of neural activation patterns (Visser et al., 2011; 2013) found no evidence that the amygdala processes the learned qualities of a merely instructed stimulus, the CS^+U.

Although all our analyses reflect between-region differences in different forms of fear processing between six identified fear-related ROIs, one could also investigate whether completely unrelated ROIs show a fear response. Although it is hard to select completely unrelated regions, we investigated whether the superior temporal gyrus (i.e., auditory cortex) shows a similar main effect of fear conditioning (as suggested by an independent reader). Specifically, we zoomed in on the most reliable effect across all cortical fear-related regions: the enhanced correlation between patterns of responses to the instructed and experienced stimulus (CS^+P) versus the neutral stimulus (CS^-). While this effect did reach significance in each of those regions, even after correcting for multiple comparisons, it did not reach significance in the superior temporal gyrus, $F(1,19) = 2.52$, $p = .129$.

Inter-CS similarities across phase. Both the CS^+P and the CS^+U, but not the CS^-, carry a representation of threat. Fear-related regions should therefore show CS^+ evoked activation patterns that are more similar to each other than to CS^- evoked activation patterns. Intriguingly, the above set of similarity analyses indicated that the amygdala's neural activation pattern to the CS^+P was not more similar to the CS^+U than it was to CS^-, whereas all other fear-related regions did show a higher similarity between their responses to CS^+P and CS^+U than between any of the two CS^+s and the CS^- (Figure 3). Moreover, while all fear-related regions, including the amygdala, showed a higher consistency (within phase, but across mini-blocks) in their neural pattern response to the CS^+P than the CS^-, only the non-amygdala fear-related regions also showed a higher consistency in their response to the CS^+U than the CS^- (Figure 4). Together,

these results are already suggestive of the idea that the amygdala is involved differently in the processing of an instructed and experienced versus a merely instructed fear contingency.

However, the central question in our analyses was whether brain regions involved in fear learning would respond differently to both CS^+s in the test phase, despite the fact that the same verbal information was given about both CS^+s. A neural trace of experience-based Pavlovian fear learning should be apparent from similarities between CS evoked activation patterns during test and the activation pattern evoked by the CS^+P during training (CS^+Ptr), i.e., where the contingency was experienced and the Pavlovian fear memory was formed. Specifically, one should expect higher similarities between CS^+Ptr patterns and the patterns evoked by the same CS during testing (i.e., CS^+Pte). If those similarities were larger than between CS^+Ptr and CSU^+te (as well as between CS^+Ptr and CS^-te), they would indicate a Pavlovian memory trace. By contrast, if CS^+PtrCS^+Pte similarities were no more pronounced than CS^+PtrCS^+Ute similarities, this would indicate a more generalized representation of threat during testing that does not retain a specific experience-based memory element. In this critical analysis, we observed a significantly higher similarity between CS^+Ptr and CS^+Pte (CS^+PtrCS^+Pte) as well as CS^+Ptr and CS^+Ute (CS^+PtrCS^+Ute), than between CS^+Ptr and CS^-te (CS^+PtrCS^+Pte vs. CS^+PtrCS^-te similarity: $F(1,19) = 19.36$, $p < .001$; CS^+PtrCS^+Ute vs. CS^+PtrCS^-te similarity: $F(1,19) = 14.27$, $p = .001$; Figure 5). Importantly, CS^+PtrCS^+Pte did not differ from CS^+PtrCS^+Ute pattern similarity, $F(1,19) = .397$, $p = .536$. These results suggest both CS^+s evoked similar threat-related processing during test. However, there was an interaction with region ($F(10,190) = 3.32$, $p = .007$; also when comparing the left and right amygdala only: $F(2,18) = 3.31$, $p = .047$). Namely, the right amygdala exhibited CS^+Pte patterns that were significantly more similar to CS^+Ptr patterns than were both CS^+Ute and CS^-te patterns (CS^+PtrCS^+Pte vs. CS^+PtrCS^+Ute similarity: $t(19) = 2.204$, $p = .040$; CS^+PtrCS^+Pte vs. CS^+PtrCS^-te similarity: $t(19) = 3.990$, $p = .001$; Figure 5f). The CS^+Ute and CS^-te patterns did not differ in their similarity to the CS^+Ptr pattern, $t(19) = 1.013$, $p = .324$. Importantly, in all other regions, the patterns of CS^+Pte and CS^+Ute were not significantly different in their

similarity to the pattern of CS^+Ptr, all *t*s(19) < 1. Hence, this analysis isolated a threat-related neural response during testing in the right amygdala that was exclusively evoked by the CS^+P, as opposed to the CS^+U, meeting our criterion for a Pavlovian trace of actually experienced CS-US pairings. Other regions appeared to register a merely instructed threat (CS^+U) in the same way as a threat that is not only instructed but also previously experienced (CS^+P).

The observation that activation in the right amygdala is more similar for CS^+Ptr and CS^+Pte than for CS^+Ptr and CS^+Ute might also reflect the fact that CS^+Ptr and CS^+Pte are visually more similar than CS^+Ptr and CS^+Ute. However, if this alternative explanation is correct, then activation for CS^+Utr and CS^+Ute should also be more similar than activation for CS^+Utr and CS^+Pte. No such difference was observed, $t(16) = .727$, $p = .476$. In fact, the similarity between CS^+Ptr and CS^+Pte was higher than that between the CS^+Utr and CS^+Ute, $t(19) = 3.502$, $p = .002$.

Further supporting the encoding of this Pavlovian fear trace in the right amygdala, we observed a relation between CS^+P specific right amygdala neural responses and the observed differential (CS^+P > CS^+U) fear response, in that the difference between CS^+PtrCS^+Pte and CS^+PtrCS^+Ute pattern similarities as depicted in Figure 5f predicted the difference in CS^+Pte and CS^+Ute fear ratings during test across participants (*Spearman's* $\rho = .465$, $p = .039$; Figure 5g). This post-hoc analysis should of course be treated with caution because our study (and its sample size) was not in first instance set up to study inter-subject correlations.

Inter-region similarities in Intra-CS similarities. If the right amygdala processes experience-based threat in a way that can be dissociated from its processing of instruction-based threat, it might also preferentially exchange that information with other threat areas. Specifically, we wondered whether CS^+P related functional connectivity of the right amygdala with the ACC would differ from that of the left amygdala, for comparison. The ACC is the region that is most prominently and consistently activated during fear conditioning studies (Fullana et al., 2015; Mechias et al., 2010) and also exhibits strong structural and functional connectivity with the amygdalae (e.g., Bissière et al., 2008; Carlson et al., 2013; Etkin et al.,

2006; Van Marle et al., 2010; Williams et al., 2006; for a review, see Kim et al., 2011). To this end, we opted to carry out voxel-pattern-informed connectivity analyses, which have recently been demonstrated to be more sensitive and reliable than standard connectivity analysis (Geerligs & Henson, 2016). To perform a similarity-based functional connectivity analysis, we used the trial-based model (Figure 2a) and Spearman correlated the trial-by-trial intra-CS similarities per CS and phase (Figure 6a) between regions for each subject separately, thereby indexing the similarities between these three regions in how a CS pattern relates to itself across time (Figure 6b-c, Kriegeskorte et al., 2008). We found that intra-$CS^{+}P$ relative to intra-$CS^{+}U$ similarity time courses from the right, but not the left amygdala, showed a higher correlation with corresponding ACC time courses (interaction between CS type and amygdala side: $F(1,19) = 5.68$, $p = .028$; follow-up comparison of correlations between intra-$CS^{+}P$ and intra-$CS^{+}U$ similarity time courses from right amygdala and ACC: $t(19) = 2.698$, $p = .014$; from left amygdala and ACC: $t(19) = .767$, $p = .453$; Figure 6b-c). These results further corroborate the conclusion that the right amygdala reacts differently to experienced and instructed fear, but also point towards an extended role of a larger network (Okon-Singer et al., 2015; Pessoa & Adolphs, 2010), by showing an increased functional connectivity with the ACC for communicating this experience-based component of fear learning.

To further illustrate this idea, and allow for a more comprehensive picture of the data, we also computed inter-region similarity matrices depicting each possible region-to-region connectivity, for each CS-type separately (see Figure 7a). As a point of reference, we further included two occipital regions, namely the lateral occipital cortex (LO) and occipital pole (Occ). Multi-dimensional scaling analyses on Figure 7c depict the relations between the six fear-related regions for both CS^{+}s in the test phase. We used two-dimensional solutions (using PROXSCAL, SPSS), as these offered the most optimal stress levels relative to the number of dimensions (i.e., the elbow in the scree plot). Most importantly, these analyses visualize how the connectivity patterns change depending on the CS type processing. For example, Figure 7c suggests a more integrated role for the right amygdala when it comes to processing the $CS^{+}P$. In fact, when

comparing the overall connectivity between all six fear related regions per CS (averaging all 15 possible connections; Figure 7b), it appeared to be enhanced for CS^{+}Pte relative to CS^{+}Ute, $t(19) = 2.633$, $p = .016$. When testing this for each region separately (its average connectivity with all other five regions for CS^{+}Pte relative to CS^{+}Ute), the only two regions showing significantly stronger connectivity during CS^{+}Pte processing were the right amygdala and ACC, $t(19) = 2.780$, $p = .012$, $t(19) = 2.541$, $p = .020$, respectively (all other regions, $t(19) < 1.91$, $p > .071$), again suggesting that these two regions and their interaction play an important role in the learning or expressing of experience-based fear.

Similarity analyses: Experiment 2

In the second experiment, we used a different, but conceptually similar procedure to investigate whether we could replicate the observation that the right amygdala (in comparison to the left amygdala) dissociated between instructed and experienced (i.e., old) versus merely instructed (i.e., new) fear contingencies. Moreover, this experiment employed stimuli belonging to different visual categories (houses versus faces) across different trials of the same condition, allowing us to study the similarity between pattern responses to the presentation of a house versus a face as a CS. The results, shown in Figure 8, hinted at a main effect of fear relevance, CS^{+} versus CS^{-}, $F(1,17) = 4.00$, $p = .062$, suggesting that object category independence (i.e., house-face pattern similarity) indexed activation of a threat representation. More importantly, there was a three-way interaction between CS type, CS novelty, and amygdala side, $F(1,17) = 5.38$, $p = .033$, that was qualified by a two-way interaction between CS type and novelty in the right amygdala, $F(1,17) = 5.36$, $p = .033$, but not the left amygdala, $F(1,17) = .133$, $p = .720$. More specifically, the right amygdala again differentiated between the processing of novel (i.e., merely instructed) and old (i.e., previously instructed and experienced) fear contingencies (Figure 8c-h). Namely, the similarity between faces and houses was higher for CS^{+}old than CS^{-} old, $t(18) = 2.653$, $p = .016$, but not for CS^{+}new than CS^{-}new, $t(18) = -.217$, $p = .831$. This result further supports our conclusion that the right amygdala carries a trace of the CS-US contingency experience made during Pavlovian fear conditioning. No other fear-related regions (displayed in

Figure 8 for illustrative purposes only) showed a similar CS type by novelty interaction (all *F*s < 1.782, *p*s > .200), again suggesting that the right amygdala was most sensitive in encoding a Pavlovian trace.

Discussion

Demonstrating a purely experience-based element in human Pavlovian fear conditioning and identifying its neural correlates has been a major goal of learning research over the past decades. While the strong phenomenological and functional homologies between human and non-human fear conditioning have always suggested language-independent processing in humans, too, previous efforts have not yielded conclusive evidence. Our new approach is not dependent on controversial methods to exclude verbal or conscious processing (Mitchell et al., 2009) and exploits recent advances in the multivariate analysis of neural signatures of fear learning and memory (Bach et al., 2011; Dunsmoor et al., 2013; Hauner et al., 2013; Li et al., 2008; Visser et al., 2011; 2013). Because of these unique features, we are able to provide much sought evidence for experience-based threat processing. Specifically, using pattern similarity analyses, our two experiments demonstrated that the right amygdala was the only fear-related region whose neural activation pattern to experienced CSs was different from its neural pattern to merely instructed CSs. Of course, our data do not imply that the right amygdala does not process instructed threat information. Rather, it appears to respond in a unique way, perhaps in the form of specialized neural ensembles, to experience-based threat information. More generally, the amygdala has also been implicated in other processes besides fear processing (Okon-Singer et al., 2015; Pessoa & Adolphs, 2010). Therefore, future studies should determine whether other separable experience-dependent neural traces in the right amygdala can be identified for other forms of learning as well.

The present findings seem to converge on those of previous fear conditioning studies that were set up to single out unconscious fear conditioning. However, as noted above, the present study was not designed to study unconscious or implicit fear conditioning. Instead, the

manipulations in the present study were made very explicit: participants were very much made aware of the instructions and the actual CS-US pairings. This way, our study tried to create conditions for instructed fear learning to override a hypothesized experience-based component to fear learning. Our results show that these instructions were successful in evoking a similar neural response to the merely instructed as compared to the instructed and experienced stimulus in most fear related regions, except in the right amygdala. It is possible that the present experience-based trace in the right amygdala is related to the one identified in previous unconscious fear conditioning studies. However, the present study cannot (and did not aim to) prove the hypothesized unconscious nature of this experience-based trace (Öhman & Mineka, 2001). Therefore, it does not distinguish between an experience-based memory trace that is generated fully automatically and unconsciously versus one that relies on conscious contingency knowledge.

The present observation of apparent hemispheric differences in threat processing in the amygdala also adds to another important question in fear research, more specifically, concerning the presence of amygdala lateralization (Baker & Kim, 2004; Sergerie et al., 2008). Our results clearly suggest that the right amygdala shows a separable neural response to actual CS-US pairings, whereas the left amygdala appears to be more susceptible to fear instructions. These findings are concordant with previous (lesion) studies suggesting that the right amygdala is associated with a fear response to experiencing negative events, while the left amygdala is more responsive to the verbally-mediated cognitive representation of fear (Funayama et al., 2001; Phelps et al., 2001).

Last, our findings also fit well with those of another recent study by Atlas and colleagues (2016). In this study, Atlas and colleagues used two CSs which were both predictive of the US, but in different phases of the experiment. Crucially, they contrasted a condition in which this reversal in contingencies was always instructed to a condition where it was not, and observed that the amygdala was more responsive to the actual (changes in) contingencies rather than the instructions that preceded those. Interestingly, their analyses did not show a hemispheric

difference, in contrast to earlier findings suggesting that the left amygdala can be responsive to instructions (present results; Funayama et al., 2001; Phelps et al., 2001). However, as also argued in their discussion, Atlas and colleagues (2016) focused on the effects of instruction in changing environments and examined dynamic learning-related responses, whereas our study employed a manipulation where the contingencies were not changing, and instructions were given full opportunity to override the hypothesized Pavlovian trace.

In sum, across two experiments, we investigated neural pattern responses in fear related regions to experience-based fear processing in the face of verbal fear instructions. Our results show that verbal instructions were successful in evoking a similar neural response in fear-related regions to merely instructed versus instructed and experienced fear stimuli, except for the right amygdala. Instead, the human right amygdala showed a Pavlovian trace, suggesting it to be more sensitive to the actual experience of CS-US contingencies.

Author Contributions

S.B., J.D.H., R.K., & M.B. designed the experiment. S.B. & J.D. programmed and conducted the experiments. S.B., J.D., K.S.L.Y., & R.K. analyzed the data. S.B., J.D.H., J.D., R.K., & M.B. wrote the paper.

References

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016). Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. Elife, 5, e15192.

Bach, D. R., Weiskopf, N., & Dolan, R. J. (2011). A stable sparse fear memory trace in human amygdala. *The Journal of Neuroscience, 31(25),* 9383-9389.

Baker, K. B., & Kim, J. J. (2004). Amygdalar lateralization in fear conditioning: evidence for greater involvement of the right amygdala. *Behavioral neuroscience, 118(1),* 15.

Bissière, S., Plachta, N., Hoyer, D., McAllister, K. H., Olpe, H. R., Grace, A. A., & Cryan, J. F. (2008). The rostral anterior cingulate cortex modulates the efficiency of amygdala-dependent fear learning. *Biological psychiatry, 63(9),* 821-831.

Carlson, J. M., Cha, J., & Mujica-Parodi, L. R. (2013). Functional and structural amygdala–anterior cingulate connectivity correlates with attentional bias to masked fearful faces. *Cortex, 49(9),* 2595-2600.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2002). Fear conditioning in humans: the influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron, 33(4),* 653-663.

Dolan, R. J., & Vuilleumier, P. (2003). Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences, 985(1),* 348-355.

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2013). Aversive learning modulates cortical representations of object categories. *Cerebral Cortex, 24,* 2859-2872.

Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron, 51(6),* 871-882.

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2015). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry.*

Funayama, E. S., Grillon, C., Davis, M., & Phelps, E. A. (2001). A double dissociation in the affective modulation of startle in humans: effects of unilateral temporal lobectomy. *Journal of Cognitive Neuroscience, 13(6),* 721-729.

Geerligs, L., & Henson, R. N. (2016). Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *NeuroImage, 135,* 16-31.

Grillon, C. (2009). D-cycloserine facilitation of fear extinction and exposure-based therapy might rely on lower-level, automatic mechanisms. *Biological psychiatry, 66(7),* 636-641.

Hauner, K. K., Howard, J. D., Zelano, C., & Gottfried, J. A. (2013). Stimulus-specific enhancement of fear extinction during slow-wave sleep. *Nature neuroscience, 16(11),* 1553-1555.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293(5539),* 2425-2430.

Kim, M. J., Loucks, R. A., Palmer, A. L., Brown, A. C., Solomon, K. M., Marchante, A. N., & Whalen, P. J. (2011). The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. *Behavioural Brain Research, 223(2),* 403-410.

Knight, D. C., Waters, N. S., & Bandettini, P. A. (2009). Neural substrates of explicit and implicit fear memory. *Neuroimage, 45(1),* 208-214.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2.

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences, 111(8),* 2871-2878.

Li, W., Howard, J. D., Parrish, T. B., & Gottfried, J. A. (2008). Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science, 319(5871),* 1842-1845.

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annual review of neuroscience, 24(1),* 897-931.

Mechias, M. L., Etkin, A., & Kalisch, R. (2010). A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. *Neuroimage, 49(2),* 1760-1768.

Mertens, G., Kuhn, M., Raes, A. K., Kalisch, R., De Houwer, J., & Lonsdorf, T. B. (2016). Fear expression and return of fear following threat instruction with or without direct contingency experience. *Cognition and Emotion, 30(5),* 968-984.

Mineka, S., & Öhman, A. (2002). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological psychiatry, 52(10),* 927-937.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32(02),* 183-198.

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393(6684),* 467-470.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences, 10(9),* 424-430.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review, 108(3),* 483.

Okon-Singer, H., Hendler, T., Pessoa, L., & Shackman, A. J. (2015). The neurobiology of emotion–cognition interactions: fundamental questions and strategies for future research. *Frontiers in human neuroscience*, 9.

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science, 15(12),* 822-828.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature neuroscience, 10(9),* 1095-1102.

Pavlov, I.P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. London: Oxford Univ. Press

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience, 11(11),* 773-783.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature neuroscience, 4(4),* 437-441.

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLOS ONE, 9(1),* e84888.

Schultz, D. H., & Helmstetter, F. J. (2010). Classical conditioning of autonomic fear responses is independent of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes, 36(4),* 495.

Sergerie, K., Chochol, C., & Armony, J. L. (2008). The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews, 32(4),* 811-830.

Tabbert, K., Merz, C. J., Klucken, T., Schweckendiek, J., Vaitl, D., Wolf, O. T., & Stark, R. (2011). Influence of contingency awareness on neural, electrodermal and evaluative responses during fear conditioning. *Social cognitive and affective neuroscience, 6,* 495-506.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23(1),* 87-102.

Van Marle, H. J., Hermans, E. J., Qin, S., & Fernández, G. (2010). Enhanced resting-state connectivity of amygdala in the immediate aftermath of acute psychological stress. *Neuroimage, 53(1),* 348-354.

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *The Journal of Neuroscience, 31(33),* 12021-12028.

Visser, R. M., Scholte, H. S., Beemsterboer, T., & Kindt, M. (2013). Neural pattern similarity predicts long-term fear memory. *Nature neuroscience, 16(4),* 388-390.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3(1),* 1-14.

Williams, L. M., Das, P., Liddell, B. J., Kemp, A. H., Rennie, C. J., & Gordon, E. (2006). Mode of functional connectivity in amygdala pathways dissociates level of awareness for signals of fear. *The Journal of neuroscience, 26(36),* 9264-9271.

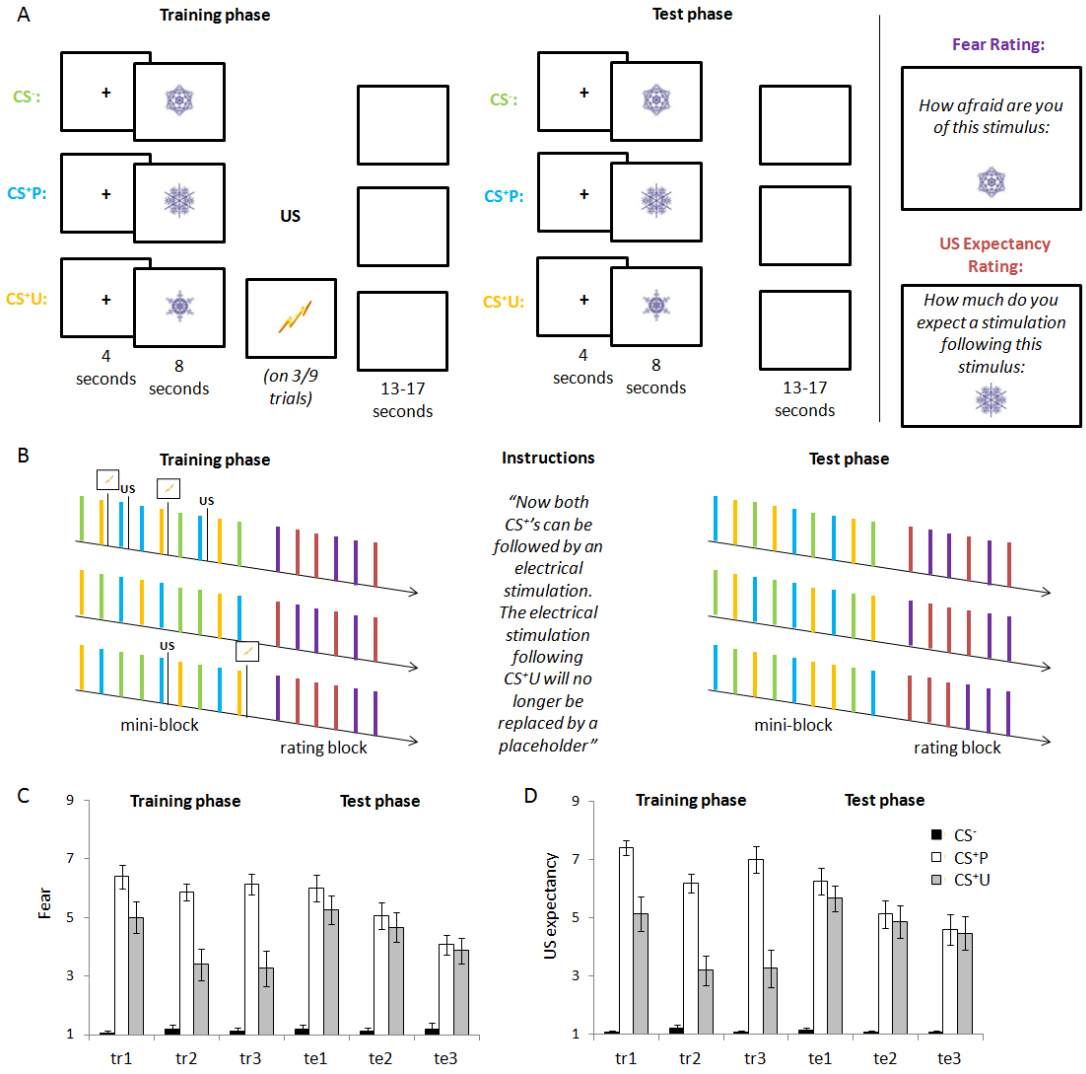

Figure 1. Procedure Experiment 1 and behavioral results. A. Experiment trials and rating screens. Each trial started with the presentation of a fixation cross, followed by the presentation of a CS. In the training phase, the CS$^+$P was occasionally followed by a US (electrical stimulation), the CS$^+$U by a US placeholder (picture of a lightning bolt), and the CS$^-$ was never followed by either a US or placeholder. In the test phase, none of the CSs was followed by a US or placeholder. Rating screens assessed participant's fear experience and US expectancy associated with each of the CSs. B. Experimental procedure and instructions. Before training, subjects were instructed that only the CS$^+$P could be followed by the US, while the CS$^+$U could only be followed by the placeholder. Before testing, subjects were told to expect USs after both CS$^+$s. Both phases consisted of three mini-blocks (where every CS was randomly presented three times), each followed by a rating block. Three out of nine CS$^+$ presentations were

followed by either the US or the US placeholder. C. Mean fear ratings. D. Mean US expectancy ratings. The error bars are ±1 standard error of the mean (SEM). tr=training; te=test.

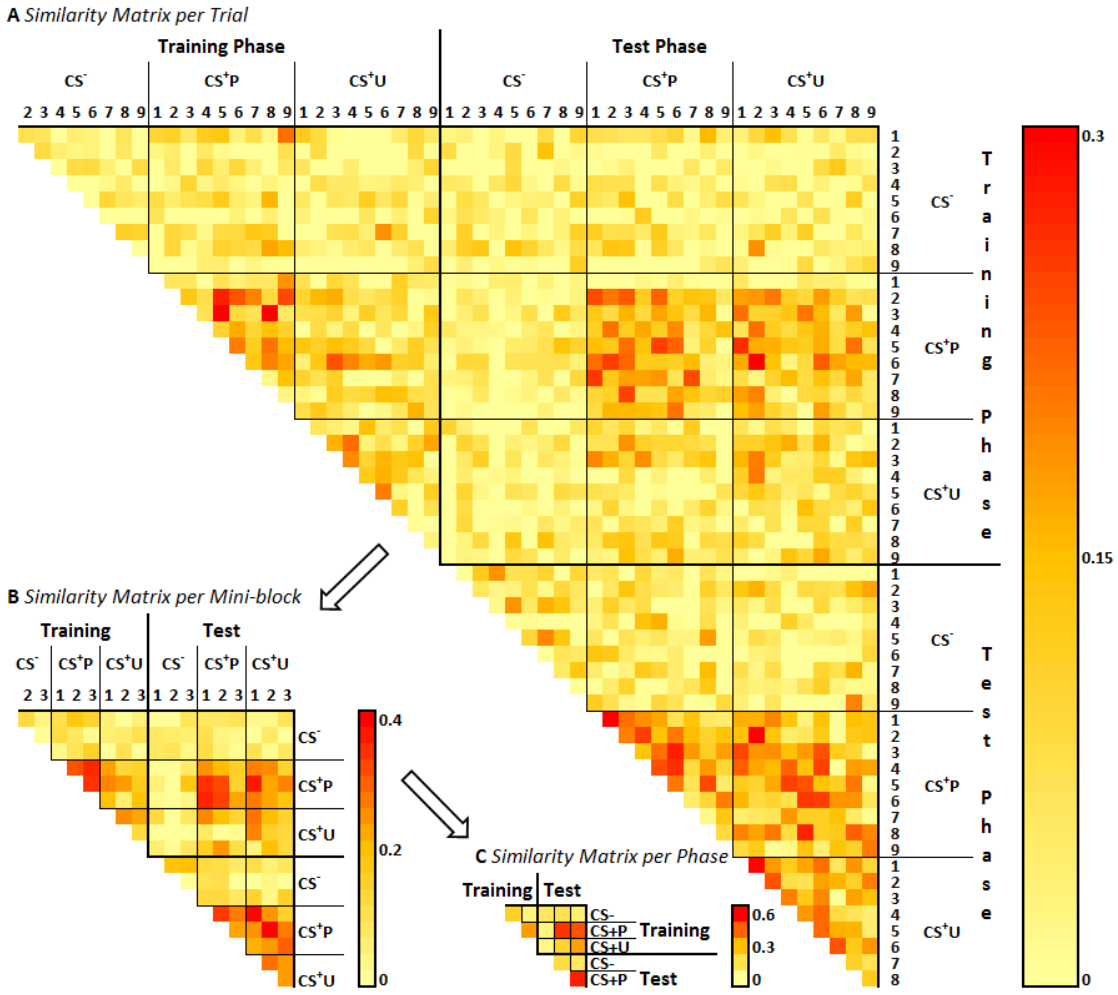

Figure 2. An example of the different types of similarity matrices in the anterior cingulate cortex (Experiment 1). The similarity matrices represent color-coded average Pearson correlation coefficients across subjects, for every possible correlation between all different CS-presentation regressors in the trial-based model (A), mini-block model (B), or the phase model (C). The vertical bars adjacent to each matrix indicates its color coding depending on the correlation coefficient.

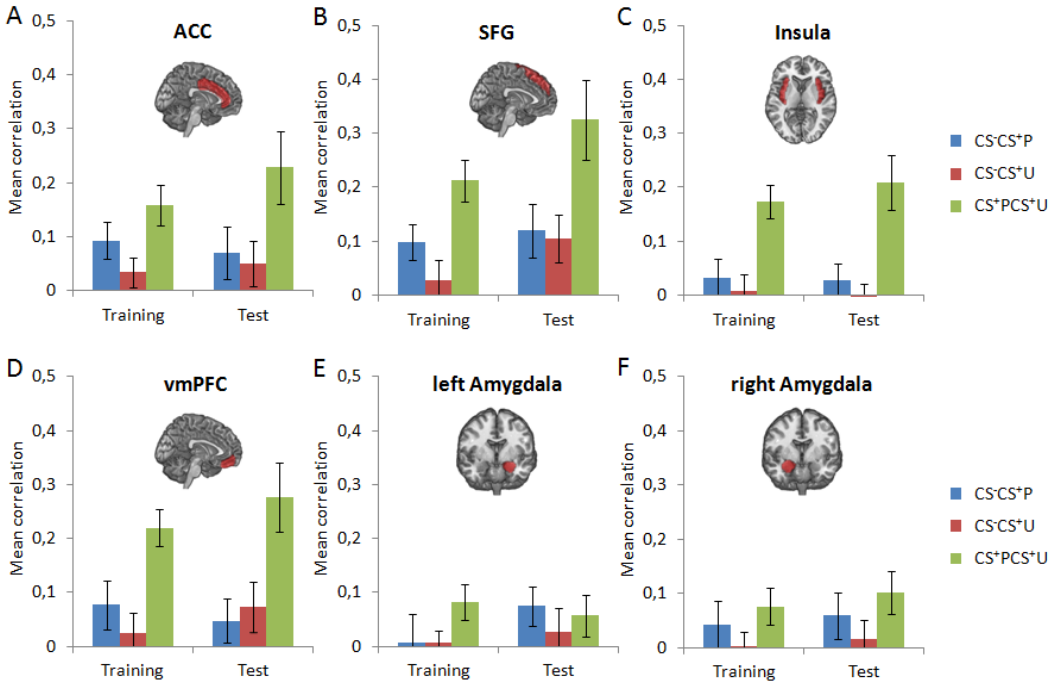

Figure 3. Inter-CS similarities per region and experimental phase (Experiment 1), based on the mini-block model (see Figure 2). More similar multi-voxel activation patterns between the two CS^+s than between a CS^+ and the CS- indicate processing of learned stimulus value in ACC, SFG, Insula, and vmPFC. The amygdala does not exhibit higher similarity between the instructed and experienced CS^+ (CS^+P) and the merely instructed CS^+ (CS^+U) relative to the similarity between these CS^+s similarities and the CS-. The error bars are ±1 SEM.

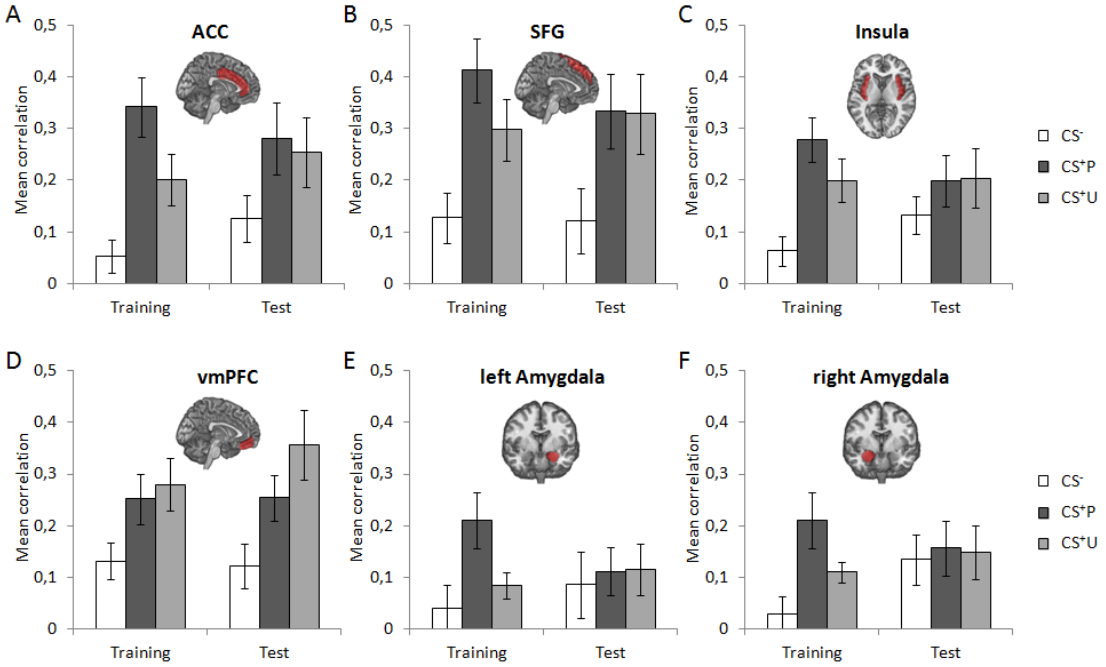

Figure 4. Intra-CS similarities from mini-block to mini-block per CS type, region and experimental phase (Experiment 1), based on the mini-block model (see Figure 2). Higher temporal consistency in intra-CS^+ than intra-CS^- similarities indicates processing of learned stimulus qualities in a given ROI. This is not observed for the merely instructed CS^+ (CS^+U) in the amygdalae. The error bars are ±1 SEM.

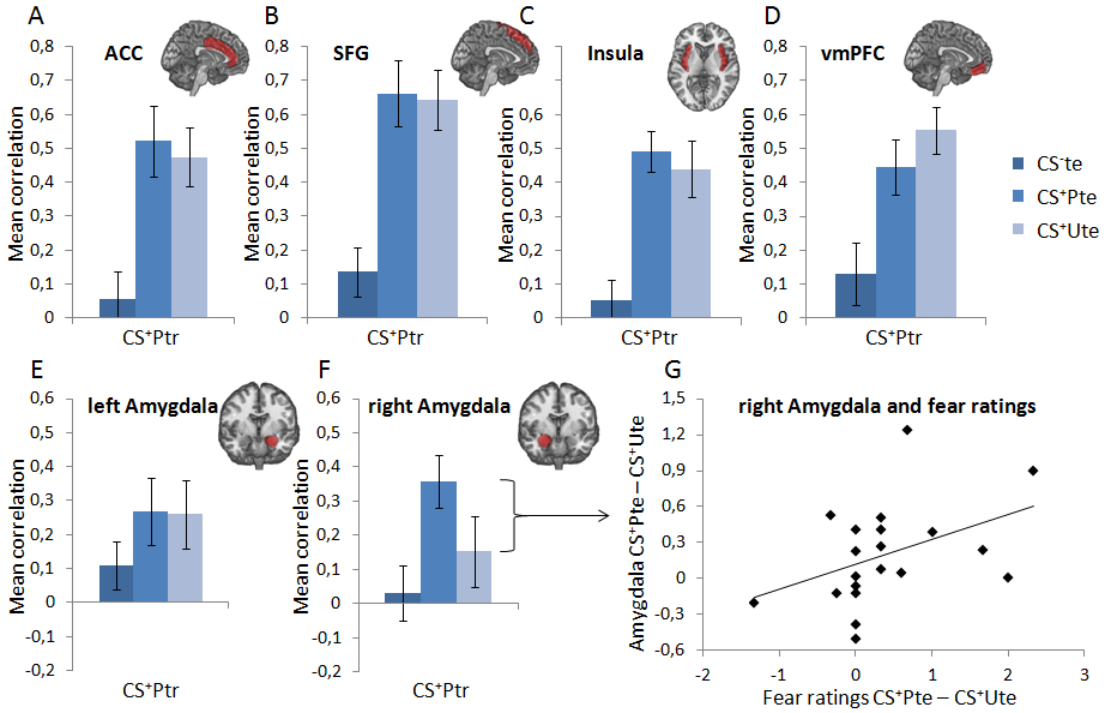

Figure 5. Comparison of inter-CS similarities between each of the three different CS types from the test phase (CS^{-}te, CS^{+}Pte, CS^{+}Ute) and the CS^{+}P pattern from the training phase (CS^{+}Ptr), based on the phase model (see Figure 2), reveals CS^{+}P specific processing of threat-related information in the right amygdala (Experiment 1): CS^{+}P associated multi-voxel activation patterns during test (CS^{+}Pte) are more similar to CS^{+}P associated patterns during training (CS^{+}Ptr) than CS^{+}U associated patterns during test (CS^{+}Ute) (F). Other regions do not show such differentiation (A-E). The error bars are ±1 SEM. G. Individual differences in the difference between CS^{+}PtrCS^{+}Pte and CS^{+}PtrCS^{+}Ute inter-CS similarities in the right amygdala were correlated with differences in fear ratings between CS^{+}P and CS^{+}U in the test phase.

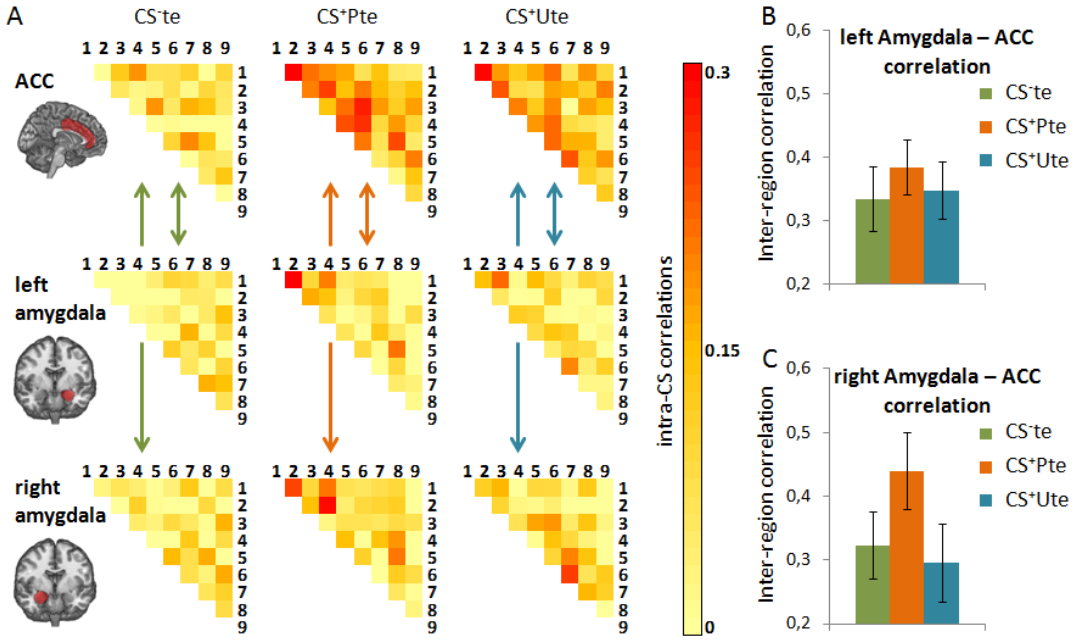

Figure 6. Inter-region similarity analyses between the ACC and left and right amygdala (Experiment 1). A shows trial-by-trial intra-CS similarity matrices from the trial-based models for each region and CS type. On this basis, Spearman correlation coefficients were calculated between each combination of the resulting trial-by-trial ACC and left and right amygdala intra-CS similarity time courses, separately for each CS. The right amygdala showed a higher inter-region similarity with the ACC for the $CS^{+}P$ specifically (B), relative to the left amygdala where no such effect was observed (C). The error bars are ±1 SEM.

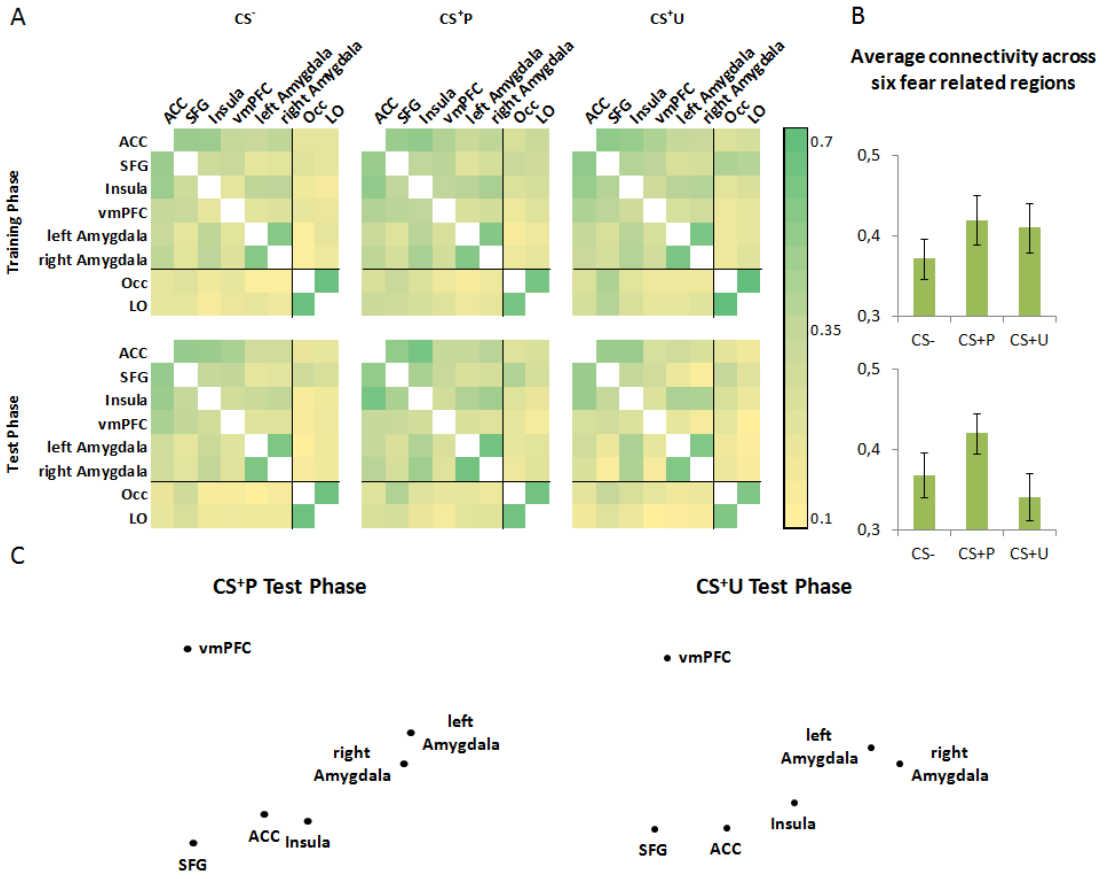

Figure 7. Inter-region similarity analyses per different CS-type and experimental phase. A. Correlations were calculated between each combination of two regions' intra-CS similarity matrices from the trial-based models (see Figure 2), per CS and phase separately, as explained in Figure 6. B. The correlations across all fear-related regions were averaged and are presented per CS type and phase separately. C. Visual two dimensional scaling depiction of the similarities between different regions for each CS^+ in the test phase separately.

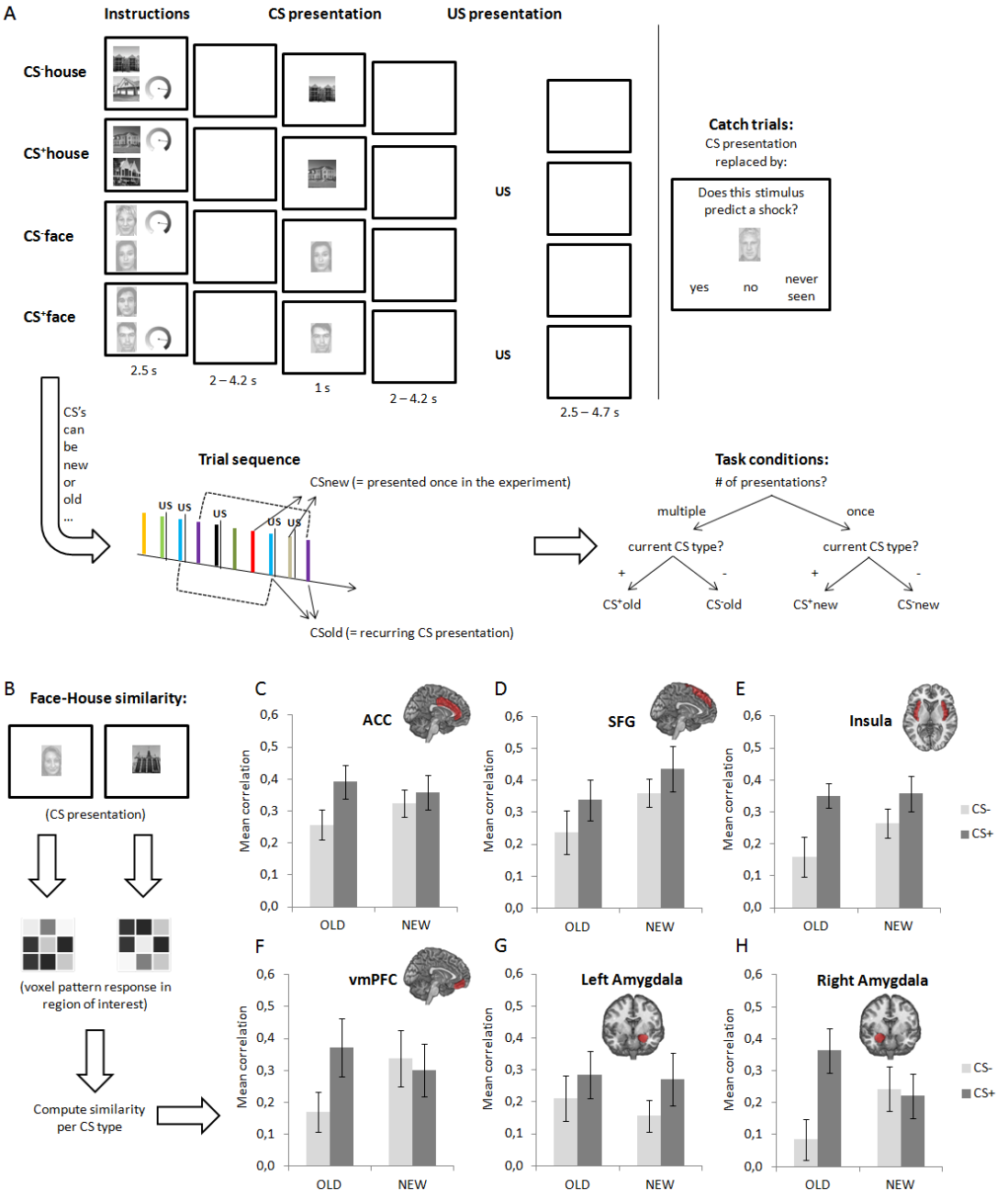

Figure 8. Procedure Experiment 2 and results. A. Each trial consisted of a fear contingency instruction and a CS presentation. The instruction indicated which of two pictures (CS^+) would be followed by an electrical US by presenting an intensity meter next to that picture. As illustrated on the lower left part of A, some CSs and instructions were recurring (OLD), others were always novel (NEW). Orthogonal to this, some instructions and subsequent CS presentation employed pictures of houses, others of faces. CS^+ presentation was always followed by a US presentation. On a small subset of trials, CS presentation was replaced by a catch question, to assure participants paid attention to the experiment. B. The similarity

analyses focused exclusively on pattern similarities between face and house pictures during CS presentations, for each CS type separately (CS^+old, CS^-old, CS^+new, and CS^-new). C-H. These analyses revealed that the right, but not the left, amygdala showed a differential response to fear relevance as a function of novelty. Namely, the pattern response for houses and faces were more similar when these denoted a CS^+ than when they indicated a CS^-. However, this difference was only present when the CSs had been instructed and experienced before (old CSs), but disappeared when they were novel. The error bars are ±1 SEM.