Title: Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning.

Authors: Anne Gabrielle Eva Collins, Michael Joshua Frank

Appendix

Supplemental Behavioral analysis

Logistic regression analysis of performance:

We analyzed performance in phase A and B using logistic regression. Regressors were #previous correct presentations of the current visual input pattern, delay (trials since the visual input pattern was last presented) and categorical variable TS1 vs. TS2. One subject with outlier regression weights, indicating a poor fit of the regression, was excluded from the group analysis.

In phase A, there was a significant effect of number of correct presentations ($t(31) = 8.3$, $p < 10\text{-}4$) and delay ($t(31) = -5.03$, $p < 10\text{-}4$) on the probability of a correct choice, but not of TS1 vs. TS2 ($t(31) = 0.21$, ns), indicating that the observed difference in learning speed could be accounted for by the larger delays for TS1 given its lower frequency. The lack of benefit for TS1 over TS2 during initial learning is thus expected given that it would take some time for subjects to discover that C0 and C1 corresponded to the same task-set (and moreover, that colors and not shapes constituted contexts that could be clustered), and hence subjects only show degraded performance due to larger delays. However, by the end of this learning phase, our model predicts that subjects should have discovered the structure of the tasks, and that C0 and C1 corresponded to the same TS, which should be revealed in subsequent presentations of novel stimuli in those contexts. In phase B, we again found a significant effect of number of correct presentations ($t(31) = 11.2$, $p < 10\text{-}4$) on the probability of correct choice, but now this was accompanied by a significant effect of TS1>TS2 ($t = 5.08$, $p < 10\text{-}4$,). Moreover the effect of delay was no longer significant ($t=0.9$, ns). Indeed, the effects of both delay and TS were significantly different between phase A and B ($t(31)=2.77$, $p = 0.009$; $t(31) = 3.32$, $p = 0.002$), whereas the effect of number of correct presentations did not ($t(31 = 0.35$, ns). This analysis confirms that, in phase B, subjects were able to treat C0 and C1 as equivalent from the outset to pool information about new stimuli across them and thus learn faster, which also means that delays between successive presentations of the same visual input pattern are less relevant.

To test more directly the sharing of information across C0 and C1, we performed an additional logistic regression within C0 and C1 trials, excluding C2 trials (Figure S1, right panel), to determine whether correct performance in one context was predictive of performance in the other context linking the same TS. Thus, when predicting the probability of a correct choice for a trial of context C_i and stimulus S_j, regressors were #previous correct trials for C_iS_j, delay, and also #previous correct trials for the current stimulus in the other context ($C_{1\text{-}i}S_j$).

As expected, there was a main effect of # previous correct trials in both phases ($t(30)>4.5$, $p < 10\text{-}4$). There was an effect of delay in phase A ($t(30)=-2.5$, $p = 0.016$) but not in phase B ($t(30)=-1.4$, ns), although here the difference was not significant ($t(30 = 0.67$, ns). Finally, there was a marginal effect of other context information in phase B ($t(30) = 1.94$, $p = 0.06$), but not in phase A ($t(30) = -1.66$, ns); critically this effect was significantly different between phases ($t(30)=2.42$, $p = 0.02$).

Behavioral analysis of EEG experiment.

EEG subjects performed two blocks of the same experiment, with non-overlapping colored-shapes. Here we report behavior for both blocks. To limit confounds due to potential knowledge of task structure, EEG analysis was only performed over first block trials.

We replicated all key behavioral findings from the behavioral experiment in the behavioral analysis of the EEG experiment.

Phase A +B

There was a significant effect in phase B in which participants showed enhanced performance for C0-C1 trials relative to C2 trials ($t = 2.16$, $p = 0.035$), indicating generalization in learning (Fig. S1 bottom).
The opposite deficit for C0-C1 trials in phase A was also observed similar to the behavioral experiment, but not significantly so ($t = -1.5$, $p = 0.13$). There was no interaction with block.

Logistic regression analysis of all phase A-B trials also replicated previous results noted above. There was a main effect of delay ($t = 2.6$, $p=0.01$) in phase A but no effect of C0-C1 vs C2 ($t = .45$, ns). In contrast, there was a strong effect of task-set in phase B ($t = 4.16$, $p<10\text{-}4$), that was significantly greater than in phase A ($t=2.7$, $p<0.01$). The delay effect did not differ between phases ($t=0.88$, ns).

Logistic regression analysis restricted to C0-C1 trials also confirmed previous results (see figure S1, right panel). There was a trend for an effect of delay ($t = -1.74$, $p =0.089$), which did not interact with phase ($t = 0.19$, ns). There was a significant effect of information from other context in phase B ($t = 2.79$, $p=0.007$), but not in phase A ($t=0.34$, ns), with a significant difference between phases ($t=2.61$, $p=0.01$).

Phase C
While transfer was not significant across both iterations of the task ($t = 0.06$), we focused on the first iteration in which subjects were naïve to the structure of the task. In the first block, we found significantly greater performance for C3 over C4 trials ($t=2.04$, $p=0.05$, Fig. S2 left).
We also replicated the results relating to the clustering action choice prior: the distribution of first action choices was significantly different from random across both blocks ($chi2 = 24$, $p<10\text{-}4$), as well as within each block (both $chi2>9.3$, $p<0.025$). TS1 action choice was significantly more likely than TS2 action choice across both blocks (binomial test; $p = 0.0012$), and within each block ($p<.05$, figure S2).

Modeling.

FRL model: details of decay and within dimension mechanisms.
We implement decay or forgetting in FRL such that for all color C, shape S, and action a, after each trial, Q estimates decay towards their initial value:

$$Q \leftarrow Q+f*(Q_0-Q).$$

Second, we implement bleed-over in dimension learning by updating not only the values of the specific C and S but also additionally updating other shapes of this trial's color or other colors of this trials shape: for all other S_i,

$$Q(C_t,S_i,a) \leftarrow Q(C_t,S_i,a) + \alpha_{dim} \times \delta;$$

and for all other C_i,

$$Q(C_i,S_t,a) \leftarrow Q(C_i,S_t,a) + \alpha_{dim} \times \delta.$$

$\alpha_{dim} < \alpha$ is a bleed-over learning rate parameter. This learning property accounts for observed low-level biases in action selection.

SRL model: details of decay and bias mechanisms.
To account for forgetting, we implement the same decay for all clusters as in FRL. To account for action selection biases, we implement additional noise in the action policy with a mixture of three terms: a softmax over $Q(Zc_t,Zs_t,.)$ with mixture weight $(1-\varepsilon)$, where Zc_t and Zs_t are the maximum a priori clusters, and a bias term with probability ε. The bias term is a mixture of two policies implementing low-level attentional biases on Color or Shape, with mixture weight ε_{CS}. They are implemented as a softmax over $mean_i(Q(Zc,Zs_i,:))$ and $mean_i(Q(Zc_i,Zs,:))$.

Model fitting was performed for both experiments and lead to similar results. Since we used the model fits for model-based analysis of the EEG data, we report fits for the first block of the EEG experiment.
We tested a class of non-structure reinforcement learning models, that included or excluded, in addition to classic RL mechanisms (softmax parameter β and learning rate α) the following features of the model described in main text:

- undirected noise (parameter ε)
- decay (parameter f)
- bias (parameter α_{dim})

Measure of fit penalized for complexity (using Akaike Information criterion - AIC) supported inclusion of all three mechanisms in the FRL model. Similar testing of multiple features' role was also performed for SRL.

In addition to AIC, we report the non-penalized measure of fit pseudo-r2: this is the log-likelihood of the observed data, normalized by log-likelihood of chance, such that a value of 0 corresponds to a chance model, whereas a value of 1 corresponds to perfect prediction.

Supplemental EEG results.

Figure S3C shows that while SRL fits subjects' behavior better than FRL, SPE and FPE share a large amount of variance. However, the amount of shared variance is not predictive of difference in Fit, ensuring that the amount by which FRL and SRL explain EEG variance, and the degree to which this predicts subjects' transfer performance cannot be explained away by worse model fit for some subjects.

Supplemental GLM effects

The results reported in the main text come from a GLM including regressors for FPE and SPE (orthogonalized against FPE, given that they are strongly correlated). It indicates additional variance explained by the SRL model within FPE-sensitive regions of interest. Here we also perform the complementary analysis, where the ROIs are obtained from a non-orthogonalized SPE regressor, and the additional effect of FRL PE is tested within those ROIS.
Results show first that very similar ROIs are obtained (compare Fig S4B and C, as well as grey and black lines in fig S4A), as expected from the high degree of shared variance between the regressors. Second, we found no additional effect of FPE over that of SPE in any of the 3 ROIs (early $t = 0.66$, medium $t = 0.8$, late $t = 1.4$, fig S5), though there was an effect if averaged across all 3 ROIs (overall: $p = 0.048$, $t = 2.07$). Moreover, only SPE was related to behavioral transfer, generalization and clustering, as reported in the main text.

Correct vs. Incorrect feedback

Our experiment used deterministic feedback. As such, and since learning occurred quickly, subjects experienced many more positive (correct) than negative (incorrect) feedback, and we did not have enough trials to perform the regression analysis of prediction error for incorrect trials (negative prediction errors). Nevertheless, for completeness, we show here feedback-locked ERPs for incorrect trials (FigS6), as well as the topography of effects at the three time points of interest.

To visualize part of the effects of the regression analysis, we also include ERPs of low and high FPE correct trials (grouped by median split)

Supplemental Figures

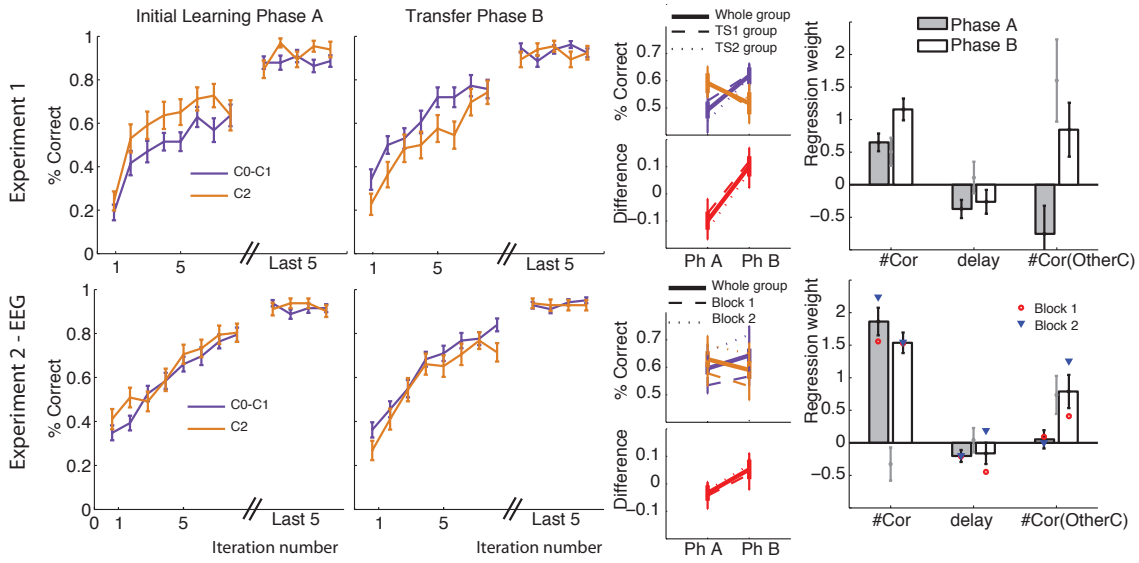


Figure S1

Figure S1: EEG Behavioral results for a priori transfer. Horizontal transfer. Left: learning curves for initial phase A and first transfer phase B. Within-cluster transfer is evident by faster learning of new S-A associations for C0/C1 than for C2 in phase B, despite slower initial learning (see text). Middle, top: summary measure over first 8 trials for each input pattern across phases shows an interaction between phase A and B on performance in TS1 vs TS2 stimuli. Middle, bottom: Difference in TS1 vs. TS2 performance increases significantly between phase A and B. Right: logistic regression weights within TS1 trials. Regressors are (#Cor): number of previous correct choices for a given stimulus and context, (delay): delay since last correct choice, and (#Cor OtherC) number of correct trials for the same stimulus in a different context corresponding to the same rule. Error bars are standard error.

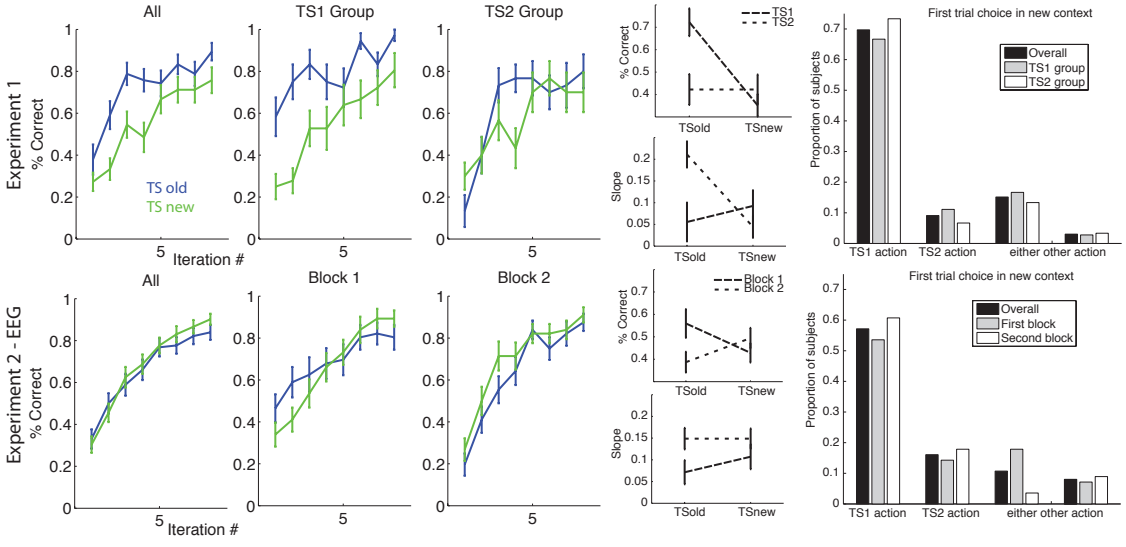


Figure S2

Figure S2: Transfer phase C. Top: Behavioral experiment. Bottom: EEG experiment.
Left: learning curves for transfer phase C. Overall learning is speeded for TS that were previously valid in old contexts. This effect is particularly evident for those subjects for which the old TS was the more popular TS1 (clustered across two contexts) than TS2. Middle panel: summary measure over first 3 trials

for each condition (TSold or TSnew). Top: average performance, bottom: slope (change in performance). Right panel: Action choice for very first trial in the transfer phase. Proportion of subjects who chose the action corresponding to the action prescribed by TS1 for that stimulus, for TS2, or for either of the other two actions. There is a very strong bias towards TS1, prior to any information concerning the new phase, despite the fact that this TS or action was no more frequent across trials. Error bars are standard error.

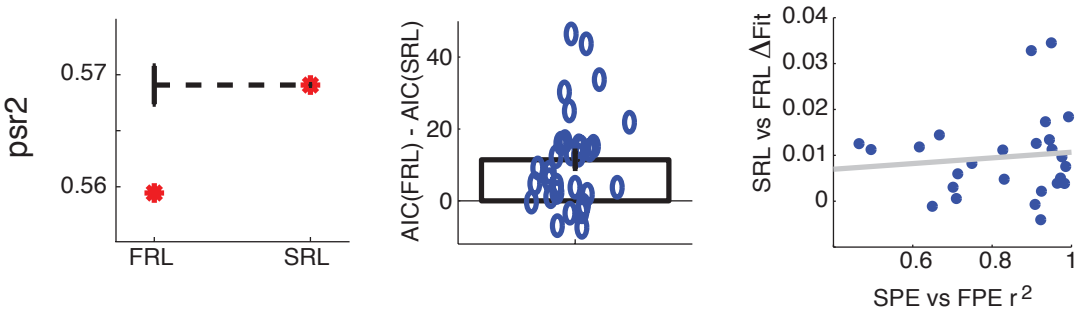


Figure S3

Figure S3: Computational modeling. **Top: Model fitting results.** Left: average pseudo-r2 for FRL and SRL model (red *), and error bars show within subject standard error. Middle: AIC was significantly lower for SRL than FRL (circles are individual subjects). Right: shared variance between SPE and FPE was high (>0.5 for all subjects), but not correlated with difference in fit.

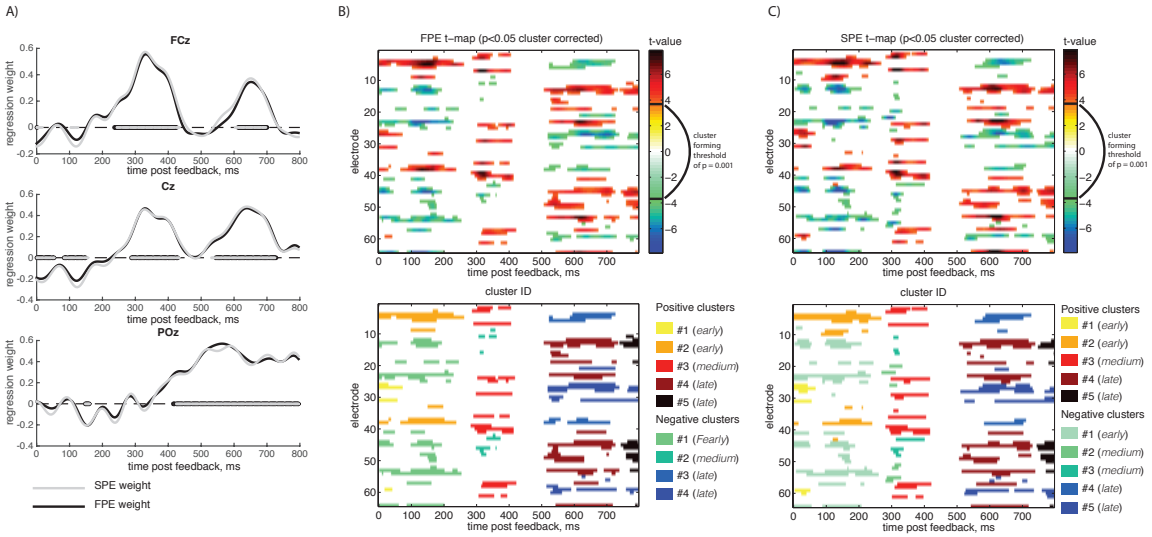


Figure S4: Comparison of FPE and SPE effects. A). Average regression weights of FPE regressor in main text GLM analysis (black). Average regression weights of SPE regressor in reversed GLM analysis (grey). Dots indicate p<.05 significance, cluster corrected. The similarity in the two patterns is due to the large amount of variance shared by FPE and SPE. Panels B and C show the same comparison, including all electrodes.

B): Clusters for EEG GLM analysis in main text, presented main Fig. 5. Top: t-values of effect of PE, thresholded for p<.05 significance, cluster corrected. Bottom: cluster id. ROI belonging was identified by temporal separation of the 3 groups of clusters.

C): Clusters for reversed GLM analysis. Top: t-values of effect of SPE, thresholded for p<.05 significance, cluster corrected. Bottom: cluster id. ROI belonging was identified by temporal separation of the 3 groups of clusters.

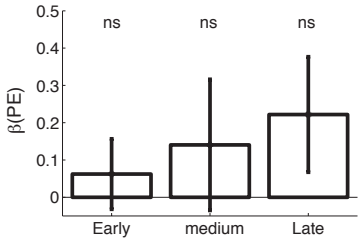


Figure S5

Figure S5: Additional effect of PE is SPE ROIs (compare to main text Fig. 6). There was no significant additional effect of PE in any of the three ROIs. Error bars are standard error of the mean. "ns" indicates a non-significant effect (p>.05).

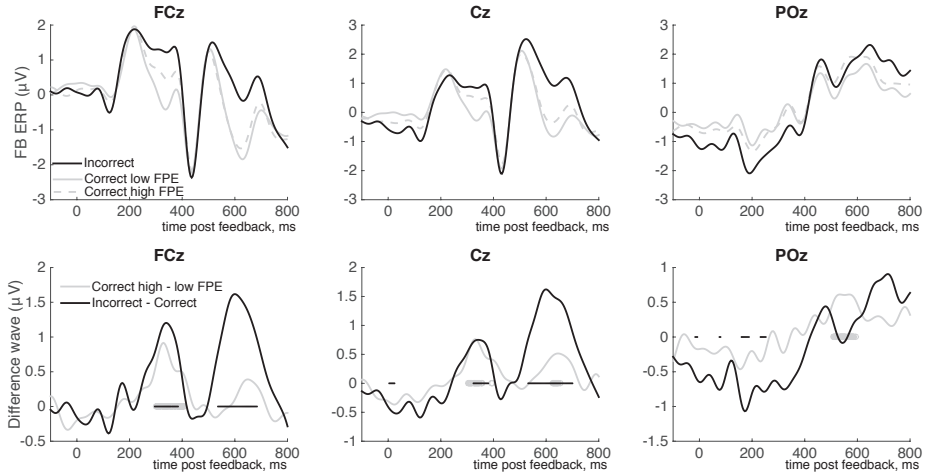


Figure S6: **FB-locked ERPs Error vs. Correct trial. Top:** ERPs for three representative electrodes locked on feedback. Black line indicates error trials, grey lines correct trials. To illustrate the regression effects plotted in fig 4 and fig S4, we median-split correct trials into high and low FPE trials. **Bottom:** Black line is the difference wave between Incorrect and Correct trials; grey line the difference between correct high and low FPE trials. Comparison with Fig S4 shows a close match between this median split analysis and the regression analysis, although the regression analysis carries more statistical power. Black/grey dots indicate when the effect is significant (p<0.001 uncorrected).

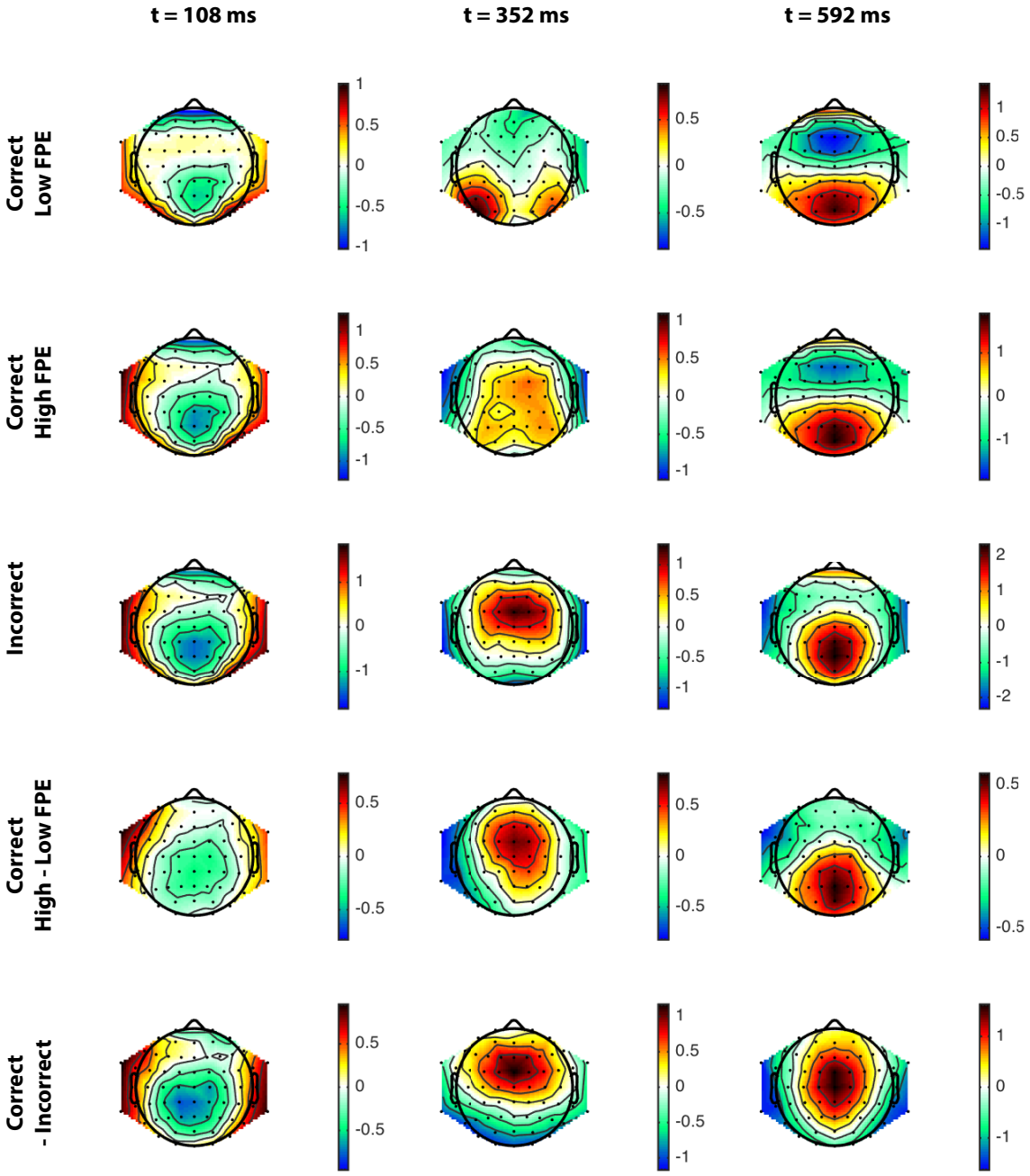
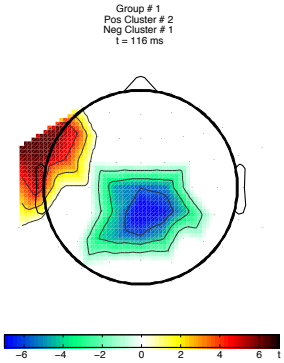


Figure S7: Topoplots of Incorrect, and Correct high and low FPE trials' activity (μV), as well as differential activity (same analysis as Figure S6). The three time points are representative of the time-space clusters observed, and are the same as Fig. 4.



Movie S1: Main text GLM-PE analysis: t-values of effect of PE, thresholded for p<.05 significance, cluster corrected. This movie presents the same data as Figure S4B, but on a scalp map, as a function of time post-feedback.