Title Page

Selective Extinction Through Cognitive Evaluation: Linking Emotion Regulation And Extinction

Authors:

Birthe Macdonald[1,2] birthe.macdonald@uzh.ch

Shannon Wake[1]

Tom Johnstone[1,3]

[1]University of Reading, Whiteknights, Reading RG6 6AL, UK

[2]University of Zurich, UFSP Dynamics of Healthy Aging, Andreasstrasse 15, 8050 Zurich, Switzerland

[3]Swinburne University of Technology, John Street, Hawthorn, Victoria 3122, Australia

Keywords: fear conditioning, fMRI, skin conductance, cognitive reappraisal

ABSTRACT

The extinction of a previously conditioned response can be modulated through cognitive processes, including feature-based information, and explicit instruction. Here we introduce a Selective Extinction through Cognitive Evaluation (SECE) task in which information is cognitively evaluated on a trial-by-trial basis to ascertain the extinction contingencies. Participants were conditioned to expect an electric shock during the presentation of one of two letters (CS+/CS-). During the SECE task, the letters were presented within words belonging to two categories, one of which indicated safety (COG-_CS+ trials), while risk of shock was maintained for the other category (COG+_CS+ trials).

Skin conductance responses indicated that participants reduced their response to COG-_CS+ trials compared to COG+_CS+ trials. Clusters in bilateral insula and anterior cingulate cortex showed activation for COG+_CS+ trials that was reduced for COG-_CS+ trials. A network of brain regions including left inferior frontal gyrus, and bilateral temporal and parietal cortices showed greater activation for COG-_CS+ versus COG+_CS+ trials. This is consistent with the semantic processing and decision making necessary to evaluate the trial contingencies.

We compared activation in the SECE task to activation in a cognitive reappraisal task in which participants were asked to attend to, or regulate their emotional reactions to affective IAPS images. This task replicated prefrontal activation seen in previous reappraisal studies. A voxel-wise conjunction analysis found no overlap between the cognitive reappraisal and the SECE task, but we did find evidence for common activation in follow-up ROI analyses, supporting the idea of common lateral prefrontal mechanisms involved in both processes.

INTRODUCTION

The ability to adaptively respond to cues in the environment that signal benefit or harm is vital for mental health and well-being. Adaptive responding requires that we not only learn to

respond to the initial association between cue and outcome, but that we can regulate such responses when the cue-outcome contingency changes. A number of mental health disorders have been associated with patients failing to adapt their behaviour to stimuli that no longer predict a previously learned consequence (Amstadter, 2008; Gross & Levenson, 1997; Gross & Muñoz, 1995; Joormann & Vanderlind, 2014; Khoury & Lecomte, 2012). Various approaches have been applied to investigate the neural mechanisms by which emotional responses, particularly to aversive stimuli, can be regulated and how their dysfunction is related to psychopathology (e.g. Dibbets, Broek, & Evers, 2015; Mochcovitch, da Rocha Freire, Garcia, & Nardi, 2014; Picó-Pérez, Radua, Steward, Menchón, & Soriano-Mas, 2017).

One way to investigate the mechanisms that support adaptive responding to stimuli that change in their ability to predict negative outcomes, is research into the extinction of conditioned fear. Through fear conditioning, a previously neutral stimulus (CS+) that is repeatedly paired with an aversive, unconditioned stimulus (US), elicits a conditioned response (CR) even when the US is no longer present (Pavlov, 1927; Rescorla, 1968; Watson & Rayner, 1920). During fear extinction, the US is removed, and over a number of trials the CS+ ceases to elicit the CR. The extinction of conditioned CR's serves to reduce the affective response to the CS+, however, the CS-US association is not unlearnt; instead, a representation of the new contingencies (i.e. US omission) is formed which down-regulates the original CR. This is evident by the spontaneous return of conditioned fear after initial fear conditioning and extinction (Bouton, 2002; Huff, Hernandez, Blanding, & LaBar, 2009; Milad & Quirk, 2012; Schiller et al., 2008).

On a neural level it has been found that the amygdala, which receives visual information directly via the thalamus, is involved in both conditioning and extinction (Kim & Jung, 2006). Conditioned responses are also associated with increased activation in the insula and dorsal anterior cingulate cortex (dACC) via information from sensory cortices and reflecting

responses to painful stimuli as well as the increased salience of the CS+ (Etkin, Egner, & Kalisch, 2011; Mechias, Etkin, & Kalisch, 2010). Further, the vmPFC has regularly been found to be involved in extinction, processing the changing contingencies between conditioning and extinction phases (Milad & Quirk, 2002; Milad et al., 2007; Morgan, Romanski, & LeDoux, 1993; Morgan & LeDoux, 1995; Phelps, Delgado, Nearing, & LeDoux, 2004). However, in a recent meta-analysis Fullana et al (2018) report that the vmPFC is not consistently activated in extinction. This result, alongside the finding that the amygdala is also not always found to be active during conditioning and extinction, may reflect inconsistencies in experimental design across studies (Morriss, Hoare, & van Reekum, 2018) related to different conceptualisations of the underlying mechanisms.

Although conditioning and extinction have often been described in terms of perceptual and reflexive associative processes, a compelling argument has been made that a cognitive component is also involved (Lovibond, 2004). This idea is supported by several studies investigating Conditioned Inhibition (Rescorla, 1969), Learned Safety (Pollak et al., 2010), Reversal Learning (Atlas, 2019; Rolls, Critchley, Mason, & Wakeman, 1996), and Occasion Setting (Holland, 1992; Trask, Thrailkill, & Bouton, 2017) all of which result in CR reduction in a safe condition. During occasion setting, the CR is modulated by a feature or context that predicts the presence or absence of the US (Holland, 1992; Trask et al., 2017). Explicit safety signals preceding a previously reinforced CS+ can reduce CR in low anxious human participants (Grillon & Ameli, 2001), and CR's can be reduced by explicit instruction that the CS+ will no longer be reinforced (Luck & Lipp, 2016). Similarly, CR's can be reduced via explicit cognitive strategies (Delgado, Nearing, Ledoux, & Phelps, 2008), in that CR's can be cognitively modulated through the use of conditioned safety signals or overt instructions. Evidence suggests that the neural circuitry involved in such adaptive responding, including the vmPFC, overlaps with extinction circuitry (Atlas, 2019; Delgado et al., 2008;

Rolls et al., 1996; Zhou et al., 2019). Hartley & Phelps (2010) have extended this concept to emotion regulation through reappraisal, an elaborate cognitive process of changing the meaning of emotional stimuli or scenarios (Jackson, Malmstadt, Larson, & Davidson, 2000; Johnstone, van Reekum, Urry, Kalin, & Davidson, 2007; Urry et al., 2006). Hartley & Phelps suggest that at least some of the cognitive processes (i.e. context-dependent updating of stimulus value leading to the reduction of an initial response), as well as the brain regions involved in these processes (i.e. the amygdala, insula, vmPFC) are shared with extinction of conditioned fear. In summary, it has been shown that extinction can be invoked and modulated by the presence of additional stimuli or contexts, or by direct instruction.

The principal aim of this study was to extend this process of context dependent updating to include not just one feature that is simply present or absent, but classes of features that require further cognitive evaluation on a trial-by-trial basis to predict the likelihood of the US occurring. As in real-life situations, participants were required to consider all the information presented to them to predict whether the US is likely to occur. Therefore, this novel design extends previous, signal-, feature- or instruction-based extinction by including an additional cognitive component.

In selective extinction through cognitive evaluation (SECE), participants are first conditioned to expect an electric shock following the presentation of a CS+, in this case, letters. In a subsequent phase, the same letters are presented again, this time briefly surrounded by words belonging to two distinct categories (COG+/COG-). Participants are told that one of these word categories will be “safe” (COG-), i.e. no shock is expected when this word category is presented while the other category is “dangerous” (COG+), i.e. a risk of shock is to be expected when this word category is presented. In order to work out whether an electric shock is likely to occur, participants must a) be aware of the conditioning contingencies, i.e. which letter was the CS+ in the conditioning phase, b) work out which word category is

dangerous when the first shock is presented during the SECE phase, and c) apply this knowledge to each subsequent trial. In contrast to occasion setting, conditioned inhibition, and instructed extinction, in the SECE task repeated cognitive evaluation of the CS and its context is necessary to predict the likelihood of electric shock on a trial-by-trial basis. This also has an advantage of allowing for repeated observations and therefore increased statistical power.

Research Questions 1 & 2: Does a reduction in the CR occur when participants need to explicitly cognitively evaluate additional information to predict whether the US will occur, and which regions of the brain are involved in this process?

We predicted that *COG-_CS+* (i.e. *safe*) trials would result in down-regulation of the CR, whereas the CR would be maintained during *COG+_CS+* (*dangerous*) trials. On a neural level, we hypothesized that this would result in increased activation in regions related to semantic processing and decision making during safe (COG-_CS+) compared to COG+_CS+ trials, representing the repeated cognitive and affective evaluation of the additional information, and the implementation of new stimulus-affect contingencies. We expected brain regions associated with affective responses such as the insula and dACC to be more active during COG+_CS+ than during *COG-_CS+* trials.

Research Question 3: Is there evidence of overlap in neural activation associated with the SECE task and activation associated with a matched cognitive reappraisal task?

After the SECE task, we also ran a shortened version of a cognitive reappraisal emotion regulation task. Although the cognitive reappraisal task was matched to the SECE task as closely as possible in terms of timing and number of trials, the two tasks differ greatly in complexity and demand on participants. Our intention was not a direct comparison between the SECE and the cognitive reappraisal task, but rather to ascertain whether there was overlap

in activation in the same participants, as would be expected by Hartley & Phelps' (2010) suggestion that extinction and cognitive reappraisal share common neural circuits.

Materials and Methods

Participants

20 (8 male) participants were recruited via the University of Reading Research Panel and through university-wide emails. Participants were eligible if they were between 18-55 years old, right-handed, and had never been diagnosed with a psychological disorder. All participants were screened for their MRI suitability and provided informed consent. Mean age was 27 years (range: 18 – 44). Participants received £10 and a picture of their brain to thank them for their participation. This study was run in accordance with the Declaration of Helsinki (1991, p.1194) and reviewed and approved by the University of Reading Research Ethics Committee.

One participant's SCR data had to be excluded from the analysis of both tasks due to excessive noise introduced by the MRI scanner that masked true responses. Two participants' data were excluded from the analysis of the SECE task because they did not show any SCR's to the US (electric shock) during the task. This resulted in a final sample size of 17 for the SCR analysis of the SECE task.

One participant did not wish to participate in the cognitive reappraisal task, resulting in a sample size of 19 for the fMRI, and 18 for the SCR analysis of the reappraisal task.

SECE Task

Conditioning Phase

During the first phase of this experiment one of two letters (B and T) was paired with an electric shock (10 pulses at 100hz) 50% of the time. The letter that served as the CS+ was

counterbalanced between participants. Conditioning consisted of 30 trials: 10 CS-, 10 reinforced CS+, and 10 non-reinforced CS+. Each trial lasted for 4000ms, and reinforced CS+ trials co-terminated with an electric shock. ITI's were jittered between 2000 and 6000ms throughout the task.

SELECTIVE EXTINCTION THROUGH COGNITIVE EVALUATION (SECE) PHASE

The SECE phase of the task was carried out in a 2x2 design (CS+ vs CS-, COG- vs COG+ word category, see Table 1). 20 unreinforced trials of each condition were presented COG+_CS+, COG-_CS+, COG+_CS-, COG-_CS-), with an additional 20 reinforced COG+_CS+ trials. Descriptive statistics of the word categories are presented in Table 2 and 3. In an additional 20 trials (10 CS+, 10 CS-), the letters were shown by themselves to allow us to check if the words influenced physiological and neural activation in general, compared to no-word trials. Finally, to ensure that participants were paying attention, 2 words were shown that did not belong to either word category. When presented with these words participants were instructed to respond with a button press (see Table 3). The trial structure is shown in Figure 1. For the first 500ms of each trial, the letter was presented on its own. The word then appeared around it (with the letter in the correct place but emphasized in upper case) for 1000ms. The word then disappeared and the letter remained on screen by itself for an additional 4000ms. The SECE phase was comprised of a total of 126 trials, the first 4 of which were non-reinforced. During reinforced trials the electric shocks were delivered so that they co-terminated with the trial. Only non-reinforced trials (i.e. trials with no electric shock) were included in the analyses.

Letters representing the CS+ and CS- as well as "safe" (COG-) and "dangerous" (COG+) categories were counterbalanced across participants, and stimuli were presented in pseudo-random order, with no condition occurring more than twice in a row.

Stimuli were presented using EPrime 2.0 via a fibre-optic goggle system, screen resolution 1024 x 768 (NordicNeuroLab AS, Bergen, Norway).

Table 1. Design of the SECE task

Conditioning		**SECE**		**Letters only**
		COG- *(animals or plants)*	COG+ *(animals or plants)*	*no further reinforcement*
CS+	10 (+ 10 reinforced)	20	20 (+ reinforced)	10
CS-	10	20	20	10
		Additional trials: 1 trial of each category at the beginning of the SECE phase, 2 trials with words that were *not* animals or plants		

Table 2. Means and Standard Deviations in word frequency based on the British National Corpus (Kilgariff, 1997). No significant effect was found of letter (t(38) = 1.51, p = 0.14) or of category (t(38) = -1.41, p = 0.17).

		Animal	**Plant**	**Total**
		Mean (SD)	Mean (SD)	Mean (SD)
T	Mean (SD)	204 (175.09)	246.1 (256.11)	225.05 (214.61)
B	Mean (SD)	84.2 (61.18)	198.8 (144.1)	141.5 (122.75)
Total	Mean (SD)	144.1 (141.67)	222.45 (203.71)	

Table 3. List of words used in the SECE task, with correct ratings of category and typicality for that category.

word	correct category rating	typicality rating
Banana	1	4.39
Basil	1	5.56
Bean	1	4.56
Birch	0.89	4.33
Broccoli	1	4.28
cabBage	0.94	4.89
cucumBer	1	4.33
mulBerry	0.94	3.56
raspBerry	1	4.33
roseBush	0.94	5.44
Average	0.971	4.567
carnaTion	0.83	4.33
carroT	1	4.5
letTuce	1	5
minT	1	4.61
palmTree	1	5

poTato	1	4.39
tomaTo	1	4.78
Tulip	1	5.72
Turnip	1	3.78
waTercress	1	4.39
Average	0.983	4.65
blackBird	1	5.28
blueBird	1	4.61
Buzzard	0.94	4.22
cariBou	0.72	3.22
coBra	1	4.72
gerBil	0.78	4.61
lamB	1	6.5
mockingBird	1	4.33
roBin	1	5.22
zeBra	1	6.11
Average	0.944	4.882
caTerpillar	1	4.61
cheeTah	1	5.44
osTrich	1	5.33
panTher	1	5.56
parroT	1	5.44
pheasanT	0.94	4.67
sTingray	1	3.78
sTork	0.94	4.56
Tiger	1	6.39
Turkey	1	5.33
Average	0.988	5.111
wardroBe		
plaTe		

Note: Prior to this study, 10 participants (who did not later take part in the study) were asked to rate whether each word represented an animal or a plant. Column 2 shows the average correctness of this rating. Participants were also asked how typical of this category they felt each word was. This was rated on a 7-point scale: 1 = “not at all typical”, 2 = “very typical”, 3 = “quite typical”, 4 = “neither typical nor untypical”, 5 = “quite typical”, 6 = “very typical”, “7 = “extremely typical”.

INSTRUCTIONS

For the SECE task, participants were told that they would see two letters during the first phase, one of which would be associated with the risk of electric shock. They were then told

that after a short break, they would see the same letters again, and this time a word would appear briefly. Most of these words would belong to two distinct categories: plants or animals, and one of these categories was safe, meaning that regardless of the letter presented, they would not be shocked on such trials. Participants were asked to determine which word category was safe during this phase, and to keep the contingencies in mind throughout. They were informed that they would be able to work out the task contingencies very quickly. They were also asked to focus on whether they thought they might receive an electric shock or not during each trial. Participants were not given any additional information about the letter only trials. Finally, participants were told to press a button on the rare occasion that they saw a word that did not belong to one of the two categories.

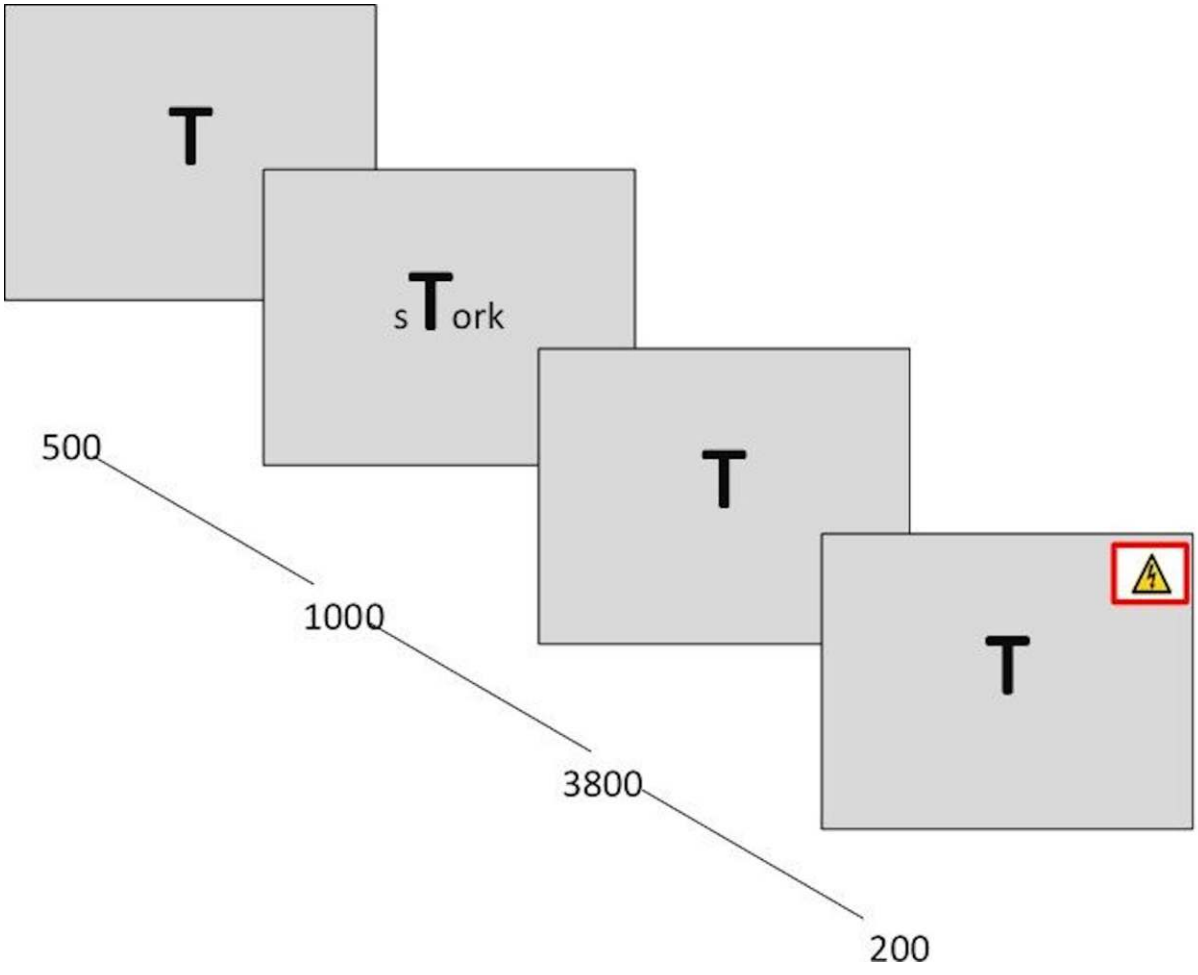

Figure 1. Example of a reinforced sece trial. Letters initially appeared on the screen by themselves for 500ms before the word appeared for 1000ms with the letter emphasized in bold and larger font. The word then disappeared and the letter

remained on screen by itself for a further 4000ms. In case of a reinforced COG+_CS+ trial, the trial co-terminated with the electric shock.

QUESTIONNAIRES

Questionnaires assessed trait anxiety (State Trait Anxiety Inventory, STAI, (Spielberger, Gorsuch, & Lushene, 1970), tendency to worry (Penn State Worry Questionnaire, PSWQ, (Meyer, Miller, Metzger, & Borkovec, 1990), emotion regulation capacity (Emotion Regulation Questionnaire, ERQ, (Gross & John, 2003) and intolerance of uncertainty (Intolerance of Uncertainty Scale, IUS, (Buhr & Dugas, 2002).

ELECTRIC SHOCK

Electric shocks were delivered through 2 Ag-AgCl electrodes attached to participants' right middle and ring fingers and connected to an ADInstruments Powerlab 26T Isolated Stimulator. The intensity was determined by the participant prior to the task through a procedure described below.

COGNITIVE REAPPRAISAL TASK

The basic design of the cognitive reappraisal task is described in detail in various publications (Jackson et al., 2000; Johnstone et al., 2007; Urry et al., 2006), and consists of a series of trials in which affectively valenced pictures are presented, with participants instructed to either decrease the emotional impact of the picture using reappraisal, or simply attend to the picture. In this study, 40 negative and 20 neutral images from the International Affective Picture System (IAPS, (Lang, Bradley, & Cuthbert, 2008)) were used. We selected highly negative and highly arousing pictures for this task (both ratings 1 – 9, 1= extremely negative/not at all arousing, 9= extremely positive/extremely arousing). The mean valence rating for the negative images was 1.99, SD = 0.26, arousal mean = 5.95, SD = 0.74. The average valence rating for the neutral images was 5.1, SD = 0.37, arousal mean = 3.46, SD = 0.46. Images were presented in pseudo-random order, ensuring that no more than 3 negative

images, and no more than 1 neutral image were presented in a row, and the order was counterbalanced across subjects. During 10 presentations of the negative images, participants were asked to decrease their emotional response to the image (Negative/Decrease), during the other 10 negative images, they were asked to maintain their attention to the image. During the presentation of the neutral images, participants were always asked to maintain their attention to the image.

The timing of the cognitive reappraisal task was amended to closely match that of the SECE task. Trials lasted for a total of 6000 ms and instructions were presented 1000 ms into the presentation of the picture through a set of Siemens (Siemens, Malvern, PA, USA) headphones for 1000 ms. The ITI was jittered between 2000 and 6000ms. Trials were presented through EPrime 2.0 in a pseudo-random order via a fibre-optic goggle system (NordicNeuroLab AS, Bergen, Norway) and the order of instructions was counterbalanced across subjects.

PHYSIOLOGICAL DATA COLLECTION

Skin conductance response (SCR) data was recorded at 1000 Hz with an ADInstruments PowerLab 26T and PowerLab ML116 SCR module using 2 Ag-AgCl electrodes on the distal phalanges of the middle and ring fingers of the participant's non-dominant hand (Cacioppo, Tassinary, & Berntson, 2007) . A low constant-voltage AC excitation of 22mVrms at 75 Hz was passed through the electrodes, which was converted to DC before being digitized and stored.

FMRI DATA COLLECTION

MRI data were collected on a 3T Siemens Trio MRI scanner with 12-channel head matrix coil at The University of Reading Centre for Integrative Neuroscience and Neurodynamics (CINN). Functional scans consisted of a t2*-weighted gradient echo, echoplanar imaging

(EPI) sequence (37 interleaved transverse slices, phase encoding P to A, 3 mm thickness, 128*128 matrix; 192 mm field of view; TR: 2000ms, TE: 30ms, Flip Angle: 90°; 904 whole-brain volumes for the SECE task, 484 whole-brain volumes for the instructed regulation task). A high-resolution structural image was also acquired using an MPRAGE sequence (176 x 1 mm slices,1*1mm voxels, TE: 2.9 ms, TR: 2020 ms, TI:1100 ms, FOV: 250 mm, Flip Angle: 90°). Fieldmaps to be used to correct for magnetic field distortion were acquired using a gradient echo sequence (P to A, 3*3*3mm voxel size, TE1: 5.19ms, TE2: 7.65ms, TR: 400ms, FOV: 192mm, Flip Angle: 60°).

PROCEDURE

Informed consent was obtained from each participant before completing the MRI screening form and the questionnaires.

Prior to entering the scanner, the stimulator electrodes were attached to participants' fingers, a shock at very low intensity (0.5mV) was delivered and the intensity was increased in steps of 0.5mV. After each shock, the participant was asked to rate the sensation on a scale of 1 ("not painful at all") to 10 ("extremely painful"). When they reached 8 on this scale, the experimenter reduced the intensity of the shock by 1 step and informed that this was the intensity the shock would remain at for the duration of the experiment (procedure based on (Delgado et al., 2008). Subsequently, sensors to collect skin conductance were attached.

Because the SECE task was the primary focus of the study, participants always completed the SECE task first, minimising potential habituation (neural and physiological) for this task. After this task, they were given a brief break before completing the instructed emotion regulation task. The high resolution T1-weighted scan was completed last. After the scanning session, participants were asked about the contingencies of the task, i.e. which letter was the CS+ and which category was dangerous, and their answers were recorded. They then filled in

a questionnaire to assess how they felt during each type of trial as well as how they felt throughout the task (7-point Likert scales, 1 = “not stressed”, 7 = “extremely stressed”, as well as 1 = “not bored” to 7 = “extremely bored”, and 1 = “not sleepy” to 7 = “extremely sleepy”).

Participants were then verbally debriefed and given a debrief sheet to take home with them.

Statistical Analysis

SCR data analysis

Data was visually checked for motion artefacts and these were removed manually. SCR data was filtered using a median filter with a width of 3 to remove artefacts resulting from the electric shock, and a bandpass filter with range of 0.01Hz to 1Hz (Johnstone, T. (2017, September 8). Psychophysiology Analysis Software. Retrieved from osf.io/4wsm3). SCR data was analysed using a MATLAB script that detected maximum deflection from a 2 second pre-trial baseline using a window of 7 seconds from trial onset. The mean value for each condition and participant was calculated and imported into SPSS for statistical analysis. Residuals were normally distributed, therefore non-transformed data was analysed.

fMRI data analysis

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). The following pre-statistics processing was applied; motion correction using MCFLIRT(Mark Jenkinson, Bannister, Brady, & Smith, 2002) ; non-brain removal using BET (Smith, 2002) ; spatial smoothing using a Gaussian kernel of FWHM 5mm; grand-mean intensity normalisation of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s), and B0 fieldmap unwarping (Mark Jenkinson, 2003).

Single subject analysis

A fixed effects general linear model was used to analyse individual subject data. Regressors were created for each condition by convolving a stimulus boxcar function with the standard FSL gamma function. Temporal derivatives were included in this glm. Motion estimates were added as regressors to control for head displacement. Trials that included an electric shock were modelled separately (1 regressor for this condition) to account for variance but not subsequently included in analyses.

Registration to a standard space was performed using a two stage procedure with FLIRT (M Jenkinson & Smith, 2001; Mark Jenkinson et al., 2002). The mean functional volume for each participant was registered to the individual's high resolution structural image using 6 degree of freedom (DOF) BBR white matter boundary mapping. In a second step the individual's high resolution structural image was normalised to the Montreal Neurological Institute (MNI) template brain using a 12 DOF affine transformation. These two transformations were combined and used for subsequent registration of that participant's contrast images to MNI space before higher-level group analysis.

Group analysis

Comparison of contrasts across participants was carried out using Mixed Effects (FMRIB's Local Analysis of Mixed Effects, FLAME 1) with automatic outlier de-weighting and Random Field-based cluster thresholding. Results were corrected for multiple comparisons with a familywise error of $p<0.05$.

Comparison between The SECE task and a cognitive reappraisal task

In this comparative analysis we aimed to find brain areas that were activated in both the LSCCE- and the instructed emotion regulation task. We used fslmaths to multiply the thresholded, binarised activation maps from the two contrasts of interest (COG-_CS+ > COG+_CS+ and Negative Decrease > Negative Attend). This way, voxels that were activated

in one but not the other contrast would be multiplied by 0 and result in no activation in the resulting map, and voxels that were activated in both tasks would retain activation in the conjunction map.

Because the two prefrontal clusters resulting from the two contrasts of interest (left IFG in the SECE task, and left vlPFC in the instructed emotion regulation task) were spatially very close together, we further investigated their activation during both tasks. For each cluster, we extracted the mean % signal change from both the task we originally found them to be active in, and the other task (so for both the left IFG cluster from the SECE task, and the left vlPFC cluster from the instructed emotion regulation task). These values were then compared using SPSS.

RESULTS

SECE TASK

QUESTIONNAIRES

All participants reported being aware of the exact contingencies of the task and correctly identified the CS+ as well as the dangerous word category.

Participants' responses about how stressed they felt during each type of trial were analysed in 2 x 2 (CS (CS+ vs CS-) x trial type (COG- vs COG+)) repeated measures ANOVA. We found a significant effect of CS ($F(1,19) = 107.5$, $p < 0.001$, partial $\eta^2=0.85$), trial type ($F(1,19) = 102.64$, $p < 0.001$, partial $\eta^2=0.84$), and a significant CS x trial type interaction ($F(1,19) = 16.1$, $p < 0.001$, partial $\eta^2=0.46$). Responses relating to CS+ trials were higher than those relating to CS- trials, suggesting that conditioning was effective. Follow up t-tests revealed that participants felt significantly more stressed during COG+_CS+ (5.2) than COG-_CS+ trials (Mean = 2.6, $t(19) = 9.58$, $p < 0.001$, mean difference = 2.6, CIlower = 2.34, upper = 2.86, Cohen's d = 2.2), and during COG+_CS- (2.0) than COG-_CS- trials (Mean = 1.2, $t(19) = 3.1$, $p = 0.006$, mean difference = 0.8, CIlower = 0.53, upper = 1.07, Cohen's d =

0.88). There was a significant effect of CS, too, the difference between COG+_CS+ and COG-_CS+ trials was larger than that between COG+_CS- and COG-_CS- trials (t(19) = 4.01, p = 0.001; Table 4 for means).

Table 1. Questionnaire score means. Subjective stress levels as well as boredom and sleepiness were assessed on a 7-point likert scale going from 1 (not at all stressed/bored/sleepy) to 7 (extremely stressed/bored/sleepy).

	Mean (CI)	Cronbach's Alpha
Shock Intensity	4.78 (±3.28)	
Subjective stress level during -		
- COG+_CS+	*5.2 (±0.26)*	
- COG-_CS+	*2.6 (±0.26)*	
- COG+_CS-	*2 (±0.27)*	
- COG-_CS-	*1.2 (±0.27)*	
Boredom	3.6 (±1.9)	
Sleepiness	4 (±2.05)	
STAI	39.55 ± (8.7)	0.88
PANAS positive	28.45 (±6.6)	0.9
PANAS negative	13.4 (±3.3)	0.78
ERQ Reappraisal	31 (±9.9)	0.79
ERQ Suppression	14.9 (±6.5)	0.81
PSWQ	46.7 (±17.8)	0.96
IUS	59 (±20.2)	0.95

SKIN CONDUCTANCE

To check whether an aversive response occurred when participants briefly saw words of both categories with the CS embedded prior to the start of the extinction phase, a paired t-test was performed on the first 2 CS+ and CS- trials of the extinction phase, both including one word of each category prior to any US being delivered (participants were not aware which category was safe and which was dangerous at this stage). There was no significant difference between the two CS' (t(16) = 1.14, p = 0.27, Cohen's d = 0.57, however, the means show the expected direction (CS+ = 0.15, ; CS- = 0.08, Mean difference = 0.07, CIlower = -0.01, upper = 0.15). For the SECE phase overall, a paired t-test showed that SCR was significantly higher during

(non-reinforced) COG+_CS+ trials (M = 0.17 μS, SD = 0.15), than COG-_CS+ trials (M = 0.09 μS, SD = 0.09, t(18) = 3.3, p = 0.002, one-tailed, Cohen's d = 0.89, see Table 5 for means).

Table 2. Mean SCR in in μSiemens, CIs and significance tests. p < 0.01. CI's were calculated pairwise for the main contrasts of interest.**

Conditioning	***CS- Mean***	***CS+ Mean***	***95% Confidence Interval***	***F (1,16)***	***p***
Early	0.033	0.09	±0.1	0.39	0.54
Late	0.16	0.17	±0.05	0.5	0.49
Reappraisal Baseline	***CS- Mean***	***CS+ Mean***		***t(16)***	***P***
	0.08	0.15	±0.07	0.95	0.35
Reappraisal	***COG- Mean***	***COG+ Mean***		***t (16)***	***p***
CS-	0.1	0.13	±0.023	-0.7	0.47
CS+	0.09	0.17	±0.032	-3.3	0.002**
Letter Only	***CS- Mean***	***CS+ Mean***		***t(16)***	***p***
	0.04	0.09	±0.033	-1.8	0.09

Such a finding is consistent with participants being able to successfully decrease their emotional arousal to the CS+ when they identify, based on the word category, that there is no risk of receiving an electric shock.

To establish that the above effect was the consequence of successful regulation of the CS+ response when in the context of the safe word category, and not simply learning a new direct association between the word category and US, we compared SCR during COG+_CS- trials, (M = 0.13 μS, SD= 0.09), and COG-_CS- trials, (M = 0.10 μS, SD = 0.1). SCR was not significantly different between these two conditions, (t(17) = 0.7, p = 0.49, Cohen's d = 0.16, see Figure 2). Such a finding indicates that the dangerous word category had an impact on emotional arousal when paired with CS+ stimuli, but not with CS-.

Trials in which the letters were presented by themselves, without reinforcement, were analysed with a separate t-test. This showed no significant difference in the in SCR amplitude

(Means (SD): CS+ = 0.1 (0.09); CS- = 0.04 (0.12); t(17) = 1.1, p = 0.09, one-tailed, Cohen's d = 0.37, see Table 5).

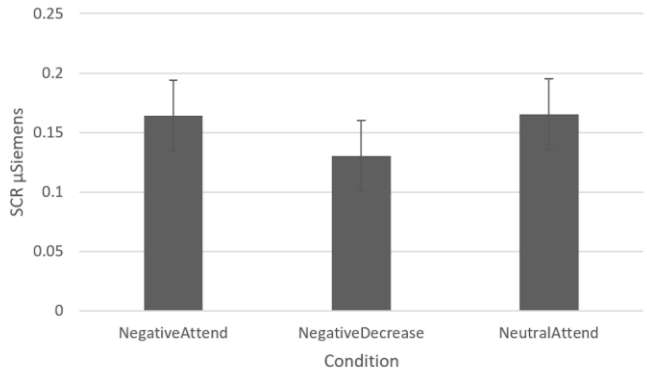

Figure 2. SCR associated with COG-_CS+ and COG+_CS+ and cs- trials. Error bars represent within subject confidence intervals.

fMRI Results

COG+_CS+ > COG-_CS+

COG+_CS+ trials compared to COG-_CS+ trials revealed activation in left insula and right anterior cingulate (ACC, see Table 6). Contrary to our hypothesis, no significant amygdala activation was found.

COG-_CS+ > COG+_CS+

The opposite contrast revealed increased activation in left inferior frontal gyrus (IFG), bilateral inferior temporal gyrus (ITG) and right superior parietal cortex during safe (COG-_CS+) compared to COG+_CS+ trials (see Table 36, Figure 3). Including participants' questionnaire scores (STAI, PSWQ, IUS, ERQ) as covariates in the analyses did not have an effect on these results.

Table 3. Significant clusters of activation in the main contrasts of interest, COG+_CS+ > COG-_CS+; and COG-_CS+ > COG+_CS+. All clusters survived cluster based thresholding at 2.3. Peak coordinates are presented in mni space.

Contrast	Anatomical Region	Hemisphere	Cluster size (mm^3)	X	Y	Z
COG+_CS+ > COG-_CS+	Insula	Left	2376	-40	8	6

	Anterior Cingulate Cortex	Right	2152	4	10	42
COG-_CS+ > COG+_CS+	Inferior Temporal Gyrus	Left	5608	-54	-54	-10
	Superior Parietal Lobule	Right	5456	30	-54	62
	Lateral Occipital Cortex	Left	4512	-26	-68	52
	Middle Frontal Gyrus and inferior frontal gyrus	Left	3352	-42	32	20
	Inferior Temporal Gyrus	Right	3144	52	-62	-18
	Precentral Gyrus	Right	3112	4	-26	66
	Cerebellum	Right	3032	6	-78	-26

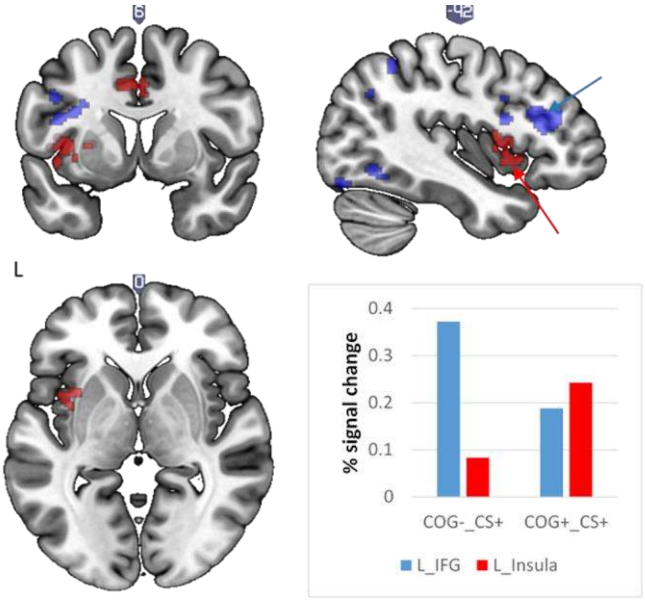

Figure 3. Neural activation observed during the SECE task: COG-_CS+ > COG+_CS+ (blue) as well as COG+_CS+ > COG-_CS+ (red). Areas that showed increased activation during safe compared to COG+_CS+ trials included left inferior frontal gyrus, bilateral inferior temporal gyrus and superior parietal cortex. Areas that showed increased activation to dangerous compared to safe trials included the left insula and dorsal anterior cingulate. The bar graph (c)) shows the percent signal change in the left ifg and the left insula.

COGNITIVE REAPPRAISAL TASK

QUESTIONNAIRES

Questionnaire data was analysed using a repeated measures ANOVA with 3 levels (trial type, Neutral Attend vs Negative Attend vs Negative Decrease). We found a significant effect of trial type ($F(2,18) = 36.86$, $p < 0.001$, partial $\eta^2=0.67$). Follow up t-tests revealed that participants felt significantly more stressed during Negative Attend than Negative Decrease trials ($t(18) = 4.44$, $p < 0.001$, Cohen's $d = 0.75$), and significantly more stressed during Negative Decrease than Neutral Attend trials ($t(18) = 4.31$, $p < 0.001$, Cohen's $d = 1.07$, see Table 7 for means and CI's).

All participants reported being able to see the images and using the correct strategy to regulate their emotion when asked to do so.

Table 4. Means of participants' stress level during each trial type and how bored and sleepy they felt throughout the task. Subjective stress levels as well as boredom and sleepiness were assessed on a 7-point likert scale going from 1 (not at all stressed/bored/sleepy) to 7 (extremely stressed/bored/sleepy).

	Mean (CI)
Subjective stress levels during -	
Neutral Attend	*3.16 (±0.28)*
Negative Decrease	*4.47 (±0.4)*
Negative Attend	*6.32 (±0.4)*
Bored	*2.58 (±1.89)*
Sleepy	*3.42 (±2.5)*

SCR RESULTS

To investigate whether an effect of emotional arousal exists between negative and neutral images, a two-tailed paired t test was conducted to compare SCR between negative attend and neutral attend conditions. Inconsistent with our predictions, there was no significant difference ($t(17) = -0.038$, $p = 0.97$, Cohen's $d = 0.095$) in SCR when participants were asked to attend to negative images, ($M = 0.16$ μS,), compared to when they were asked to attend to neutral images, ($M = 0.17$ μS, within participants CIlower = -0.1, upper = 0.44, see Figure 4).

To investigate whether SCR was reduced as participants intentionally decrease negative emotion through reappraisal, a one-tailed paired t test was conducted to compare SCR during negative attend trials with SCR during negative decrease trials. Results indicated that SCR was only marginally greater ($t(17) = 1.5$, $p = 0.08$, one-tailed, Cohen's $d = 0.33$), when participants attended to their emotional response to negative images, ($M = 0.16$ μS, CI lower = 0.137, upper = 0.183), compared to when they aimed to decrease their emotional arousal by reinterpreting the image with a better outcome ($M = 0.13$ μS, within participants CI lower = 0.107, upper = 0.143). Such a finding is not a clear indication of regulatory success as participants attempt to reduce their emotional arousal (see Figure 4).

Including participants' questionnaire scores (STAI, PSWQ, ERQ, IUS) in these analyses as covariates did not affect these results.

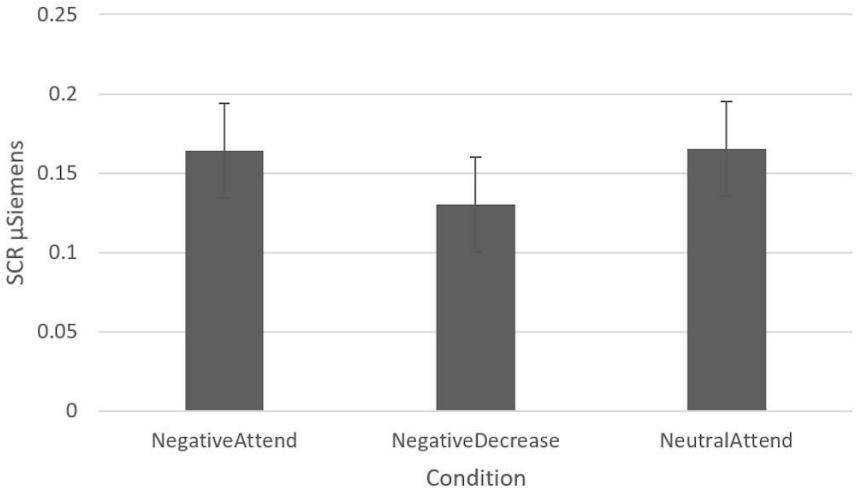

Figure 4. SCR in μSiemens to the different conditions in the instructed emotion regulation task. Error bars represent within subject confidence intervals.

fMRI results

Negative Decrease > Negative Attend

For this contrast we found activation in a network including left IFG, left middle temporal gyrus (MTG) and bilateral dorsolateral prefrontal cortex (dlPFC, see Table 8, Figure 5).

For the opposite contrast (Negative Attend > Negative Decrease), activation was found in a

network including bilateral insula and superior parietal cortex.

On the basis of previous research that used this task, an additional regions of interest (ROI) analysis was conducted using bilateral amygdala ROI masks and a small volume correction. This revealed a cluster in left amygdala (see Table 8, Figure 5).

Participants' questionnaire scores (STAI, PSWQ, ERQ, IUS) did not have an effect on these results.

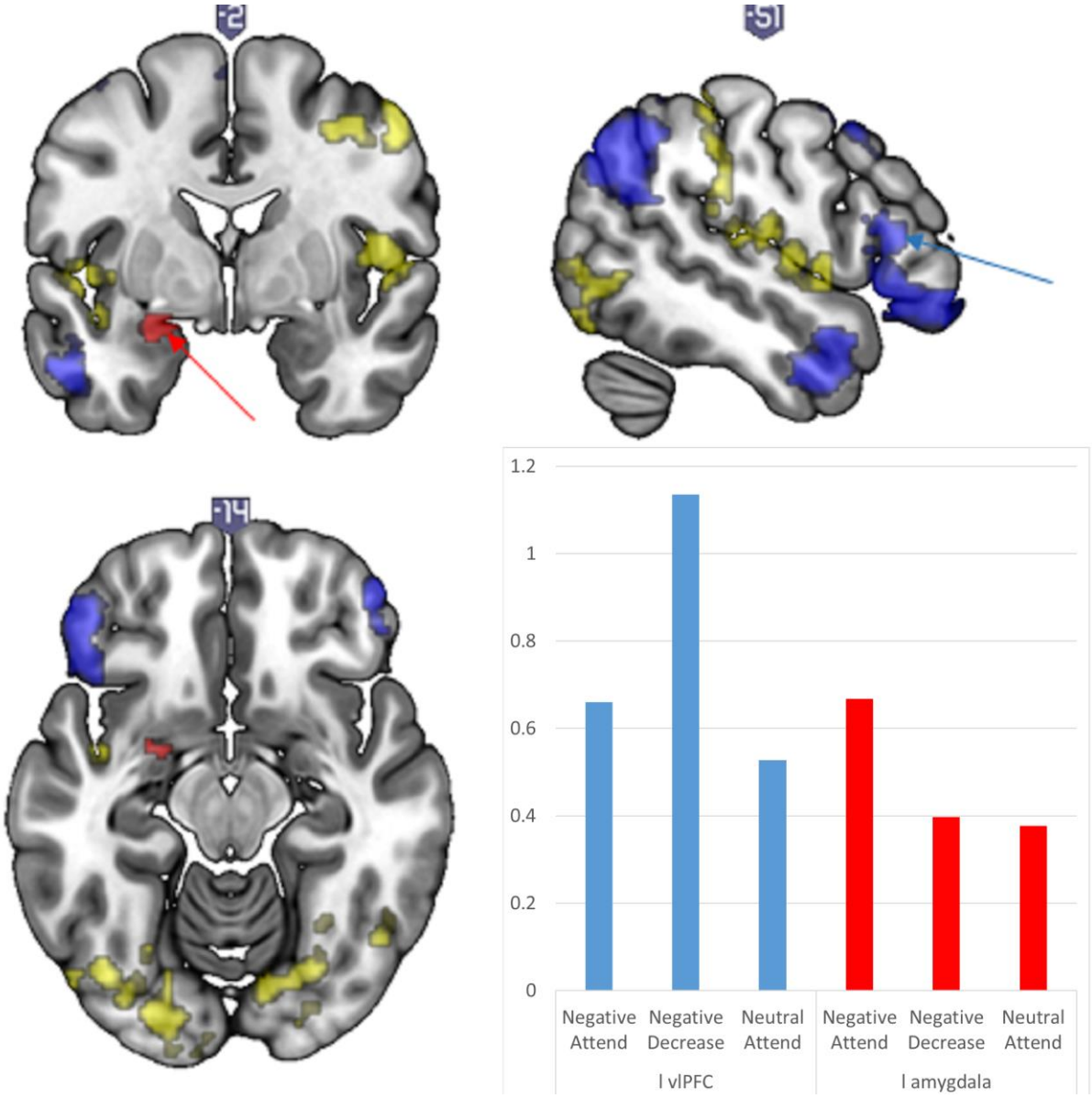

Figure 5. Activation in the instructed emotion regulation task. Negative Attend > Negative Decrease is shown in yellow, amygdala activation from an roi analysis on this contrast is shown in red. Blue shows clusters that were active in the Negative Decrease > Negative Attend including vlPFC, inferior temporal gyrus, and lateral parietal cortex.

Table 5. Peak activation in the instructed emotion regulation task. All clusters survived cluster based thresholding at 2.3. Peak coordinates are presented in MNI space.

Contrast	Anatomical Region	Hemisphere	MNI Coordinates (x,y,z)	Cluster size (mm^2)
NegativeDecrease > NegativeAttend	Middle frontal gyrus	left	-44, 8, 54	23592
	Lateral occipital cortex	left	-50, -58, 44	12760
	Lateral occipital cortex	right	58. -54, 38	7496
	Inferior frontal gyrus	left	-48, 40, -12	5912
	Superior frontal gyrus	right	20, 26, 62	4552
	Middle frontal gyrus	Right	2, 14, 50	2984
	Frontal pole	Right	46, 48, -12	2848
	Inferior temporal gyrus	left	-48, 0, -36	2576
NegativeAttend> NegativeDecrease	Lateral occipital cortex	left	-24, -90, 32	51256
	Parietal Operculum into central operculum and insular cortex	left	-60, -28, 18	17416
	Parietal Operculum into central operculum	right	60, -28, 22	8296
	Precentral gyrus	right	48, 8, 30	3336
ROI analysis Amygdala				
NegativeAttend> NegativeDecrease	Dorsal amygdala	left	-22,-2,-16	360

OVERLAP BETWEEN SECE AND INSTRUCTED REGULATION TASK

Analysis of the group level activation maps from both tasks revealed no overlapping prefrontal areas. The mean within-task condition % signal change differences from the two lateral PFC clusters (i.e. IFG COG-_CS+ - COG+_CS+, IFG Negative Decrease – Negative Attend; vlPFC COG-_CS+ - COG+_CS+, vlPFC Negative Decrease – Negative Attend) were extracted and analysed in a 2 x 2 (cluster (left IFG vs left vlPFC) x task ([COG-_CS+ > COG+_CS+] vs [NegDec > NegAtt])) repeated measures ANOVA. We found significant main effects of cluster ($F(1,19)=24.88$, $p<0.001$, partial $\eta^2 = 0.57$) but not of task

($F(1,19)=2.35$, $p = 0.14$, partial $\eta^2 = 0.11$) and a significant cluster x condition interaction ($F(1,19)=26.6, p<0.001$, partial $\eta^2 = 0.58$). Post hoc t-tests revealed that the left IFG showed a significant difference between COG-_CS+ and COG+_CS+ ($t(19) = 5.22$, $p< 0.001$) but not between between Negative Decrease and Negative Attend appraisal trials ($t(19) = 0.45$, $p = 0.66$), and that the left vlPFC showed a significant difference between Negative Decrease and Negative Attend appraisal trials ($t(19) = 5.22$, $p < 0.001$) but not between COG-_CS+ and COG+_CS+ trials ($t(19) = 0.43$, $p = 0.67$). This confirms the voxelwise results i.e. each cluster showed differential activation within the task it resulted from but not during the other task. Further inspection of the means shows, however, that the left IFG is activated during both the Negative Decrease and Negative Attend conditions in the instructed emotion regulation task (see Table 9 for means, Negative Attend: $t(19) = 6.52$, $p < 0.001$, 95% CI lower = 0.25, CI upper = 0.48; Negative Decrease: $t(19) = 5.74$, $p < 0.001$, 95% CI lower = 0.25 CI upper = 0.52).

Table 6. Mean % signal change for the left ifg and left vlpfc clusters in the 4 conditions of interest for this comparison.

	Left IFG **Mean % signal change**	**Left vlPFC** **Mean % signal change**
COG-_CS+	0.37	0.03
COG+_CS+	0.18	0.005
Negative Attend	0.37	0.59
Negative Decrease	0.38	1.02

DISCUSSION

The aim of this study was to investigate the selective extinction of conditioned responses through cognitive evaluation of conditioned stimuli embedded in different contexts. Participants were conditioned to expect a risk of electric shock during the presentation of one of two letters. In a subsequent phase, words belonging to two distinct categories briefly appeared. One word category signalled a safe trial, while the other category signalled continued risk of electric shock (i.e. a dangerous trial). Thus, participants had to cognitively

evaluate the information given in each trial in order to determine the risk of receiving an electric shock.

As predicted, we found increased skin conductance in response to COG+_CS+ compared to COG-_CS+ trials as well as corresponding increased activation in bilateral insula and right dorsal ACC. Notably, this pattern was not seen in any CS- trials, regardless of the word category. The insula and dACC have previously been associated with the perception and anticipation of pain as well as other salient stimuli (Brooks, Nurmikko, Bimson, Singh, & Roberts, 2002; Legrain, Iannetti, Plaghki, & Mouraux, 2011; Porro et al., 2002), and with the processing of threat and the generation of physiological responses (Kalisch & Gerlicher, 2014; Mechias et al., 2010). The insula is also one of the brain regions commonly found during negative emotional processing in instructed emotion regulation paradigms (Buhle et al., 2014; Kohn et al., 2014). In addition, activation in dACC has been associated with threat processing in both instructed and uninstructed conditioning studies (Kalisch & Gerlicher, 2014; Mechias et al., 2010). Taken together, these results indicate that the conditioned affective response evoked in anticipation of an electric shock was maintained during COG+_CS+ trials, and reduced during COG-_CS+ trials.

During COG-_CS+ compared to COG+_CS+ trials we found increased activation in left MFG and IFG, bilateral temporal gyrus and right parietal cortex. During these trials, participants were first presented with the previously established CS+, which was then briefly presented within a word belonging to a "safe" category (COG-), before the CS+ alone was shown again. Thus, participants could evaluate the risk of shock by categorising the word, remembering the category contingencies, and using the contextual information to decide whether each trial was COG- or COG+. Activation in MFG, IFG and ITG has been observed during language processing and semantic working memory tasks (Ferstl, Neumann, Bogler, & von Cramon, 2008; Nee et al., 2013), as well as during emotion regulation studies when

participants use a predefined strategy to reappraise an affective response (Buhle et al., 2014; Kohn et al., 2014). When the conditioned stimulus is presented embedded within a word from the safe category (COG-), increased activation in these brain regions may reflect a greater amount of processing required to regulate or reverse the previously learnt contingencies. Right superior parietal lobule and precentral gyrus have been shown to be involved in motor inhibition (Thoenissen, Zilles, & Toni, 2002). A tentative explanation might be that this reflects the inhibition of a freezing response to the anticipation of an electric shock, though this would require further more direct evidence.

In summary, participants were able to selectively reduce the CR trial by trial, on the basis of cognitively evaluated additional information. This is consistent not only with successful classic extinction of conditioned fear by removal of the US, but also of other forms of fear reduction, including occasion setting, and instructed extinction (Holland, 1992; Luck & Lipp, 2016). During occasion setting, a feature is added to a previous CS+ which signals that the US will not occur. Similarly, in SECE, information (i.e. the words) were added to the CS' which signalled safety from the US under certain circumstances (i.e. when the word category had been learnt to be safe). During instructed extinction participants are informed that the US will no longer occur, and the CR is reduced. Similarly, in SECE, participants were informed that the US would no longer occur when the presented word belonged to one of the categories (but not which). This resulted in a selective CR reduction when the presented word belonged to that category, thus, participants were able to learn the category contingencies and selectively utilize that knowledge on a trial-by-trial basis.

Participants also completed an instructed emotion regulation task. Compared to previous studies, we modified this paradigm by shortening the trials to make the design as similar as possible to the SECE tasks. Even though our adaptation of the instructed emotion regulation task was shorter than previous versions, it revealed a neural pattern of results that replicates

those found with longer versions of this task (Buhle et al., 2014; Johnstone et al., 2007; Urry et al., 2006). In particular, we found increased amygdala activation while participants were attending to the negative images compared to when they were decreasing their reaction to the negative images. During Negative/Decrease trials we found increased activation in left vlPFC and dlPFC compared to Negative/Attend trials. These results suggest that a shortened version of the instructed emotion regulation task is a viable option in developmental and clinical studies where minimising the exposure to distressing stimuli may be preferable.

Despite neural patterns of responding being consistent with findings of previous literature, we did not find the expected result of increased SCR to Negative/Attend compared to Negative/Decrease trials in this task. There may be a number of reasons for this conflicting result. The cognitive reappraisal task was always completed after the SECE task, thus, participants had been lying down in the scanner for approximately 45 minutes already before the cognitive reappraisal task started. In general, skin conductance responsivity declines over time which may have led to a reduced signal-to-noise ratio and might have contributed to this result. This decline with time is exactly why we administered the SECE task first, the maximise our sensitivity for our primary research questions. In addition, the trials in the present cognitive reappraisal task were shorter than those used in previously published studies. In contrast to the SECE task, the stimuli in the cognitive reappraisal task are more complex and may take more time to process, which may delay the electrodermal response beyond the trial analysis window. Finally, it is also possible that the relatively small number of participants we tested in this study might have resulted in a lack of power in the analysis of SCR data in this second task.

Overlap between SECE and instructed emotion regulation task

The voxelwise comparison between left frontal activation resulting from the SECE and the instructed emotion regulation task did not reveal any overlap. As the two tasks differ greatly in the cognitive and attentional demands placed on participants, this is not surprising. While participants are required to process complex images during the instructed emotion regulation task, as well as imagining a more positive outcome to decrease their initial reaction, the processes involved in the SECE task are less deliberative and the manipulation of the stimuli requires a very specific categorical decision. Thus, the network of brain areas involved is expectedly more defined and specific in the SECE task. Further investigation of extracted activations in the left prefrontal clusters revealed that the IFG cluster activated during COG-_CS+ trials (compared to COG+_CS+) was, indeed, also active during *both* Negative Attend *and* Negative Decrease conditions in the instructed emotion regulation task. This highlights one problem with standard reappraisal-based emotion regulation tasks: Cognitive evaluative processes are involved both in the initial appraisal and in the regulation of emotion, thus, comparing the conditions potentially masks brain regions that are part of this evaluative network. An advantage of the SECE task is that it can be used investigate the underlying cognitive and attentional mechanisms involved in emotion regulation with greater specificity. And if different types of affective psychopathology are associated with deficiencies in specific processes that underlie reappraisal (e.g. anxiety might be associated with different processes than depression or PTSD as suggested by (Gross & Jazaieri, 2014)), adaptations of the SECE task might be useful in identifying them.

Limitations & Extensions

The sample size in this study, although not unusual for an MRI study (Delgado et al., 2008; Milad et al., 2007; Phelps et al., 2004), is still relatively small. Because this was the first attempt to carry out an SECE, and the associated risks with MRI scans, we wanted to keep

the number of participants small before knowing whether the task works as intended, while ensuring enough power to be able to find effects. We therefore based our sample size on established large effect size for conditioning studies (e.g. Leuchs, Schneider, & Spoormaker, (2018) demonstrated an effect size of 0.75 for conditioning of SCR, a value which is close to what we observed). We hope that future studies will replicate our results and establish the SECE task in the literature.

In this study we made a decision to keep the order in which the two tasks were completed fixed for all participants. This reflects a priority being given to investigating the physiological and neural correlates of the SECE task, the primary aim of the study. As our experimental procedure was fairly long (participants spent roughly 1hr in the MRI scanner) and included different sources of negative emotion, we were concerned that some participants might choose not to complete the second task (as one participant in fact did). In addition, electrodermal activity drops over time, especially when participants lie down in an MRI scanner. Clearly the fixed order does not allow strong conclusions to be made about *differences* in brain activation between the two tasks, though as pointed out the two tasks differ substantially in the type of cognitive and perceptual processing involved, and so are not directly comparable in any case. The instructed regulation task was included to look for *common* regions of activation that would indicate that at least some cognitive processes might be involved in both tasks, as has been suggested (Hartley & Phelps, 2010).

We also made some design choices with the SECE task that influence the interpretation of the results. We did not explicitly tell participants which word category was safe from the beginning. Extinction by instruction is a commonly employed design choice and has been shown to be effective (Lovibond, 2004). In this instance we wanted to maximise the extent to which participants evaluated the combination of CS and word category, rather than responding by rote. We also hoped to be able to examine differences in responding over time,

reflecting the learning of the contingencies, however, as discussed above, we did not find any effects of time.

The SECE task is designed to be adaptable to investigate the involvement of a range of cognitive or attentional processes (e.g. CS could be presented in different spatial locations to engage either spatial working memory or spatial attention, CS might be embedded in contexts with personal relevance to the participants, requiring engagement of autobiographical memory). Future studies might investigate a range of these processes to arrive at a more controlled understanding of how different cognitive processes are involved in adaptive emotion regulation.

ACKNOWLEDGEMENTS

The authors wish to thank Shan Shen for assistance with scanning, and reviewers of a previous version of this manuscript for their constructive comments.

FUNDING

This research was supported by a UK Medical Research Council Doctoral Training Studentship (MR/J003980/1) to the first author.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to report.

AUTHOR CONTRIBUTIONS

Birthe Macdonald: Study design, data acquisition, analysis and interpretation, manuscript draft, agreement to be accountable for all aspects of the work

Shannon Wake: data acquisition and analysis, critical manuscript revision, approval of final manuscript version, agreement to be accountable for all aspects of the work

Tom Johnstone: Study design, data interpretation, critical manuscript revision, approval of final version, agreement to be accountable for all aspects of the work

Data Accessibility Statement

The data used in this study, along with all relevant materials and information will be uploaded to a suitable open science repository.

Abbreviations

A: Anterior

ACC: Anterior Cingulate Cortex

Ag-AgCl: Silver-Silver Chloride

ANOVA: Analysis of Variance

CINN: Centre for Integrated Neuroscience and Neurodynamics

CR: Conditioned Response

CS: Conditioned Stimulus

COG+_CS+: Letter: CS+, word category indicating threat of electric shock

COG-_COG-: Letter: CS+, word category indicating NO threat of electric shock

COG+_CS-: Letter: CS-, word category indicating threat of electric shock (in CS+ condition, no shocks given when letter: CS-)

COG-_CS-: Letter: CS-, word category indicating NO threat of electric shock

dACC: dorsal Anterior Cingulate Cortex

dlPFC: dorsolateral Prefrontal Cortex

ERQ: Emotion regulation Questionnaire

EPI: Echoplanar Imaging

FOV: Field of View

MFG: Middle Frontal Gyrus

MPRAGE: Magnetization Prepared Rapid Acquisition Gradient Echo

MRI: Magnetic Resonance Imaging

IFG: Inferior Fontal Gyrus

ITG: Inferior Temporal Gyrus

IUS: Intolerance of Uncertainty Scale

P: Posterior

PSWQ: Penn State Worry Questionnaire

ROI: Region of Interest

SCR: Skin Conductance Response

SECE: Selective Extinction through Cognitive Evaluation

STAI: State Trait Anxiety Inventory

STG: Superior Temporal Gyrus

TE: Echo Time

TI: Inversion Time

TR: Repetition Time

US: Unconditioned Stimulus

vmPFC: ventromedial Prefrontal Cortex

vlPFC: ventrolateral Prefrontal Cortex

REFERENCES

Amstadter, A. (2008). Emotion regulation and anxiety disorders. *Journal of Anxiety Disorders*, *22*(2), 211–221. https://doi.org/10.1016/j.janxdis.2007.02.004

Atlas, L. Y. (2019). How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. *Current Opinion in Behavioral Sciences*, *26*, 121–129. https://doi.org/10.1016/J.COBEHA.2018.12.008

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry*, *52*(10), 976–986. https://doi.org/10.1016/S0006-3223(02)01546-9

Brooks, J. C. W., Nurmikko, T. J., Bimson, W. E., Singh, K. D., & Roberts, N. (2002). fMRI of thermal pain: effects of stimulus laterality and attention. *NeuroImage*, *15*(2), 293–301. https://doi.org/10.1006/nimg.2001.0974

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., … Ochsner, K. N. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cerebral Cortex (New York, N.Y. : 1991)*, *24*(11), 2981–2990. https://doi.org/10.1093/cercor/bht154

Buhr, K., & Dugas, M. J. (2002). The intolerance of uncertainty scale: psychometric properties of the English version. *Behaviour Research and Therapy*, *40*(8), 931–945. https://doi.org/10.1016/S0005-7967(01)00092-4

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Handbook of Psychophysiology* (3rd ed.). New Yor, NY: Cambridge University Press.

Delgado, M. R., Nearing, K. I., Ledoux, J. E., & Phelps, E. A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron*, *59*(5), 829–838. https://doi.org/10.1016/j.neuron.2008.06.029

Dibbets, P., Broek, A. van den, & Evers, E. A. T. (2015). Fear conditioning and extinction in anxiety- and depression-prone persons. *Memory*, *23*(3), 350–364. https://doi.org/10.1080/09658211.2014.886704

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93. https://doi.org/10.1016/j.tics.2010.11.004

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581–593. https://doi.org/10.1002/hbm.20422

Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., … Harrison, B. J. (2018). Fear extinction in the human brain: A meta-analysis of fMRI studies in healthy participants. *Neuroscience & Biobehavioral Reviews*, *88*, 16–25. https://doi.org/10.1016/j.neubiorev.2018.03.002

Grillon, C., & Ameli, R. (2001). Conditioned inhibition of fear-potentiated startle and skin conductance in humans. *Psychophysiology*, *38*(05), 807–815.

Gross, J. J., & Jazaieri, H. (2014). Emotion, Emotion Regulation, and Psychopathology: An Affective Science Perspective. *Clinical Psychological Science*, *2*(4), 387–401. https://doi.org/10.1177/2167702614536164

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362.

Gross, J. J., & Levenson, R. W. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology*, *106*(1), 95–103. https://doi.org/10.1037/0021-843X.106.1.95

Gross, J. J., & Muñoz, R. F. (1995). Emotion Regulation and Mental Health. *Clinical Psychology: Science and Practice*, *2*(2), 151–164. https://doi.org/10.1111/j.1468-2850.1995.tb00036.x

Hartley, C. A., & Phelps, E. A. (2010). Changing Fear: The Neurocircuitry of Emotion Regulation. *Neuropsychopharmacology*, *35*(1), 136–146. https://doi.org/10.1038/npp.2009.121

Holland, P. C. (1992). Occasion Setting in Pavlovian Conditioning. In D. L. Medin (Ed.), *Psychology of Learning and Motivation* (Vol. 28, pp. 69–125). https://doi.org/10.1016/S0079-7421(08)60488-0

Huff, N. C., Hernandez, J. A., Blanding, N. Q., & LaBar, K. S. (2009). Delayed Extinction Attenuates Conditioned Fear Renewal and Spontaneous Recovery in Humans. *Behavioral Neuroscience*, *123*(4), 834–843. https://doi.org/10.1037/a0016511

Jackson, D. C., Malmstadt, J. R., Larson, C. L., & Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology*, *37*(4), 515–522. https://doi.org/10.1111/1469-8986.3740515

Jenkinson, M, & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Jenkinson, Mark. (2003). Fast, automated, N-dimensional phase-unwrapping algorithm. *Magnetic Resonance in Medicine : Official Journal of the Society of Magnetic Resonance in Medicine /*

Society of Magnetic Resonance in Medicine, *49*(1), 193–197. https://doi.org/10.1002/mrm.10354

Jenkinson, Mark, Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Johnstone, T., van Reekum, C. M., Urry, H. L., Kalin, N. H., & Davidson, R. J. (2007). Failure to regulate: counterproductive recruitment of top-down prefrontal-subcortical circuitry in major depression. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(33), 8877–84. https://doi.org/10.1523/JNEUROSCI.2063-07.2007

Joormann, J., & Vanderlind, W. M. (2014). Emotion Regulation in Depression: The Role of Biased Cognition and Reduced Cognitive Control. *Clinical Psychological Science*, *2*(4), 402–421. https://doi.org/10.1177/2167702614536163

Kalisch, R., & Gerlicher, A. M. V. (2014). Making a mountain out of a molehill: On the role of the rostral dorsal anterior cingulate and dorsomedial prefrontal cortex in conscious threat appraisal, catastrophizing, and worrying. *Neuroscience & Biobehavioral Reviews*, *42*, 1–8. https://doi.org/10.1016/j.neubiorev.2014.02.002

Khoury, B., & Lecomte, T. (2012). Emotion Regulation and Schizophrenia. *International Journal of Cognitive Therapy*, *5*(1), 67–76. https://doi.org/10.1521/ijct.2012.5.1.67

Kim, J. J., & Jung, M. W. (2006). Neural circuits and mechanisms involved in Pavlovian fear conditioning: a critical review. *Neuroscience and Biobehavioral Reviews*, *30*(2), 188–202. https://doi.org/10.1016/j.neubiorev.2005.06.005

Kohn, N., Eickhoff, S. B., Scheller, M., Laird, A. R., Fox, P. T., & Habel, U. (2014). Neural network of cognitive emotion regulation--an ALE meta-analysis and MACM analysis. *NeuroImage*, *87*, 345–355. https://doi.org/10.1016/j.neuroimage.2013.11.001

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* Gainesville, FL: University of Florida.

Legrain, V., Iannetti, G. D., Plaghki, L., & Mouraux, A. (2011). The Pain Matrix Reloaded The Pain Matrix Reloaded. A Salience Detection System for the Body. *Progress in Neurobiology*, *93*(1), 111–124.

Leuchs, L., Schneider, M., & Spoormaker, V., I. (2018). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology*, *56*(1). https://doi.org/10.1111/psyp.13283

Lovibond, P. F. (2004). Cognitive processes in extinction. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *11*(5), 495–500. https://doi.org/10.1101/lm.79604

Luck, C. C., & Lipp, O. V. (2016). Instructed extinction in human fear conditioning: History, recent developments, and future directions. *Australian Journal of Psychology*, *68*(3), 209–227. https://doi.org/10.1111/ajpy.12135

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. https://doi.org/10.1016/j.neuroimage.2009.09.040

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495. https://doi.org/10.1016/0005-7967(90)90135-6

Milad, M. R., & Quirk, G. J. (2002). Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature*, *420*(6911), 70–74. https://doi.org/10.1038/nature01138

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annual Review of Psychology*, *63*, 129–151. https://doi.org/10.1146/annurev.psych.121208.131631

Milad, M. R., Quirk, G. J., Pitman, R. K., Orr, S. P., Fischl, B., & Rauch, S. L. (2007). A Role for the Human Dorsal Anterior Cingulate Cortex in Fear Expression. *Biological Psychiatry*, *62*(10), 1191–1194. https://doi.org/10.1016/j.biopsych.2007.04.032

Mochcovitch, M. D., da Rocha Freire, R. C., Garcia, R. F., & Nardi, A. E. (2014). A systematic review of fMRI studies in generalized anxiety disorder: Evaluating its neural and cognitive basis. *Journal of Affective Disorders*, *167*, 336–342. https://doi.org/10.1016/j.jad.2014.06.041

Morgan, M A, Romanski, L. M., & LeDoux, J. E. (1993). Extinction of emotional learning: contribution of medial prefrontal cortex. *Neuroscience Letters*, *163*(1), 109–113.

Morgan, Maria A., & LeDoux, J. E. (1995). Differential contribution of dorsal and ventral medial prefrontal cortex to the acquisition and extinction of conditioned fear in rats. *Behavioral Neuroscience*, *109*(4), 681.

Morriss, J., Hoare, S., & van Reekum, C. M. (2018). It's time: A commentary on fear extinction in the human brain using fMRI. *Neuroscience & Biobehavioral Reviews*, *94*, 321–322. https://doi.org/10.1016/j.neubiorev.2018.06.025

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A Meta-analysis of Executive Components of Working Memory. *Cerebral Cortex*, *23*(2), 264–282. https://doi.org/10.1093/cercor/bhs007

Pavlov, I. (1927). *Conditioned reflexes.* New York, NY, US: Dover.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron*, *43*(6), 897–905. https://doi.org/10.1016/j.neuron.2004.08.042

Picó-Pérez, M., Radua, J., Steward, T., Menchón, J. M., & Soriano-Mas, C. (2017). Emotion regulation in mood and anxiety disorders: A meta-analysis of fMRI cognitive reappraisal studies. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *79*, 96–104. https://doi.org/10.1016/j.pnpbp.2017.06.001

Pollak, D. D., Rogan, M. T., Egner, T., Perez, D. L., Yanagihara, T. K., & Hirsch, J. (2010). A translational bridge between mouse and human models of learned safety. *Annals of Medicine*, *42*, 127–13. https://doi.org/10.3109/07853890903583666

Porro, C. A., Baraldi, P., Pagnoni, G., Serafini, M., Facchin, P., Maieron, M., & Nichelli, P. (2002). Does Anticipation of Pain Affect Cortical Nociceptive Systems? *J. Neurosci.*, *22*(8), 3206–3214.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*(1), 1–5.

Rolls, E. T., Critchley, H. D., Mason, R., & Wakeman, E. A. (1996). Orbitofrontal Cortex Neurons: Role in Olfactory and Visual Association Learning. *JOURNALOF NEUROPHYSIOLOGY*, *75*(5).

Schiller, D., Cain, C. K., Curley, N. G., Schwartz, J. S., Stern, S. A., LeDoux, J. E., & Phelps, E. A. (2008). Evidence for recovery of fear following immediate extinction in rats and humans. *Learning & Memory*, *15*(6), 394–402. https://doi.org/10.1101/lm.909208

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155. https://doi.org/10.1002/hbm.10062

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Thoenissen, D., Zilles, K., & Toni, I. (2002). Differential involvement of parietal and precentral regions in movement preparation and motor intention. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(20), 9024–9034.

Trask, S., Thrailkill, E. A., & Bouton, M. E. (2017). Occasion setting, inhibition, and the contextual control of extinction in Pavlovian and instrumental (operant) learning. *Behavioural Processes*, *137*, 64–72. https://doi.org/10.1016/j.beproc.2016.10.003

Urry, H. L., van Reekum, C. M., Johnstone, T., Kalin, N. H., Thurow, M. E., Schaefer, H. S., … Davidson, R. J. (2006). Amygdala and ventromedial prefrontal cortex are inversely coupled during regulation of negative affect and predict the diurnal pattern of cortisol secretion among older adults. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(16), 4415–25. https://doi.org/10.1523/JNEUROSCI.3215-05.2006

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, *3*(1), 1–14. https://doi.org/10.1037/h0069608

Zhou, F., Geng, Y., Xin, F., Li, J., Feng, P., Liu, C., … Becker, B. (2019). Human Extinction Learning Is Accelerated by an Angiotensin Antagonist via Ventromedial Prefrontal Cortex and Its Connections With Basolateral Amygdala. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2019.07.007