Published in final edited form as:

Neuroscience. 2018 May 15; 378: 11–21. doi:10.1016/j.neuroscience.2016.04.003.

Neural correlates of success and failure signals during neurofeedback learning

Joaquim Radua, MD, BStatPhD^{1,2,3}, Teodora Stoica⁴, Dustin Scheinost, PhD⁴, Christopher Pittenger, MD, PhD^{5,6,7}, and Michelle Hampson, PhD⁴

¹FIDMAG Germanes Hospitalàries - CIBERSAM, Sant Boi de Llobregat, Barcelona, 08830, Spain

²Institute of Psychiatry, Psychology and Neuroscience, King's College London, SE5 8AF, UK

³Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, 171 77, Sweden

⁴Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT, 06510, USA

⁵Department of Psychiatry, Yale School of Medicine, New Haven, CT, 06510, USA

⁶Department of Psychology, Yale University, New Haven, CT, 06510, USA

⁷Child Study Center, Yale School of Medicine, New Haven, CT, 06510, USA

Abstract

Feedback-driven learning, observed across phylogeny and of clear adaptive value, is frequently operationalized in simple operant conditioning paradigms, but it can be much more complex, driven by abstract representations of success and failure. This study investigates the neural processes involved in processing success and failure during feedback learning, which are not well understood. Data analyzed was acquired during a multisession neurofeedback experiment in which ten participants were presented with, and instructed to modulate, the activity of their orbitofrontal cortex with the aim of decreasing their anxiety. We assessed the regional blood-oxygenation-level-dependent response to the individualized neurofeedback signals of success and failure across twelve functional runs acquired in two different magnetic resonance sessions in each of ten individuals. Neurofeedback signals of failure correlated early during learning with deactivation in the precuneus/posterior cingulate and neurofeedback signals of success correlated later during learning with deactivation in the medial prefrontal/anterior cingulate cortex. The intensity of the latter deactivations predicted the efficacy of the neurofeedback intervention in the reduction of anxiety. These findings indicate a role for regulation of the default mode network during feedback learning, and suggest a higher sensitivity to signals of failure during the early feedback learning and to signals of success subsequently.

Address for correspondence: Dr. Joaquim Radua, FIDMAG Germanes Hospitalàries, Carrer del Dr. Antoni Pujadas 38, Sant Boi de Llobregat, 08830 Spain, Telephone: (+34) 936529999, FAX: (+34) 936400268, jradua@fidmag.com.

Conflict of Interest:

The authors declare no competing financial interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

Failure; feedback; functional magnetic resonance imaging (fMRI); learning; neurofeedback; success

1. INTRODUCTION

Feedback-driven learning is observed across phylogeny and is of clear adaptive value. When a child learns to ride a bicycle, for example, movements leading to successful weight balance are reinforced, whereas the opposite applies to movements leading to instability. Similarly, biofeedback enables learned control over physiological activity (Association for Applied Psychophysiology and Biofeedback et al., 2007). For example, a visual feedback representation of heart rate allows a subject to try different strategies to decrease it. There is evidence that biofeedback is effective for symptom reduction in some clinical conditions, including pain (Nestoriuc et al., 2008, Glombiewski et al., 2013), epileptic seizures (Tan et al., 2009), and attention deficit-hyperactivity disorder (Arns et al., 2009), as well as for improving motor performance after stroke (Stanton et al., 2011).

Real-time functional magnetic resonance imaging (rt-fMRI) neurofeedback adapts this strategy to enable volitional control over regional brain activity, rather than a peripheral physiological parameter (Stoeckel et al., 2014). Regional brain activity is monitored using fMRI and is provided as a feedback signal to the subject, who learns, over time, to control it. Although rt-fMRI neurofeedback is a relatively new technique, research using this technique has expanded dramatically in recent years, and many labs around the world now have or are currently collecting rt-fMRI neurofeedback data sets. As rt-fMRI neurofeedback studies use a wide variety of learning paradigms and collect data in many different clinical groups, re-analyses of these data sets could potentially provide valuable information about the neural mechanisms of learning and how they vary across contexts and populations. However, the potential to explore the neurobiological substrates of learning in these data sets has been underappreciated. As we hope to illustrate here, rt-fMRI neurofeedback provides a rich source of data for studying the neural basis of feedback learning, because brain activity is necessarily monitored (and recorded) throughout learning. Here we examine brain activity patterns associated with success (positive feedback) and failure (negative feedback) signals during a recently-described rt-fMRI neurofeedback experiment (Scheinost et al., 2013).

Several previous, non-neurofeedback studies have contrasted brain responses to positive and negative feedback during learning tasks (Nieuwenhuis et al., 2005, Marco-Pallares et al., 2007, van Duijvenvoorde et al., 2008, Peters et al., 2014). Results from these studies have implicated a number of different brain areas including the striatum, fronto-parietal regions and the anterior and posterior cingulate cortex. These studies have used binary signals of success and failure, rather than continuous signals, and have reported those brain areas responding more to success than failure or vice-versa. In many real-world feedback learning scenarios (such as motor learning and socio-emotional development), feedback is not provided in a binary manner, but in a continuous, dimensional manner. Thus, for the sake of ecological validity, examining brain patterns associated with the processing of continuous

success/failure signals is important. In addition, the continuous nature of success/failure signals in our neurofeedback paradigm enables us to identify brain activity patterns associated specifically with success and failure rather than a contrast between the two, which the previous feedback learning studies have been limited to reporting.

We designed a neurofeedback intervention to help individuals reduce contamination anxiety by learning to control a region of their orbitofrontal cortex (OFC) whose activation was associated with that anxiety. Participants were presented with a graphical readout of OFC activity in real time and were instructed, in intermingled blocks, to increase or decrease it during presentation of visual stimuli, some of which provoked contamination anxiety. As described previously (Scheinost et al., 2013), the subjects learned to control OFC activity, and this successful neurofeedback learning was associated with reduced anxiety and with changes in brain functional connectivity.

To investigate the neural substrates of success and failure neurofeedback signals, we analyzed whole-brain fMRI data acquired during rt-fMRI neurofeedback sessions to identify correlates of the neurofeedback signals. As very little is known about the neural substrates of continuous success and failure signals during feedback learning, we took an exploratory approach and conducted whole-brain analyses to identify relevant brain areas. However, based on previous studies (Nieuwenhuis et al., 2005, Marco-Pallares et al., 2007, van Duijvenvoorde et al., 2008, Koralek et al., 2012, Garrison et al., 2013, Peters et al., 2014), we hypothesized a potential involvement of a few regions such as the medial prefrontal cortex (MePFC)/anterior cingulate cortex, the precuneus/posterior cingulate, the inferior/middle frontal gyrus and the striatum.

2. EXPERIMENTAL PROCEDURES

2.1 Neurofeedback experiment and data acquisition

This study analyzes the fMRI data acquired during a neurofeedback experiment detailed elsewhere (Hampson et al., 2012, Scheinost et al., 2013). In that experiment, 20 healthy participants with high but subclinical contamination anxiety (scoring 8 or greater on the Obsessions and Washing Compulsions Subscale of the Padua Inventory-Washington State University Revision (PI-WSUR) (Burns et al., 1996)) underwent a multisession fMRI-based neurofeedback protocol while lying in the MRI device. Ten subjects received real neurofeedback, while the others received sham neurofeedback (in which the feedback signal was uncorrelated with their brain activity). The latter group was required in order to distinguish the therapeutic effects of neurofeedback from other aspects of the intervention, such as viewing many contamination related images. Only data from the real neurofeedback subgroup are analyzed here, as no neurofeedback learning occurred under the sham condition. All participants consented in accordance with a protocol reviewed and approved by the Yale University Human Research Protection Program.

Participants were presented with contamination-related photographs and instructed to either increase or decrease the activity of their OFC, which was presented as a line below the photograph (Figure 1). Each block lasted for 26 seconds, so that each ~4.5min fMRI run comprised 3 pairs of increase/decrease neurofeedback blocks, plus a total of 4 neutral-

photographs resting blocks before, between and after the pairs. Acquisition parameters were optimized for detection of OFC signal as follows: 1.5T Siemens Sonata scanner, T2*-sensitive gradient-recalled single shot echo-planar pulse sequence, TE 30ms, TR 2000ms, FOV 200mm, 3.1mm isotropic voxels, 31 axial-oblique AC-PC aligned slices covering all the OFC and most of the brain above.

Before neurofeedback, each individual underwent an MRI session in which the subregion of his/her OFC most involved in contamination anxiety was localized and participated in a Reappraisal Strategy Development session to develop initial individualized cognitive strategies to increase or decrease the anxiety; these strategies constituted a starting point for trial-and-error learning during the neurofeedback sessions. Subjects then underwent two neurofeedback sessions, which comprised a T1 structural MRI sequence, six neurofeedback runs with increase and decrease neurofeedback blocks being presented in a counterbalanced order, and other MRI sequences used for other analyses (Scheinost et al., 2013).

2.2 Evaluation of the behavioral response

A behavioral measure of each subject's ability to control contamination anxiety was collected before the first neurofeedback session, between the first and the second sessions, and several days after the second session (Hampson et al., 2012, Scheinost et al., 2013). Subjects were instructed to control their anxiety while viewing 25 contamination-related images and to report how much anxiety they experienced in response to each image on a 1–5 scale. The images shown before and after the neurofeedback sessions were different, to avoid habituation, but were matched in terms of the level of contamination anxiety they induced in pilot experiments (Hampson et al., 2012).

This behavioral measure of control over contamination anxiety provided 731 data measures (3 measures X 10 subjects X 25 images = 750 responses, with 19 missing values). These responses were modeled as a function of the assessment session in which they were collected (before the first, between the first and the second, or after the second neurofeedback session). Subject was included as a random factor in order to take the personal effect into account. A multiple imputation approach was used to randomly impute the 19 (2.5%) missing responses based on the session and the subject. R commands lmer and mi from packages lme4 and mi were used to conduct these analyses (Su et al., 2011, Bates et al., 2014). Note that such analysis is more powerful than the simple t-tests used in the previous publication (Scheinost et al., 2013).

2.3 First-level fMRI analysis

fMRI data processing was carried out using FEAT, part of the FMRIB's Software Library (FSL, www.fmrib.ox.ac.uk/fsl). Standard pre-processing steps with default parameters comprised: a) motion correction (Jenkinson et al., 2002); b) non-brain removal (Smith, 2002); c) spatial smoothing using a Gaussian kernel of FWHM 5mm; d) grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor; and e) high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting with $\sigma=50$ s). Runs with moderate to high motion (defined as more than 0.3mm average relative displacement) were discarded.

A first-level, time-series statistical analysis accounting for local autocorrelation (Woolrich et al., 2001) was separately carried out for each run of each participant. The following independent variables were included in the analysis: neurofeedback blocks, intensity of success signals, and intensity of failure signals. All regressors were convolved with a gamma function and included with their temporal derivatives; note that these derivatives are commonly used in FSL to account for regional timing differences in the hemodynamic response function and were not further analyzed (Boynton et al., 1996). The regressor 'neurofeedback blocks' was only included to remove brain activity changes of no interest such as the overall brain activation associated with contamination-related visual perception and its associated anxiety, or the brain changes related to the efforts and strategies used to modulate the OFC activity.

The contrast 'success' (intensity of success signals) was coded as the magnitude of the increase of the neurofeedback line during increase blocks or the magnitude of the decrease of the line during decrease blocks (Figure 2), and it was intended to capture changes of neural activity related to the processing of neurofeedback signals of success. Note that an increase of the line during an increase block, or a decrease during a decrease block, was perceived as a success by the individuals.

Similarly, the contrast 'failure' (intensity of failure signals) was coded as the magnitude of the decrease of the neurofeedback line during increase blocks or the magnitude of the increase of the line during decrease blocks (Figure 2), and it was intended to capture changes of neural activity related to the processing of neurofeedback signals of failure. The use of continuous quantitative variables ensured that the two contrasts were not simply the complements of each other.

Following individual subject regression analysis, functional statistical images were registered to MNI space with the following algorithm. First, the structural images acquired during the same session scan had the non-brain tissue removed and were visually inspected to detect any acquisition or brain-extraction artifact. Second, the functional statistical images were optimally registered to the structural images using the default, recommended Boundary-Based-Registration (BBR) method (Greve and Fischl, 2009), which is based on the white-matter boundaries and the fact that it is expected to see reliable changes in intensity across these boundaries in the functional images, whereas the grey-matter boundaries tend to be less reliable. Third, the structural images were registered to the standard MNI space. And fourth, the transformation resulting from this registration was applied to the functional statistical images. Registered functional images were also visually inspected to detect gross first-level artifacts or erroneous registrations.

2.4 Statistical analysis of imaging data

Data from first-level fMRI analyses were introduced into a nested mixed-effects model accounting for within-run variance, within-subject between-run variance and between-subject variance (Beckmann et al., 2003, Woolrich et al., 2004); the analysis also included a voxelwise estimation of the probability that a subject is an outlier and downweighting accordingly (Woolrich, 2008). Results were voxel-based thresholded at $p < 0.05$ corrected for multiple comparisons at the cluster-size-level (i.e. only those clusters with more than 546

-success- or 521 -failure- voxels, corresponding to corrected $p < 0.05$, were considered statistically significant; smoothness for the GRF analysis was empirically obtained for each contrast but in both cases $\sqrt{\det(\lambda)}$ was about 0.17) (Worsley, 2001).

Clusters of statistical significance for the success and failure contrasts were further analyzed. Specifically, the mean blood-oxygen-level dependent (BOLD) response of each cluster was extracted and introduced into: a) a paired t-test to contrast success vs. failure responses, i.e. to confirm that a region was more (de)activated to success than to failure or vice versa (note that the whole-brain analyses separately assessed the responses to success and to failure but not their comparison); b) a paired t-test to assess whether differences could exist between responses during increase and during decrease neurofeedback blocks, e.g. response to success being higher when successfully decreasing the OFC activity during decrease blocks than when successfully increasing the OFC activity during increase blocks; c) a multiple regression to test the effects of the scanning session (first vs. second) and run (1 to 6 for each session), e.g. response to success being progressively higher in time; and d) a regression to test the relationship with the improvement of subjective anxiety, e.g. individuals with higher responses to success having higher improvements of the contamination-related anxiety. The last two analyses included visual inspection of the data to assess the shape of the residuals and detect potential outliers, which were formally established based on Cook's distance $> 4/\sqrt{n-2}$ and removed; there was one outlying run in c) and one outlying subject in d), and their removal was associated to only minor changes in the results. P-values were Bonferroni-corrected for the number of clusters.

3. RESULTS

The neurofeedback intervention was effective: the mean anxiety response to contamination-related images presented during the assessment sessions (outside the magnet) decreased both during the first session (from 3.13 to 2.80; mixed-effects model with multiple imputation of missing responses: $t_{463} = -3.1$, $p = 0.002$) and during the second session (from 2.80 to 2.55; $t_{461} = -2.3$, $p = 0.02$).

Only one fMRI run from one participant had to be excluded due to excessive motion, with the other 119 runs being included in the fMRI analyses. Brain responses to the success and failure signals are detailed in Table 1 and displayed in Figure 3.

The most extensive responses were observed in the visual cortex (extending to fusiform gyri and cuneus), in the form of negative correlations with the intensity of the changes in the neurofeedback line. This relationship was found in both the success and the failure contrasts, indicating that it did not depend on the valence of the change of the line (success vs. failure) but only on its intensity. That said, it must be noted that the overlap was only partial, as decreases of activity in response to success signals largely extended to cuneus, whereas those in response to failure signals largely extended to fusiform gyri.

The MePFC showed deactivations specifically with signals of success. These deactivations were initially weak, but they grew progressively during the runs of the first session and were substantially stronger during the second session (Figure 4a). The individual mean intensity

of these MePFC deactivations was correlated with the individual reduction of the contamination-anxiety after the neurofeedback intervention (Figure 5).

A similar but opposite pattern was observed in the precuneus/posterior cingulate cortex (Pc/PCC), which showed deactivations specifically associated with failure feedback. These deactivations progressively diminished during the runs of the first neurofeedback session and were substantially weaker during the second session (Figure 4b).

The right superior temporal/postcentral gyri showed a success-related deactivation similar to that of MePFC, although weaker and invariant across learning sessions. Likewise, the right inferior/middle frontal gyri as well as the middle cingulum/supplementary motor area showed failure-related deactivations similar to that of Pc/PCC, although again weaker and invariant across learning sessions.

The response in right inferior/middle frontal gyri showed a larger response during increase blocks, though this trend was not significant after Bonferroni correction for multiple comparisons (uncorrected $p=0.050$). None of the other regions showed an effect of block type, whereas differences between success and failure reached or approached statistical significance in all non-visual clusters (all $p<0.1$ for success, all $p<0.01$ for failure).

4. DISCUSSION

Specific findings

We aimed to detect the brain networks associated with the processing of continuously valued neurofeedback signals. To this end, we analyzed the data acquired during a series of neurofeedback runs in which participants were trained to modulate their OFC activity in order to control their anxiety. Importantly, subjects were instructed to both increase and decrease OFC activity during the task and received neurofeedback under both conditions, which permitted statistical dissociation of the valence of the neurofeedback signal (success vs. failure) and the changes in OFC activity (and its associated anxiety).

We found deactivations of the MePFC in response to signals of success and deactivations of the Pc/PCC in response to signals of failure. A similarly intriguing dissociation as a function of emotional valence was seen in a prior study examining activity in these brain areas during self-reflection: the MePFC was reported to activate more when thinking about hopes and aspirations, while the Pc/PCC was reported to activate more when thinking about duties and obligations (Johnson et al., 2009). The anteroventral part of the MePFC cluster partly overlapped with the medial part of the OFC target (Figure 3). However, it must be noted that there was a delay between a decrease of the OFC activity and the model of BOLD signal changes associated with a decrease of the neurofeedback line (including processing time, and more importantly, hemodynamic delay), and that these decreases were dissociated from 'success' because the experiment included both increase and decrease blocks.

These regions are hubs of the default mode network, a brain system that is typically more active when individuals are not focused on the external environment (Greicius et al., 2003, Buckner et al., 2008). Deactivation of this network may be a prerequisite for assessment of

feedback signals, which requires focusing on the ‘external environment’ (the feedback signals). An association of feedback learning with reduced activity in these area is consistent with a recent meta-analysis of the brain response to prediction error in both instrumental and Pavlovian studies, which found responses in the same regions (as well as in the striatum), apparently with only minor differences depending on the type and valence of the conditioning (Garrison et al., 2013). The lack of striatal response in our study may be related to the striatal activation being associated not with success or failure *per se* but rather with prediction error, which we cannot examine without some model of the subjects’ expectations at each time point.

Interestingly, deactivations of the Pc/PCC progressively diminished during the runs of the first neurofeedback session, at the same rate that deactivations of the MePFC progressively grew (Figure 4). This progressive reduction of the Pc/PCC and increase of the MePFC responses may be due to a progressive loss of the sensitivity to signals of failure and growth of the sensitivity to signals of success early in the learning process. Our behavioral data show that learning continues during the second neurofeedback session, when activity in the MePFC and Pc/PCC has already reached its asymptote (anxiety ratings declined from 3.13 to 2.80 during the first session and to 2.55 during the second session), and the conjoint analyses of fMRI and behavioral data show that mean MePFC deactivations across runs predicted the individual reduction of the contamination-anxiety after the neurofeedback intervention (Figure 5). These observations suggest an interesting model. Early in learning, subjects may be more oriented towards negative feedback, associated with Pc/PCC deactivation; but successful learning may require a shift in orientation towards positive feedback and MePFC deactivation. Thus, the early phase of learning may consist of this shift in orientation (and corresponding feedback-specific deactivation patterns), and learning, indexed by reduced contamination anxiety during the behavioral test sessions, may commence (or at least proceed more efficiently) after this shift is accomplished. This occurred mid-way through the first neurofeedback session, for most subjects.

Beyond MePFC and Pc/PCC, the most extensive responses to neurofeedback signals were observed in the visual cortex, which showed a negative correlation with the magnitude (but not direction) of changes in the neurofeedback signal. This correlation may be due to the varying difficulty in the assessment of the changes in the neurofeedback line in order to know the *direction* of the signal (up vs. down). The direction of the change is clearly more difficult to assess when the change is small (i.e. when the new line segment is mostly horizontal) than when the change is large, so greater attention to visual processing is required with small neurofeedback signals than with large neurofeedback signals. An alternate explanation is that large changes in the neurofeedback signal may tend to grab the attention of subjects, causing them to change their focus from the high complexity photographic image to the simple line graph, and thus large changes in the neurofeedback signal could result in reduced visual processing. According to either of these interpretations, these visual responses would be visual-specific and more related to the perception of the signals than to the learning process. Another explanation, at least regarding the fusiform gyrus, might be related to the emotional response to feedback signals, as the fusiform gyrus has been reported to activate in response to disgusting scenic stimuli and to deactivate in response to either happy or sad stimuli (Radua et al., 2013).

A last set of responses to neurofeedback signals comprised deactivations in the right superior temporal/postcentral gyri with success and in the right inferior/middle frontal gyri and the middle cingulum/supplementary motor area with failure. These could be associated with strategy switching, or a relaxation of strategy maintenance occurring in response to the failure.

Relationship to Broader Literature on Feedback Learning

The study of the neural substrates of success and failure during feedback learning has spanned modalities. A large electroencephalography (EEG) literature has examined negative deflections in EEG signal associated with error processing, referred to as event-related negativity (ERN/Ne) e.g., (Falkenstein et al., 2000). Although EEG findings can be difficult to relate to specific brain areas due to limited spatial resolution, it has been suggested that ERN is generated in a fronto-central region (Falkenstein et al., 1991). Whether the ERN is related to error processing or more general performance monitoring is a matter of debate (Carter et al., 1998) but it is robust phenomenon associated with feedback learning.

In the functional MRI literature, several prior studies have investigated the brain responses to positive vs. negative feedback during associative, rule, or timing learning tasks (Falkenstein et al., 2000, Nieuwenhuis et al., 2005, Marco-Pallares et al., 2007, van Duijvenvoorde et al., 2008, Peters et al., 2014). However, we are not aware of previous neuroimaging studies investigating the neural response to varying intensities of signals of success and signals of failure during feedback learning. Prior studies had categorical success/failure events (corresponding to positive and negative feedback respectively) and their analyses were contrasts of different event types (e.g. positive vs. negative feedback). In our study, conversely, the success/failure signals were of varying levels of intensity, and thus the regressor representing success was not simply the complement of the regressor representing failure (and vice-versa). This allowed us to investigate the brain responses to success and failure without performing an implicit or explicit contrast between event types. Thus, this is the first study to our knowledge reporting separate brain responses to success signals and to failure signals (rather than the contrast between the two) during feedback learning. Although the results of this study are not directly comparable to prior studies for this reason, the prior studies did implicate similar brain areas including the MePFC/anterior cingulate cortex (Nieuwenhuis et al., 2005, Marco-Pallares et al., 2007, van Duijvenvoorde et al., 2008, Peters et al., 2014), the precuneus/posterior cingulate (Nieuwenhuis et al., 2005, Marco-Pallares et al., 2007, van Duijvenvoorde et al., 2008, Peters et al., 2014) and the inferior/middle frontal gyrus (van Duijvenvoorde et al., 2008, Peters et al., 2014), as well as several other regions not identified here such as the striatum (Diedrichsen et al., 2005). The overlap between brain areas implicated in this study and brain areas implicated in prior feedback learning studies suggests that processing of success and failure during neurofeedback may be similar to the processing of success and failure in other feedback learning contexts, but more work is needed to confirm this possibility.

A recently published meta-analysis of real-time fMRI neurofeedback datasets examined brain areas activated specifically during neurofeedback training (Emmert et al., 2016). Although this meta-analysis looked at activation during neurofeedback in general, and thus

did not differentiate activation related to success or failure per se, it is interesting to note that precuneus/posterior cingulate was identified as a consistently deactivated brain area (unfortunately, the medial prefrontal region area we identified was not included in this meta-analysis due to inconsistent coverage across studies).

Strengths and Limitations of Current Study

Strengths of this study include the use of two sessions and twelve runs per participant, which substantially reduced intra-subject variance and allowed the analysis of the response to neurofeedback as a function of the run number and session (see Figure 4). Also, we used a nested mixed-effects model that accounts for within-run variance, within-subject between-run variance and between subject variance, as compared to other models where e.g. the first and the last run would be modeled as identical. The study had graded rather than binary neurofeedback, which in addition to adding ecological validity, permitted disentanglement of responses to success and failure signals from one another. Finally, the neurofeedback paradigm included both 'increase' and 'decrease' neurofeedback blocks, which permitted the disentanglement of the responses to the neurofeedback signals and the changes that result from the act of control itself. For example, fMRI data acquired during blocks in which subjects were instructed to decrease OFC activity would be difficult to interpret in isolation, as changes in brain activity correlating to successful OFC deactivation would represent an admixture of the response to a neurofeedback signal of success and the intended reduction in OFC activation. To ensure that there was no effect of block type (increase versus decrease) on the deactivations, we compared the response in each cluster between block types and found no evidence of such effects except in right inferior/middle frontal gyri. Thus, other than perhaps this region, all brain regions identified in this study deactivated to success (or failure) across both block types, despite the fact that during the two types of blocks success (or failure) was associated with opposite directions of change in OFC activity. These deactivations are therefore specific to success (or failure), independent of changes in OFC activity (and contamination anxiety). Furthermore, differences between success and failure reached or approached statistical significance in all non-visual clusters, indicating specificity to success (or failure) processing rather than something more general such as attentional orientation.

Limitations must be highlighted. First, the sample of participants was small, which may limit our statistical power to detect weak effects. This may explain why no increases of activity correlating with success or failure were detected. Second, the study was conducted using neurofeedback on individuals with high-contamination-related anxiety, and thus results cannot be directly extrapolated to feedback on other populations. Third, the acquisition scheme, which was optimized for the OFC signal analyzed during the neurofeedback intervention, discarded the most superior part of the brain, as well as the most inferior part of the cerebellum, and therefore no responses in these subregions could be detected. Fourth, this study was only intended to capture changes of neural activity following the perception of each neurofeedback signal, and was thus unable to detect more global experiences caused by the cumulative effect of the previous signals. Future studies may focus on these experiences using more complex approaches such as a moving average of the last neurofeedback signals. Fifth, we cannot rule out the effects of attention or arousal

on the observed responses which could be strongly associated with success or failure. However, it must be also noted that the responses to both success and failure were detected when the line was going both up and down, and thus their differences should only be due to differences between the responses to success and failure and not to a line increase or decrease, and the attention/arousal that may induce. Sixth, we used a relatively liberal cluster-forming threshold, which may be associated to a low spatial specificity (Woo et al., 2014) but it is probably necessary in small samples in order to detect weak diffuse signals. Seventh, we could not detect BOLD changes in the striatum, which was one of the hypothesized regions (Koralek et al., 2012). We cannot discard that such effect may have been detected in a larger sample. Also, the use of a relatively liberal cluster-forming threshold may lead to an underrepresentation of smaller localized structures, though this does not seem to be the reason here because we repeated the analysis with more conservative thresholds obtaining no other results. As noted above, the striatum may be specifically related to prediction error (which we did not assess) rather than success or failure processing per se. Finally, it must be noted that the analysis conducted in this study may represent only one of the potential analytic approaches. Other analyses could be based, for example, on the feedback prediction error.

Conclusions

To sum up, our key finding is that success and failure signals during neurofeedback learning correlate with deactivation of portions of the default mode network. Specifically, the Pc/PCC deactivates with signals of failure in the first stages of the learning process, and the MePFC deactivates with signals of success in subsequent stages. Furthermore, the deactivation in MePFC is correlated with learning, suggesting that the role of this region in processing success signals is not just incidental, but rather causally related to feedback learning. Further research is encouraged to confirm these observations and to examine how these brain areas interact with other parts of the brain during feedback learning.

More generally, this study illustrates the potential value of reanalyzing of rt-fMRI neurofeedback datasets to examine the neural substrates underlying learning. Given the wide range of populations and learning paradigms used in rt-fMRI neurofeedback studies, these datasets represent a rich resource for studying human feedback learning that have to date been underexplored.

Acknowledgments

Data collection was supported by R21 MH090384 (M. Hampson) and the broader research program was supported by R01s MH095789 and MH100068 (M. Hampson). J. Radua received grant support from Instituto de Salud Carlos III - Subdirección General de Evaluación and the European Regional Development Fund (personal grants Río Hortega CM11/00024 and Miguel Servet CP14/00041, and projects PI11/01766 and PI14/00292 integrated into the National Plan for Research, Development and Innovation 2008–2011 and 2013–2016), D. Scheinost from NIH (T32 DA022975), and C. Pittenger from the Nancy Taylor Foundation for Chronic Disease.

References

Arns M, de Ridder S, Strehl U, Breteler M, Coenen A. Efficacy of neurofeedback treatment in ADHD: the effects on inattention, impulsivity and hyperactivity: a meta-analysis. *Clinical Eeg and Neuroscience*. 2009; 40:180–189. [PubMed: 19715181]

- Association for Applied Psychophysiology and Biofeedback, Biofeedback Certification International Alliance, International Society for Neurofeedback and Research. About Biofeedback. 2007. <http://www.aapb.org/>
- Bates, D., Maechler, M., Bolker, B., Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–6. 2014. <http://CRAN.R-project.org/package=lme4>
- Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in fMRI. *Neuroimage*. 2003; 20:1052–1063. [PubMed: 14568475]
- Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*. 1996; 16:4207–4221. [PubMed: 8753882]
- Buckner RL, Andrews-Hanna JR, Schacter DL. The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci*. 2008; 1124:1–38. [PubMed: 18400922]
- Burns GL, Keortge SG, Formea GM, Sternberger LG. Revision of the Padua Inventory of obsessive compulsive disorder symptoms: distinctions between worry, obsessions, and compulsions. *Behav Res Ther*. 1996; 34:163–173. [PubMed: 8741724]
- Carter CS, Braver TS, Barch DM, Botvinick MM, Noll D, Cohen JD. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*. 1998; 280:747–749. [PubMed: 9563953]
- Diedrichsen J, Hashambhoy Y, Rane T, Shadmehr R. Neural correlates of reach errors. *J Neurosci*. 2005; 25:9919–9931. [PubMed: 16251440]
- Emmert K, Kopel R, Sulzer J, Bruhl AB, Berman BD, Linden DE, Horovitz SG, Breimhorst M, Caria A, Frank S, Johnston S, Long Z, Paret C, Robineau F, Veit R, Bartsch A, Beckmann CF, Van De Ville D, Haller S. Meta-analysis of real-time fMRI neurofeedback studies using individual participant data: How is brain regulation mediated? *Neuroimage*. 2016; 124:806–812. [PubMed: 26419389]
- Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr Clin Neurophysiol*. 1991; 78:447–455. [PubMed: 1712280]
- Falkenstein M, Hoormann J, Christ S, Hohnsbein J. ERP components on reaction errors and their functional significance: a tutorial. *Biol Psychol*. 2000; 51:87–107. [PubMed: 10686361]
- Garrison J, Erdeniz B, Done J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*. 2013; 37:1297–1310. [PubMed: 23567522]
- Glombiewski JA, Bernardy K, Hauser W. Efficacy of EMG- and EEG-Biofeedback in Fibromyalgia Syndrome: A Meta-Analysis and a Systematic Review of Randomized Controlled Trials. *Evid Based Complement Alternat Med*. 2013; 2013:962741. [PubMed: 24082911]
- Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc Natl Acad Sci U S A*. 2003; 100:253–258. [PubMed: 12506194]
- Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*. 2009; 48:63–72. [PubMed: 19573611]
- Hampson M, Stoica T, Saksa J, Scheinost D, Qiu M, Bhawnani J, Pittenger C, Papademetris X, Constable T. Real-time fMRI biofeedback targeting the orbitofrontal cortex for contamination anxiety. *J Vis Exp*. 2012
- Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. 2002; 17:825–841. [PubMed: 12377157]
- Johnson MK, Nolen-Hoeksema S, Mitchell KJ, Levin Y. Medial cortex activity, self-reflection and depression. *Social cognitive and affective neuroscience*. 2009; 4:313–327. [PubMed: 19620180]
- Koralek AC, Jin X, Long JD 2nd, Costa RM, Carmena JM. Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. *Nature*. 2012; 483:331–335. [PubMed: 22388818]
- Marco-Pallares J, Muller SV, Munte TF. Learning by doing: an fMRI study of feedback-related brain activations. *Neuroreport*. 2007; 18:1423–1426. [PubMed: 17712267]

- Nestoriuc Y, Martin A, Rief W, Andrasik F. Biofeedback treatment for headache disorders: a comprehensive efficacy review. *Appl Psychophysiol Biofeedback*. 2008; 33:125–140. [PubMed: 18726688]
- Nieuwenhuis S, Slagter HA, von Geusau NJ, Heslenfeld DJ, Holroyd CB. Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*. 2005; 21:3161–3168. [PubMed: 15978024]
- Peters S, Braams BR, Raijmakers ME, Koolschijn PC, Crone EA. The Neural Coding of Feedback Learning across Child and Adolescent Development. *J Cogn Neurosci*. 2014
- Radua J, Sarro S, Vigo T, Alonso-Lana S, Bonnin CM, Ortiz-Gil J, Canales-Rodriguez EJ, Maristany T, Vieta E, McKenna PJ, Salvador R, Pomarol-Clotet E. Common and specific brain responses to scenic emotional stimuli. *Brain Struct Funct*. 2013
- Scheinost D, Stoica T, Saksa J, Papademetris X, Constable RT, Pittenger C, Hampson M. Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Translational Psychiatry*. 2013; 3:e250. [PubMed: 23632454]
- Smith SM. Fast robust automated brain extraction. *Human Brain Mapping*. 2002; 17:143–155. [PubMed: 12391568]
- Stanton R, Ada L, Dean CM, Preston E. Biofeedback improves activities of the lower limb after stroke: a systematic review. *J Physiother*. 2011; 57:145–155. [PubMed: 21843829]
- Stoeckel LE, Garrison KA, Ghosh S, Wighton P, Hanlon CA, Gilman JM, Greer S, Turk-Browne NB, deBettencourt MT, Scheinost D, Craddock C, Thompson T, Calderon V, Bauer CC, George M, Breiter HC, Whitfield-Gabrieli S, Gabrieli JD, LaConte SM, Hirshberg L, Brewer JA, Hampson M, Van Der Kouwe A, Mackey S, Evins AE. Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *Neuroimage Clin*. 2014; 5:245–255. [PubMed: 25161891]
- Su YS, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*. 2011; 45:1–31.
- Tan G, Thornby J, Hammond DC, Strehl U, Canady B, Arnemann K, Kaiser DA. Meta-analysis of EEG biofeedback in treating epilepsy. *Clinical Eeg and Neuroscience*. 2009; 40:173–179. [PubMed: 19715180]
- van Duijvenvoorde AC, Zanolie K, Rombouts SA, Raijmakers ME, Crone EA. Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *Journal of Neuroscience*. 2008; 28:9495–9503. [PubMed: 18799681]
- Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*. 2014; 91:412–419. [PubMed: 24412399]
- Woolrich M. Robust group analysis using outlier inference. *Neuroimage*. 2008; 41:286–301. [PubMed: 18407525]
- Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*. 2004; 21:1732–1747. [PubMed: 15050594]
- Woolrich MW, Ripley BD, Brady M, Smith SM. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage*. 2001; 14:1370–1386. [PubMed: 11707093]
- Worsley, KJ. Statistical analysis of activation images. In: Jezzard, P., et al., editors. *Functional MRI: an introduction to methods*. Oxford: Oxford University Press; 2001.

HIGHLIGHTS

- Signals of failure correlated early during learning with deactivation in precuneus.
- Signals of success correlated later with deactivation in anterior cingulate cortex.
- Intensity of the latter predicted the efficacy of the neurofeedback intervention.

Figure 1. Real example of the neurofeedback software screen

Footnote to Figure 1: The screen showed contamination-related or neutral photographs and the neurofeedback line, which indicated the activity of the OFC of the participant.

Figure 2. Real example of neurofeedback line and the derived fMRI regressors

Footnote to Figure 2: The line in the top ('neurofeedback line') was extracted during feedback learning from orbitofrontal cortex (OFC) activity and presented to the participant at realtime below the contamination-related photographs (Hampson et al, 2012; Scheinost et al, 2013). Individuals were instructed to increase their OFC activity during the 'increase neurofeedback blocks' (shown in red) and to decrease it during the 'decrease neurofeedback blocks' (shown in blue). Increases of the neurofeedback line during the 'increase blocks' or decreases of the line during the 'decrease blocks' represented success and are shown in green (middle). Conversely, decreases of the neurofeedback line during the 'increase blocks' or increases of the line during the 'decrease blocks' represented failure and are shown in maroon (bottom). The analysis included 'block' (only included to remove brain activity changes of no interest such as those related to contamination-related visual perception and its associated anxiety or to the efforts and strategies used to modulate the OFC activity), 'success' and 'failure' regressors convolved with a gamma function (solid lines) and their temporal derivatives (dotted lines), which are commonly used in FSL to account for regional timing differences in the hemodynamic response function and were not further analyzed (Boynton et al., 1996).

Figure 3. Deactivations and negative correlations associated with success and failure

Footnote to Figure 3: A) Brain regions in which deactivations were associated with success signals. B) Brain regions in which deactivations were associated with failure signals. Both panels show negative correlations with the intensity of the signals. No activations or positive correlations were detected. The area where there was greatest overlap across subjects in their OFC target is shown as a multicolor bull's eye.

Figure 4. Effects of session (first vs. second) and run (1 to 6 in each session) on the success-related deactivations in medial prefrontal/anterior cingulate and failure-related deactivations in precuneal/posterior cingulate cortices

Footnote to Figure 4: The curve represents the fitted response (with 95% confidence interval) as a function of the session and the run. See Table 1 for coefficients and significance levels of the fitted regression.

Figure 5. Correlation between individual mean medial prefrontal/anterior cingulate success-related deactivation and change in subjective anxiety after the neurofeedback intervention

Footnote to Figure 5: The curve represents the fitted change in subjective anxiety (with 95% confidence interval) as a function of the individual mean deactivation. See Table 1 for the coefficient of correlation and the significance level.

Table 1

Brain responses associated with the success and failure neurofeedback signals.

	Cluster		Increase vs decrease blocks (b_c)	Effects of session and run ($b_{d,e}$)			Correlation with improvement (b_e)
	Number of voxels	P	Breakdown (number of voxels) (a_i)	Session	Run (1 st session)	Run (2 nd session)	
Activations with success signals (<i>none</i>)							
Deactivations with success signals							
<i>Bilateral visual extending to precuneus (peak at MNI -20, -64, 22)</i>	2359	2e-8	Cuneus (652) Precuneus (370) Calcarine (352) Lingual (304) Superior occipital (290) Fusiform (112)	n.s.	n.s.	n.s.	n.s.
<i>Medial prefrontal (peak at MNI 2, 46, -8)</i>	767	0.005	Medial prefrontal (547)	n.s.	-12.37 (p=0.002) (<i>f</i>)	-2.37 (p=0.006) (<i>f</i>)	-0.85 (p=0.004) (<i>f</i>)
<i>R superior temporal/postcentral (peak at MNI 42, -22, 20)</i>	668	0.01	R superior temporal (187) R postcentral (174) R Rolandic operculum (161)	n.s.	n.s.	n.s.	n.s.
Activations with failure signals (<i>none</i>)							
Deactivations with failure signals							
<i>Bilateral visual extending to fusiform (two clusters)</i>							
<i>R visual/fusiform (peak at MNI 8, -92, -6)</i>	2050	1e-7	Calcarine (338) R fusiform (451) R lingual (445) R middle occipital (125) R superior occipital (123) R inferior occipital (123)	n.s.	n.s.	n.s.	n.s.
<i>L visual/fusiform (peak at MNI -26, -70, -14)</i>	1695	1e-6	L fusiform (737) L lingual (284) L inferior occipital (171) L cerebellum (135)	n.s.	n.s.	n.s.	n.s.
<i>Precuneus and posterior cingulum (peak at MNI 6, -52, 6)</i>	915	0.0009	Precuneus (456) Posterior cingulum (172)	n.s.	12.57 (p<0.001) (<i>f</i>)	2.31 (p=0.003) (<i>f</i>)	n.s.
<i>R inferior extending to middle</i>	874	0.001	R inferior frontal (459) R middle frontal (276)	-2.27 (p=0.050)	n.s.	n.s.	-0.71 (p=0.020)

	Cluster		Increase vs decrease blocks (<i>b,c</i>)	Effects of session and run (<i>b,d,e</i>)			Correlation with improvement (<i>b,e</i>)	
	Number of voxels	<i>P</i>		Breakdown (number of voxels) (<i>a</i>)	Session	Run (1 st session)		Run (2 nd session)
<i>frontal (peak at MNI 40,34, -12)</i>								
<i>Middle cingulum extending and R supplementary motor area (peak at MNI -4, -16,42)</i>	862	0.001	Middle cingulum (322) R supplementary motor area (294) R superior frontal (118)	n.s.	n.s.	n.s.	n.s.	

(*a*) Labels based on AAL Atlas; regions with less than 100 voxels not shown; findings are bilateral unless otherwise indicated.

(*b*) Statistics based on the means of the clusters obtained in the main analysis.

(*c*) Paired t-test comparing increase and decrease blocks.

(*d*) Multiple regression by session, run number in first session and run number in second session, with non-significant regressor backwards removed.

(*e*) Removal of outliers (Cook's distance > 4/sqrt(n-2)).

(*f*) Statistically significant after Bonferroni-correction for the number of clusters.