Medial Prefrontal Cortex Activity during the Extinction of Conditioned Fear: An Investigation Using Functional Near-Infrared Spectroscopy

Anne Guhn, Thomas Dresler, Tim Hahn, Andreas Mühlberger, Andreas Ströhle, Jürgen Deckert, Martin J. Herrmann

Department of Psychiatry, Psychosomatics and Psychotherapy, University of Würzburg, Würzburg, Germany

Key Words
Extinction learning · Fear conditioning · Medial prefrontal cortex · NIRS

Abstract
The majority of fear conditioning studies in humans have focused on fear acquisition rather than fear extinction. For this reason only a few functional imaging studies on fear extinction are available. A large number of animal studies indicate the medial prefrontal cortex (mPFC) as neuronal substrate of extinction. We therefore determined mPFC contribution during extinction learning after a discriminative fear conditioning in 34 healthy human subjects by using functional near-infrared spectroscopy. During the extinction training, a previously conditioned neutral face (conditioned stimulus, CS+) no longer predicted an aversive scream (unconditioned stimulus, UCS). Considering differential valence and arousal ratings as well as skin conductance responses during the acquisition phase, we found a CS+ related increase in oxygenated haemoglobin concentration changes within the mPFC over the time course of extinction. Late CS+ trials further revealed higher activation than CS− trials in a cluster of probe set channels covering the mPFC. These results are in line with previous findings on extinction and further emphasize the mPFC as significant for associative learning processes. During extinction, the diminished fear association between a former CS+ and a UCS is inversely correlated with mPFC activity – a process presumably dysfunctional in anxiety disorders.

Introduction
Fear is an aversive emotional state which at moderate levels proves biologically useful by enabling effective detection of threat and automatic activation of defensive behaviour [1]. Anxiety disorders such as post-traumatic stress disorder, panic disorder and phobias are characterized by increased fear levels that might contribute to a generally impaired ability of fear extinction [2–4]. In order to model the development and maintenance of anxiety disorders, learning theories – notably conditioned fear reactions and their extinction – have been widely applied and particularly validated with regard to the effectiveness of exposure-based treatment in psychotherapy [2].
In a classical pavlovian fear conditioning paradigm, an initially neutral stimulus, such as a tone, is paired with an aversive event (unconditioned stimulus, UCS), e.g. an electric shock, and comes to elicit the so-called conditioned response (CR, e.g. freezing) itself after several pairings. In the absence of the UCS, the CR gradually disappears in response to the conditioned stimulus (CS, tone), i.e. its amplitude and frequency decrease [2]. This decrease of the CR is referred to as extinction and is thought to be the experimental foundation of exposure therapy applied to anxiety patients. During extinction learning, the acquisition of inhibitory memories is assumed to compete with excitatory memories formed during fear conditioning and thereby suppress the CR [3, 4]. These interactions are neuroanatomically mirrored in subcortical as well as cortical structures which have been closely investigated in a variety of lesion studies [e.g. 5, 6] as well as single-cell recording [e.g. 7, 8] and stimulation studies [e.g. 9, 10] in animals. Above all, the amygdala has been shown to be notably involved in the expression and acquisition of conditioned fear [11], also in humans [12]. In intracellular in vivo recordings in rats, Rosenkranz et al. [7] demonstrated enhanced activity of the lateral amygdala while presenting conditioned affective stimuli whereas bilateral amygdala lesions prevented the acquisition of CR [13, 14]. Lesions to the medial prefrontal cortex (mPFC) on the other hand have been shown to generate an increased resistance to extinction as well as a high rate of spontaneous recovery while the acquisition of fear CR remained unaffected [6, 15, 16]. During extinction learning a rapid switch in the activity of two distinct populations of basal amygdala neurons seems to be essential for activating behavioural alterations [8]. According to Herr et al. [8], these ‘extinction neurons’ are bi-directionally connected with mPFC neurons that might mediate the consolidation of extinction memory. The ‘fear neurons’ in turn seem to be depressed if the predictive ability of the CS for danger is weakened through successful extinction training [5]. Paralleling these findings, mice that underwent extinction training in an investigation using fluorodeoxyglucose displayed elevated prefrontal cortex activity. Moreover, mice with higher prefrontal activity showed less CR and in turn mice not receiving extinction training demonstrated significantly more stable amygdala activity [17].

Although the neural mechanisms of extinction learning in humans are less well characterized than for animals [12], the general pattern of brain activation during fear and extinction learning seems to be essentially the same [18]. A variety of pavlovian conditioning studies in humans actually found amygdala involvement during fear conditioning [19–27]. Accordingly, mPFC activity could have been associated with extinction learning [20, 23, 24, 26, 28] and further linked with decreasing amygdala activity [23, 24, 29–31]. However, although extinction is thought to be crucial for understanding and improving psychotherapy, in 2009 Sehmeyer et al. [12] merely found seven studies directly focussing on neurobiological correlates of extinction learning. Frequently but even inconsistently found was prefrontal engagement. More recently these findings were enriched by studies concentrating on trait anxiety in fear extinction. These studies revealed comparable results to the aforementioned animal studies [7, 10] relating to mPFC-amygdala coupling. Anxious participants displayed enhanced amygdala activity during extinction learning that correlated negatively with mPFC involvement, indicating delayed inhibitory learning or rather generally reduced extinction [26, 28]. Irrespective of extinction learning, in anxious subjects, Indovina et al. [32] demonstrated an insufficient recruitment of the prefrontal cortex to down-regulate fear in a safety context, and Bishop et al. [33] showed reduced mPFC activity when anticipating threat. Taken together, these findings clarify extinction as a form of new learning which is hence prone to behavioural instability and emphasise the importance of a better understanding of extinction mechanisms. To date, the limited number of available studies on mPFC activity and extinction learning in humans impedes gathering insights into the mode of action in psychotherapy.

The present investigation therefore focused on contribution of the mPFC during extinction learning in a classical discriminative fear conditioning paradigm by using functional near-infrared spectroscopy (fNIRS). NIRS is an optical imaging method to non-invasively and in vivo investigate tissue such as the brain, muscle and others. It enables measuring concentration changes in oxygenated (O_2Hb) and deoxygenated (HHb) haemoglobin which are accompanied by increases in cerebral blood volume [34]. To our knowledge, fNIRS has never been used to investigate contribution of the mPFC to extinction learning. However, fNIRS has been successfully applied to measure changes in O_2Hb concentration within the mPFC during emotional tasks [e.g. 35]. The frontal positioning of the NIRS probe set enables the investigation of medial Brodmann areas 9 and 10 bilaterally extending to the dorsolateral prefrontal cortices and is hence covering the mPFC. Participants in the present study performed a fear conditioning paradigm in which one of two neutral faces (CS) was paired with an aversive scream (UCS). Immedi-
ately after establishing the fear conditioning they underwent an extinction training in which the originally neutral stimulus was repeatedly presented without the UCS. By dividing the so-called within-session extinction phase into an early and late component, we hypothesised increasing mPFC activity to the CS+ as described by others [25, 26, 36].

**Methods**

**Participants**

Thirty-five healthy volunteers (17 females, 18 males; mean age 24.7 years, standard deviation (SD) 3.32, range 20–32 years) participated in the study. All subjects were screened for current mental health using the German version of the Mini International Neuropsychiatric Interview (MINI [37]) and for right-handedness according to the Edinburgh Handedness Inventory [38] before the experiment. We further assured that all females used oral monophasic contraceptives and that they were not in their pill-off phase when participating in the experiment in order to exclude changes in hormonal levels which have been demonstrated to influence conditioned fear acquisition as well as extinction recall [39]. Psychology students were excluded to exclusively investigate paradigm-naive volunteers. For this reason, 1 male participant (age 31 years) was not considered for further data analyses because of familiarity with the procedure. Subjects were reimbursed with 7 Euros for participation in an experimental setting lasting 60 min.

They were recruited through online advertisement and gave written informed consent in accordance with the Declaration of Helsinki in their most recent version from 2008. All procedures were approved by the ethical review board of the medical faculty of the University of Würzburg (Protocol ID 151/10) and were performed in the facilities of the department of Psychiatry Psychosomatics and Psychotherapy of the University of Würzburg.

**Experimental Paradigm**

The differential fear conditioning paradigm investigated in the present study consisted of three experimental phases (habituation, acquisition and extinction). Two colour photographs of neutral male faces selected from the NimStim set of facial expressions [40] served as conditioned stimuli (CS) and a scream of 95 dB adapted from the International Affective Digital Sounds [41] was used as UCS. During the habituation, each stimulus was presented 8 times without the UCS. The following acquisition phase comprised 30 trials in total, i.e. 15 CS− and 15 CS+ trials in which 12 CS+ were paired with the UCS and the remaining 3 CS+ trials rested unpaired (reinforcement rate 80%) in order to decelerate the acquisition of conditioned fear and to extend its extinction. The CS− was never paired with the UCS. The extinction phase consisted of 18 CS− and 18 CS+ presentations without the UCS. Faces were presented for 4,000 ms and counterbalanced as CS+ and CS− to each subject, so that both faces were equally often selected as CS+ and CS−. The scream lasted 1,380 ms and appeared in a jittered time interval of 0–1,000 ms after the CS+ offset (fig. 1). Inter-trial intervals ranged from 10 to 16 s and consisted of a white fixation cross on a black screen. CS presentations were randomized within all three experimental phases.

**SCR and NIRS Measurements**

We assessed skin conductance responses (SCR) to CS+ and CS− during the whole experiment in order to ensure effective fear conditioning as well as extinction. SCR is regarded as an index for emotional responses associated with automatic arousal [43] and therefore an indicator of successful conditioning. It was assessed with two Ag/AgCl electrodes attached to the thenar eminence of the subjects’ left palm. SCR were recorded using a GSR sensor.
required with a sampling rate of 10 Hz and transformed into values of oxygen absorption of near-infrared light due to O$_2$Hb and HHb by using fNIRS. fNIRS measurements are based on differential absorption of near-infrared light due to O$_2$Hb and HHb concentration changes that arise through neurovascular coupling mirroring the metabolic demands of the nervous system. Illuminating the brain surface through the intact scalp and skull, near-infrared light reflected from deep tissue layers is received by a photodetector that is fixed some centimetres apart from the light emitter. fNIRS measurements are comfortable for the subjects because of fewer motion restrictions and no noise disturbance; it has a high temporal resolution (<1 s) and can be easily combined with other neuroimaging techniques or physiological measurements [44]. Further, more detailed information about the fundamentals of fNIRS is provided elsewhere [e.g. 45, 46]. We opted to restrict the fNIRS measurement to the extinction phase for two reasons: first, our experiment focused particularly on extinction learning and not fear acquisition, and second, we know from experience that subjects in fNIRS settings exceeding a time period of 20 min without a break might feel more and more uncomfortable. fNIRS signals were measured with the continuous-wave system ETG-4000 (Hitachi Medical Co., Tokyo, Japan) using a 3 × 11 channel array of optodes consisting of 16 photodetectors and 17 light emitters resulting in 52 channels in total. The ETG-4000 operates with two different wavelengths (695 ± 20 and 830 ± 20 nm) and its frequency is modulated for wavelengths and channels to prevent crosstalk. In order to reliably position the probe set, the lowest-row centre optode is typically placed on the Fpz position at the frontal region of the head extending symmetrically towards positions T3 and T4 according to the International 10–20 system for EEG electrode placement [47]. The interoptode distance of 30 mm enables measurements approximately 15–25 mm beneath the scalp [48]. Signals were acquired with a sampling rate of 10 Hz and transformed into values for changes in the concentration of O$_2$Hb.

Statistical Methods

All statistical analyses were performed using PASW Statistics 18 (SPSS Inc., Chicago, Ill., USA) and Matlab software (Version 7; MathWorks Inc., Natick, Mass., USA). Whenever we had directed hypotheses, one-tailed tests at a significance level of p < 0.05 were performed (otherwise two-tailed). Valence and arousal ratings were analysed separately using repeated-measures analyses of variance (ANOVA) with two within-subject factors: stimulus (CS+, CS–) and phase (habituation, acquisition, extinction). For the repeated ratings derived from the acquisition and extinction phases, values were averaged. In case of significant stimulus × phase interactions post hoc t tests were performed. Non-sphericity was considered applying Greenhouse-Geisser correction.

Before performing statistical analyses for the SCR data, we log-transformed all peak amplitudes (SCR + 1) to normalize the distribution and further scored responses < 0 μs as zero in order to adequately characterize non-responses to the CS. Afterwards we separated the existing 18 CS+ and 18 CS– extinction trials into early (n = 9) and late (n = 9) responses to compare SCR and fNIRS parameters. SCR data were analysed with repeated-measures ANOVA with stimulus (CS+, CS–) and phase (habituation, acquisition, extinction, latency) as within-subject factors. A significant interaction of both factors was further assessed with one-tailed post hoc t tests at a significance level of p < 0.05 due to our directed hypotheses in fear and extinction learning. In relation to our assumptions concerning the fNIRS signal changes, we expected that successful extinction will be indicated by a decrease in SCR to the CS+ during the time course of extinction training.

Analogous to the procedure for SCR analyses, fNIRS signals were divided into an early and late phase, each consisting of nine trials. Because of our interest in signal changes occurring in response to the CS+ onset and accordingly the anticipated UCS onset, all trials were time-locked to the jitter mean, i.e. 4,500 ms after CS+ onset, and screened for artefacts. O$_2$Hb changes were preprocessed by applying a low-pass filter of 0.5 Hz and a cosine filter correcting for low-frequency signal drifts. In a next step, functional data were modelled by four regressors (CS+ early, CS– early, CS+ late, CS– late) (online supplementary fig. 1; see www.karger.com/doi/10.1159/000337002). Events per condition were further modelled as 6 functions and convolved with a gaussian hemodynamic response function at a peak time of 6.5 s. Time series were analysed by applying a general linear model approach [49] using Matlab Version 7 software (Mathworks Inc.). The resulting β estimates per condition and subject served as parameter set for subsequent testing. According to our hypothesis of an increasing O$_2$Hb concentration towards early and late extinction trials for CS+ compared to CS–, we determined differential β values for CS+ and CS– each and contrasted these differences by using paired t tests: [(CS+late – CS+early) – (CS–late – CS–early)]. Correction for multiple comparisons across probe set channels were performed by using a cluster permutation approach. Specifically, we compared the cluster size of significantly active channels (at p < 0.1 for each channel) to the distribution of cluster sizes expected under the null hypothesis (adapted for the 2-d fNIRS case from Wager et al. [50]). To obtain the null distribution, we performed 10,000 permutation tests across all channels given a single channel p value < 0.1. Activation was thus considered significant if the probability of obtaining this cluster size under the null hypothesis was p < 0.05. According to a probabilistic map (http://www.jichi.ac.jp/brainlab/virtual_regE.html#AnatomLabel) we provide MNI coordinates (x, y, z) of significant fNIRS channels to allow for integration of our results across imaging methods. SCR and fNIRS data were tested for significant correlations using Pearson’s correlation coefficient and one-tailed tests due to our expectations of negative correlations between SCR and O$_2$Hb within fNIRS channels or even channel clusters during the extinction phases. For exploratory purposes, we examined the influence of expectancy and SCR as well as fNIRS data to consider the influence of prediction error (two-tailed tests).
Results

As expected, subjects rated the UCS as quite unpleasant (mean 7.8, SD 1.3, range 4–10) indicating that the scream of 95 dB was aversive enough to induce fear conditioning. All 34 participants reported awareness of the CS-UCS contingency at the end of the acquisition phase and displayed a significant linear decrement of the UCS expectancy ratings during the extinction phase (linear trend test: $F(1, 33) = 24.53, p < 0.001$) from 78% after the first third, over 62% after the second third to 52% at the end of the experiment. Concerning the valence and arousal ratings, the repeated-measures ANOVA yielded main effects for stimulus (valence: $F(1, 33) = 29.37, p < 0.001$; arousal: $F(1, 33) = 43.25, p < 0.001$) and phase (valence: $F(1.5, 52.7) = 15.33, p < 0.001$; arousal: $F(1.5, 50.8) = 13.98, p < 0.001$) as well as significant stimulus × phase interactions (valence: $F(1.3, 42.3) = 23.49, p < 0.001$; arousal: $F(1.3, 41.4) = 34.02, p < 0.001$). As expected, post hoc t tests revealed that CS+ and CS– were equally evaluated after the habituation phase, both as neutral and sparsely arousing, but were differentially rated during conditioning. Herein, CS+ ratings were significantly lower in valence ($t_{33} = 6.51, p < 0.001$) and higher in arousal than for CS– ($t_{33} = 7.52, p < 0.001$). Comparing acquisition and extinction phase, the decrement for CS+ arousal and the increase for CS+ valence became significant (arousal: $t_{33} = 7.02, p < 0.001$; valence: $t_{33} = 6.5, p < 0.001$) although the differential ratings persisted during extinction learning (valence: $t_{33} = 5.55, p < 0.001$; arousal: $t_{33} = 6.13, p < 0.001$; fig. 2).

For SCR analyses we had to exclude 2 female subjects who did not display any fluctuations in their responses to either CS or UCS across the whole experiment. Analyses of the remaining sample of 32 subjects revealed similar results as for subjective ratings indicating a successful conditioning during acquisition and additionally suggesting extinction learning within the last experimental phases (fig. 3). The 2 × 4 repeated-measures ANOVA showed significant main effects for stimulus ($F(1, 31) = 15.82, p < 0.001$) as well as phase ($F(3, 93) = 6.74, p < 0.001$) and again a significant interaction ($F(3, 93) = 4.51, p = 0.005$). Post hoc t tests demonstrated significantly higher SCR amplitudes to CS+ than CS– during acquisition ($t_{31} = 4.87, p < 0.001$), which significantly diminished during the extinction phases (paired t test for SCR to CS+ during acquisition compared to early extinction: $t_{31} = 4.15, p < 0.001$, and late extinction: $t_{31} = 3.77, p < 0.001$). Differences between CS+ and CS– remained significant during the time course of extinction (early extinction: $t_{31} = 1.93, p = 0.032$; late extinction: $t_{31} = 2.5, p = 0.009$). However, SCR amplitudes during the extinction phase returned to habituation level (early extinction: $t_{31} = 0.55, p = 0.293$; late extinction: $t_{31} = 0.69, p = 0.248$) and further did not change significantly through early and late extinction trials ($t_{31} = 0.33, p = 0.746$).

For the fNIRS data we did not have to exclude channels from analyses due to little motion artefacts in the

Fig. 2. Valence and arousal ratings. Assessment by using the Self-Assessment Manikin (SAM [42]). Significantly different values for CS+ and CS– per experimental phase (mean + SEM) are depicted as *** to indicate a significance level of $p < 0.001$. 

**Neuropsychobiology 2012;65:173–182**
frontal region as expected. We found four significant probe set channels for which our analysed t-contrast \([((CS+ \text{ late} - CS+ \text{ early}) - (CS– \text{ late} - CS– \text{ early}))\] revealed significant results, that is the difference between early and late trials was larger for CS+ than CS– (channel 35: \(t_{33} = 2.26, p = 0.016 [x = 27, y = 68, z = 9]\); channel 27: \(t_{33} = 2.22, p = 0.017 [x = –13, y = 68, z = 20]\); channel 26: \(t_{33} = 1.92, p = 0.032 [x = 15, y = 68, z = 21]\); channel 47: \(t_{33} = 1.79, p = 0.042 [x = 15, y = 71, z = 3]\)). Eleven other channels revealed significant results by trend and are mentioned for completeness (channel 20: \(t_{33} = 1.64, p = 0.055\); channel 16: \(t_{33} = 1.62, p = 0.057\); channel 24: \(t_{33} = 1.62, p = 0.058\); channel 36: \(t_{33} = 1.61, p = 0.059\); channel 19: \(t_{33} = 1.58, p = 0.062\); channel 21: \(t_{33} = 1.57, p = 0.066\); channel 5: \(t_{33} = 1.57, p = 0.063\); channel 37: \(t_{33} = 1.53, p = 0.068\); channel 29: \(t_{33} = 1.5, p = 0.072\); channel 45: \(t_{33} = 1.42, p = 0.083\); channel 51: \(t_{33} = 1.4, p = 0.086\)). Ten of the aforementioned probe set channels resulted in one single significant cluster (\(p < 0.03\); channels 35, 27, 26, 47, 16, 24, 36, 5, 37, 45); no other cluster reached the significance threshold. Figure 4 pictures mean \(\beta\) values for this cluster by separating into CS+ and CS– as well as early and late trials according to our analysed t-contrast (see above). The cluster is being composed of a significant increase of CS+ trials during the early and late extinction phase (\(t_{33} = 1.89, p = 0.034\)) and significant exceeding \(O_2\) Hb values between CS+ and CS– trials during the late extinction phase (\(t_{33} = 2.2, p = 0.018\)). We neither found differences between CS– trials across the two phases (\(t_{33} = 0.9, p = 0.187\)) nor between CS+ and CS– early trials (\(t_{33} = 0.23, p = 0.41\); fig. 4, 5).

As depicted in figure 5b, the hemodynamic responses approximately started according to the expected UCS onset (around 5,000 ms) and revealed no differences during CS presentation. Hemodynamic responses starting at the expected UCS onset during extinction learning seem to reflect a prediction error, i.e. the expected UCS did not occur.

**Correlations**

Correlation analyses concerning SCR and \(O_2\) Hb in the sample of 32 subjects revealed a negative correlation between mean cluster \(O_2\) Hb values for late CS+ trials and the difference score between early and late SCR to CS+ (\(r = –0.327; p = 0.034\)). Subjects who displayed decreasing SCR to CS+ from early to late trials as it is expected during successful extinction, showed higher \(\beta\) values within late extinction phase indicating higher activity within the mPFC. In order to adequately compare SCR and fNIRS data during the same time interval, we additionally performed SCR peak detection during the extinction phase in a later interval between 5.5 and 8.5 s after CS presentation. This segment corresponded to the analysed fNIRS segment, which was assumed to mark the hemodynamic response function onset predicting the UCS. Even if we changed the analysed SCR segment in this way, the negative correlation between the difference score of late and early SCR responses and the cluster activity during the
late extinction phase remained significant ($r = -0.324$; $p = 0.035$). The exploratory analyses concerning the recurrence of the UCS revealed that only the second out of three expectancy ratings correlated significantly with SCR to CS+ trials within the early ($r = 0.37$, $p = 0.037$) and late extinction phase ($r = 0.4$, $p = 0.022$), suggesting that subjects who subjectively tend to resist to extinction learning show appropriately higher SCR values during the extinction phases and herewith demonstrate less well extinction learning than participants who did report more certainty towards the disappearance of the UCS.

**Discussion**

In the present study, 34 healthy subjects underwent a fear conditioning paradigm with two neutral faces as CS and a loud, aversive scream as UCS to examine the time course of extinction learning by analysing concentration changes in $O_2$Hb across early and late extinction trials. SCR and valence as well as arousal ratings were assessed to ensure successful conditioning.

We found significantly different valence and arousal ratings for CS+ and CS– trials as well as SCR data after the habituation phase. CS+ presentations evoked lower valence ratings and appropriately higher arousal ratings for CS+ than CS– as well as higher SCR amplitudes. During the extinction phase, fNIRS data displayed a significant increase in response to CS+ trials from early to late extinction within one cluster of 10 probe set channels covering the mPFC. The cluster activity elicited by CS+ trials further exceeded CS– trials during late extinction while $\beta$ values for CS– showed no significant difference across both extinction phases. To our knowledge, this is the first study investigating fear extinction by using fNIRS. This optical imaging method is restricted to the cortical surface and therefore cannot directly be compared to methods with higher spatial resolution, e.g. fMRI. Nonetheless, our findings are in accordance with previous imaging results confirming an mPFC contribution during within-session extinction [e.g. 25, 26, 36].

In order to critically review our results, we have to mention some inconsistencies of the data and will follow to discuss these aspects in light of the current literature. First one might argue that valence and arousal ratings as well as SCR data did not reflect successful extinction learning. We did find strong conditioning effects for all variables, but irrespective of the UCS absence during the extinction phase, participants continued to rate the former CS+ as significantly more unpleasant and more arousing than the CS–. However, there are other studies that found this kind of resistance to extinction in verbal reports [e.g. 25–28], and besides, UCS expectancy ratings showed a constantly decreasing expectancy across the extinction phase, reflecting that our participants did unlearn the CS-UCS association. Differential verbal ratings for CS+ and CS– might reflect lasting aversiveness of the UCS that prolonged throughout extinction trials, but on
the physiological level we do see an altered fear processing. SCR levels during early and late extinction decreased significantly from acquisition and further reached the habituation level as it is defined for successful extinction learning. Above all, at the rate of extinction and acquisition the number of trials is comparable to former studies [23, 27] or even contains a higher number of extinction trials [25–28].

Another constraint of our study is related to the mPFC activity we associated with extinction learning. A recently published review by Etkin et al. [51] argued that mPFC activity during extinction learning might reflect remnants of fear conditioning because studies on fear appraisal and sympathetic arousal also found mPFC engagement while generating fear responses. If mPFC activity would indeed reflect an explicit threat evaluation, one would have expected a decrease in activity from early to late trials contrarily to the increasing mPFC activity we found in the present study. Beta values for CS+ and CS− also started on an equal level during early extinction, we thus argue that mPFC activity in our study reflects extinction learning rather than a fear response. This is in line with the already explained successful induction of extinction, imminent in our UCS expectancy ratings and SCR data. Moreover, correlations between SCR and O₂Hb values emphasise the expected top-down control executed by the mPFC as subjects who exhibited a greater SCR decrement from early to late extinction phase also revealed higher β values and thereby more activity in the mPFC during the late CS+ condition.

A study investigating fear conditioning as a form of prediction error learning does also confirm our assumption. Spoormaker et al. [52] examined CS+ trials in which no UCS was administered and found increased activity in ventromedial, dorsolateral and orbitofrontal regions as neuronal correlates of this so-called negative prediction error. The absence of negative consequences therefore seems to be associated with prefrontal engagement that would also fit explanations of fear extinction [52]. This might also explain the timing of our hemodynamic response function. The temporal gap between CS+ and UCS presentations enabled us to examine the onset of the expected neuronal response towards the anticipated UCS during the extinction phase rather than the CS onset. The mentioned negative prediction error is existent if the UCS did not occur against one’s expectation. This mPFC-coupled learning process could only start in the absence of the UCS and not in the beginning of the CS presentation. Linnman et al. [53] investigated neuronal responses on shock delivery in a fear conditioning paradigm and found increased engagement of the dorsal anterior cingulate cortex, a region corresponding to the mPFC, during the non-delivery of an expected UCS. This finding fits our results as well, although it restricts comparability to available studies on fear extinction that did not provide such temporal information about CS and UCS. Future studies taking these differences into account would certainly contribute to a better understanding of temporal interactions such as the functional mPFC-amygdala coupling.

One major limitation of our study is the restricted application of fNIRS during the extinction phase. We discussed our result of significant mPFC activity as successful extinction learning, but in fact we cannot strictly obviate mPFC contribution during the acquisition phase. Numerous animal studies highlighted prefrontal contribution during extinction learning, i.e. when the CR is already acquired. Moreover, a systematic review about neuroimaging literature on human fear conditioning by Sehlmeyer et al. [12] did not find support for mPFC involvement during fear conditioning. Thus, it appears reasonable to restrict our fNIRS measurement exclusively to the extinction phase.

Secondly, we have to admit that we did not investigate mPFC activity during extinction retention, i.e. 24 h after the initial fear conditioning. The present study was not intended to compare mPFC engagement on the acquisition and recall of extinction. We well know that the ventromedial prefrontal cortex frequently found in animal studies is involved in recall of extinction rather than the initial acquisition [16], on the other hand there are the aforementioned studies in humans that found mPFC activity already during within-session extinction. It might be possible that long-term storage of extinction memory is supported in other for example more dorsal situated brain regions as Gottfried and Dolan [24] already speculated. Future studies therefore have to consider the consolidation of extinction memory by implementing a second extinction training after a delay period.

A third limitation relates to context changes which might have occurred through attaching the fNIRS probe set. In this connection we have to consider context dependency of extinction learning suggesting a return of fear by presenting the formerly CS+ again in the initial context, i.e. without the fNIRS probe set. We would again like to stress the fact that we kept all other parameters constant during the short break for attaching the probe set to minimize context effects. However, future studies using fNIRS for investigations of prefrontal activation during extinction learning could overcome this limitation by implementing shorter fear conditioning paradigms to assess all experimental phases.
Conclusions

The present study revealed increasing mPFC activation to CS+ trials during extinction that was different from that for a CS− which displayed no change across early and late extinction learning. Based on these findings, we propose mPFC activity during extinction learning to reflect better regulation of CR expression. The increase of prefrontal contribution from early to late extinction trials seems to be associated with changes in the associative significance of CS+ and UCS. Increasing associative strength might thereby rely on amygdala activity, decreasing associative strength appears to be inversely correlated with mPFC activity. Patients suffering from anxiety disorders or even high trait-anxious subjects have been characterized by increased CR to threat cues and reduced extinction [54]. Thus, they show deficient associative learning and accordingly deficient recruitment of amygdala and mPFC [26, 28].

Future studies have to examine cortical-subcortical interactions in more detail to ascertain strategies to affect mPFC activity in the treatment of anxiety disorders. Combined methods such as fNIRS and fMRI would further provide complementary results [44] by offering both high temporal as well as high spatial resolution. Here, a more precise definition of mPFC subregions involved in extinction learning might open up prospects to strengthen prefrontal areas, and transcranial magnetic stimulation could for instance be such a tool [55]. Transcranial magnetic stimulation is as restricted to the cortical surface as fNIRS. The use of fNIRS for mapping the prefrontal cortex is therefore not contradictory by searching innovative treatment options for facilitating extinction learning or even exposure therapy.

Acknowledgments

The authors would like to thank Juliana Rost for her support in participant recruiting and data acquisition. We also thank Michael M. Plichta for his ideas in data analyses and modelling, Evelyn Glotzbach who helped us with analysing the SCR data as well as Wilma Harnisch for calibrating our sound system.

This publication was funded by the German Research Foundation (DFG, SFB TRR 58, C04 project). The funding source had no role in study design, data collection and analyses, decision to publish, or preparation of the manuscript.

References

11 Pape HC, Pare D: Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiol Rev 2010; 90: 419–463.

Medial Prefrontal Cortex Activity during the Extinction of Conditioned Fear

Neuropsychobiology 2012;65:173–182