EEG activity predictive of learning through feedback

Matthew H. Danyluik*[1], Sucheta Chakravarty*[2], and Jeremy B. Caplan[3,4]

[1]Integrated Program in Neuroscience, McGill University, Montreal, QC, Canada
[2]Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA
[3]Department of Psychology, University of Alberta, Edmonton, AB, Canada
[4]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, AB, Canada

Abstract

Most of what we know about neural mechanisms of incremental learning through feedback comes from descriptive, univariate analyses. Here we go one step further, seeking brain activity that is not just statistically reliable (potentially small-but-significant) but can track such learning at the item level, taking a classifier-based approach to narrow in on basic neural encoding processes. Participants (N = 45) learned 48 word-value mappings through trial-and-error. First we checked whether established EEG markers of feedback processing, the feedback-related negativity (FRN) and frontal midline theta activity (FMT), are in fact predictive of trial-to-trial learning of the current item— and they were (above chance, but not by much), validating the behavioural relevance of those features. Next we asked whether there might be considerably more information about encoding on single trials beyond these statistically robust, regular signals. Indeed, multivariate classifiers (LDA and SVM), incorporating signal-features beyond the FRN and FMT, predicted learning more substantially and exceeding previous performance on episodic recognition using the same basic approach (Chakravarty et al., 2020). Time-frequency spectral features produced better classifications (AUC ∼0.7) than time-domain features. Finally, a possible shortcut due to accuracy varying systematically with trial number could not explain away classification success. In sum, FRN and FMT are not just descriptive of feedback-driven learning but also a bit predictive— but are the tip of the iceberg (subject-specific, spatiotemporal features) uncovered by the multivariate classifiers. This extends current classifier-based approaches to brain activity from episodic memory to incremental, feedback-driven learning.

Keywords: Feedback-driven learning; classifier analyses; feedback-related negativity; frontal midline theta; verbal memory; subsequent memory effect

*Shared first authors. Corresponding author: Jeremy B. Caplan, jcaplan@ualberta.ca. Department of Psychology, University of Alberta, Edmonton, AB, Canada
Acknowledgements. The authors thank Tobias Sommer for helpful comments on the manuscript. Supported in part by Alberta Innovates and the Natural Sciences and Engineering Research Council of Canada.

Statements and Declarations

Funding. Supported in part by Alberta Innovates and the Natural Sciences and Engineering Research Council of Canada.

Conflicts of interest/Competing interests. Not applicable.

Availability of data and material. Data will be provided upon reasonable request of the authors.

Code availability. Code will be provided upon reasonable request of the authors.

Authors' contributions. The work was conceived, designed, conducted and written collaboratively with all authors working cohesively as a group. M.D. carried out the bulk of the data-analyses presented here.

Ethics approval. The procedures were approved by a University of Alberta ethical review board.

Consent to participate. All authors have consented to participate.

Consent for publication. All authors have consented to publish this manuscript.

Introduction

We often learn incrementally through feedback, where we evaluate the outcomes of our actions to optimize future behaviour (Sutton & Barto, 2018). Our overarching goal was to investigate learning-related brain activity using an approach that weds the subsequent-memory effect approach to identifying brain activity that differentiates actual memory outcome, with classifier methods that enforce the requirement that such activity be predictive rather than simply descriptive. Next we briefly summarize electroencephalographic (EEG) research on feedback processing, describe our learning task that has a greater learning demand than typical tasks, and describe the two-stage (univariate followed by multivariate) approach we apply to identify brain activity that underlies feedback-driven learning.

Learning from feedback

Feedback-driven learning is thought to be mediated by the calculation of a reward prediction error (RPE)— the difference between an action's expected and obtained outcome (Schultz et al., 1997). Accordingly, studying RPE-linked brain activity may inform our understanding of how the brain incrementally updates action expectations, where reward shapes learning. Extensive research using univariate methods has repeatedly replicated two features of interest that may reflect RPE: the feedback-related negativity and frontal-midline theta activity. For the first stage of our two-stage approach, we investigate these two features, through a classification lens.

First, the *Feedback-Related Negativity* (FRN) is an event-related potential (ERP) found at frontocentral electrodes around 200–350 ms post-feedback, where FRN amplitude is lower while processing negative than positive feedback (Bellebaum & Daum, 2008; Frank et al., 2005; Goyer et al., 2008; Hajcak et al., 2006; Holroyd et al., 2004; Marco-Pallares et al., 2011; Nieuwenhuis et al., 2002; Pfabigan, Seidel, et al., 2015; Walsh & Anderson, 2011; Wu & Zhou, 2009; Yeung et al., 2004). A leading account of the FRN suggests that, as the FRN is sensitive to whether feedback was expected, it could act as an RPE detector as well, enabling error-driven learning (Holroyd & Coles, 2002). As such, some have reported that FRN magnitude is related to learning outcomes (i.e., that it is more negative during errors which are later corrected; Hester et al., 2008; van der Helden et al., 2010), though others have failed to replicate this finding (Chakravarty et al., in review; Luft et al., 2014; van de Vijver et al., 2011). We also note that the FRN window overlaps with that of a reward positivity (a positive deflection during rewarding feedback; e.g., Proudfit, 2015), and that it is unclear whether the FRN effect is driven by a greater positive deflection while processing rewarding feedback, or a greater negative deflection while processing errors (Holroyd et al., 2008). We will not conclude on either side of this debate, although for simplicity, we use the older and more common acronym, FRN, to refer to this particular measure of EEG voltage.

Second, *Frontal Midline Theta* (FMT) is characterized by elevated theta-band (4–8 Hz) power over frontocentral regions peaking around 300 ms post-feedback. Some have suggested that FMT is an "error detector" which appears following any negative feedback (Bernat et al., 2015; Cohen et al., 2007; Marco-Pallares et al., 2008), while others have reported that FMT is modulated by outcome magnitude or probability (Christie & Tata, 2009; Hajihosseini & Holroyd, 2013), suggesting that it is sensitive to RPE (Cavanagh et al., 2010; Mas-Herrero & Marco-Pallares, 2014). FMT has also been explicitly linked to

learning, with some reporting that it shows subsequent memory effects, or greater power during the study of items that are later remembered than later forgotten (e.g., Chen & Caplan, 2017; Fell et al., 2011; Klimesch, 1999; Klimesch et al., 2010).

To date, researchers have studied brain activity during feedback processing in two main paradigms. Many employ a probabilistic reward learning task (Bernat et al., 2015; Cavanagh et al., 2010; Christie & Tata, 2009; Cohen et al., 2007; Gehring & Willoughby, 2002; Goyer et al., 2008; Hajcak et al., 2007; Marco-Pallares et al., 2008), where on each trial, participants bet on which of a small set of stimuli (typically 2 to 4) will give a greater reward when chosen. Participants can then use feedback for their selection to optimize future bets on the stimuli later in the task. Others use a time estimation task (Gehring et al., 1993; Hajihosseini & Holroyd, 2013; P. Li et al., 2016; Luu et al., 2004; Miltner et al., 1997; Nieuwenhuis et al., 2005; van de Vijver et al., 2011), where participants aim to press a button a specific amount of time (e.g., one second) after a cue, and receive feedback on how accurate their guess was to guide their future actions.

Note that neither task offers a clear and substantial learning objective for participants to achieve, which is the task-regime we are interested in here. In probabilistic reward learning, participants often do not learn a fixed word-value association, but instead learn to choose the stimulus that is *most likely* to give a reward, with a reward probability of, for example, 80%. In time estimation, feedback typically scales with participant performance, where better–performing participants must be increasingly accurate to achieve the same reward as in previous trials. As such, in both cases, it is possible for a participant to learn the task well, but still receive negative feedback from their actions. Accordingly, such paradigms are well suited for studying the neural correlates of processing *unexpected* feedback, but not so much feedback which helps achieve a concrete learning goal.

Our task of interest. To address such a related, but different question, in the present study, we analyze a reward-driven verbal learning task (initially developed and investigated behaviourally in Chakravarty et al., 2019) where participants learned to associate each of 48 words with a binary response— either to choose the word and receive a reward (in the case of words we call "high value"), or to avoid it (in the case of words we call "low value") and receive the same. Participants learned these word-value mappings through trial-and-error over 16 training cycles, aiming to maximize reward earned over the course of the task (for similar approaches, see Arbel et al., 2014; Arbel et al., 2013; Chase et al., 2011; Ernst and Steinhauser, 2012).

A classifier-based subsequent-memory-effect approach

A major goal of memory research is to identify how brain activity supports memory performance: success versus failure. Measures of brain activity are, by necessity, observational, making it difficult to establish causality, and differentiate necessary versus sufficient activity. Indeed, much brain activity that is observed during memory tasks may be unrelated to memory success and even unrelated to the memory task altogether. However, progressively refining the approach to data analysis can bring us increasingly closer to identifying brain activity that drives memory performance and thus should eventually give us insights into the brain-activity basis of memory. One of the most important such major breakthroughs was the trial-sorting method introduced by Sanquist et al. (1980) and named the "subsequent-memory effect" by Wagner et al. (1998). The subsequent-memory effect

contrasts brain activity measured while people study items (e.g., words) that they later remembered versus items that they later forgot. The idea is that of all the brain activity present, a subset of that might reflect processes during encoding that lead to better memory. Those should be more present during later-remembered than later-forgotten trials. A major step forward, subsequent memory effects still typically average across trials and across participants, and thus identify activity that while robust by some definition, may be selected more for reliability across subjects and trials than for their causal role in memory outcome.

The field of machine-learning has offered tools to take the subsequent memory effect one step further. This perspective views conventional analyses as descriptive. This can yield results that are statistically robust but with considerable overlap in the signals that differentiate classes, such as subsequently remembered from subsequently forgotten items, which can be insufficient to make non-negligible predictions about learning or memory on a trial-to-trial level. Machine-learning classification methods have been used to identify signal that not only exhibits descriptive effects but can in fact substantially predict class identity, including avoiding overfitting training data. This approach has been applied to predicting subsequent memory in tasks like recognition and free recall (e.g., Arora et al., 2018; Chadwick et al., 2010; Chakravarty et al., 2020; Halpern et al., 2023; Höhne et al., 2016; Mirjalili et al., 2022; Weidemann & Kahana, 2021), as well as with functional magnetic resonance imaging (e.g., Watanabe et al., 2011).

Those prior studies typically examined episodic memory tasks where a stimulus, usually a word, is studied only once prior to the memory test. To our knowledge, a classifier-based subsequent-memory effect approach has not previously been used to study incremental, feedback-driven learning. That was our main goal here. Learning may occur at any time during the task, but unlike those typical episodic memory tasks, in a feedback-driven task, it is only at the time that the feedback stimulus, itself, is presented, that the participant obtains information about whether or not to adjust their knowledge about the stimulus. We therefore presume that in our task, when an item's value is learnt, brain activity that reflects learning of the item occurs while, or soon after, the feedback is viewed, with the consequence that the next trial with that stimulus is more likely to be responded correctly. Thus, we take a classifier-based approach to study EEG data recorded during feedback processing to evaluate the brain activity linked to learning concrete word-value associations. We build on our previous conventionally analyzed event-related potential study with this task (Chakravarty et al., in review).

One complication is that applications of classifier analyses of brain activity often start with data-driven feature-selection and the classifier hyperparameters and algorithms vary across studies. Although justifiable when the aim is to maximize classifier success, this can make it hard to evaluate classifier performance across studies, and raises the possibility that there might be some selective reporting in the field. With this perspective in mind, Chakravarty et al. (2020) took a two-stage approach aimed to minimize researcher degrees of freedom and obtain a more baseline level of classifier success, albeit probably more like a lower-limit. Developing a classifier-based subsequent-memory effect analysis of a conventional episodic, verbal, old/new recognition task, Chakravarty and colleagues started by asking whether frequently reported univariate features might already have some predictive power to classify subsequent memory (subsequent hits versus misses). They then applied

the simplest classifiers, linear discriminant analysis (LDA) and support vector machines (SVM) to signal that was first downsampled, not by data-driven feature-selection but by sampling a subset of electrodes and by time-binning voltage. We take the same approach to our feedback-driven, incremental learning task. First, we evaluate the predictive power of the highly replicated univariate measures of feedback processing in the EEG, the FRN and FMT, and then we turn to multivariate classifier analyses.

Finally, multivariate classifier-based analyses of EEG activity have typically used spectrographic features, or power as a function of electrode and time (e.g., Arora et al., 2018; Halpern et al., 2023; Y. Li et al., 2024; Mirjalili et al., 2022; Noh et al., 2014; Weidemann & Kahana, 2021). To our knowledge, Chakravarty et al. (2020) was the first to report classification success leaving the signal in its original time domain (electrode by time-bin). Because Fourier Transforms (and wavelet transforms) are linear, it seems intuitive that a linear classifier should have access to linear combinations of features and potentially exploit the same information embedded in the signal whether in the time-domain or spectrographic (time-frequency) form. We thus compared classifiers acting on the time-domain alone versus the time-frequency domain to check this. To foreshadow, the time-frequency features led to greater classifier success, which we revisit in the Discussion.

Overview

We take a two-stage approach to identifying learning-related brain activity during a feedback-driven task. First we re-consider robust univariate signals, the FRN and FMT, through the lens of classification, to ask whether these already might offer substantial predictive power. Second, we look beyond those univariate features and ask whether subject-specific, multivariate classifiers can identify more learning-predictive information in the EEG.

Stage 1. Consider that both the FRN and FMT are conventionally analyzed at the trial-average level. It is unclear whether either signal can index an RPE calculation and future behavioural adjustments at the level of *individual trials*. We take inspiration from the classifier-based subsequent memory effect paradigm, where researchers ask whether trial-level brain activity during item-encoding can be used to predict whether that item will later be remembered or forgotten (Arora et al., 2018; Chadwick et al., 2010; Chakravarty et al., 2020; Halpern et al., 2023; Höhne et al., 2016; Mirjalili et al., 2022; Weidemann & Kahana, 2021). We ask whether trial-level FRN and FMT magnitude can predict whether word values will be learned on a given trial, as indicated by future responses when that word is presented. Specifically, following the signal detection theory approach taken by Chakravarty et al. (2020) for single-trial episodic item-recognition, we trace the receiver operating characteristic (ROC) curve using single-trial FRN/FMT magnitude to predict correct responses, quantifying predictive power using the area under the curve (AUC).

Stage 2. Next, suspecting that univariate signals do not describe the full complexity of feedback processing, we ask whether multivariate classifiers incorporating broader spatial patterns of time and time-frequency domain data can predict learning to a greater degree. To minimize researcher degrees of freedom (Chakravarty et al., 2020) we apply linear classifiers to predict successful behavioural adjustments from pre-selected spatial, temporal, and spectral EEG features.

Thus, in a cognitively demanding paradigm with a concrete learning objective, we examine whether established univariate signals and broader multivariate activity measured during feedback processing can index successful item-value learning.

Methods

Data and conventional analyses are reported by Chakravarty et al. (in review) where the full methodological details are reported. We summarize them here. The procedures were approved by a University of Alberta ethical review board.

Participants

In total, 68 introductory psychology students at the University of Alberta participated in exchange for partial course credit. All participants spoke English as their first language and had normal or corrected-to-normal vision. Written consent was obtained prior to the experiment. We performed a manual inspection of the EEG, and abnormal channels were removed, for example, flat/no signal, channels affected by high frequency/line noise, head motion, etc. We followed the standard guidelines for our system: 256 channel HCGSN nets from EGI, recording done with Netstation. We also used EEGLAB's kurtosis method to further remove bad channels. Further, ICA was used to remove artifacts due to eye blinks, muscle noise, etc., removing approximately 5–10 ICs per participant. Exclusion of participants based on EEG data was based on a combination of factors: 1) too many channels (∼30% or more) removed due to noise, 2) a cluster of channels removed from a specific region, 3) poor quality of the ICs, and 4) many missing event triggers. Consequently, data from 10 participants were excluded: 5 due to corrupted data files, 3 due to excessive EEG noise, and 2 due to missing event triggers. In the current study, an additional 11 participants were excluded for achieving less than 80% average accuracy in the last four training cycles (cycles 13 to 16, see below). Another 2 participants were excluded, as their average spectrogram power values were an order of magnitude greater than the across-participants average, leaving 45 participants available for the current analysis.

Stimuli

Briefly, stimuli were a set of 48 words, assigned to one of two feedback outcomes: a 10-point reward (high-value) or a 1-point reward (low-value). At the beginning of each participant's session, half of the words were randomly categorized as high-value, with the rest categorized as low-value. Word values were to be learned through trial-and-error feedback over multiple word presentations, where the participants aimed to select high-value words to maximize reward over time.

Task

Participants completed a word-choice task where, on each trial, they chose between one of the 48 word stimuli and a fixed non-word string ("HHHHH") presented side-by-side on a computer screen (Figure 1). The left/right position of the word relative to the string was randomized across trials. Participants were instructed that the task included high- and low-value words and that, if they encountered a high-value word, choosing the word

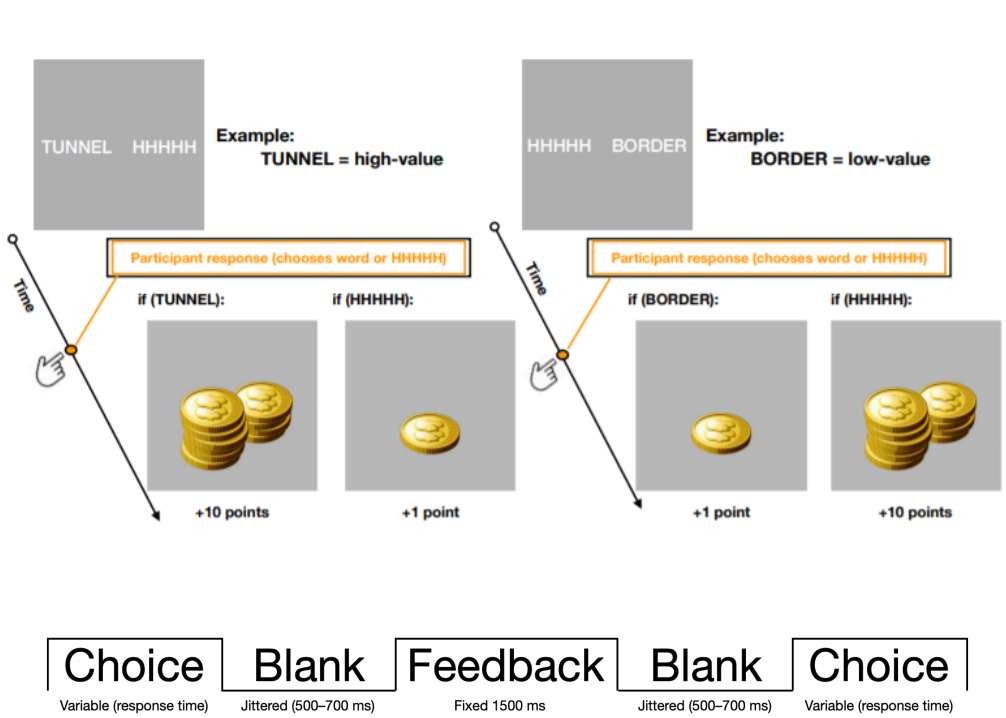

Figure 1

A schematic of the task. a) Each grey box represents what a participant saw on the screen at a given stage of the task. Note that the alternative option, "HHHHH," was identical on each trial, sometimes on the left and sometimes on the right. It served as a placeholder to give the participant a way to "not-choose" the word. Reproduced from Chakravarty et al. (in review). b) Timeline of the procedure.

would give 10 points and choosing the string would give 1 point (the opposite was true for low-value words— selecting the word gave 1 point, while selecting the string gave 10 points).

Each of the 48 words was presented once each per cycle over 19 cycles, where the word order was random within each cycle. Participants were instructed that their aim was to maximize the total points earned across cycles. After the task, points were converted to a monetary bonus using a conversion rate of $0.0006 per point rounded up to the nearest $0.25 CAD; participants could earn up to $5 CAD.

On a given trial, the word and string remained visible until the participant made a keyboard response: pressing "Q" for the stimulus on the left, and "P" for the stimulus on the right. Feedback then appeared with a jittered onset time of 500–700 ms after the response was made, remaining visible for 1500 ms. For the 10-point reward, an image of a pile of coins was displayed, while for the 1-point reward, an image of a single coin was presented (Figure 1). Participants could also see their total points earned so far during each feedback presentation. After feedback, the next trial started with a jittered onset of 500–700 ms.

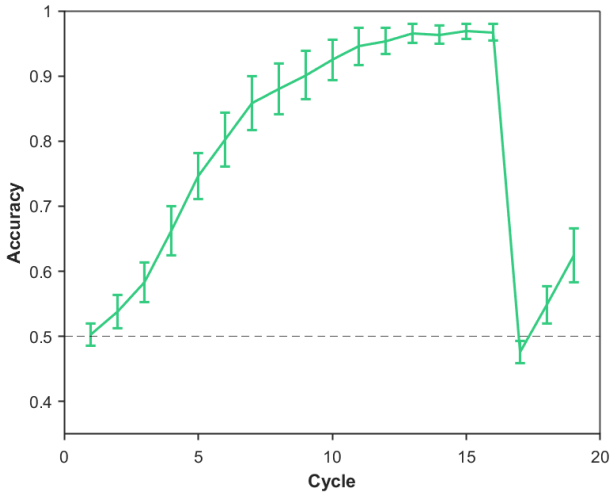

Figure 2

Average accuracy across participants by cycle. Participants accurately learned word values before the reversal at cycle 17, with average accuracy nearing 100%. Error bars represent 95% confidence intervals.

Finally, left out of the current analyses (but reported in Chakravarty et al., in review), on cycle 17, a surprise reversal occurred, where half of the word values were randomly switched. We focus on the training stage, where participants successfully learned stable word-value associations (Figure 2; full behavioural analyses are also reported in Chakravarty et al., in review).

EEG Recording

EEG activity was recorded using a 256 channel Geodesic Sensor Net (Electrical Geodesics, Inc.). Recordings were sampled at 500 Hz, amplified at a gain of 1000, and referenced to the mid-central electrode Cz. Electrode impedance was kept below 50 kΩ. Recordings were re-referenced to the channel average and digitally band-pass filtered between 0.1–40 Hz. Visible artifacts such as eyeblinks were corrected using Independent Component Analysis, and recordings from noisy channels were rejected and interpolated using nearby electrodes. We also rejected any trial epochs (0–1500 ms post-feedback) where recordings deviated by more than 300 μV from baseline (–200–0 ms), or any which included point-to-point voltage fluctuations of more than 25 μV, resulting in the removal of $M(SD) = 2.20 \pm 7.38$ trials (range=[0, 48]). Pre-processing was done using custom MATLAB scripts in conjunction with the EEGLAB library (Delorme & Makeig, 2004).

Classification Problems

Predicting subsequent accuracy

First we asked whether brain activity during feedback could predict subsequent behavioural adjustments in the training phase of the task (i.e., the first 16 cycles). Specifically,

we analyzed whether word responses (correct/10 points vs. incorrect/1 point) in cycle $N+1$ could be predicted from brain activity associated with the word (its feedback-processing) in cycle N.

For cycles 1 to 15, we labelled each trial/word as "subsequently correct" or "subsequently incorrect" according to whether the participant responded correctly or incorrectly to that word in the following cycle (cycles 2 to 16, respectively). Specifically, correct responses were marked as those giving a 10-point reward, which could be achieved by choosing either the word (for high-value words) or the string (for low-value words).

Then we aimed to classify subsequently correct from subsequently incorrect trials based on brain activity during feedback processing (see below). Analyses were performed separately based on the feedback received in cycle N (correct trials only, incorrect trials only, or all trials), allowing us to ask whether the brain activity associated with the processing of more rewarding (10 point) and less rewarding (1 point) feedback could predict subsequent accuracy equally well.

Predicting item-value acquisition

As participants had a 50% chance of correctly responding to a given word, even the subsequently correct responses could have been lucky guesses. Accordingly, we established a more conservative criterion to determine the one trial when a word's value was ostensibly acquired. We marked that trial as "acquired" if, following that trial, a participant only responded correctly to *that same word* on the remainder of the training cycles. We further required that "acquired" trials were followed by at least 3 correct responses for that word, thus eliminating words that might have been acquired during the final 3 cycles.[1] With this approach, there was at most one "acquired" trial for a given word.

Note that we use the term "acquired" with the aim of isolating the trial where the participant stably determined a word's value. As such, the rest of the trials for a given word were labelled as "unchanged", including those both before and after the word was "acquired". However, even with our stricter criterion, it is still unclear *exactly* when a word value was acquired. If a participant produced an incorrect response followed by 10 consecutive correct responses, it is possible that the word was acquired following that incorrect response (i.e., the participant learned from their error), or following one of the first few correct responses (i.e., the participant learned from an unexpected reward). Accordingly, we arranged two classification problems considering both situations. First, we assumed that item value acquisition only took place following correct-response, and labelled each correct trial as "acquired" (a correct response followed by nothing but 3+ consecutive correct responses to that same word) or "unchanged" (all other correct trials). Next, we assumed that acquisition only took place following errors, and followed the same procedure, but for incorrect trials instead.

[1] While learning presumably occurred in the last 3 cycles, we deemed that there were not enough consecutive correct responses for us to ascertain this from the behavioural data so here we excluded items that might have been acquired in cycles 14–16).

Planned comparisons

We analyzed two well-established markers of feedback processing, the FRN and FMT. Before asking whether behavioural adjustments could be *predicted* from the FRN or FMT, we followed a conventional trial-average approach and evaluated the mean difference in either signal between our conditions of interest (subsequently correct/subsequently incorrect and acquired/unchanged).

To quantify the FRN, for a given trial, we extracted EEG epochs from –200–1500 ms relative to feedback onset. Epochs were baseline-corrected relative to the –200–0 ms baseline period. FRN magnitude was defined as the average baseline-corrected voltage at the frontocentral electrode FCz (channel 15) from 200–350 ms following feedback onset. We chose this window based on our rough reading of FRN papers. There is a good argument that the FRN is better measured toward the earlier end of this window, whereas P200 and P300 effects may "contaminate" the later portion of this window. In a follow-up analysis, Chakravarty et al. (in review) confirmed a robust FRN effect using a window of 190–230 ms, suggesting that a shorter, earlier window would also capture the effect. However, here we stick with our initial window to reduce degrees of freedom. The contamination by other ERP effects is a problem, but rather than post-hoc adjust the window, we turn to the multivariate classifiers to obtain the (more) full picture of learning-related brain activity. Our larger goal with this approach is in fact to avoid the process of trying to isolate specific putative signals that in fact overlap spatially and temporally.

To quantify FMT, EEG recordings were transformed to the time-frequency domain using custom MATLAB scripts, which included functions from the EEGLAB (Delorme & Makeig, 2004) and Better OSCillation detection (BOSC) method (Whitten et al., 2011) libraries. For each participant, a Morlet wavelet transform was applied to the continuous EEG recording made at each electrode throughout the entire task. The mother wavelet had a width of 6 cycles, and the 18 daughter wavelets had frequencies logarithmically spaced between 1.0–19.0 Hz. Wavelet power was log-transformed before further analyses. For each trial, FMT was defined as the power averaged across frequencies sampled between 4–8 Hz over the 200–450 ms time window following feedback onset (Luu et al., 2004).

Trial-averaged FRN and FMT measures for the two pairs of conditions (subsequently correct/subsequently incorrect and acquired/unchanged) were compared across subjects using paired-samples t-tests. Based on previous literature, we expected greater FRN magnitude (lower voltage) and higher FMT log-power for subsequently correct than subsequently incorrect trials. Likewise, we expected greater FRN magnitude and FMT power for acquired than unchanged.

We repeated the above analysis for several other comparisons, where the same procedure was followed, but for different conditions (e.g., subsequent accuracy following correct trials only). For all statistical tests, an alpha value of 0.05 was used to determine significance.

Classifier analysis

Separate classifier analyses were performed for each participant, and then performance is summarized across participants.

Univariate classifiers

To ask if the FRN and FMT could predict subsequent accuracy and item value acquisition, we followed the signal detection theory approach (Green & Swets, 1966) previously used by Chakravarty et al. (2020) for old/new recognition. With the set of trial-wise FMT magnitudes, we set sliding classification thresholds, where trials above the threshold were classified as subsequently correct and trials below as subsequently incorrect. A true positive rate and false positive rate were calculated for each threshold based on the true trial labels, and an ROC curve was drawn plotting the rates against each other for each threshold. AUC, calculated with MATLAB's perfcurve.m function, measured classifier success. AUC$>$ 0.5 would support our prediction that FMT should be more pronounced when participants received feedback that informed their future responses, while AUC$<$ 0.5 would suggest the opposite. We applied the same procedure for predicting acquired and unchanged trials and likewise for the FRN. Note that for the FRN, we set successive thresholds corresponding to increasingly *negative* voltage. Accordingly, we could interpret AUC$>$ 0.5 as indicating greater FRN magnitude in subsequently correct or acquired trials, matching our interpretation of AUC for FMT.

Multivariate classifiers

Next we considered whether multivariate EEG activity during feedback processing could predict subsequently correct/incorrect and acquired/unchanged trials.

Features. We conducted our multivariate analysis in two parts. We first classified time-domain features (voltage) and compared this with classification of time-frequency features (wavelet spectrogram).

Recall that we sampled from 256 electrodes at 500 Hz, yielding a large number of features per trial which risks raising the computational cost for little gain. As such, we restricted our analysis to a smaller feature subspace. Specifically, we chose 6 electrodes[2] roughly covering the scalp (Figure 3) and averaged the signal from these electrodes into 5 non-overlapping 300 ms time bins[3] from feedback onset to 1500 ms post-feedback. We extracted a total of 30 features (6 electrodes $\times$ 5 time bins) per trial for our time-domain classifiers.

For the same set of electrodes and time windows described above, we calculated the average power in each of 4 frequency bands: delta (1.0, 1.2, 1.4, 1.7, 2.0, 2.4, 2.8, 3.4 Hz), theta (4.0, 4.8, 5.7, 6.7, 8.0 Hz), alpha (9.5, 11.3 Hz), and beta (13.5, 16.0, 19.0 Hz). This

[2]It is common for researchers using EGI's dense-array caps (which by default offer 256 electrodes with minimal setup time) to analyze averages of clusters of neighbouring electrodes,with the idea that that should increase the signal-to-noise ratio. In our lab's experience, at least with a sufficient amount of data, cluster-averages produce results very similar to single electrodes. Our goal in this manuscript is not to help the classifier perform the best it can, but rather, to minimize researcher degrees of freedom and characterize classifier success using rather naïve methods. For this reason, we continue with our lab's standard, analyzing single electrodes rather than cluster-averages. We note that a future research project aimed at developing more optimal classifiers would do well to compare cluster-averages to single electrode subsampling. We also note that at least indirectly, signal quality was influenced by the full 256 electrodes because they all contributed to the ICA-driven artifact correction procedures.

[3]Chakravarty et al. (2020) used 100 ms bins; 300 ms time bins were chosen here to better compare with the time-frequency features, for which 100 ms bins would be severely oversampled

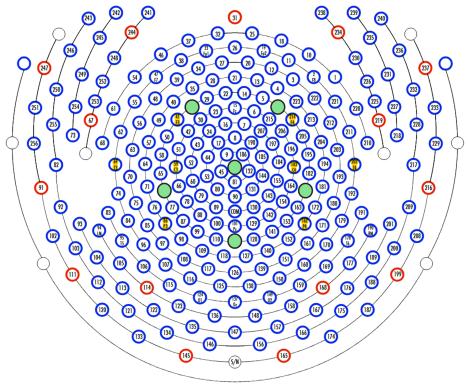

Figure 3

Electrode locations on the 256-channel Geodesic Sensor Net. The 6 electrodes used to generate the features for our multivariate classifiers are highlighted in green.

produced a total of 120 features (6 electrodes × 5 time samples × 4 frequency bands) per trial for our time-frequency domain classifiers.

Algorithms. To minimize researcher degrees of freedom, we followed the classifier approach that Chakravarty et al. (2020) applied to single-trial item-recognition. Specifically, we compared two simple linear classifiers, Fisher's linear discriminant analysis (LDA; Fisher, 1936) and support vector machine (SVM; Cortes and Vapnik, 1995) with a linear kernel. For LDA, we used the fitcdiscr.m function (with the Gamma regularization parameter set to 0.5), and for SVM, we used the fitcsvm.m function (with the BoxConstraint tolerance parameter set to 0.5), both from MATLAB's Statistics and Machine Learning toolbox (Martinez et al., 2017). Classifier success was measured by calculating the AUC of the ROC curve for classifier score (i.e., classifier confidence that a trial belonged to the positive class, subsequently correct or acquired).

Cross-validation. For each analysis, classifiers were trained and tested under a 5-fold cross-validation scheme.[4] The trial set was randomly partitioned into 5 subsets/folds, stratified such that the proportion of trials belonging to each of the two classes was relatively preserved across them. In each iteration, a classifier was trained using data from 4 folds and tested on the remaining 1. Each fold was used as the test set exactly once. AUC was averaged across the 5 test folds to measure overall performance.

Post-hoc feature analysis. For each LDA classifier where AUC was above chance, we visualized the feature weights, or the relative importance of each feature for classification (the DeltaPredictor property of the classifier object). Specifically, we re-scaled and averaged weights across participants, then plotted across time and space.

Performance evaluation

For both univariate and multivariate classifiers, we used one-sample t-tests to measure performance across participants: the set of participant-wise AUC values for a given classification problem was tested against chance (AUC = 0.5).

[4]This follows the approach taken by Chakravarty et al. (2020) but with fewer folds to compensate for having relatively fewer trials here

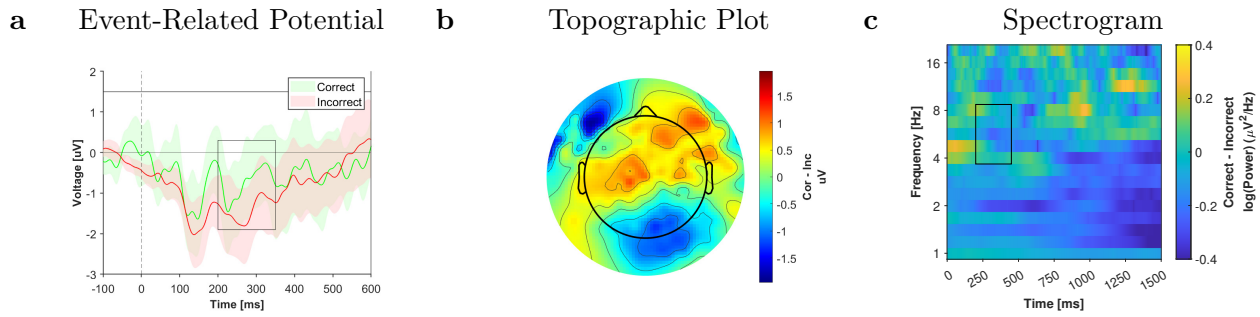

Figure 4

(a) Event-related potentials during the reversal cycle (cycle 17) when half the word values were switched. The FRN is evidence in a more negative voltage at FCz during the window 200–350 ms when processing feedback indicating the response was incorrect than correct. Error bars are timepoint-wise 95% confidence intervals based on standard error of the mean. (b) Topographic plot of mean voltage at all electrodes, spline-interpolated, across the 200–350 ms time window. (c) Spectrogram plotting the difference in log(power) between feedback indicating correct and incorrect responses at FCz as a function of frequency and time. The rectangle indicates the time window (200–450 ms) and frequency range (4–8 Hz) used to quantify FMT.

Results

Before reporting the main analyses, we check whether there is robust evidence of the presence of both the FRN and FMT in our task, in conditions where they are typically found, during cycle 17 when half the item–response mappings (item values) were switched after participants had learnt most of the items (Figure 2). We expected more negative voltage (FRN) during incorrect (mostly switched trials, violating expectation) than correct (mostly non-switched feedback trials, matching expectation) and more theta-band power during switched than non-switched feedback trials (FMT). The FRN showed a significant effect in the expected direction (Figure 4a), paired-samples $t(41) = 3.18$, $p = 0.003$, and with frontocentral topography, consistent with prior FRN reports (Figure 4b). The FMT effect (Figure 4c) was not significant in this contrast, $t(41) = -0.689$, $p = 0.495$.

Planned comparisons for the FRN and FMT

Subsequent accuracy

First, we compared trial-averaged FRN and FMT between subsequently correct and incorrect trials, where the FRN was significantly greater (i.e., FRN voltage was more negative) for the subsequently correct condition (Figure 5, Table 1). This relationship held when restricting our analysis to trials where a correct response was given, but not for trials where an incorrect response was given. Meanwhile, FMT power was significantly greater for subsequently correct trials, and again, this was only true when the whole set of trials or only correct trials were used (Figure 5, Table 1). Note, however, that power was broadly increased in the subsequently correct condition, not selectively for FMT frequencies or that

Signal	All	Correct	Incorrect
FRN	$\mathbf{t = -3.82}$, $\mathbf{p < 0.001}$	$\mathbf{t = -3.83}$, $\mathbf{p < 0.001}$	$t = -0.088$, $p = 0.931$
FMT	$\mathbf{t = 2.12}$, $\mathbf{p = 0.039}$	$\mathbf{t = 2.37}$, $\mathbf{p = 0.022}$	$t = 0.545$, $p = 0.588$

Table 1

Paired-samples t-tests ($df = 44$) of the difference in FMT power or FRN voltage between subsequently correct and subsequently incorrect trials across participants. Columns represent whether all trials in cycle N were considered, or only correct/incorrect trials. Boldface indicates a significant difference, $p < 0.05$.

Signal	Correct	Incorrect
FRN	$\mathbf{t = 2.87}$, $\mathbf{p = 0.006}$	$t = 1.30$, $p = 0.202$
FMT	$\mathbf{t = -3.51}$, $\mathbf{p = 0.001}$	$t = -1.61$, $p = 0.114$

Table 2

Paired-samples t-tests ($df = 44$) of the difference in FMT power or FRN voltage between acquired and unchanged trials across participants. Columns represent whether we assumed item value acquisition occurred following correct or only incorrect-response.
Boldface indicates a significant difference, $p < 0.05$.

time window (Figure 5).

Item value acquisition

As elaborated in the Methods, we followed up with a classification problem that was designed to be more targeted toward the precise (inferred) trial when each given item was learned, which we call "acquired." This was compared to prior and subsequent trials which we call "unchanged." This comparison was again significant only following correct but not incorrect trials (Figure 6, Table 2). Curiously, the directions of the effects were reversed from the subsequent-accuracy comparisons.

Summary of planned comparisons

The trial-averaged FRN and FMT were more pronounced for subsequently correct trials, as expected, and unchanged trials, which we did not anticipate. This held for correct, but not incorrect current feedback trials. The descriptive effects are thus confirmed. Next we test whether these signals are predictive.

Predicting subsequent accuracy

First, we asked whether the FRN, FMT, or multivariate brain activity predicted subsequent accuracy.

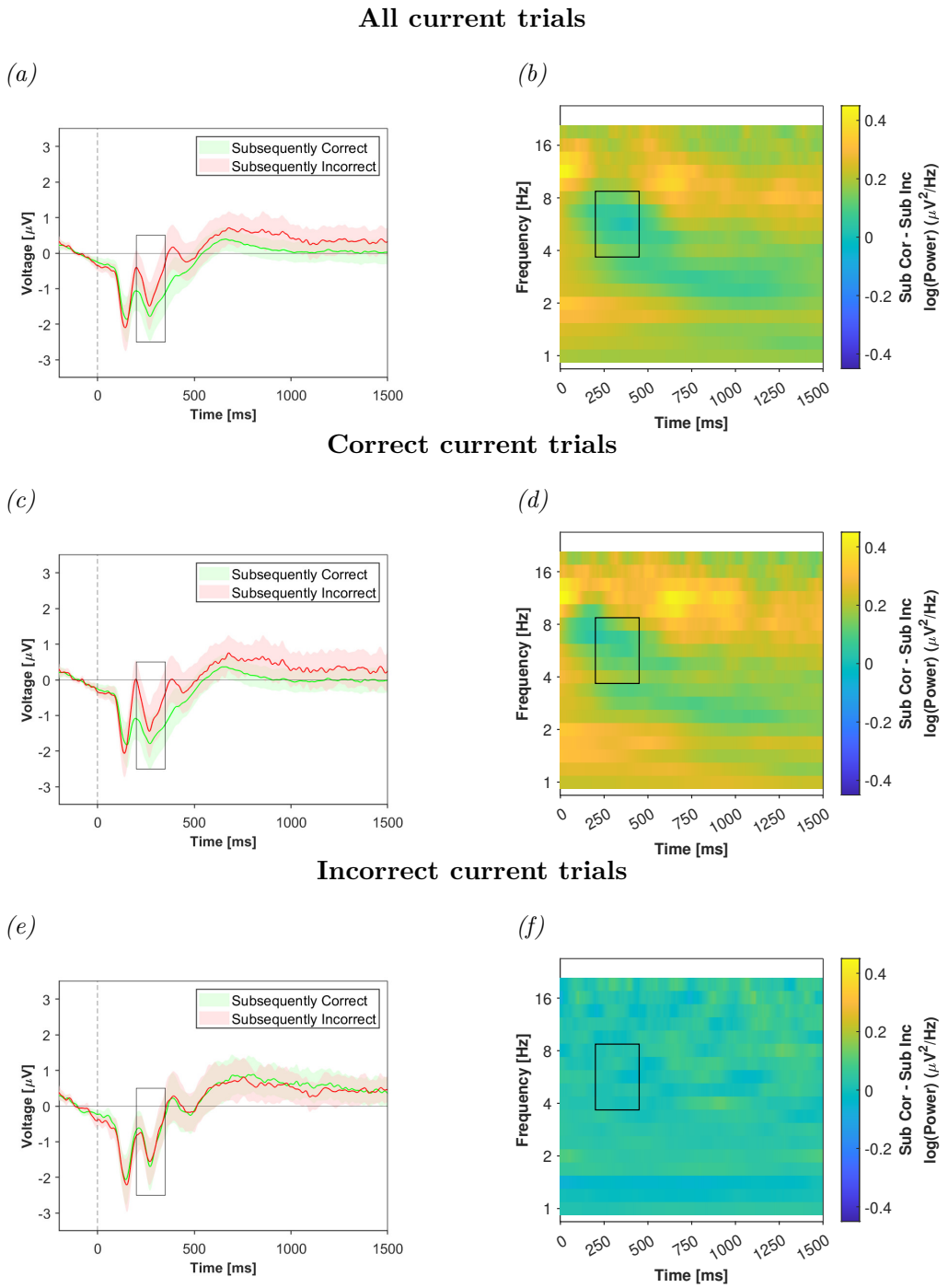

Figure 5

a,c,e) Event-related potential plots (left) at FCz for subsequently correct and subsequently incorrect trials across participants. Black rectangles denote the FRN time window. Error bars are timepoint-wise 95% confidence intervals based on standard error of the mean. b,d,f) Spectrograms of the power difference at FCz between the average subsequently correct and subsequently incorrect trials across participants. The black rectangles denote the FMT frequencies and time window. a,b) All trials are included. c,d) Only trials that were currently correct are included. e,f) Only trials that were currently incorrect are included. Sub=Subsequently.

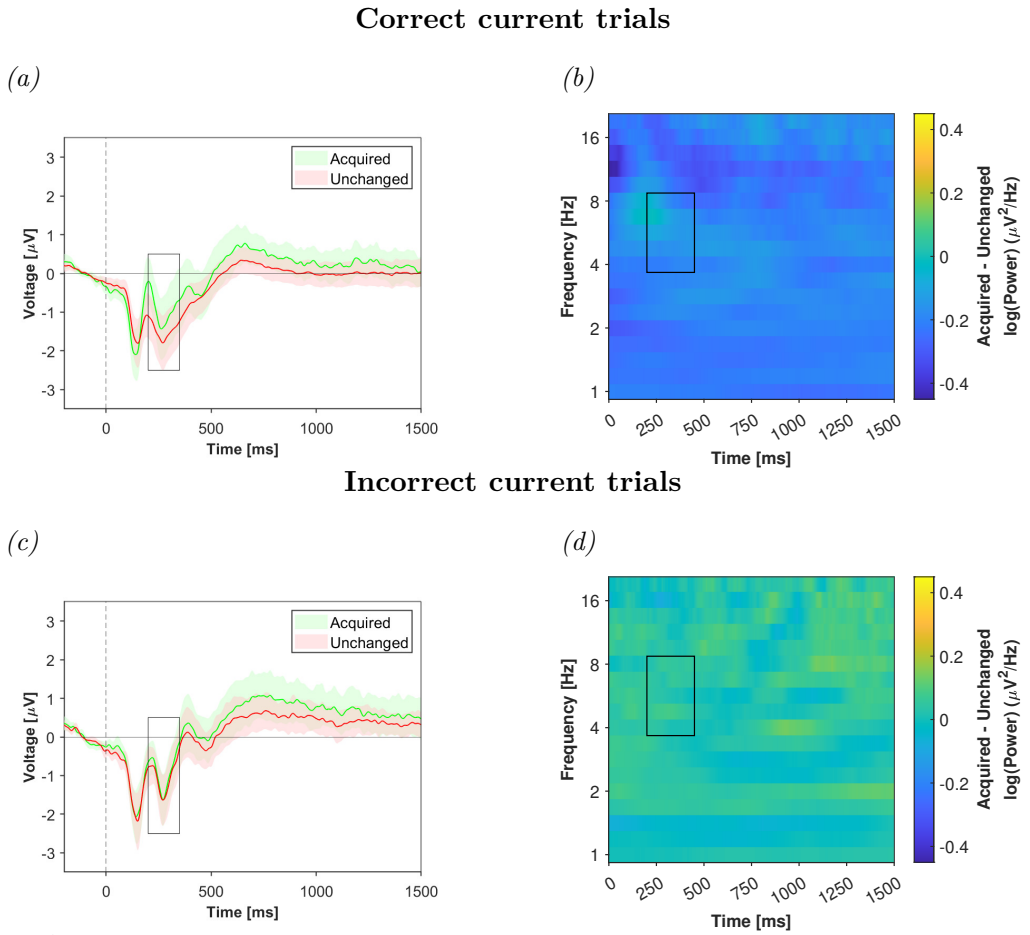

Figure 6

a,c) Event-related potential plots (left) for for acquired and unchanged trials across participants. Black rectangles denote the FRN time window. Error bars are timepoint-wise 95% confidence intervals based on standard error of the mean. b,d) Spectrograms showing the power difference at FCz between the average acquired and unchanged trials across participants. Black rectangles denote the FMT frequencies and time window. a,b) All trials are included. c,d) Only trials that were currently correct are included. Sub=Subsequently.

Classifier	All	Correct	Incorrect
FRN	**0.541 [0.521 0.562]** **t = 4.08, p < 0.001**	**0.554 [0.531 0.578]** **t = 4.71, p < 0.001**	0.492 [0.477 0.506] $t = -1.14$, $p = 0.261$
FMT	0.524 [0.497 0.551] $t = 1.82$, $p = 0.075$	**0.534 [0.501 0.567]** **t = 2.09, p = 0.043**	0.504 [0.483 0.525] $t = 0.37$, $p = 0.717$
LDA (T)	**0.609 [0.582 0.636]** **t = 8.14, p < 0.001**	**0.612 [0.581 0.643]** **t = 7.25, p < 0.001**	0.494 [0.471 0.518] $t = -0.48$, $p = 0.633$
SVM (T)	**0.573 [0.552 0.594]** **t = 6.95, p < 0.001**	**0.573 [0.549 0.596]** **t = 6.20, p < 0.001**	0.496 [0.472 0.520] $t = -0.34$, $p = 0.738$
LDA (T-F)	**0.716 [0.690 0.743]** **t = 16.51, p < 0.001**	**0.711 [0.681 0.741]** **t = 14.10, p < 0.001**	0.517 [0.494 0.541] $t = 1.47$, $p = 0.148$
SVM (T-F)	**0.643 [0.616 0.670]** **t = 10.64, p < 0.001**	**0.641 [0.610 0.671]** **t = 9.16, p < 0.001**	0.512 [0.492 0.532] $t = 1.22$, $p = 0.229$

Table 3

Predicting subsequent accuracy. Mean AUC values, with 95% confidence intervals and paired-samples t-tests ($df = 44$), for each subsequent accuracy classification problem. Columns represent whether all trials in cycle N were considered, or only correct/incorrect trials. Classifiers included univariate measures (FMT or the FRN), and two multivariate classifiers (SVM or LDA) informed by either time (T) or time-frequency (T-F) domain data. Boldface indicates significantly different than chance (0.5), $p < 0.05$.

Univariate classification

Adding to the descriptive effects, AUC values from the FRN classifiers were significantly above chance across subjects when using all trials and when we restricted the analysis only to trials for which the current trial was correct (Figure 7a and Table 3). FMT predicted subsequent accuracy when restricted to correct current trials, though AUC was only marginally significant (Figure 7a and Table 3).

Multivariate classification using time-domain features

Next, we attempted to classify subsequent accuracy using LDA and SVM informed by multivariate time-domain features. Both LDA and SVM predicted subsequent accuracy when trained on all trials or correct (current) trials, but not when trained on incorrect current trials only (Figure 7a and Table 3). Multivariate classifiers were more successful than the FRN: for example, when predicting from all trials, LDA and SVM AUC values were significantly greater than those from the FRN ($t(44) = 5.22$, $p < 0.001$ and $t(44) = 2.63$, $p = 0.012$, respectively).

LDA feature maps (Figure 8) show that in both cases, the classifiers emphasized

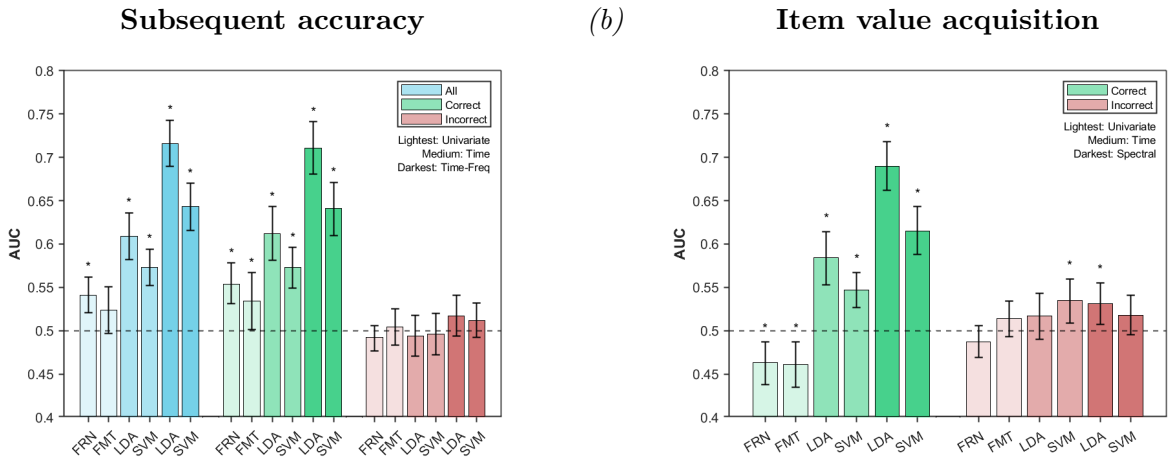

Figure 7

*Predicting subsequent accuracy and item value acquisition. Mean AUC values, with 95% confidence intervals, for each subsequent accuracy and item value acquisition classification problem. Blue, green, and red bars indicate classifiers trained on all, correct, and incorrect trials, respectively. Classifiers (indicated by brightness) included univariate measures (FMT or FRN; lightest), and two multivariate classifiers (SVM or LDA; medium) informed by either time or time-frequency domain data (darkest). * significantly different than chance (0.5), $p < 0.05$. The dashed line denotes chance AUC (0.5).*

all electrodes in the second time bin and distal electrodes across several time windows—overlapping with, but not exclusive to, the spatiotemporal window for the FRN.

Multivariate classification using time-frequency features

We conducted the same multivariate analyses with time-frequency domain data. The same pattern of classifier-success emerged, where LDA and SVM predicted subsequent accuracy following correct or all feedback, but not incorrect-response alone (Figure 7a and Table 3). Time-frequency classifiers were significantly more successful than those informed by time domain data: for example, both LDA ($t(44) = 7.36$, $p < 0.001$) and SVM ($t(44) = 5.28$, $p < 0.001$) performed significantly better when making predictions from all trials. The same relationship held when predicting from correct trials ($t(44) = 6.19$, $p < 0.001$ for LDA, and $t(44) = 4.00$, $p < 0.001$ for SVM).

Figures 9 and 10 show the feature weights for the above-chance classifiers. In both cases, the classifiers emphasized early latencies (in the 1–300 ms range), the beta and delta frequency bands, and distal electrodes. Heavily weighted features did not overlap with the spatial, temporal, or frequency characteristics of FMT.

Summary of predicting subsequent accuracy

While both the FRN and FMT differentiated between subsequently correct and incorrect trials at the trial-average level, it was not known whether the signals would be

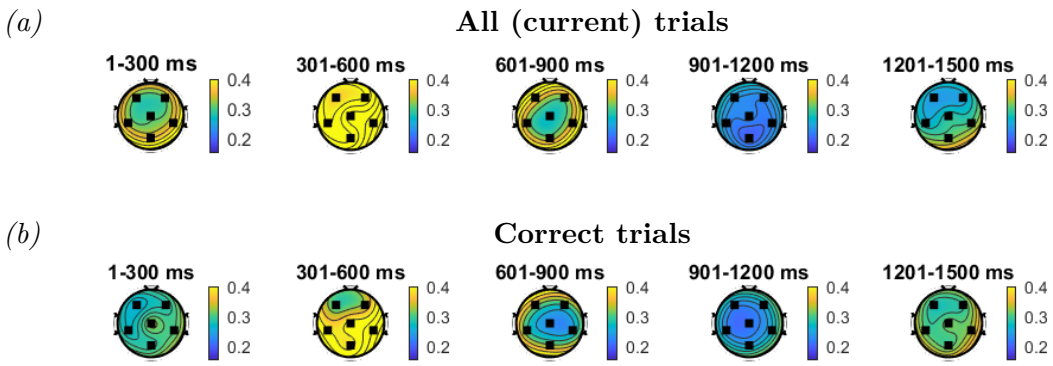

Figure 8

Predicting subsequent accuracy from all or correct trials: LDA time domain feature weights. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

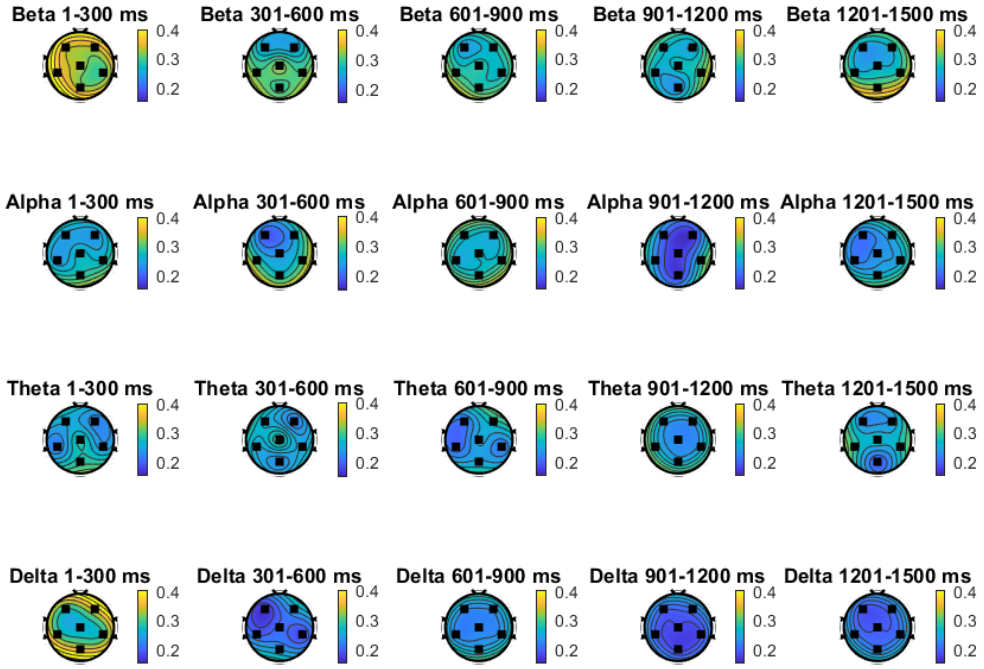

Figure 9

Predicting subsequent accuracy from all trials: LDA time-frequency feature weights. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

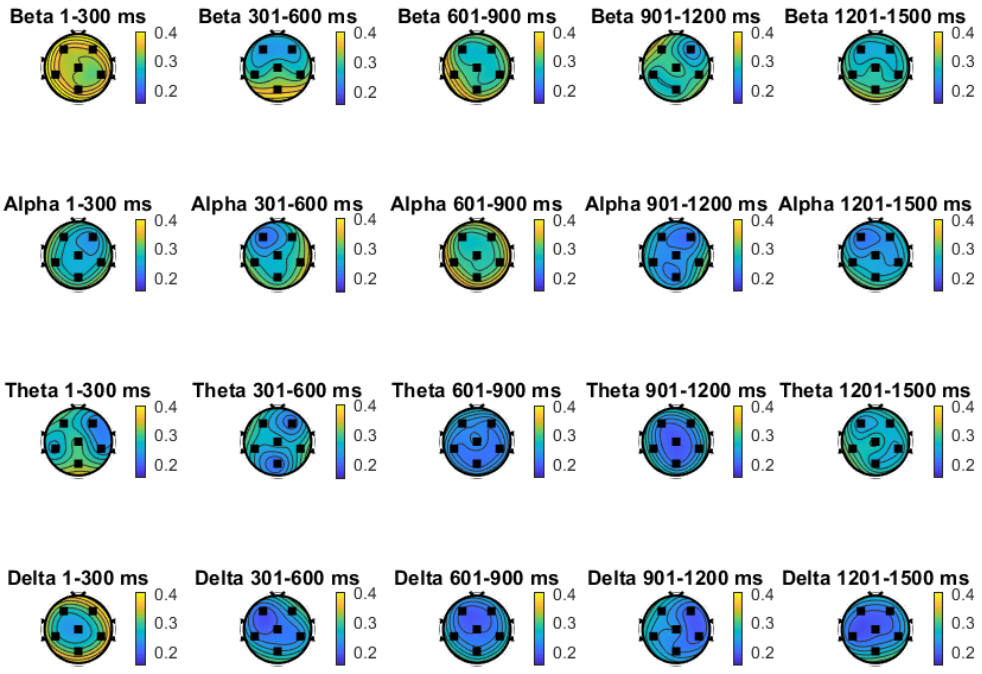

Figure 10

Predicting subsequent accuracy from correct trials: LDA time-frequency feature weights. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

predictive at the level of individual trials. From our results, the FRN was confirmed as a predictive signal. FMT was also predictive, but it was less robust.

Moving beyond univariate signals, the multivariate classifiers incorporating broad patterns of brain activity predicted subsequent accuracy to a substantial degree, if informed by trials where all or correct-response was given. Classifiers performed best when informed by time-frequency data, with AUC around 0.7 in the best cases. We could also predict subsequent accuracy using time domain data to a lesser degree, with some improvement over using the FRN alone. Importantly, we observed that activity extending beyond the FRN and FMT (though likely including the former) did, in fact, inform classifier decisions.

Predicting item value acquisition

As before, we next report the more selective analysis targeting the trial at which a given item's value was presumably first learned, predicting "unchanged" versus "acquired" trials: the transition points where word values became ostensibly known for the rest of training.

Classifier	Correct	Incorrect
FRN	**0.463 [0.438 0.487]** **t = −3.02, p = 0.004**	0.487 [0.469 0.506] $t = -1.34$, $p = 0.186$
FMT	**0.461 [0.435 0.487]** **t = −2.99, p = 0.005**	0.514 [0.493 0.534] $t = 1.36$, $p = 0.180$
LDA (T)	**0.584 [0.553 0.614]** **t = 5.57, p < 0.001**	0.517 [0.490 0.543] $t = 1.25$, $p = 0.216$
SVM (T)	**0.547 [0.527 0.567]** **t = 4.78, p < 0.001**	**0.535 [0.509 0.560]** **t = 2.74, p = 0.009**
LDA (T-F)	**0.690 [0.662 0.718]** **t = 13.78, p < 0.001**	**0.531 [0.507 0.555]** **t = 2.61, p = 0.012**
SVM (T-F)	**0.615 [0.588 0.643]** **t = 8.39, p < 0.001**	0.518 [0.495 0.541] $t = 1.54$, $p = 0.130$

Table 4

Predicting item value acquisition. Mean AUC values, with 95% confidence intervals and paired-samples t-tests (df = 44), for each item value acquisition classification problem. Columns represent whether we assumed item value acquisition took place following correct or incorrect-response. Classifiers included univariate measures (FMT or the FRN) and two multivariate classifiers (SVM or LDA) informed by either time (T) or time-frequency (T-F) domain data. Boldface indicates significantly different than chance (0.5), $p < 0.05$.

Univariate classification

The FRN predicted item value acquisition following correct-response (Table 4 and Figure 7b), but the effect was in the opposite direction of what we expected (i.e., greater FRN voltage predicted unchanged trials; significantly different than chance according to a two-tailed test). This matched what we observed in our ERPs, where the average FRN for the unchanged condition was more negative (Figure 5). Meanwhile, the FRN could not predict acquired trials following errors, consistent with results from our subsequently correct/incorrect analysis.

The set of FMT AUC values was significantly *below* chance for correct trials (Table 4 and Figure 7), suggesting that reduced FMT activity could predict item value acquisition. As with the FRN, FMT held no predictive value following incorrect-response.

Multivariate classification using time-domain features

Using multivariate time domain data, LDA and SVM both predicted acquired trials from correct-response (Table 4 and Figure 7b). Compared to their corresponding subsequent accuracy classifiers, LDA performance was significantly lower ($t(44) = -2.41$, $p = 0.020$),

while there was a trend towards the same for SVM ($t(44) = -1.92$, $p = 0.060$). For incorrect trials, SVM AUC values were marginally (but still significantly) above chance, whereas LDA AUC values were not. These results mirror the subsequent accuracy analysis, where time-domain features only tracked behavioural adjustments after rewarding feedback.

The average LDA feature maps (Figure A1) show a pattern remarkably similar to the feature maps for the subsequent accuracy classifiers (Figure 8), with LDA emphasizing the second time window (301–600 ms) and a broad range of electrodes, again overlapping with the FRN.

Multivariate classification using time-frequency features

When informed by time-frequency domain data, LDA and SVM predicted item value acquisition to a substantial degree from correct-response (Table 4 and Figure 7). Again, time-frequency classifiers performed significantly better than their time-domain counterparts ($t(44) = 6.38$, $p < 0.001$ for LDA, $t(44) = 4.86$, $p < 0.001$ for SVM). Classifiers performed similarly to their counterparts predicting subsequent accuracy from correct trials ($t(44) = -1.50$, $p = 0.140$ for LDA, $t(44) = -1.67$, $p = 0.102$ for SVM). As before, AUC dropped when classifiers were trained with incorrect trials only, with SVM at chance performance and LDA marginally above.

LDA feature weights for classifiers informed by correct trials (Figure A2) show a pattern that mirrors that from the subsequent accuracy classifiers (Figure 10), where LDA emphasized non-central electrodes, early time windows, and the delta and beta frequency bands, again failing to overlap with FMT.

Summary of predicting item value acquisition

Our classifiers predicted acquired and unchanged trials to about the same degree as they predicted subsequent accuracy, with AUC again approaching 0.7 in some cases. As with the subsequent accuracy analysis, time-frequency data informed predictions to a greater degree than time domain data, and activity beyond the FRN and FMT informed classifier decisions. For the univariate analyses, increased FRN and FMT magnitude successfully predicted *unchanged* trials, which we did not expect *a priori*. The direction of this effect matched our initial trial-averaged results, however, suggesting that both signals could offer information at the individual trial level. Finally, our classifiers were much more successful when we assumed that item value acquisition took place only following rewarding feedback.

In establishing the item value acquisition approach, we aimed to "de-noise" the behavioural data by assuming that acquisition only took place when indicated by a consistent stream of correct responses. While these acquisition labels were likely still noisy to some degree, we expected that classifiers informed by such labels would be less influenced by random guesses. However, our classifiers performed similarly across the subsequent accuracy and item value acquisition problems.

Addressing effects of class imbalance

Because participants were more accurate over time (Figure 2), there were more subsequently correct trials in later cycles and more acquired trials in earlier cycles. Alongside this trial imbalance, we also observed a systematic shift in EEG activity across cycles

(Figure 11), whether due to drift (where the average value of the signal changes as the task progresses) or some systematic change in brain activity over the session. So, a "lazy" classifier might be able to use some measure of this shift as a way to roughly infer task progress (cycle number), then take advantage of the relationship between cycle number and the class imbalance to simply make predictions based on inferred cycle— ignoring true learning-relevant brain activity in the process.

First we quantified the class imbalance across cycles. Then we asked whether the imbalance was related to classifier success. Finally, we assessed how our results changed when we restricted classifiers from accessing cycle number. We used the subsequent accuracy classifiers as a test case, since this classifier problem started out better-powered than the item value acquisition classifier.

Establishing the cycle imbalance confound

First, to quantify how well a classifier could perform if it were only exploiting the cycle imbalance, we traced ROC curves for each participant using each trial's cycle number. Concretely, each trial was labelled according to which cycle it belonged to. Then, we set each cycle number as a threshold, with each trial belonging to a cycle at this threshold or above classified as "subsequently correct." As more subsequently correct trials appeared in later cycles, AUC values were well above chance, as shown in Table 5. In both classification problems where classifiers consistently performed above chance (predicting subsequent accuracy from all or correct trials), the cycle imbalance was substantial, with AUC values around 0.8. Meanwhile, in the problem where classifiers performed around chance (where only incorrect trials were used), the cycle number confound was present, but to a lesser degree (AUC = 0.591).

Next, we wondered whether our classifiers performed best on participants whose data was most skewed by the cycle imbalance (i.e., participants who succeeded at the task earlier on). For each classifier problem, across participants, we correlated classifier AUC values with AUC values quantifying the class imbalance, with results shown in Table 5. For the two problems where classifier AUC was well above chance, predicting from all or correct trials, multivariate classifier success was related to the cycle imbalance, regardless of whether time or time-frequency data was used. Interestingly, this relationship was not present for the FRN or FMT univariate classifiers, despite the FRN achieving clearly above-chance classifier performance. Meanwhile, there was no relationship between cycle imbalance and classifier success for any classifiers trained on incorrect trials only (Figure A5), just as none of these classifiers initially performed above chance.

Correcting for the cycle imbalance confound

Because we found that cycle-related class-imbalance was related to classifier success, we considered the possibility that our classifiers may have used learning-irrelevant brain activity to predict learning outcomes. To check for the potential effects of this class imbalance confound,[5] we asked how performance would change if we trained classifiers on data which

[5]The go-to method to address class imbalance is synthetic minority oversampling technique (SMOTE; Chawla et al., 2002, also used by Chakravarty et al., 2020), where synthetic trials are computed from existing trials from the smaller class. However, the class imbalance of concern here is within each given cycle. SMOTE would be inappropriate, for example, when the smaller class had just one, or even two trials, which would

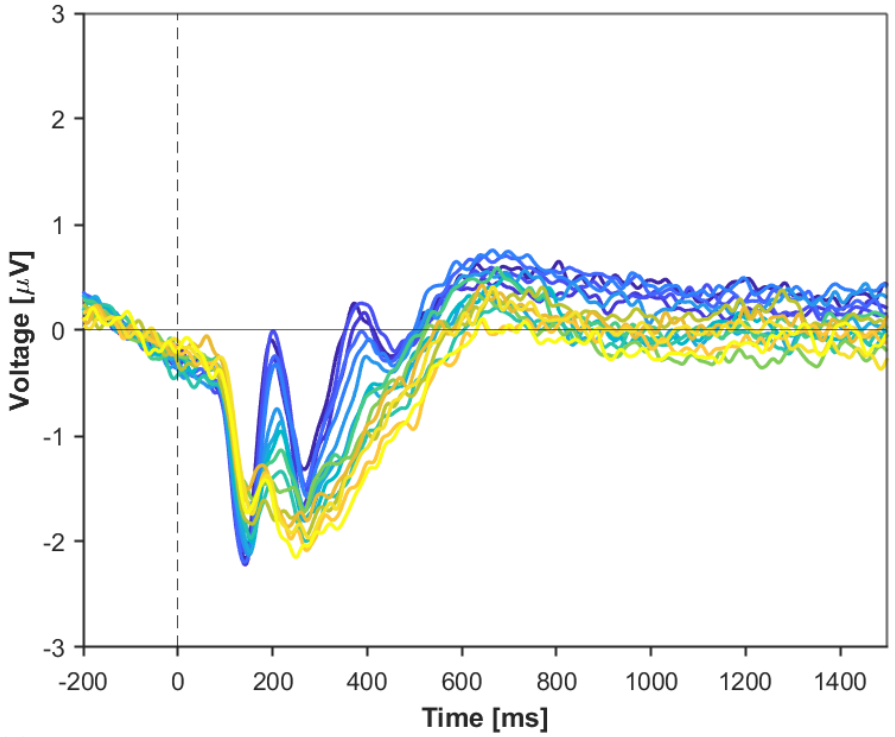

Figure 11

ERP at FCz across participants for each of the 16 learning cycles. Dark blue ERPs represent earlier cycles, while light yellow ERPs represent later cycles.

offered no overlapping information between cycle number and subsequent accuracy. We did this by subsampling the larger class (subsequently accurate versus subsequently inaccurate), similar to how Y. Li et al. (2024) addressed accuracy-based class-imbalance as a function of serial position in their free-recall task. If the classifiers could still succeed under these conditions, they must be doing so without using cycle number as a shortcut for predicting behavioural adjustments. (But note that the converse is unclear; if classifier success goes to chance in this analysis, that could be due to the reduced number of included trials leading to noisy ROCs, so these should be viewed more as a lower-limit on classifier success).

For each participant, we determined how many trials from a given cycle belonged to either class. Next, within each cycle, we randomly shuffled the trials from the most represented class (usually subsequently correct) and discarded the first N trials from this list, with N corresponding to the difference between the number of over- and under-represented trials within the cycle. When all trials had the same label for a given cycle, the entire cycle was discarded. This gave a "balanced" dataset including an equal number of trials from both classes within each cycle.

We trained our classifiers on 80% of this balanced data, and tested on the remaining 20% and all data discarded in the cycle balancing procedure (the 80/20 split ensured that

force us to exclude even more data.

Classifier	All	Correct	Incorrect
Cycle Number AUC	**0.817 [0.793 0.841] t = 26.72, p < 0.001**	**0.823 [0.795 0.852] t = 22.84, p < 0.001**	**0.591 [0.568 0.613] t = 8.22, p < 0.001**
FRN AUC Corr	$r = 0.072$, $p = 0.636^{\dagger}$	$r = 0.035$, $p = 0.818^{\dagger}$	$r = 0.058$, $p = 0.707$
FMT AUC Corr	$r = 0.069$, $p = 0.653$	$r = 0.099$, $p = 0.516^{\dagger}$	$r = 0.186$, $p = 0.221$
LDA AUC Corr (T)	$\mathbf{r = 0.42}$, $\mathbf{p = 0.004}^{\dagger}$	$\mathbf{r = 0.40}$, $\mathbf{p = 0.007}^{\dagger}$	$r = 0.22$, $p = 0.144$
SVM AUC Corr (T)	$r = 0.29$, $p = 0.050^{\dagger}$	$r = 0.22$, $p = 0.140^{\dagger}$	$r = 0.16$, $p = 0.302$
LDA AUC Corr (T-F)	$\mathbf{r = 0.80}$, $\mathbf{p < 0.001}^{\dagger}$	$\mathbf{r = 0.74}$, $\mathbf{p < 0.001}\dagger$	$r = 0.17$, $p = 0.267$
SVM AUC Corr (T-F)	$\mathbf{r = 0.72}$, $\mathbf{p < 0.001}^{\dagger}$	$\mathbf{r = 0.73}$, $\mathbf{p < 0.001}^{\dagger}$	$r = 0.04$, $p = 0.809$

Table 5

Assessing the class imbalance confound for predicting subsequent accuracy (see illustrative plots in the Appendix, Figures A3–A5). First row: mean AUC values [95% confidence intervals], and results from t-tests against chance (df = 44), for tracing ROC curves only using cycle number for each subsequent accuracy classification problem. Columns: whether all trials in cycle N were considered, or only correct/incorrect trials. Other rows: the results of correlating (df = 43) cycle number-derived AUC values classifier-derived AUC values across participants. Classifiers included univariate measures (FMT or the FRN) and two multivariate classifiers (SVM or LDA) informed by either time (T) or time-frequency (T-F) domain data. Boldface indicates statistical significance, $p < 0.05$. † indicates that the original classifier AUC values were above chance (0.5), $p < 0.05$.

trials from both classes were present in the test set). Since trial counts dropped substantially (Table 6), we excluded any participant whose training set did not include at least 20 observations of each class. Finally, we repeated the 80-20 split 5 times and averaged classifier AUC values across iterations. Note that we restricted this analysis to our multivariate classifiers, since the confound did not appear to influence our univariate results (i.e., FRN and FMT AUC values were unrelated to cycle imbalance).

Results (Table 7 and Figure 12) found that AUC dropped substantially, unsurprising given the reduction in data analyzed, but remained above chance when predicting subsequent accuracy from all trials, regardless of the features or classifier used. For correct trials, only LDA informed by time domain data retained above-chance performance. As before, our classifiers could not make predictions following errors.

The time domain feature maps for classifiers trained on all and correct trials (Figure A6 and Figure A7, respectively) both before and after balancing the training trials, emphasized some overlapping features, including the second time bin and distal electrodes throughout the feedback processing window. For classifiers trained on time-frequency data from all trials (Figure A8), the feature maps again showed some similarities, where both

Trials	All	Correct	Incorrect
Before	718.38	584.44	133.93
After	216.36	98.76	97.82

Table 6

Average trial counts before and after balancing trials within each cycle for subsequent accuracy classification problems.

Classifier	All	Correct	Incorrect
LDA (T)	**0.529 [0.507 0.550]** **$t(43) = 2.71$, $p = 0.001$**	**0.533 [0.512 0.555]** **$t(39) = 3.11$, $p = 0.003$**	0.497 [0.477 0.516] $t(42) = -0.32$, $p = 0.749$
SVM (T)	**0.527 [0.508 0.546]** **$t(43) = 2.91$, $p = 0.006$**	0.520 [0.500 0.541] $t(39) = 2.00$, $p = 0.052$	0.497 [0.481 0.513] $t(42) = -0.38$, $p = 0.704$
LDA (T-F)	**0.529 [0.509 0.549]** **$t(43) = 2.94$, $p = 0.005$**	0.513 [0.484 0.541] $t(39) = 0.91$, $p = 0.368$	0.506 [0.485 0.527] $t(42) = 0.61$, $p = 0.548$
SVM (T-F)	**0.523 [0.507 0.539]** **$t(43) = 2.91$, $p = 0.006$**	0.515 [0.492 0.538] $t(39) = 1.29$, $p = 0.204$	0.503 [0.485 0.520] $t(42) = 0.31$, $p = 0.757$

Table 7

Predicting subsequent accuracy with balanced trials. Mean AUC values [95% confidence intervals] and paired-samples t-tests for each subsequent accuracy classification problem, with training trials balanced within each cycle. Columns: whether all trials in cycle N were considered, or only correct/incorrect trials. Classifiers included SVM or LDA informed by either time (T) or time-frequency (T-F) domain data. Boldface indicates significantly different than chance (0.5), $p < 0.05$.

emphasized the beta band and distal electrodes.

Summary of checking the influence of class imbalance

Class imbalance changing with cycle number was established as a potential confound for interpreting classification-success, at least for the multivariate classifiers. In many cases, our classifiers performed best on participants who quickly achieved high accuracy during training (i.e., participants whose data were most skewed by cycle number). The confound had the greatest impact on our best-performing classifiers.

However, correlations between classifier success and the cycle confound were not present for all successful classifiers (e.g., the univariate FRN classifier). Further, when we explicitly prevented our classifiers from being informed by cycle number during training, several classifiers retained above-chance performance, even though the classifier success rate reduced, as expected due to the reduction in the number of training trials. Finally, the spatiotemporal characteristics of the classifying features have considerable overlap with and

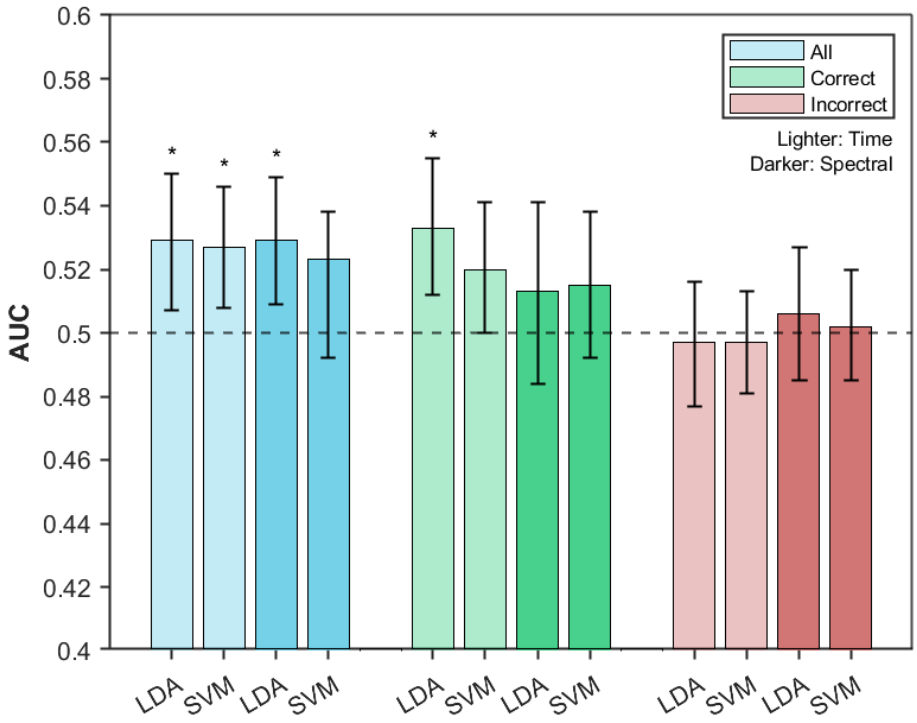

Figure 12

*Predicting subsequent accuracy after balancing trial counts within each cycle (compare with Figure 7a). Mean AUC values, with 95% confidence intervals. Blue, green, and red bars indicate classifiers trained on all, correct, and incorrect trials, respectively. Classifiers (indicated by brightness) included univariate measures (FMT or FRN; lightest), and two multivariate classifiers (SVM or LDA; medium) informed by either time or time-frequency domain data (darkest). * significantly different than chance (0.5), $p < 0.05$. The dashed line denotes chance AUC (0.5).*

without our approach to addressing class balance, which argues against the classification having been purely due to classifiers finding a shortcut and inferring trial number only. This suggests that our initial classifiers likely captured some genuine learning-relevant brain activity.

Discussion

The feedback-related negativity (Bellebaum & Daum, 2008; Cohen et al., 2007; Walsh & Anderson, 2011) and frontal midline theta activity (Bernat et al., 2015; Hajihosseini & Holroyd, 2013; van de Vijver et al., 2011) are widely reported neural signals associated with feedback processing. Moving beyond prior descriptive analyses, we explicitly evaluated the *predictive* power of the two signals, determining whether their purported roles at the trial-average level would generalize to predicting feedback-driven learning at the single-trial level. We then leveraged multivariate classifiers to ask whether the single-trial

EEG contains additional learning-relevant information beyond either signal.

First, we established that both the FRN and FMT were more pronounced for the average subsequently correct than subsequently incorrect trial, as expected (though the subsequently correct condition was characterized by broadband power increases beyond FMT). Then, analyzing the predictive power of the signals at the trial level, we found that the FRN and FMT could significantly predict subsequent accuracy, though only during feedback following correct responses (with FMT AUC values only marginally above chance). Despite these findings, both signals were less pronounced when a word was followed by a series of consecutive correct responses ("acquired" trials) at both the trial-average and single-trial levels. Feature weights for multivariate classifiers suggested that signals overlapping with the FRN, but not FMT, were relevant for discriminating between trials. Our results validate the FRN as a predictive signal, but are less clear about FMT.

In a parallel analysis of the same dataset, Chakravarty et al. (in review) aimed to study the FRN at points where a meaningful amount of learning occurred, studying the first training cycle and the participant-specific "steepest" cycle which preceded their largest increase in accuracy. The FRN indexed feedback magnitude, as expected, but only differentiated subsequently correct and incorrect trials when restricted to correct (current) trials only. Likewise, at the trial level, the FRN could only *predict* subsequent accuracy following correct responses. To boost power, we expanded our current analysis to the full set of training cycles, with the trade-off of including many trials late in the task which were unlikely to index learning. Despite these changes, we found the same result— that the FRN could only predict subsequent accuracy following relatively more rewarding feedback, demonstrating the robustness of this finding.

The leading account of the FRN suggests that it indexes an RPE calculation following *errors*, inheriting an assumption from much of the reinforcement learning literature that learning primarily occurs following errors (Sutton & Barto, 2018). By extension, if the FRN indexes RPE, it should predict learning outcomes primarily following errors or when reward is less than expected (Holroyd & Coles, 2002). However, our results alongside Chakravarty et al. (in review) do not clearly support such an account, at least for this task, because the FRN was only able to predict learning following relatively *rewarding* feedback, namely, correct-response trials but not errors. One possible explanation for this difference may be related to our task having a high learning demand. In this context, perhaps feedback stating that the just-given response was incorrect is harder to integrate (for example, requiring the participant remember both the response they made *and* that it was wrong) than feedback confirming that the response was correct. Also, because of the challenging nature of word-value learning, many correct responses may have been "exploratory" or "hypothesis-testing" decisions by the participant, so that the role of the feedback may accordingly function differently with respect to learning than in other tasks in which the participant has a more confident mental model of the task. However, consider that the way in which feedback was given during the task may have influenced our findings: participants could not make *absolute* errors, only relative ones, as incorrect responses still gave a 1-point reward. As such, we cannot explicitly comment on whether participants did or did not learn from errors in our task. Otherwise, relative classifier performance in either condition may have simply been a product of trial count: after epoch rejection, participants averaged 630.5 correct trials and 135.5 incorrect trials during the 16 training cycles.

Beyond the negative-going deflection thought to characterize the FRN, the FRN time window also includes a positive deflection sensitive to rewarding feedback (Holroyd et al., 2008). Importantly, Chakravarty et al. (in review) concluded that the apparent sensitivity of the FRN to value acquisition may have been largely driven by the magnitude of this reward positivity. Visual inspection of our ERP plots (Figure 5) suggests the same, where the most prominent differences between conditions appear limited to the positive deflections early in the FRN time window, rather than the negative deflections later on. Our results alongside those from Chakravarty et al. (in review) suggest that the negative-going deflection thought to characterize the FRN may not drive feedback-guided learning, and that signal sensitive to feedback magnitude (i.e., the reward positivity) may be more relevant for behavioural outcomes, converging with Krigolson (2017) and with individual sensitivity to reward covarying with voltage during reward trials more than non-reward trials (Proudfit, 2015). In this view, the negativity is produced by a N2 peak that is relatively invariant in response to feedback, whereas the reward positivity reflects positive reward processing, so that gain-reward trials are the ones that evoke meaningful reward-related activity during the FRN/RewP time window. This could be why using the EEG signal, especially the FRN measure, succeeded in predicting trial-level learning on correct-feedback trials but not error-feedback trials.

Other explorations of the FRN in similar reward-driven verbal learning paradigms have produced mixed results regarding its status as an RPE detector. Chase et al. (2011) supported the FRN as an RPE-sensitive signal, since it was most pronounced when word-value associations were unexpectedly violated. However, inconsistent with this, Arbel et al. (2014) reported that the FRN did not dampen when feedback was presented for already learned word-value associations (as expected if the FRN is sensitive to reward probability), and Ernst and Steinhauser (2012) found that FRN did not separate error trials based on whether the error was corrected later in the task.

More generally, the literature linking the FRN (and FMT) to feedback-driven learning is centered on tasks without a clear and substantial learning objective. These include probabilistic reward learning tasks, where participants learn to choose the stimulus with the greatest reward probability (Bernat et al., 2015; Cohen et al., 2007; Marco-Pallares et al., 2008), and time estimation tasks, where feedback scales with task performance (P. Li et al., 2016; Luu et al., 2004; van de Vijver et al., 2011). Accordingly, our work adds to a growing literature suggesting that the FRN and FMT may not generalize as learning indices in tasks with a more explicit learning target (Arbel et al., 2014; Chakravarty et al., in review; Chase et al., 2011; Ernst & Steinhauser, 2012), suggesting that they may be best understood as markers of feedback processing itself, rather than signals which explicitly guide goal-oriented future behaviour.

Following our univariate analyses, we expected that classifiers informed by multivariate brain activity could predict learning outcomes more successfully. Accordingly, our multivariate classifiers predicted subsequently correct and acquired trials better than the FRN and FMT, despite our intentional steps to limit overfitting and researcher degrees of freedom, including pre-defining classifier features and sticking closely to the methods used by Chakravarty et al. (2020) in a standard verbal recognition task. While our methods were parsimonious, above-chance AUC values approached 0.7 in the best-case scenarios—well within the range of values reported by others predicting memory outcomes from EEG

activity using linear techniques (Arora et al., 2018; Y. Li et al., 2024; Noh et al., 2014).

However, one challenge had to be addressed. Trial counts were skewed by cycle number, where participants were more accurate as the task progressed, leading to more subsequently correct and unchanged trials in later cycles. As such, we wondered whether our classifiers were taking advantage of the relationship between cycle number and trial labels to simply make predictions based on inferred task progress, instead of being informed strictly by learning-related brain activity. This confound appeared to influence classifier success, as classifiers performed best on data from subjects who learned more quickly in the task. After we balanced the numbers of each trial type within each cycle, explicitly preventing classifiers from accessing information related to task progress, AUC values dropped substantially, indicating that the original classification AUCs may have been inflated by the trial-imbalance confound, but this drop is also expected due to the reduced number of trials going into the analyses. However, many AUC values remained above chance, despite the greatly reduced trial count (and by extension noisier classification problem). Crucially, the feature maps before and after this procedure were similar for the same classification problems, echoing findings reported by Chakravarty et al. (2020) using SMOTE, that although class imbalance could theoretically pose a major problem, it seems not to have a large effect on classifier-based analyses of EEG. This suggests our initial classifiers must have been informed by genuine learning-relevant brain activity. Our approach to trial balancing within learning cycles could be used by researchers studying similar multi-trial paradigms to evaluate the robustness of their classifier success.

Of our successful classifiers, those predicting from time domain data emphasized distal electrodes throughout the feedback processing window, and all electrodes 301–600 ms post-feedback. Interestingly, this is the time range where one of the most robustly reported subsequent-memory effects is reported, typically in episodic item-recognition tasks (Karis et al., 1984). Known as the late positive component or even possibly a P300, this activity has been linked to shallow levels of processing during the study phase (Fabiani et al., 1986; Karis et al., 1984), which may relate to its role here. Meanwhile, successful time-frequency domain classifiers emphasized beta and delta activity, along with an array of non-central electrodes and time bins throughout the feedback processing window. In both cases, features beyond the FRN and FMT windows were relevant for making classifier decisions, suggesting that broad, multivariate EEG activity is necessary to describe the full scope of feedback processing.

Finally, as indicated in the introduction, we were curious about why published classifier analyses typically use spectrographic (power in the time-frequency domain) features and rarely the original time-domain signal (which is the natural starting point from the perspective of event-related potentials). Broadly speaking, classifications with more features produced better performance (univariate $<$ time-domain $<$ spectrographic domain). So it could very well simply be that inputting more features monotonically increases the amount of behaviourally relevant information available to the classifier, an idea to be tested in the future. However, the degree to which spectrographic features outperformed the time domain is still somewhat surprising. Because Fourier and wavelet transforms are linear—values are weighted sums of the original time-domain values (voltages)— LDA and SVM would both seem to be well suited to exploiting the same information whether the feature space is expressed in the time or time-frequency domain. In other words, at least from

a naïve perspective, the spectrogram should not contain any more information than the time-domain signal, and arguably less, because the squaring operation to compute power values throws away information about the sign of the signal. But our classifiers did, in fact, succeed better with spectrographic than time-domain features. We can only speculate, but perhaps this is because wavelet power is relatively insensitive to phase, or precise-timing information. The spectrographic features may simply be more resilient to the jitter of activity features across trials. It seemed obvious to us to start with time-domain features before moving on to spectrographic features, in part because of the intuitive argument we expressed about spectrographic features being composed of the original time-domain features. While other groups may have had very justified independent reasons to skip to the time-frequency domain, we wonder if some of the missing time-domain classifiers are in researchers' "file-drawers." Perhaps they were less robust or not clearly above chance, and this led some researchers to leave behind the time domain and move on to the more robust time-frequency domain.

A few limitations should be noted. First, our feedback stimulus indicating a correct choice response was larger than that indicating an incorrect choice. Pfabigan, Sailer, and Lamm (2015) showed that larger stimuli evoke larger FRNs. With this in mind, we constructed our comparisons and classifications such that the stimulus displayed at the time that brain activity was analyzed was always the same. Hence, for example, correct-feedback processing was compared with correct-feedback processing depending on the accuracy of the subsequent trial with a given item (and likewise for incorrect). This concern does not present a confound for the results we present here.

Second, a stronger test of the predictive information contained within the feedback-processing EEG signal would be to keep a held-out set of data to apply a trained classifier only at the very final stage. This is particularly important when researchers casually explore the hyperparameters of the classifiers, as compellingly demonstrated by Skocik et al. (2016). However, importantly, we are not optimizing hyperparameters. Moreover, cross-validation is within-subject. In within-subjects applications of classifiers in cognitive neuroscience, held-out data is quite rare. The reason is largely due to the limited number of trials per subject (compared to between-subjects classification problems, for example). Our paradigm produces more trials than most memory EEG studies, and even so, we are up against lower limits for trial counts which become quite pronounced in some cases. That said, with a creative approach to task design, potentially many experimental sessions per participant (although that may also introduce more challenging variability), it may be feasible to hold-out data at the single-subject level in future studies. For the work reported here, we have minimized the possibility of overfitting at the hyperparameter level by sticking closely to our previous hyperparameters and avoiding exploring them.

Taken together, our classifiers successfully predicted feedback-driven learning in a cognitively demanding trial-and-error learning task. The most successful classifiers were informed by multivariate brain activity extending beyond the feedback-related negativity (which held some predictive power, likely driven by an overlapping reward positivity) and the frontal midline theta (which held less), suggesting that the bulk of feedback-relevant brain activity extends beyond established univariate signals. In this regard, the FRN, and particularly FMT, may be best characterized as markers of feedback processing itself, rather than integrative signals which explicitly guide future learning.

Data and Code Availability. Data and analysis code are available from the authors upon reasonable request.

Author Contributions. Each author contributed substantially to all aspects of the research and the manuscript.

Declaration of Competing Interests. The authors declare no competing interests.

References

Arbel, Y., Murphy, A., & Donchin, E. (2014). On the utility of positive and negative feedback in a paired-associate learning task. *Journal of Cognitive Neuroscience*, *26*(7), 1445–1453.

Arbel, Y., Goforth, K., & Donchin, E. (2013). The good, the bad, or the useful? the examination of the relationship between the feedback-related negativity (frn) and long-term learning outcomes. *Journal of Cognitive Neuroscience*, *25*(8), 1249–1260.

Arora, A., Lin, J., Gasperian, A., Maldjian, J., Stein, J., Kahana, M., & Lega, B. (2018). Comparison of logistic regression, support vector machines, and deep learning classifiers for predicting memory encoding success using human intracranial EEG recordings. *Journal of Neural Engineering*, *15*(6).

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience*, *27*(7), 1823–1835.

Bernat, E. M., Nelson, L. D., & Baskin-Sommers, A. R. (2015). Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, *52*, 626–637.

Cavanagh, J. F., Frank, M. J., Klein, T. J., & Allen, J. J. B. (2010). Frontal theta links prediction errors to behavioural adaptation in reinforcement learning. *NeuroImage*, *112*, 341–352.

Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology*, *20*, 544–547.

Chakravarty, S., Chen, Y. Y., & Caplan, J. B. (2020). Predicting memory from study-related brain activity. *Journal of Neurophysiology*, *124*(6), 2060–2075.

Chakravarty, S., Fujiwara, E., Madan, C. R., Tomlinson, S. E., Ober, I., & Caplan, J. B. (2019). Value bias of verbal memory. *Journal of Memory and Language*, *107*, 25–39.

Chakravarty, S., Ober, I., Madan, C. R., Chen, Y. Y., Fujiwara, E., & Caplan, J. B. (in review). The feedback-related negativity and reward prediction error in trial-and-error learning of many stimuli.

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioural adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*(4), 936–946.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Y. Y., & Caplan, J. B. (2017). Rhythmic activity and individual variability in recognition memory: Theta oscillations correlate with performance whereas alpha oscillations correlate with ERPs. *Journal of Cognitive Neuroscience*, *29*(1), 183–202.

Christie, G. J., & Tata, M. S. (2009). Right frontal cortex generates reward-related theta-band oscillatory activity. *NeuroImage*, *48*, 415–422.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and eeg spectra. *NeuroImage*, *35*(7), 968–978.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Delorme, A., & Makeig, B. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, *134*, 9–21.

Ernst, B., & Steinhauser, M. (2012). Feedback-related brain activity predicts learning from feedback in multiple-choice testing. *Cognitive, Affective, and Behavioural Neuroscience*, *12*, 323–336.

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and recall in an incidental memory paradigm. *Psychophysiology*, *23*(3), 298–308.

Fell, J., Ludowig, E., Staresina, B. P., Wagner, T., Kranz, T., & Elger, C. E. e. a. (2011). Medial temporal theta/alpha power enhancement precedes successful memory encoding: Evidence based on intracranial EEG. *Journal of Cognitive Neuroscience*, *31*, 5392–5397.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, *47*, 495–501.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system of error detection and compensation. *Psychological Science*, *4*, 385–390.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*, 2279–2282.

Goyer, J. P., Woldorff, M. G., & Huettel, S. A. (2008). Rapid electrophysiological brain responses are influenced by both valence and magnitude of monetary rewards. *Journal of Cognitive Neuroscience*, *20*, 2058–2069.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912.

Hajihosseini, A., & Holroyd, C. B. (2013). Frontal midline theta and n200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology*, *50*, 550–562.

Halpern, D. J., Tubridy, S., Davachi, L., & Gureckis, T. M. (2023). Identifying causal subsequent memory effects. *Proceedings of the National Academy of Sciences, USA*, *120*(13), e2120288120.

Hester, R., Barre, N., Murphy, K., Silk, T. J., & Mattingley, J. B. (2008). Human medial frontal cortex activity predicts learning from errors. *Cerebral Cortex*, *18*(8), 1933–1940.

Höhne, M., Jahanbekam, A., Bauckhage, C., Axmacher, N., & Fell, J. (2016). Prediction of successful memory encoding based on single-trial rhinal and hippocampal phase information. *Neuroimage*, *139*, 127–135.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., & Coles, M. G. (2004). Dorsal anterior cingulate cortex shows fmri response to internal and external error signals. *Nature Neuroscience*, *7*, 497–498.

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*, 688–697.

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the von Restorff effect. *Cognitive Psychology*, *16*(2), 177–216.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A systematic review and analysis. *Brain Research Reviews*, *29*, 169–195.

Klimesch, W., Freunberger, R., & Sauseng, P. (2010). Oscillatory mechanisms of process binding in memory. *Neuroscience and Biobehavioural Reviews*, *34*, 1002–1014.

Krigolson, O. E. (2017). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology*, *132*(Part B), 175–183.

Li, P., Baker, T. E., Warren, C., & Li, H. (2016). Oscillatory profiles of positive, negative and neutral feedback stimuli during adaptive decision making. *International Journal of Psychophysiology*, *107*, 37–43.

Li, Y., Pazdera, J. K., & Kahana, M. J. (2024). EEG decoders track memory dynamics. *Nature Communications*, *15*, 2981.

Luft, C. D. B., Takase, E., & Bhattacharya, J. (2014). Processing graded feedback: Electrophysiological correlates of learning from small and large errors. *Journal of Cognitive Neuroscience*, *26*(5), 1180–1193.

Luu, P., Tucker, D. M., & Makeig, S. (2004). Frontal midline theta and the error-related negativity: Neurophysiological mechanisms of action regulation. *Clincal Neurophysiology*, *115*, 1821–1835.

Marco-Pallares, J., Cucurell, D., Cunillera, T., Garcia, R., Andres-Pueyo, A., Munte, T. F., & Rodriguez-Fornells, A. (2008). Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia*, *46*, 241–248.

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*(6), 852–860.

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory data analysis with MATLAB*. Chapman; Hall/CRC.

Mas-Herrero, E., & Marco-Pallares, J. (2014). Frontal theta oscillatory activity is a common mechanism for the computation of unexpected outcomes and learning rate. *Journal of Cognitive Neuroscience*, *26*(3), 447–458.

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798.

Mirjalili, S., Powell, P., Strunk, J., James, T., & Duarte, A. (2022). Evaluation of classification approaches for distinguishing brain states predictive of episodic memory performance from electroencephalography abbreviated title: Evaluating methods of classifying memory states from EEG. *NeuroImage*, *247*(118851).

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G., Holroyd, C. B., & Kok, A. (2002). A computational account of altered error processing in older age: Dopamine and the error-related negativity. *Cognitive Affective & Behavioural Neuroscience*, *2*, 19–36.

Nieuwenhuis, S., Slagter, H. A., Von Geusau, N. J. A., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: Differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168.

Noh, E., Herzmann, G., Curran, T., & de Sa, V. R. (2014). Using single-trial EEG to predict and analyze subsequent memory. *Neuroimage*, *84*, 712–723.

Pfabigan, D. M., Sailer, U., & Lamm, C. (2015). Size does matter! perceptual stimulus properties affect event-related potentials during feedback processing. *Psychophysiology*, *52*(9), 1238–1247.

Pfabigan, D. M., Seidel, E.-M., Paul, K., Grahl, A., Sailer, U., Lanzenberger, R., Windischberger, C., & Lamm, C. (2015). Context-sensitivity of the feedback-related negativity for zero-value feedback outcomes. *Biological Psychology*, *104*, 184–192.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459.

Sanquist, T. H., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology*, *17*(6), 568–576.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H., & Wyble, B. (2016). *I tried a bunch of things: The dangers of unexpected overfitting in classification* [Preprint on bioRxiv].

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction (2nd edition)*. MIT Press.

van de Vijver, I., Ridderinkhof, K. R., & Cohen, M. X. (2011). Frontal oscillatory dynamics predict feedback learning and action adjustment. *Journal of Cognitive Neuroscience*, *23*(12), 4106–4121.

van der Helden, J., Boksem, M. A., & Blom, J. H. (2010). The importance of failure: Feedback-related negativity predicts motor learning efficiency. *Cerebral Cortex*, *20*(7), 1596–1603.

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., Rosen, B. R., & Buckner, R. L. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science*, *281*, 1188–1191.

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences*, *108*(47), 19048–19053.

Watanabe, T., Hirose, S., Wada, H., Katsura, M., Chikazoe, J., Jimura, K., Imai, Y., Machida, T., Shirouzu, I., Miyashita, Y., & Konishi, S. (2011). Prediction of subsequent recognition performance using brain activity in the medial temporal lobe. *NeuroImage*, *54*, 3085–3092.

Weidemann, C. T., & Kahana, M. J. (2021). Neural measures of subsequent memory reflect endogenous variability in cognitive function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Whitten, T. A., Hughes, A. M., Dickson, C. T., & Caplan, J. B. (2011). A better oscillation detection method robustly extracts EEG rhythms across brain states: The human alpha rhythm as a test case. *NeuroImage*, *54*(2), 860–874.

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research*, *1286*, 114–122.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, *111*, 931–959.

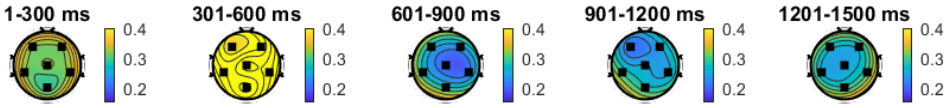

Figure A1

Predicting item value acquisition from correct trials: LDA time domain feature weights. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

Appendix

In this Appendix we present additional topographic plots and additional analyses of the effects of cycle imbalance.

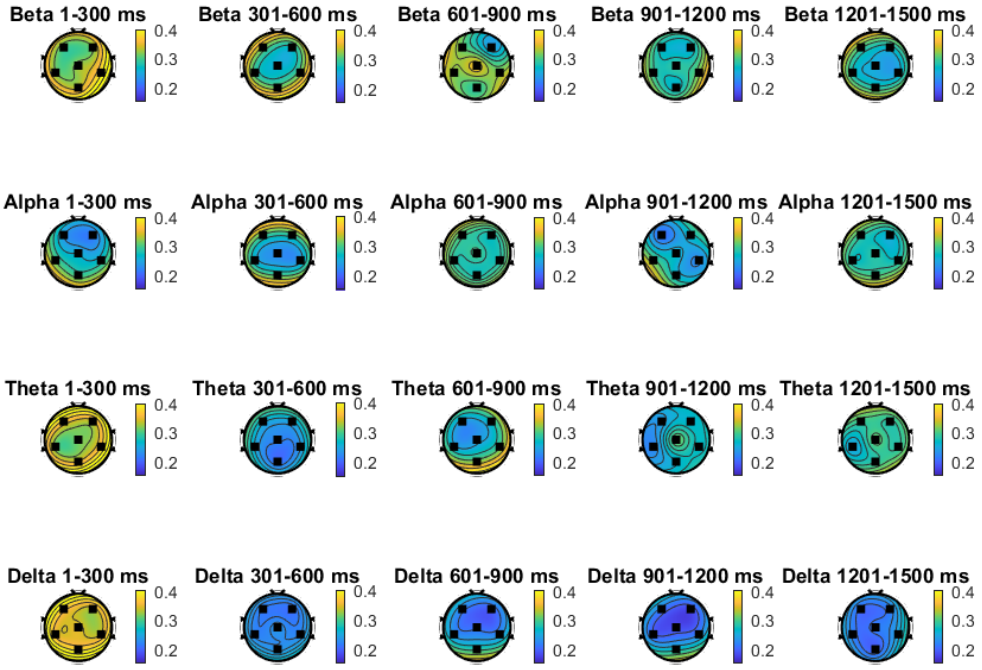

Figure A2

Predicting item value acquisition from correct trials: LDA time-frequency feature weights. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

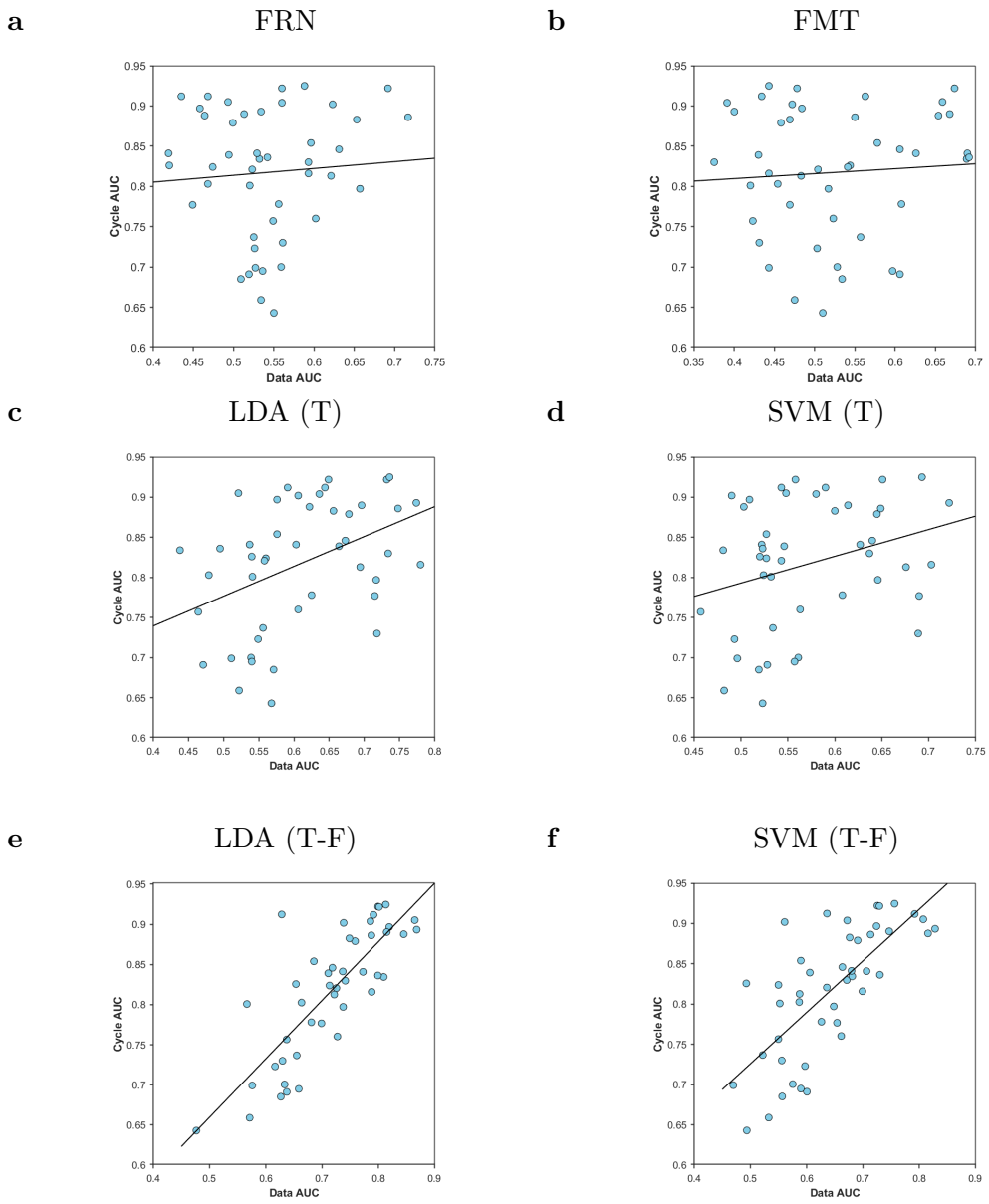

Figure A3

Cycle imbalance confound: predicting subsequent accuracy from all trials. Scatter plots show the relationship between AUC values derived from predicting subsequently correct trials from EEG data and cycle number. Each point represents one participant. See Table 5 for r and p values.

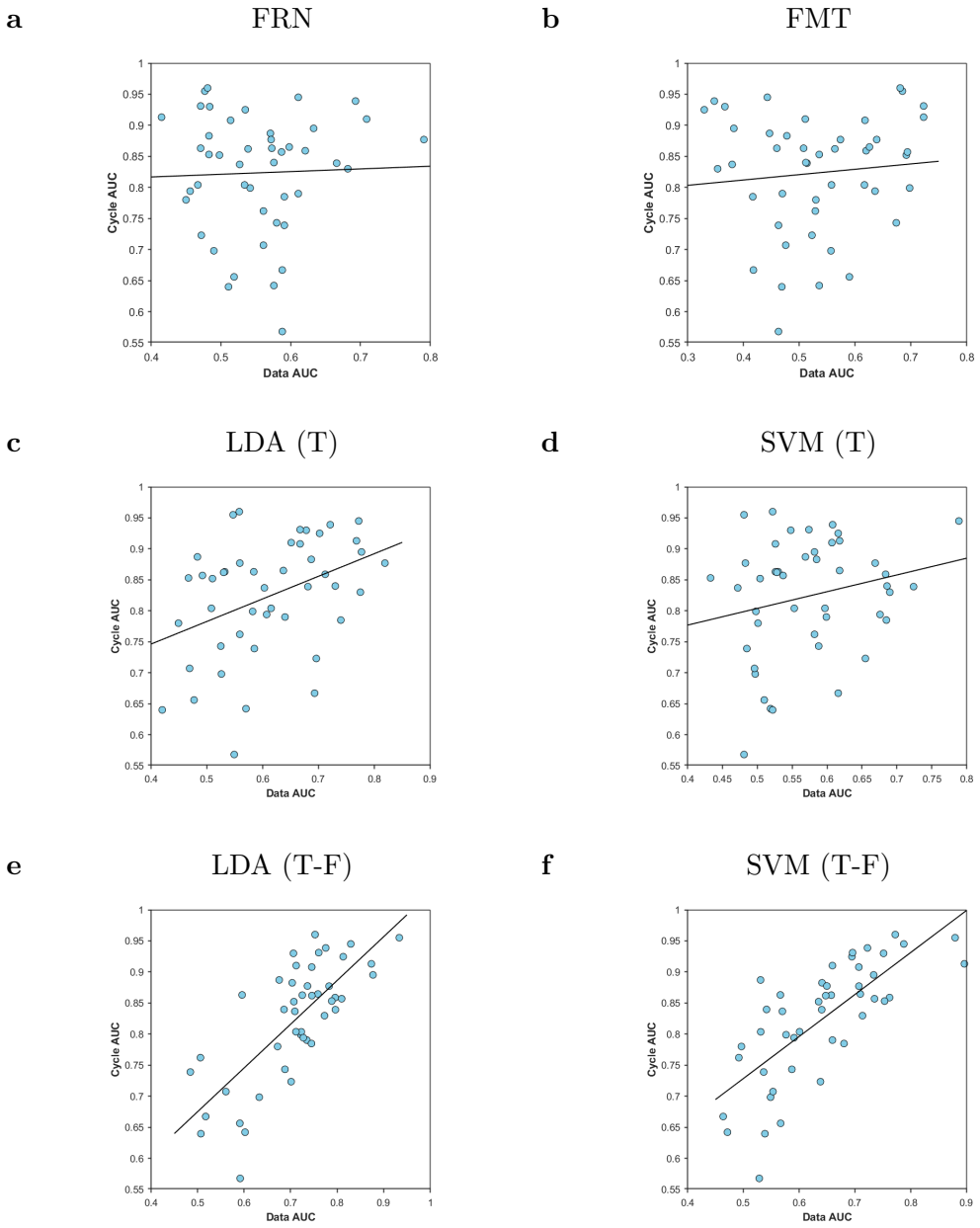

Figure A4

Cycle imbalance confound: predicting subsequent accuracy from correct trials. Scatter plots show the relationship between AUC values derived from predicting subsequently correct trials from EEG data and cycle number. Each point represents one participant. See Table 5 for r and p values.

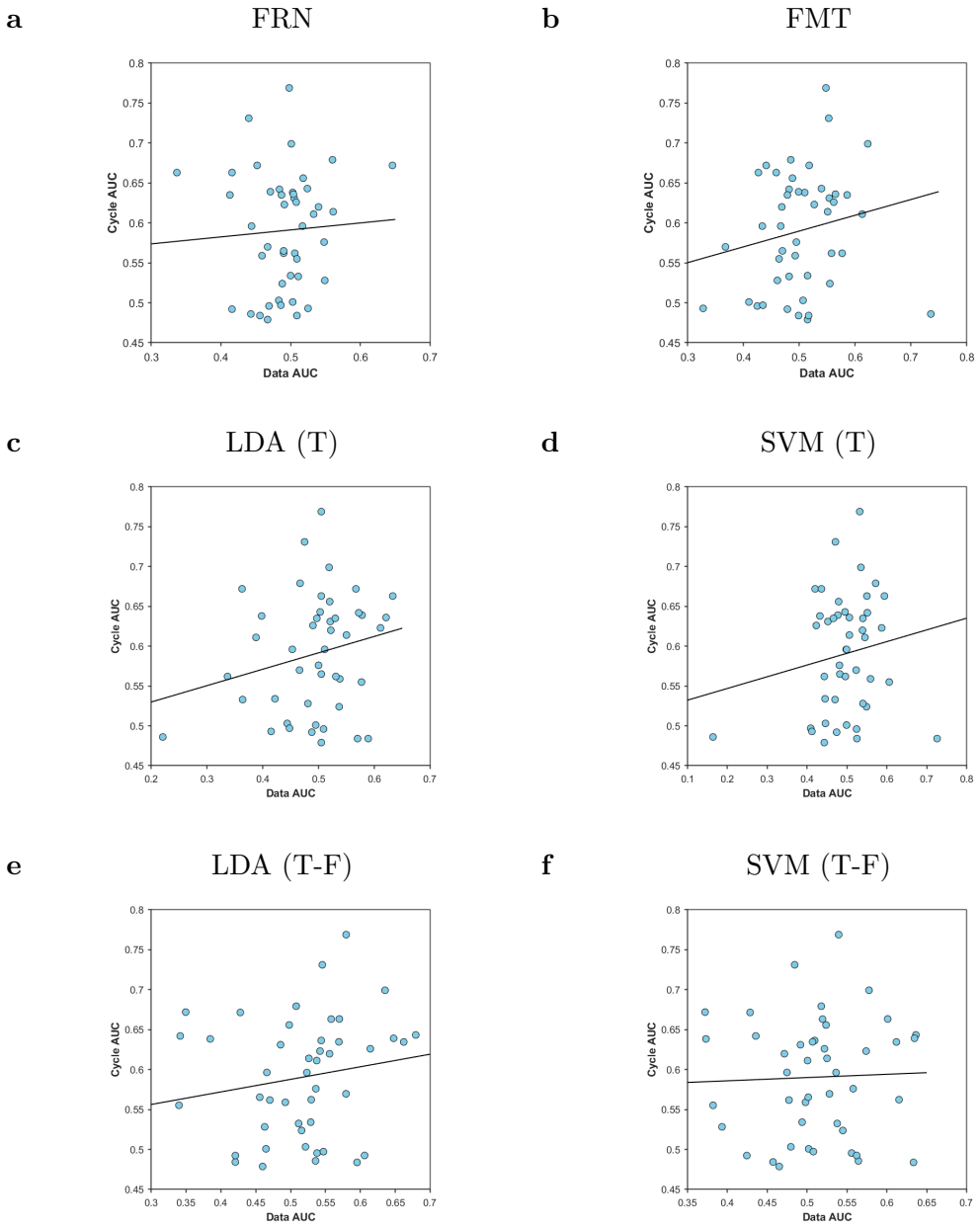

Figure A5

Cycle imbalance confound: predicting subsequent accuracy from incorrect trials. Scatter plots show the relationship between AUC values derived from predicting subsequently correct trials from EEG data and cycle number. Each point represents one participant. See Table 5 for r and p values.

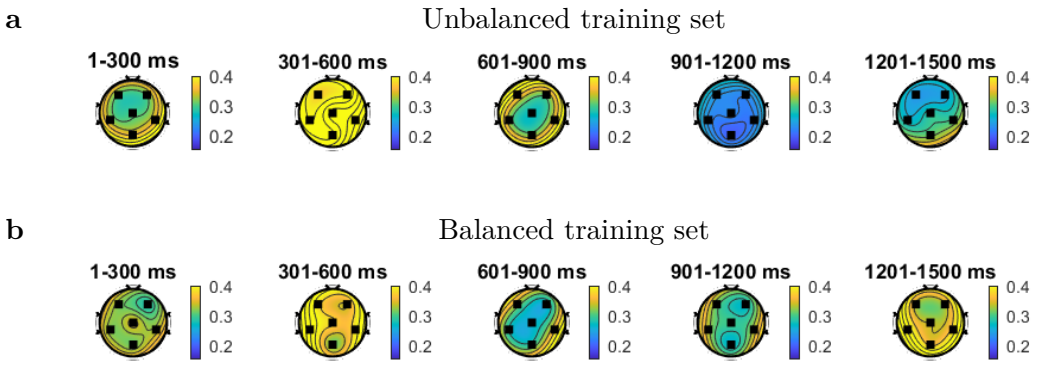

Figure A6

Predicting subsequent accuracy from all trials: LDA time domain feature weights before and after balancing training trials within each cycle. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

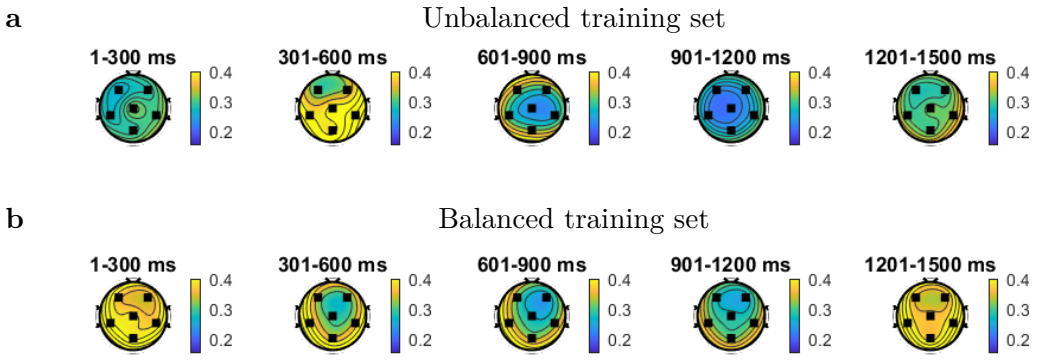

Figure A7

Predicting subsequent accuracy from correct trials: LDA time domain feature weights before and after balancing training trials within each cycle. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.

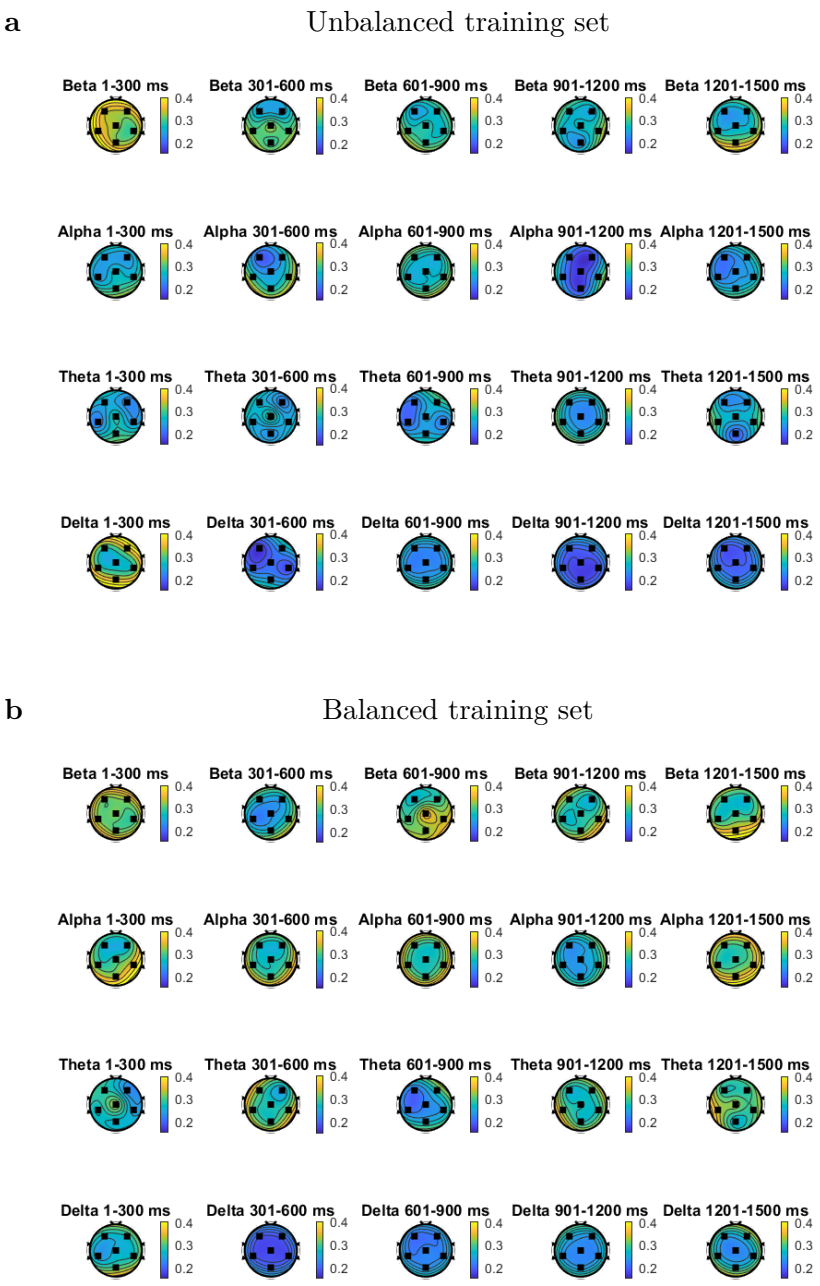

Figure A8

Predicting subsequent accuracy from all trials: LDA time-frequency domain feature weights, before and after balancing training trials within each cycle. Topographies show MATLAB's DeltaPredictor value, indicating the degree to which a feature influenced classification, at various scalp sites. Plots are interpolated using a spline method.