An Investigation of Awareness and Metacognition in Neurofeedback with the Amygdala Electrical Fingerprint

Madita Stirner, Guy Gurevitch, Nitzan Lubianiker, Talma Hendler, Christian Schmahl, Christian Paret

ARTICLE INFO

Keywords: Amygdala, Awareness, Confidence, Metacognition, Neurofeedback, Self-regulation

ABSTRACT

Awareness theory posits that individuals connected to a brain-computer interface can learn to estimate and discriminate their brain states. We used the amygdala Electrical Fingerprint (amyg-EFP) - a functional Magnetic Resonance Imaging-inspired Electroencephalogram surrogate of deep brain activation - to investigate whether participants could accurately estimate their own brain activation. Ten participants completed up to 20 neurofeedback runs and estimated their amygdala-EFP activation (depicted as a thermometer) and confidence in this rating during each trial. We analysed data using multilevel models, predicting the real thermometer position with participant rated position and adjusted for activation during the previous trial. Hypotheses on learning regulation and improvement of estimation were not confirmed. However, participant ratings were significantly associated with the amyg-EFP signal. Higher rating accuracy also predicted higher subjective confidence in the rating. This proof-of-concept study introduces an approach to study awareness with fMRI-informed neurofeedback and provides initial evidence for metacognition in neurofeedback.

1. Introduction

Can humans be aware of their own brain states; i.e., can they perceive how strongly certain brain regions are activated at a given point in time? The answer to this intriguing question has not only academic relevance for the study of brain-mind relationship, but could inform the design of better brain-computer interfaces (BCIs) with practical implication for clinical application and cognitive training. The dawn of real-time functional Magnetic Resonance Imaging (fMRI) fueled BCI research and brought neuroscience-based treatment of mental disorders within reach (Thibault et al., 2018). Many different neural markers have been used for feedback to teach individuals with mental disorders to regulate their brains, called neurofeedback, including Blood Oxygenation Level Dependent (BOLD) changes of brain regions involved in emotion (Linhartová et al., 2019). Notwithstanding the growing literature to show feasibility and clinical utility, research on the principles that mediate neurofeedback learning is still limited. Individuals learn...
voluntary brain self-regulation when they receive contingent reinforcement for mental actions that are causally related to changes in the brain (Black et al., 1977; Caria, 2016). Additionally, individuals may become aware of brain states when they learn to distinguish between mental events that are correlated with changes in feedback and those that are not – a process called discrimination learning (Gaume et al., 2016). Two-process theory assumes reinforcement learning and discrimination learning to operate interactively when individuals practice neurofeedback (Gaume et al., 2016; Lacroix, 1986). Research using Electroencephalography (EEG) neurofeedback showed that individuals can learn to discriminate cortical activation markers such as neural frequency, cortical potentials and motor classifiers (Frederick, 2012; Frederick et al., 2016, 2019; Kotchoubey et al., 2002; Schurger et al., 2017). Although this work showed that individuals can estimate changes of neural activation with significant accuracy, it did not investigate metacognition of brain states, i.e., whether individuals were conscious about the accuracy of their estimations. Metacognitive skills are relevant to the study of awareness, but are still largely overlooked in BCI theory and neglected in empirical neurofeedback research altogether (Muñoz-Moldes & Cleeremans, 2020). Furthermore, it is unclear how existing research would generalize to fMRI-informed BCIs, which arguably enable more precise and flexible targeting of cortical and subcortical neural circuits (Lubianiker et al., 2019). To fill the knowledge gap, we focus here on amygdala neurofeedback. The amygdala guides emotional learning (LaBar et al., 1995), modulates behavior based on affective appraisals (Kuhn et al., 2020), and steers peripheral physiological responding (Inman et al., 2018). The amygdala is hypo-reactive in mood disorder and hyper-reactive in trauma-related disorders (Schulze et al., 2019) and anxiety disorders (Etkin & Wager, 2007), putting amygdala neurofeedback in the focus of recent research activity (Paret & Hendler, 2020). To allow for repeated sessions and yet precision of target, we recorded brain activation as indicated by the amygdala Electrical Fingerprint (amyg-EFP) signal; an EEG-substrate optimized to correlate with the amygdala BOLD signal (Meir-Hasson et al., 2016). Previous research showed that the amyg-EFP reliably predicts the amygdala BOLD signal and can be used for neurofeedback training in healthy persons (Keynan et al., 2016, 2019). With the amyg-EFP we could administer a high neurofeedback dose without resource-consuming fMRI scanning, which is an advantage for a multi-session study design like ours.

Participants downregulated the amyg-EFP signal in up to 20 neurofeedback runs. After each regulation trial, they rated on a continuous scale how much they believed that the amyg-EFP was activated. Importantly, participants did not receive feedback on the actual activation state until after the rating. The rating accuracy, i.e., the difference between the rating and the real amyg-EFP activation change, was used in the primary analysis of awareness. In addition to the main study, we report the results from a pilot study to inform the development of the final experimental design.

Although the major goal of this study was proof-of-concept, we had two a priori hypotheses. Our first a priori hypothesis H1 concerned the acquisition of control: Participants learn to regulate the amyg-EFP signal over the course of 20 neurofeedback runs. The second a priori hypothesis H2 concerned our main research question: Participants improve rating accuracy of their brain activity over the course of 20 neurofeedback runs. Amyg-EFP ratings were complemented by ratings of subjective confidence about the rating to directly assess metacognition. We used this data to further investigate an exploratory hypothesis H3: Higher rating accuracy of brain activity is associated with higher subjective confidence.

A priori hypotheses including the statistical analysis plan had been preregistered online before data acquisition (Table 1). Pre-registered analyses are labeled ‘confirmatory’ in the methods section below, while complementary analyses that were not preregistered are called ‘exploratory’. Changes to the preregistered protocol can be found in the text and are summarized in the Supplement (Table S1). For reasons of conciseness, we report data from preregistered questionnaires in the Supplement (Table S7) and do not further address it in the paper. To maximize transparency and to facilitate reproducibility, we provide primary research data and analysis code online (see Table 1) and we provide the Consensus on the Reporting and Experimental Design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf) Checklist (Ros et al., 2020) in the Online Supplement.

| Category          | Component                | Description                                                                 | URL                          |
|-------------------|--------------------------|------------------------------------------------------------------------------|------------------------------|
| Code              | Data processing and      | Includes code to aggregate and pool primary research data and data analysis   | osf.io/                      |
|                   | analysis                 | code. The R code used for statistical data analysis can be directly accessed | 4rbt8/                       |
|                   | Data descriptive        | Primary research data to describe sample such as age, sex, etc.              | osf.io/                      |
|                   | participant data        |                                                                               | esv2p                        |
|                   | Data event log files    | Primary research data from experiments                                       | osf.io/                      |
|                   | Data post-processed,    | Primary experiment data, structured as event lists that are easier to read    | osf.io/                      |
|                   | non-transformed data    |                                                                               | d1v5h                        |
|                   | Data pooled data tables | Primary experiment data; pooled event lists over all participants            | osf.io/                      |
|                   | Data aggregated         | Experiment data aggregated across participants                              | osf.io/                      |
|                   | event data              |                                                                               | dnrpa                        |
|                   | Project main page       | Table of content and information how to cite dataset                         | osf.io/                      |
|                   | Preregistration         |                                                                               | grxez                        |
2. Methods

2.1. Participants

Participants were recruited via announcements on university notice boards and advertisements in social media as well as on the website of the Central Institute of Mental Health (CIMH), Mannheim, Germany. Participants had to self-report no current mental diagnosis, no history of mood or psychotic disorders, and no intake of psychotropic drugs for eligibility. We assessed six participants for the pilot study and 14 participants for the main study. One participant had to be excluded from the pilot study because she fulfilled the a priori defined exclusion criterion of Beck’s Depression Inventory II (BDI-II; Beck et al., 1996) score > 13 suggesting mild depression, resulting in a convenience sample of five healthy participants (3 female; mean age $M = 28.67$ years, $SD = 8.48$ years, 3 students). Two participants from the main study dropped out after the first session; one told us that she did not have time to continue, the other person did not give an explanation. Another two participants had to be excluded because they received less than a minimum of five training sessions due to lockdown measures related to the COVID-19 pandemic. Thus, we achieved the pre-registered (convenience) sample size of ten (3 female, 7 students). The mean age was 27.3 years ($SD = 5.54$). One participant reported diagnosis of social phobia and attention deficit hyperactivity disorder in the past.

2.2. General procedure

Participants attended a maximum of 10 neurofeedback sessions with 2 runs per session, in 1–2 sessions per week. Participants were seated in front of a computer monitor wearing the EasyCap from the Brain Products GmbH (Gilching, Germany). For EEG data acquisition, the BrainAmp MR amplifier (Brain Products GmbH, Gilching, Germany) was used; EEG electrodes were sintered Ag/AgCl ring electrodes. Electrodes were positioned according to the standard 10/20 system; the reference electrode was placed between Fz and Cz. Online calculation of amygdala-EFP amplitude was done using MATLAB R2019 software (MATLAB, 2018) based on data from the Pz channel as described elsewhere (Meir-Hasson et al., 2016). Impedances of the ground, reference, and Pz electrode did not exceed 10 kΩ. The sampling rate of the raw EEG data was 250 Hz and it was recorded using the OpenViBE Acquisition Server (Renard et al., 2010). This experiment was realised using Cogent 2000 for stimulus presentation with MATLAB, developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging.

Fig. 1. Experimental design and feedback thermometer (main study). A. Overview of the structure of the employed paradigm. R blocks are ‘Regulation-only blocks’; S + R blocks are ‘Self-estimation and Regulation’ blocks. B. Example of the thermometer display for the downregulation condition. The word “regulate” appearing below the feedback display cued a neurofeedback trial.
Neuroscience.

This research was conducted in accordance with the declaration of Helsinki and was approved by the Ethics Committee of the Medical Faculty Mannheim of the University of Heidelberg. The experiments were conducted at the CIMH in Mannheim, Germany. All participants provided informed written consent before participation and were compensated with a book voucher of 20 € for participation. Participants were debriefed at the end of the experiment.

2.3. Trial structure

The duration of a neurofeedback run was 16 mins; there was a short break between two runs of a session. A neurofeedback run started with a three minutes baseline assessment (black cross displayed on grey background) and was followed by a ‘Regulation-only block’ (R block) and a ‘Self-estimation and Regulation block’ (S + R block), which are explained in detail below. The run ended with another R block (Fig. 1A). Participants had to downregulate the amyg-EFP in order to “charge” the thermometer-like feedback display (Fig. 1B). The range of the feedback display was 1–12 bars.

R block: Two continuous-feedback trials were presented with the instruction to regulate (trial duration = 30 s). Participants were shown the written instruction “regulate” below the feedback display. Feedback was given continuously and was updated every three seconds. Between trials, a black cross was displayed for 12 s. The R block was to acquaint participants with the feedback and data were not analyzed.

S + R block: Such block was composed of 12 downregulation trials with brain state ratings and intermittent feedback (also known as end-of-block feedback; Fig. 2). After a regulation phase cued with the word “regulate” below the blank feedback display (18 s), participants rated their brain activation. They pressed the right and left button on the keyboard to “charge” and “uncharge” the thermometer-like feedback display. The initial thermometer position (i.e., the number of bars) was set to 1 bar. The final position after 9 s was logged as the response. Then, the intermittent feedback was presented, marked with an X on the thermometer, together with the rated thermometer position (number of bars) for 6 s. That way, participants could evaluate the accuracy of the rated thermometer position relative to the real thermometer position represented by the intermittent feedback. Confidence ratings were included in the first four trials and in the last four trials of the S + R block. Participants were asked to rate their confidence in the brain state rating they had just made (“How confident are you that you are correct?”). Confidence was probed with a four-point Likert-scale (1 = “not at all” [confident] to 4 = “very confident”; verbal anchor “somewhat” [confident] was displayed between 2 and 3) and participants responded with the left and right keyboard button. The initial position was set to 1 and the final position after 9 s was logged as response.

2.4. Online feedback calculation

The current amyg-EFP value was normalized to the baseline (i.e., 3 min amyg-EFP recording before each run) for continuous feedback, using \( z_{ij} = \frac{x_{ij} - \bar{y}}{\sigma_p} \) with \( x_{ij} = \) continuous feedback value i of trial j, \( x_{ij} = \) amyg-EFP value i of trial j, \( \bar{y} = \) baseline mean and \( \sigma_p = \) baseline SD. Intermittent feedback was determined as personal effect size \( d_i = \frac{\bar{x}_j - \bar{y}}{\sigma_{pooled}} \) with \( d_j = \) intermittent feedback value of trial j, \( \bar{x}_j = \) mean of trial j, \( \bar{y} = \) baseline mean and \( \sigma_{pooled} = \) pooled SDs (Standard Deviations) of trial j and baseline (Paret et al., 2019). One bar corresponded to \( d = 0 \) and the thermometer maximum of 12 bars corresponded to an amyg-EFP change of \( d = -3 \) SD from baseline.

2.5. Participant instruction

To avoid potential measurement artefacts, participants were instructed to keep their eyes open and to look straight ahead, not to blink more than usual and to sit calmly. Participants were told to regulate brain activation so that the thermometer-like feedback display would “become full” (reminiscent of a fully charged battery display). They were not informed about the targeted brain function nor were they given an explicit strategy to regulate brain activation right away. If participants did not make progress in regulation until

![Fig. 2. Procedure in S + R blocks. Overview of an exemplary S + R trial. Participants were first presented the blank thermometer alongside the command to “regulate”. They then had to indicate their self-estimation, in some trials followed by a confidence rating, before being presented the feedback together with their estimation.](image-url)
the third or fourth session, they were told to put themselves in a meditative state to regulate brain activity. The decision to provide participants with strategies was based on visual inspection of neurofeedback success as well as the participant’s verbal strategy report when asked at the end of the run. Strategy instructions were given if the experimenter observed that the participant was not able to increase the number of bars for most of the time and when the reported strategies were unrelated to what we assumed to be instrumental. We did not set a formalized decision criterion. All participants but one were given strategy instructions on the third or fourth training session in the main study (see Table 2).

### 2.6. Statistical analysis

#### 2.6.1. General analytic approach

Collected data had a nested structure where repeated observations were nested in participants. We employed multilevel regression analysis to account for the nested structure. Note that this approach is different from the preregistered analysis of covariance (ANCOVA) approach (see Table S1 for more details). Analyses are based on S + R blocks.

Multilevel models (MLMs) were built stepwise. As a first step, we modelled random intercepts per person (intercept-only model) and calculated intraclass correlations (ICCs). Intercept-only models are referenced ‘model A’ below (e.g. for hypothesis X, the intercept-only model would be called: HX.A). The ICC is calculated by relativizing the amount of variance in the outcome that is due to interindividual differences on the overall variance. Thus, the ICC quantifies the proportion of variance that can be attributed to differences between participants. Next, predictor variables were included stepwise, indicated by consecutive numbering of models (for hypothesis X: HX.B, HX.C, etc.). Predictors coding time were centred on the first run (\(x_{\text{centered}} = (x - \bar{x})\)) and the predictors ‘variance’, ‘real thermometer position’ and ‘prior’ were centered on the participant mean (\(x_{\text{centered}} = (x - \bar{x})\)) (see below for predictor definitions).

For stepwise inclusion of predictors, we established the following decision criteria:

1. For every fixed effect predictor, we modelled the corresponding random slope in a subsequent step.
2. Nested models whose only difference consisted in the inclusion of a random slope were compared using the Likelihood Ratio Test (LRT). If the LRT did not indicate improvement of model fit by including the random slope and/or the estimated variance of the random slope was zero, we removed the random slope from the model. An exception was the inclusion of random slopes for fixed effects predictors in an interaction term. In this case, we followed Heisig and Schaeffer (2019) and included the respective random effects predictors where interaction terms of fixed effects were modelled in order to account for dependent observations and thus to prevent anticonservative statistical inference.

Using this algorithm we arrived at the ‘final’ model (for hypothesis X: HX.final) which was used to test the hypothesis. Formulas of the final models are reported in the Supplement.

Models were fitted using the Restricted Maximum Likelihood (REML) method to prevent biased variance estimates. For comparing the model fit of two nested models via LRT, models were refitted with the Maximum Likelihood method. All reported model estimates are based on REML if not specified otherwise.

Quantile-quantile plots of model residuals were inspected visually. If they indicated heteroskedasticity, we used robust standard errors (SEs) to assess statistical inference. If we do not report otherwise, statistical inference testing is based on robust SEs. Following suggestions by Pustejovsky and Tipton (2018), the “bias-reduced linearization” adjustment was employed when calculating cluster-robust variance-covariance matrices of type cluster-randomized 2 (CR2). Interference testing of fixed effects estimates was done using a small sample correction; p-values and degrees of freedom were corrected based on Satterthwaite approximation. The null-hypothesis was rejected when statistical tests surpassed the \(p < 0.05\) criterion. Following Nakagawa et al. (2017), marginal and conditional (Pseudo-) \(R^2\) for (generalized) linear mixed models were

| Subject number | Trials | Runs | Sessions |
|----------------|--------|------|----------|
| 1 \(^a\)       | 240    | 20   | 10       |
| 2              | 216    | 18   | 9        |
| 3 \(^a\)       | 216    | 18   | 9        |
| 4 \(^a\)       | 216    | 18   | 9        |
| 5 \(^a\)       | 144    | 12   | 6        |
| 6 \(^a\)       | 120    | 10   | 5        |
| 7 \(^a\)       | 120    | 10   | 5        |
| 8 \(^a\)       | 228    | 19   | 10       |
| 9 \(^a\)       | 240    | 20   | 10       |
| 10 \(^a\)      | 120    | 10   | 5        |
| Total          | 1,860  | 155  | 78       |

Subject #2: run #1 and run #2 had to be excluded due to a measurement artefact. Subject #8: run #4 was lost due to technical problems.

\(^a\) Strategy instruction in session 3 or 4.
calculated to assess model fit. Marginal and conditional $R^2_{GLMM}$ are measures of explained variance and are therefore comparable to the coefficient of determination $R^2$. While the marginal $R^2_{GLMM}$ reflects the variance explained by the fixed effects of a MLM, the conditional $R^2_{GLMM}$ reflects the variance explained by the entire model, i.e., by both fixed and random effects.

2.6.2. Confirmatory analyses

**Hypothesis H1, improvement of downregulation:** We assessed linear improvement of downregulation with the predictor ‘run’ on the outcome ‘real thermometer position’. The level 1 (L1) unit was run nested in participants. The random intercept only-model (H1.A) was built first. Then, the predictors were included in the following order: centred ‘run’ was entered first as a fixed effect, in a subsequent step it was additionally included as random effect (H1.B and H1.C, respectively). A significant positive slope of ‘run’ was hypothesized as this would mean that participants improved in downregulating the amyg-EFP signal over time. The multilevel equation for the resulting final model can be found in the Supplement (Formula S1).

**Hypothesis H2, improvement of rating accuracy:** The absolute (i.e., unsigned) difference between ‘real thermometer position’ and ‘rated thermometer position’ (i.e., ‘difference rating real’) was used to evaluate H2 on a run-wise basis. The smaller the difference, the better the rating accuracy. Learning should be reflected by a linear decrease of the difference across runs. The L1 unit was run nested in participants. The centred variance of ‘real thermometer position’ per run (i.e., the variance of the feedback thermometer across trials) was added to the random-intercept model (H2.A), first as fixed effect (H2.B) and then additionally as random effect (H2.C). Including the feedback variance accounts for ‘freezing’ of the feedback display, e.g., due to floor/ceiling effects, where the amyg-EFP signal is out of the displayed range. If the proportion of such trials is high, participants can move the cursor to the maximum/minimum of the rating scale and would be correct, although the rating would not reflect true brain activation. In the final steps, centred ‘run’ was entered (H2.D, H2.E). A significant negative slope of ‘run’ was hypothesized as this would mean that participants improved in self-estimation accuracy over time. The multilevel equation for the resulting final model can be found in the Supplement (Formula S2a).

2.6.3. Exploratory analyses

To achieve a deeper understanding of how participants rate amyg-EFP activity, we investigated how much the participants’ ratings are guided by other sources than brain state awareness. We assumed that the feedback from the previous trial could serve to approximate the feedback on the next trial. In other words, feedback from the previous trial was treated as prior to inform the rating on the following trial, which was inspired by a similar approach introduced by Schurger et al. (2017). To evaluate how much participants would rely on the prior vs. other sources of information (such as brain state awareness) when they rate brain activation, we ran a MLM with the rated thermometer position as the outcome and the real thermometer position and the prior as predictors. To differentiate the exploratory analysis from the confirmatory test of hypothesis H2, the following models are labeled H2ex. The L1 unit was trial nested in participants. A random intercept-only model (H2ex.A) was computed first. Next, the centred predictor ‘real thermometer position’ was entered (H2ex.B, H2ex.C). Then, the centred prior was included (H2ex.D, H2ex.E). To account for a potential increase in potential improvement in the accuracy of brain state ratings over time. Additionally, the random effects for both predictors ‘session’ and ‘real thermometer position’ were entered (H2ex.F). The equation of the resulting final model H2ex.final can be found in the Supplement (Formula S2b).

Furthermore, we investigated whether confidence ratings (the outcome) increases as rating accuracy improves, i.e., as the difference between rated and real thermometer position decreases linearly (the predictor). The L1 unit was run nested in participants. First, a random intercept-only model (H3.A) was assessed with confidence as the outcome, aggregated per run (i.e., ‘confidence’). To control for possible differences in confidence between participants from the pilot study vs. main study, we included a dummy variable coding the type of experimental paradigm with the dummy predictor ‘paradigm’, which was entered in the second step (H3.B). The outcome ‘confidence’ could be biased by low variance of the feedback thermometer as discussed above (see section Confirmatory analyses – Hypothesis H2, improvement in rating accuracy). Therefore, next, we included the (standardized) feedback variance (H3.C, H3.D). Then, the absolute (i.e., unsigned) difference between the rated and the real thermometer position (‘difference rating real’) was entered (H3.E, H3.F). Predictor variables were z-standardized to account for differences in thermometer resolution between the pilot and the main study. The MLM formula of the resulting model H3.final can be found in the Supplement (Formula S3a).

All statistical analyses were conducted using R (R Core Team, 2020). For linear regression analyses, the package “sandwich” was used to compute HCCMs (Zeileis, 2004; Zeileis et al., 2020), and the package “parameters” was employed for processing linear model parameters (Lüdecke et al., 2020). Multilevel regression analyses were conducted using the package ‘lm4’ for fitting MLMss (Bates et al., 2015). The package “piecewiseSEM” was used to calculate marginal and conditional $R^2_{GLMM}$ (Lefcheck, 2016). The package “clubSandwich” was used to calculate cluster-robust SEs and inference tests based on cluster-robust SEs for MLMs (Pustejovsky, 2021).

2.7. Changes to preregistered study protocol

This manuscript has a few changes to the original research plan that were required after preregistration. This includes minor changes of eligibility criteria and the adoption of an MLM instead of ANCOVA approach for statistical analysis of hypothesis H2. Correlative hypotheses on questionnaire outcomes were not investigated due to subject dropout and the resulting low sample size. More details are found in the Supplement, Table S1.
3. Results

We assessed a preliminary version of the experimental design in a pilot study with 5 participants. The major difference to the main study was that the participants had to upregulate the amyg-EFP in half of the trials (condition A), while they had to downregulate in the other half of the trials (condition B). Conditions A and B were presented in semi-randomized order during each session. The design and analysis procedure of the pilot study is described in the Supplement (Methods S1, Fig. S1, Table S2). We found that four of five participants had a positive slope estimate for predictor ‘run’ in condition A and two of five had a positive slope estimate in condition B (note that a positive slope in condition B indicates learning downregulation). No slope estimate was significantly different from zero for any individual (Fig. S2, Fig. S3, Table S3). To follow up the negative results we conducted post-hoc correlation analyses to better understand how switching between upregulation and downregulation trials within runs may relate to regulation success. For that purpose we submitted the real thermometer position condition A and condition B trials to a correlation analysis. The Pearson correlation coefficients for all participants had a negative sign and ranged from \( r = -0.07 \) to \( r = -0.49 \) (Table S4). Four of five coefficients were significantly different from zero. In essence, this series of single-case studies suggests that changes of feedback in the desired direction were either achieved in one condition or the other, but rather not in both. Review of this preliminary data suggested that the experimental design was feasible. We abandoned the upregulation condition to decrease complexity and to increase the number of downregulation-trials for the main study. The decision to go with the downregulation condition was based on the lack of previous studies investigating amyg-EFP upregulation and a solid basis of studies showing feasibility of amyg-EFP downregulation (Fruchtman-Steinbok et al., 2021; Goldway et al., 2019; Keynan et al., 2016, 2019; Meir-Hasson et al., 2016).

In the main study, ten participants contributed 1,860 trials (\( M = 186 \) trials per participant; \( SD = 52.84 \)) in total to the analyses (Table 2). Data from the first and second run from one participant had to be excluded because of heavy EEG artefacts. One run of another participant was lost due to technical problems. Statistical analyses to test the \( a \) priori hypotheses H1 and H2 were carried out with the remaining 155 runs. The mean number of runs per participant was 15.5 (\( SD = 4.40 \)).

3.1. Do participants improve downregulation (H1)?

The ICC of the intercept-only model was \( \rho = 0.32 \). Thus, approximately 32% of variance in the real thermometer position was based on individual differences, indicating that the multilevel structure should be taken into account (Table 3). Seven of ten participants had a positive random slope estimate for ‘run’, indicating downregulation learning. However, the fixed effect ‘run’ did not significantly predict the outcome of ‘real thermometer position’ \( (\gamma_{10,\text{final}(\text{run})} = 0.09 \ (SE = 0.07), \ p = .222) \) in the MLM, favouring the null-hypothesis that participants did not learn downregulation (Fig. 3).

3.2. Do participants improve rating accuracy (H2)?

45% of variance in differences between rated and real thermometer position (i. e., the outcome) was due to differences between individuals, supporting the use of MLMs to account for interindividual differences (Table 4). The centred predictor ‘variance’ was significantly associated with the outcome ‘difference rating real’ \( (\gamma_{10,\text{final}(\text{variance}_{\text{realthermpos}}}) = 0.22 \ (SE = 0.03), \ p < .001 \). That is, higher variance in the feedback was associated with lower accuracy in ratings. Contrary to the hypothesis, the fixed slope estimate for ‘run’ did not significantly predict the outcome \( (\gamma_{20,\text{final}(\text{run})} < 0.01 \ (SE = 0.01), \ p = .555; \ Fig. 4) \). Accordingly, we could not confirm that rating accuracy improves over time.

### Table 3
Multilevel modelling of the effect of time (‘run’) on downregulation performance (‘real thermometer position’) in the main study.

| Models | H1.A | H1.B | H1.C = H1.final |
|--------|------|------|----------------|
| Fixed effects | b [95% CI] | b [95% CI] | b [95% CI] |
| Intercept (\( \gamma_{00} \)) | 1.80 [0.88, 2.73] ** | 0.90 [−0.11, 1.91] | 1.04 [0.15, 1.93] * |
| ‘Run’ (\( \gamma_{10} \)) | – | 0.12 [−0.06, 0.29] | 0.09 [−0.05, 0.23] |
| Random effects | VC | VC | VC |
| Residual (\( \sigma^2 \)) | 4.03 | 3.71 | 2.78 |
| Intercept (\( \tau_{0} \)) | 1.90 | 1.62 | 0.84 |
| ‘Run’ (\( \tau_{1} \)) | – | – | 0.03 |
| Model summary | Marginal R² | 0.07 | 0.04 |
| Conditional R² | 0.32 | 0.35 | 0.57 |
| Number of estimated parameters | 3 | 4 | 5 |

\( N_{\text{observations}} = 155, N_{\text{participants}} = 10. \) CI, confidence interval; VC, variance components. Parameter estimates are unstandardized. All predictors are centred. SEs for CIs are robust. Inference testing is based on Satterthwaite approximation.

* \( p < .05 \). ** \( p < .01 \).
3.3. Can participants predict the real thermometer position (H2ex)?

Approximately 36% of variance in the outcome ‘rated thermometer position’ was due to variance between participants (Table 5). Both predictors ‘real thermometer position’ ($\gamma_{10,\text{final}}(\text{real thermometer position}) = 0.14 (SE = 0.05), p = .019$) and ‘prior’ ($\gamma_{20,\text{final}}(\text{prior}) = 0.48 (SE = 0.02), p < .001$) significantly predicted the outcome ‘rated thermometer position’. This shows that ratings were indeed based on the feedback received in the current and in the previous trial (Fig. 5). For the interaction term real thermometer position $\times$ session, Fig. 5 shows that the regression coefficient was higher (= the slope was steeper) for late vs. early sessions. However, the interaction was not statistically significant ($\gamma_{40,\text{final}}(\text{real thermometer position}\times\text{session}) = 0.01 (SE = 0.01), p = .276$, Fig. 5), therefore an increased influence of ‘real thermometer position’ on ratings across sessions could not be demonstrated.
3.4. Are participants more confident when they rate brain activity more accurately (H3)?

We start with a description of the confidence rating profiles of our participants: On average, participants endorsed moderate confidence levels ($M = 2.64$, $SD = 0.8$). However, confidence rating profiles (i.e., the relative proportion of rated confidence level)
differed markedly between individuals (Fig. S4). Within individuals, confidence was relatively stable and did not significantly change over time, as indicated by a non-significant fixed effect estimate for ‘run’ ($\gamma_{10, \text{conf}} (\text{run}) = 0.01$ (SE = 0.01), $p = .417$, Table S5 and Fig. S5).

A significant correlation between confidence and accuracy of brain state ratings would support that participants can reflect on cognitive processes underlying amyg-EFP prediction. Confidence should be high when the difference between ‘rated thermometer position’ and ‘real thermometer position’ is low. As this research question was fully explorative, we used both samples (i.e., from the pilot study and from the main study; from the pilot study we analyzed trials from the downregulation condition for consistency reasons) to increase power, which resulted in an extended sample of 15 participants. The multilevel analysis was based on a total of 254 observations (note that confidence was assessed in 66% of trials in S+R blocks). Interindividual differences explained a high amount of variance (57%, model H3.A in Table 6).

In accordance with the hypothesis that higher rating accuracy (i.e., lower difference between rated and real brain activity) would go along with higher confidence, ‘difference rating real’ significantly predicted ‘confidence’ in the expected direction ($\gamma_{20, \text{final}} (\text{difference rating real}) = -0.32$ (SE = 0.10), $p = .009$; Fig. 6). As we speculated that participants would be more confident with less feedback variance between trials, we included this predictor into the model. ‘Variance’ showed a non-significant positive trend to predict ‘confidence’ ($\gamma_{10, \text{final}} (\text{variance real thermometer position}) = 0.14$ (SE = 0.07), $p = .078$). The latter

### Table 5

Multilevel modelling predicting brain state self-estimation (‘rated thermometer position’).

| Models            | H2ex.A       | H2ex.B       | H2ex.C       | H2ex.D       | H2ex.F = H2ex.final |
|-------------------|--------------|--------------|--------------|--------------|--------------------|
| Fixed effects     |              |              |              |              |                    |
| Intercept ($\tau_0$) | 2.37 [1.38, 3.37] | 2.37 [1.38, 3.36] | 2.37 [1.38, 3.36] | 2.37 [1.38, 3.36] | 2.31 [1.38, 3.25] |
| ‘Real thermometer position’ ($\gamma_{10}$) | – | 0.62 [0.39, 0.86] | 0.42 [0.27, 0.58] | 0.19 [0.11, 0.26] | 0.14 [0.04, 0.23] |
| ‘Prior’ ($\gamma_{20}$) | – | – | – | 0.49 [0.44, 0.53] | 0.48 [0.43, 0.52] |
| ‘Session’ ($\gamma_{30}$) | – | – | – | – | 0.01 [–0.05, 0.07] |
| ‘Real thermometer position’ x ‘session’ ($\gamma_{40}$) | – | – | – | – | 0.01 [–0.01, 0.03] |
| Random effects    | VC           | VC           | VC           | VC           | VC                 |
| Residual ($\sigma^2$) | 4.61 | 2.44 | 2.14 | 1.57 | 1.55 |
| Intercept ($\tau_0$) | 2.54 | 2.55 | 2.55 | 2.55 | 2.55 |
| ‘Real thermometer position’ ($\tau_1$) | – | – | 0.06 | 0.02 | 0.01 |
| ‘Session’ ($\tau_3$) | – | – | – | – | – |
| Model summary     |              |              |              |              |                    |
| Marginal $R^2$    | 0.36         | 0.66         | 0.64         | 0.75         | 0.75 |
| Conditional $R^2$ | 0.36         | 0.66         | 0.64         | 0.75         | 0.75 |
| Number of estimated parameters | 3 | 4 | 5 | 6 | 9 |

$N_{\text{observations}} = 1,860, N_{\text{participants}} = 10$. CI, confidence interval; VC, variance components. Parameter estimates are unstandardized. All predictors are centred. SEs for CIs are robust. Inference testing is based on Satterthwaite approximation.

* $p < .05$. ** $p < .01$. *** $p < .001$.

![Multilevel Model H2ex.final: Brain State Rating - Marginal Effects](image)

Fig. 5. Exploratory multilevel model: prediction of brain state rating. Based on model H2ex.final. Display of the marginal effect of ‘real thermometer position’ on ‘rated thermometer position’ for different values of fixed effect of centred ‘session’ (i.e., 1st session, 5th session, and 10th session) and different values of fixed effect of centred ‘prior’ ($M, M \pm 1$ SD) on population level. Light shadings around graph lines represent 95% CIs of predicted values; CIs in the figure are based on uncorrected SEs.
### Table 6
Multilevel modelling of the effects of paradigm, ‘variance’ and rating accuracy ('difference rating real') on participants' confidence ratings ('confidence') in the combined sample.

| Models | H3.A | H3.B | H3.C | H3.E | H3.F = H3.final |
|--------|------|------|------|------|-----------------|
| Fixed effects | b [95% CI] | b [95% CI] | b [95% CI] | b [95% CI] | b [95% CI] |
| Intercept (γ₀₀) | 2.61 [2.29, 2.92] | 2.94 [2.28, 3.61] | 2.94 [2.28, 3.61] | 2.94 [2.34, 3.55] *** | 2.89 [2.32, 3.45] *** |
| ‘Paradigm’ (γ₁₀) | – | –0.51 [−1.25, 0.23] | –0.51 [−1.24, 0.22] | –0.53 [−1.22, 0.16] | –0.54 [−1.18, 0.11] |
| ‘Variance’ (γ₁₀) | – | – | –0.04 [−0.13, 0.06] | 0.10 [−0.04, 0.23] | 0.14 [0.00, 0.27] |
| ‘Difference rating real’ (γ₂₀) | – | – | – | 0.26 [−0.46, −0.06] | −0.32 [−0.52, −0.12] |

Random effects | VC | VC | VC | VC | VC |
Residual (σ²) | 0.28 | 0.28 | 0.28 | 0.25 | 0.23 |
Intercept (τ₀) | 0.37 | 0.33 | 0.33 | 0.32 | 0.26 |
‘Difference rating real’ (τ₂) | – | – | – | – | 0.06 |
Model summary | Marginal R² | 0.57 | 0.59 | 0.59 | 0.63 | 0.67 |
Conditional R² | 0.57 | 0.59 | 0.59 | 0.63 | 0.67 |
Number of estimated parameters | 3 | 4 | 5 | 6 | 7 |

N<sub>observations</sub> = 254, N<sub>participants</sub> = 15. CI, confidence interval; VC, variance components. Parameter estimates are unstandardized. All predictors are standardized except for the dummy predictor. SEs for CIs are robust. Inference testing is based on Satterthwaite approximation.

. p < .10. * p < .05. ** p < .01. *** p < .001.

---

**Fig. 6.** Exploratory multilevel model: prediction of confidence rating. Based on model H3.final. Display of the marginal effect of self-estimation accuracy ('difference rating real') on 'confidence' for different values of centred 'variance' (M, M ± 1 SD) on population level. Light shadings around graph lines represent 95% CIs of predicted values; CIs in the figure are based on uncorrected SEs.
rebutted the concern raised above and showed, on the contrary, that higher feedback variance was associated with higher confidence, although not with statistical significance. The standardized predictors ‘difference rating real’ and ‘variance’ correlated with $r = 0.61$. Hence, bias introduced by correlated predictors is probably low (Schielezeth et al., 2020). Participants from the two experiment versions did not differ significantly ($\gamma_{\text{final}}(\text{paradigm}) = -0.54 (SE = 0.33), p = .139$). Note that ‘paradigm’ was included for theoretical reasons and the sample sizes per group were too small to expect significant differences. In a complementary analysis with participants from the main study only, the same trend for the fixed effect ‘difference rating real’ emerged with $\gamma_{\text{manifinal}}(\text{difference rating real}) = -0.36 (SE = 0.17), p = .077$, corroborating that higher confidence was associated with higher rating accuracy, although not surpassing the statistical threshold for significance (Supplement, Formula S3c and Table S6).

4. Discussion

Brain state awareness is a fundamental concept in awareness theory of neurofeedback but is largely neglected by empirical research (Munoz-Moldes & Cleeremans, 2020). This study was conducted to show the feasibility to investigate the self-estimation of brain states with fMRI-inspired EEG neurofeedback (i.e. amyg-EFP). Furthermore, we aimed to provide initial evidence for awareness and meta-cognition with amyg-EFP neurofeedback. We administered visual neurofeedback to healthy participants, complemented by brain state ratings that were recorded before intermittent feedback was received. Assuming that higher awareness would be reflected in higher rating accuracy, we investigated how well participants were able to track activation changes of the amyg-EFP signal. In general, the analyses revealed a high degree of interindividual differences in neurofeedback control and rating tendency, accounting for a substantial part of the variance in the data. To account for interindividual differences, we used multilevel regression modelling to test our hypotheses. Hypothesis H1, which proposed that participants learn voluntary control of the amyg-EFP, had to be rejected on the group level, although a visual inspection of the learning curves suggests a tendency to improve for some participants (c.f. Fig. 3, subjects 1, 2, 4, 5, 7, 9, 10). It is possible that the complexity of the experimental design, where participants rated success and confidence before seeing the feedback, made it difficult to learn. Furthermore, there are many ways for online calculation of intermittent feedback and we cannot rule out that other success measures than the effect size-measure we had used would have resulted in better learning.

The sample size was small and variance between participants was large. Participants were instructed with strategies at the discretion of the experimenter, who based this decision on observation of learning progress and interviews with the participant. Lack of an objective criterion and/or systematic strategy instruction is a confounding factor of this study. Future studies with more participants and clear-cut criteria for strategy instruction would be helpful to further investigate the effects. Following our preregistered analysis approach, we used the difference between the rated thermometer position and the real thermometer position as an index of rating accuracy. Hypothesis H2, predicting that participants improve rating accuracy over time, had to be rejected. Notably, we found that lower feedback variance was associated with higher rating accuracy, suggesting that technical parameters influenced the dynamic of the feedback display and therewith contributed significantly to the rating.

To follow up the non-significant result, we explored how well amyg-EFP feedback, i.e., the real thermometer position, predicted the rating in a linear regression model (H2.ex). This analysis revealed that changes in feedback covaried with the ratings, suggesting that participants were able to predict changes in the amyg-EFP with significant accuracy. Importantly, the effect remained significant when we adjusted for an experience-based rating strategy that is, that individuals rate according to the feedback they had just received on the preceding trial (i.e. the prior). While the relative influence of the experience-based strategy on the rating was numerically stronger than the influence of the real thermometer position, the latter still contributed significantly to the rating. Hence, the ratings were not exclusively based on the prior and a possible explanation how participants might have been able to predict the feedback is via brain state awareness. In contrast to a priori expectations and in line with the non-significant finding from the confirmatory analysis, the accuracy of brain state ratings did not significantly improve over sessions (i.e., no evidence for learning of brain state accuracy), evidenced by a non-significant real thermometer position $\times$ session interaction. However, our study was not powered well enough to detect moderate learning effects that might have driven this interaction. Also, it has been theorized that learning to control feedback could facilitate discrimination learning (Frederick et al., 2016; Kotchoubey et al., 2002; Schurger et al., 2017) and non-significant learning to control neurofeedback in our sample could relate to non-significant improvement of rating accuracy.

Most commonly, it is believed that brain state awareness could be mediated by the interoception of autonomic nervous activity correlated with changes in neural activity (Kotchoubey et al., 2002). As electrophysiological activity of the brain is not perceivable, it has been speculated that changes in blood flow associated with changes in brain function could be perceived through extensions of receptors in the arterial walls (Kotchoubey et al., 2002). This proof-of-concept study was not designed to bring light to the potential role of interoception in predicting amyg-EFP activation. The empirical investigation of awareness has to rely on participants’ self-reports and cannot prove brain state awareness. Although we adjusted for an experience-based rating strategy, it is possible that participants used other strategies to rate brain activation accurately, which have been discussed in more depth elsewhere (Frederick et al., 2019; Kotchoubey et al., 2002). For instance, participants could learn to predict change in brain activation from the degree of effort made to control brain activation. Future studies should implement experimental control conditions that filter out the contribution of voluntary control of brain activation. In addition, transfer effects showing that participants learn to predict brain activation outside of the learning situation would help to dismantle the influence from brain state awareness.

The predictors ‘real thermometer position’ and ‘prior’ were highly correlated with $r = 0.70$. As ‘prior’ of trial $x$ equals ‘real thermometer position’ of trial $x-1$, this correlation would be expected to emerge from trial-by-trial improvement of brain self-regulation (although we did not find significant linear improvement). Furthermore, the EFP is an EEG surrogate measure of BOLD activation measured with fMRI (Meir-Hasson et al., 2016) and autocorrelations are well known in BOLD-fMRI. Thus, temporal
autocorrelation of the amyg-EFP signal may have contributed to the correlation between ‘prior’ and ‘real thermometer position’.

To shed further light on metacognition components in neurofeedback awareness, we let participants report their confidence in the rating before they received the intermittent feedback. This revealed that, when the rating accuracy was high, the confidence was too, suggesting that participants recognized how accurately they could predict feedback. The tendency to rate confidence high or low was variable between and stable within participants, emphasizing the influence of trait or state variables (c.f. Fleming & Lau, 2014). The results encourage further research to investigate how metacognition could be leveraged for training and therapy purposes with neurofeedback. As discussed above, it is unclear whether (and if, how much) participants had immediate metaknowledge about their cognitive processing of brain activation. Furthermore, the data analysis plan to assess this hypothesis was developed a posteriori. Replication of our finding is necessary to conclude that the effect truly exists.

In conclusion, we present an empirical approach to investigate awareness and metacognition of brain signals with relevance for affective processing. The studied sample size was small, although the repeated measures design with up to 20 runs resulted in a high number of observations per participant. Exploratory analyses revealed that participants knew what feedback to expect on the next trial and, at a meta-level, had insight into the accuracy of their predictions.

CRediT authorship contribution statement

**Madita Stirner:** Methodology, Formal analysis, Data curation, Writing – original draft, Investigation, Data curation, Visualization, Writing – review & editing. **Guy Gurevitch:** Software, Writing – review & editing. **Nitzan Lubianiker:** Methodology, Writing – review & editing. **Talma Hendler:** Methodology, Resources, Writing – review & editing. **Christian Schmah:** Writing – review & editing. **Christian Paret:** Conceptualization, Data curation, Supervision, Project administration, Methodology, Software, Writing – original draft, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: TH is Chief Medical Scientist of GrayMatters Health co Haifa, Israel. The other authors declare no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgements

Thanks are due to Johanna Hepp for statistical counseling, her excellent support with multilevel data analysis and critical review of the manuscript. We thank Aleksandra Tarka for support in data acquisition and analysis. This work was supported by funding from both the G.-A. Lienert foundation, through the Fellowship Personality Disorders, granted by the Asklepios Clinic and the German Society for Research and Therapy in Personality Disorders (GPs) to CP, and by internal funds.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2021.103264.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory manual*. San Antonio, TX: The Psychological Corporation.

Black, A., Cott, A., & Pavloski, R. (1977). The operant learning theory approach to biofeedback. In *Biofeedback, theory and research* (1st ed., pp. 89–127). Academic Press.

Caria, A. (2016). Self-regulation of blood oxygenation level dependent response: Primary effect or epiphenomenon? *Frontiers in Neuroscience, 10*, 117. https://doi.org/10.3389/fnins.2016.00117

Etkin, A., & Wager, T. D. (2007). Functional neuroimaging of anxiety: A meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *The American Journal of Psychiatry, 164*(10), 1476–1488. https://doi.org/10.1176/appi.ajp.2007.07030504

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, https://doi.org/10.3389/fnhum.2014.00443

Frederick, J. A. (2012). Psychophysics of EEG alpha state discrimination. *Consciousness and Cognition, 21*(3), 1345–1354. https://doi.org/10.1016/j.concog.2012.06.009

Frederick, J. A., Heim, A. S., & Dunn, K. N. (2019). Self-prompted discrimination and operant control of EEG alpha. *NeuroRegulation, 6*(2), 81–81. doi: 10.15540/nnr.6.2.81.

Frederick, J. A., Heim, A. S., Dunn, K. N., Powers, C. D., & Klein, C. J. (2016). Generalization of skills between operant control and discrimination of EEG alpha. *Consciousness and Cognition, 45*, 226–234. https://doi.org/10.1016/j.concog.2016.09.009

Fruchtman-Steinbok, T., Keynan, J. N., Cohen, A., Jaljuli, I., Mermelstein, S., Droit, G., Routledge, E., Krassnoshtein, M., Playle, R., Linden, D. E. J., & Hendler, T. (2021). Amygdala electrical-finger-print (AmygEFP) NeuroFeedback guided by individually-tailored Trauma script for post-traumatic stress disorder: Proof-of-concept. *NeuroImage: Clinical, 32*, Article 102859. https://doi.org/10.1016/j.nicl.2021.102859

Gasue, A., Vialatte, A., Mora-Sánchez, A., Ramdani, C., & Vialatte, F. B. (2016). A psychoengineering paradigm for the neurocognitive mechanisms of biofeedback and neurofeedback. *Neuroscience & Biobehavioral Reviews, 68*, 891–910. https://doi.org/10.1016/j.neubiorev.2016.06.012
