Supporting Information

Collins and Frank 10.1073/pnas.1720963115

SI Materials and Methods

Subjects. For the EEG experiment, we collected data for 40 subjects (28 female, ages 18–29), and all were included in the behavioral analyses. One subject was excluded from EEG analysis due to technical problems with the EEG cap.

Experimental Protocol.

Structure. Subjects performed a learning experiment in which they used reinforcement FB to figure out which key to press for each presented visual stimulus. The experiment was divided into 22 blocks, with new visual stimuli in each block. After stimulus presentation, subjects selected one of three keys to press with their right hand. FB indicated truthfully whether they had selected the correct action for the current stimulus. See *Trials* for more details.

Blocks. Blocks varied in the number of stimuli that participants learned concomitantly (the set size *ns*) between one and six. Specifically, the number of blocks for set sizes 1–6 were in order {3, 6, 4, 3, 3, and3}; this number was chosen to ensure at least 12 stimuli and three learning bocks per set size, with the exception of set-size 1, which was used as a control. Within a block, each stimulus was presented a minimum of 9 times and a maximum of 15 times; the block ended after $ns \times 15$ trials, or when subjects reached a performance criterion whereby they had selected the correct action for three of the four last iterations of each stimulus. Stimulus presentation was pseudorandomized. Stimuli in a given block were all from a single category (e.g., colors, fruits, or animals) and did not overlap.

Trials. Stimuli were presented centrally on the black background screen (approximate visual angle of 8°); subjects had up to 1.4 s to answer by pressing one of three keys with their right hand. Key press was followed by audiovisual FB presentation (word "Win!", ascending tone, or "loss", descending tone), with a uniformly jittered lag of 0.1–0.6 s. Failure to answer within 1.4 s was indicated by a "too slow" message. FB was presented for [0.4–0.8] s, and followed by a [0.5–0.8] s fixation cross before next trial onset.

Model Free Analysis. We analyzed behavior using a multiple logistic regression. Predictors included set size, delay (number of trials since last previous correct choice for current trial's stimulus), iterations (number of previous correct trials for current trial's stimulus), and interactions between those factors. The first two predictors were markers of WM function and were also used for the EEG multiple regression analysis. The third regressor was a marker of reward history and thus targeted the RL system. Following previous published methods, main effect predictors were transformed according to $X \rightarrow -1/X$.

Computational Modeling.

RLWM model. To better account for subjects' behavior and disentangle roles of WM and reinforcement learning, we fitted subjects' choices with our hybrid RLWM computational model. Previous research showed that this model, allowing choice to be a mixture between a classic delta rule reinforcement learning process and a fast but capacity-limited and delay-sensitive WM process, provided a better quantitative fit to learning data than models of either WM or RL alone (1, 2). The model used here is identical to the model used in ref. 3. We first summarize its key properties, following by the details:

RLWM includes two modules which separately learn the value of stimulus-response mappings: a standard incremental procedural RL module with learning rate α and a WM module that updates S-R-O associations in a single trial (learning rate 1) but is capacity-limited (with capacity K).

The final action choice is determined as a weighted average over the two modules' policies. How much weight is given to WM relative to RL (the mixture parameter) is dynamic and reflects the probability that a subject would use WM vs. RL in guiding their choice. This weight depended on two factors. First, a constraint factor reflected the a priori probability that the item was stored in WM, which depended on set size n_S of the current block relative to capacity K (i.e., if $n_S > K$, the probability that an item is stored is K/ns), scaled by the subject's overall reliance of WM vs. RL (factor $0 < \rho < 1$), with higher values reflecting relative greater confidence in WM function. Thus, the constraint factors indicated that the maximal use of WM policy relative to RL policy was $w_0 = \rho \times \min(1, K/n_S)$. Second, a strategic factor reflected the inferred reliability of the WM compared with RL modules over time: Initially, the WM module was more successful at predicting outcomes than the RL module, but because it had higher capacity and less vulnerability to delay, the RL module became more reliable with experience.

Both RL and WM modules were subject to forgetting (decay parameters ϕ_{RL} and ϕ_{WM}). We constrained $\phi_{RL} < \phi_{WM}$ consistent with WM's dependence on active memory.

Learning model details.

Reinforcement learning model. All models included a standard RL module with simple delta rule learning. For each stimulus s and action a, the expected reward $Q(s,a)$ was learned as a function of reinforcement history. Specifically, the Q value for the selected action given the stimulus was updated upon observing each trial's reward outcome r_t (1 for correct, 0 for incorrect) as a function of the prediction error between expected and observed reward at trial t:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha \times \delta_t,$$

where $\delta_t = r_t - Q_t(s,a)$ is the prediction error, and α is the learning rate. Choices were generated probabilistically with greater likelihood of selecting actions that have higher Q values, using the softmax choice rule:

$$p(a|s) = exp(\beta Q(s,a)) \Big/ \sum_i (exp(\beta Q(s,a_i)).$$

Here, β is an inverse temperature determining the degree with which differences in Q values are translated into more deterministic choice, and the sum is over the three possible actions a_i. Because we have found that within this experimental protocol, recovering β independently from the learning rate is often impractical, we fix $\beta = 100$.

Undirected noise. The softmax temperature allowed for stochasticity in choice, but where stochasticity is more impactful when the value of actions are similar to each other. We also allowed for "slips" of action ("irreducible noise," i.e., even when Q-value differences were large). Given a model's policy $\pi = p(a|s)$, adding undirected noise consists in defining the new mixture policy:

$$\pi' = (1\text{-}\varepsilon)\pi + \varepsilon U,$$

where U is the uniform random policy $[U(a) = 1/n_A, n_A = 3]$, and the parameter $0 < \varepsilon < 1$ controls the amount of noise (4–6). Ref. 7 showed that failing to take into account this irreducible noise

can render fits to be unduly influenced by rare odd data points (e.g., that might arise from attentional lapses) and that this problem is remedied by using a hybrid softmax-ε–greedy choice function as used here.

Forgetting. We allowed for potential decay or forgetting in Q values on each trial, additionally updating all Q values at each trial, according to:

$$Q \leftarrow Q + \phi(Q_0 - Q),$$

where $0 < \phi < 1$ is a decay parameter pulling at each trial the estimates of values toward initial value $Q_0 = 1/n_A$. This parameter allowed us to capture delay-sensitive aspects of WM, where active maintenance was increasingly likely to fail with intervening time and other stimuli, but also allowed us to separately estimate any decay in RL values (which is typically substantially lower than in WM).

Perseveration. To allow for potential neglect of negative, as opposed to positive FB, we estimated a perseveration parameter *pers* such that for negative prediction errors (delta < 0), the learning rate α is reduced by $\alpha = (1 - pers) \times \alpha$. Thus, values of *pers* near 1 indicate perseveration with complete neglect of negative FB, whereas values near 0 indicate equal learning from negative and positive FB.

WM. To implement an approximation of a rapid updating but capacity-limited WM, this module assumed a learning rate $\alpha = 1$ (representing the immediate accessibility of items in active memory), but included capacity limitation such that only at most K stimuli can be remembered. At any trial, the probability of WM contributing to the choice for a given stimulus is $w_{WM}(t) = P_t(WM)$. This value is dynamic as a function of experience (see below). As such, the overall policy is:

$$\pi = w_{WM}(t)\pi_{WM} + (1\text{-}w_{WM}(t))\pi_{RL},$$

where π_{WM} is the WM softmax policy, and π_{RL} is the RL policy. Note that this implementation assumes that information stored for each stimulus in WM pertains to action–outcome associations. Furthermore, this implementation is an approximation of a capacity/resource-limited notion of WM. It captures key aspects of WM such as (*i*) rapid and accurate encoding of information when low amount of information is to be stored; (*ii*) decrease in the likelihood of storing or maintaining items when more information is presented or when distractors are presented during the maintenance period; and (*iii*) decay due to forgetting. Because it is a probabilistic model of WM, it cannot capture specifically which items are stored, but it can provide the likelihood of any item being accessible during choice given the task structure and recent history (set size, delay, etc.).

Inference. The weighting of whether to rely more on WM vs. RL is dynamically adjusted over trials within a block based on which module is more likely to predict correct outcomes. The initial probability of using WM $w_{WM}(0) = P_0(WM)$ is initialized by the a priori use of WM, as defined above, $w_{WM}(0) = \rho \times min(1, K/n_S)$, where ρ is a free parameter representing the participant's overall reliance on WM over RL.

On each correct trial, $w_{WM}(t) = P_t(WM)$ is updated based on the relative likelihood that each module would have predicted the observed outcome given the selected correct action a_c; specifically:

$$\text{for WM, } p(\text{correct}|\text{stim,WM}) = w_{WM} \cdot \pi_{WM}(a_c) + (1\text{-}w_{WM})1/n_A$$

$$\text{for RL, } p(\text{correct}|\text{stim, RL}) \text{this is simply } \pi_{RL}(a_c).$$

The mixture weight is updated by computing the posterior using the previous trial's prior, and the above likelihoods, such that

$$P_{t+1}(WM) = \frac{P_t(WM) \times p(correct|stim, WM)}{P_t(WM) \times p(correct|stim, WM) + P_t(RL) \times p(correct|stim, RL)},$$

and $P_{t+1}(RL) = 1 - P_{t+1}(WM)$.

Models considered. We combined the previously described features into different learning models and conducted extensive comparisons of multiple models to determine which fit the data best (penalizing for complexity) so as to validate the use of this model in interpreting subjects' data. For all models we considered, adding undirected noise, forgetting, and perseveration features significantly improved the fit, accounting for added model complexity (*Model Comparisons*).

This left three relevant classes of models to consider:

RLF: This model combines the basic delta rule RL with forgetting, perseveration, and undirected noise features. It assumes a single system that is sensitive to delay and asymmetry in FB processing. This is a four-parameter model (learning rate α, undirected noise ε, decay ϕ_{RL}, and *pers* parameter).

RL6: This model is identical to the previous one, with the variant that learning rate can vary as a function of set size. We have previously shown that while such a model can capture the basic differences in learning curves across set sizes by fitting lower learning rates with higher set sizes, it provides no mechanism that would explain these effects, and still cannot capture other more nuanced effects (e.g., changes in the sensitivity to delay with experience). However, it provides a benchmark to compare with RLWM. This is a nine-parameter model (six learning rate α_{ns}, undirected noise ε, decay ϕ_{RL}, and *pers* parameter).

RLWM: This is the main model, consisting of a hybrid between RL and WM. RL and WM modules have a shared *pers* parameter, but separate decay parameters, ϕ_{RL} and ϕ_{WM}, to capture their differential sensitivity to delay. WM capacity is $0 < K < 6$, with an additional parameter for overall reliance on WM $0 < \rho < 1$. Undirected noise is added to the RLWM mixture policy. This is an eight-parameter model (capacity K, WM reliance ρ, WM decay ϕ_{WM}, RL learning rate α, RL decay ϕ_{RL}, undirected noise ε, and *pers* parameter).

In the RLWM model presented here, the RL and WM modules are independent, and only compete for choice at the policy level. Given our findings showing an interaction between the two processes, we also considered variants of RLWM, including mechanisms for interactions between the two processes at the learning stage. These models provided similar fit [measured by the Akaike information criterion (AIC)] to the simpler RLWM model. We chose to use the simpler RLWM model, because the more complex model is less identifiable within this experimental design, providing less reliable parameter estimates and regressors for model-based analysis.

RLWM fitting procedure. We used matlab optimization under constraint function fmincon to fit parameters. This was iterated with 50 randomly chosen starting points, to increase likelihood of finding a global rather than local optimum. For models including the discrete capacity K parameter, this fitting was performed iteratively for capacities $K = \{1,2,3,4,5\}$, using the value gave the best fit in combination with other parameters.

All other parameters were fit with constraints [0 1].

Model comparison. We used the AIC to penalize model complexity (8). Indeed, we previously showed that in the case of the RLWM model and its variants, AIC was a better approximation than the Bayesian information criterion (Schwarz, 1978) at recovering the true model from generative simulations (27). Comparing RLWM, RL6, and RLF showed that models RL6 and RLF were strongly nonfavored, with exceedance probability for RLWM of 0.95 over the whole group (10). Other single-process models were also unable to capture behavior better than RLWM.

Model simulation. Model selection alone is insufficient to assess whether the best-fitting model sufficiently captures the data. To test whether models capture the key features of the behavior (e.g., learning curves), we simulated each model with fit parameters for each subject, with 100 repetitions per subject, and then averaged to represent this subject's contribution. To account for initial biases, we assumed that the model's choice at first encounter of a stimulus was identical to the subjects, while all further choices were randomly selected from the model's learned values and policies.

Interaction Models. We tested two computational models embodying two distinct hypotheses for WM and RL interactions, that both predict the low set-size blunted RL observed experimentally.

The competitive model assumes that in low set sizes where WM is successful, it inhibits RL computations, such that the prediction error is computed normally $\delta = R - Q_{RL}$, but the update is weakened $Q_{RL} = Q_{RL} + \alpha\eta\delta$, where η indicates the degree of interference of WM in the RL computation.

The cooperative model instead assumes that in low set sizes where WM is successful, it contributes part of the reward expectation for the RL model, according to the equation: $\delta = R - [\eta Q_{RL} + (1 - \eta)Q_{WM}]$, where Q_{WM} represents reward expectations from the WM system. This RPE is then used to update RL as normal: $Q_{RL} = Q_{RL} + \alpha\delta$. Because WM learns quickly, WM contribution makes δ smaller than expected from classic RL, and thus leads to blunted RL.

Simulations for Fig. 6 were run with the following parameters for both models: $\alpha = 0.2$ and $\eta = 0$ or 0.5 for high or low WM, respectively.

We did not fit the behavior with the interaction models because this experimental design was not appropriate to capture behavioral markers of interaction during learning—as shown in Fig. 6, both models predicted the same pattern across trials in terms of choice and valuation, but only predicted differences in how their deviation from expectations changed across trials. As such, the different interaction models' parameters were not satisfactorily identifiable. Instead, assuming independence between RL and WM in the model-fitting allowed us to capture behavior well on average (Fig. 2), but also to investigate the degree to which neural signals deviated from independence in the way predicted by the cooperation vs. competition models without assuming either. Note that our other recent experimental paradigms do allow us to show evidence for RL/WM interactions in behavior, but do not distinguish between the two sorts of interactions (11). Thus, this investigation reveals the nature of the interaction in the neural signal, whereas the other paradigm shows its relevance for behavior.

EEG.

System. EEG was recorded from a 64-channel Synamps2 system (0.1–100 Hz bandpass; 500 Hz sampling rate).

Data preprocessing/cleaning. EEG was recorded continuously with hardware filters set from 0.1 to 100 Hz, a sampling rate of 500 Hz, and an online vertex reference. Continuous EEG was epoched around the FB onset (−1,500 to 2,500 ms). We used previously identified data cleaning and preprocessing method (12, 13) facilitated by the EEGlab toolbox (14): Data were visually inspected to identify bad channels to be interpolated and bad epochs to be rejected. Blinks were removed by using independent component analysis from EEGLab. The electrodes were referenced to the average across channels.

Event-related potentials. For event-related potentials (ERPs) and multiple-regression analysis, data were bandpass-filtered from 0.5 to 20 Hz, down-sampled to 125 Hz, and baselined by the mean activity between −100 and 0 ms before stimulus onset. For each subject, we performed a multiple regression at each electrode and time point within −100:700 ms around stimulus onset (101 time points) and FB onset. Because there were many fewer error than correct trials, we included only correct trials in the analysis. Scalp voltage was z-scored before being entered into the multiple-regression analysis.

In the stimulus-locked analysis, regressors of interest included z-scored set size, delay, model-derived RL expected value, and the interaction of those three regressors; regressors of no interest included reaction time (z-scored log-transformed) and z-scored trial number within block.

Further stimulus-locked analysis mostly focused on regression weights for the main three regressors, $\beta_{S\text{-}NS}$ and $\beta_{S\text{-}Delay}$, considered as markers of WM function, and $\beta_{S\text{-}Q}$, marker of RL function; which we obtained for each subject, time point, and electrode. The FB-locked analysis was identical, but with RL RPE replacing RL expected value, producing key regression weights $\beta_{F\text{-}NS}$, $\beta_{F\text{-}Delay}$, and $\beta_{F\text{-}RPE}$. Trials included in the analysis were all correct trials for which the values of the regressors were well defined, namely, trials of set size two and above, with at least one previous correct choice for the current stimulus (ensuring delay is defined).

Statistical analysis of GLM weights. We tested the significance of regression weights against 0 across subjects for all electrodes and time points. To correct for multiple comparisons, we performed cluster-mass correction by permutation testing with custom-written matlab scripts, following the method described (4). Cluster formation threshold was for a t test significance level of $P = 0.001$. Cluster mass was computed across space–time, and only clusters with greater mass than maximum cluster mass obtained with 95% chance permutations were considered significant, with 1,000 random permutations.

Corrected ERPs. To plot corrected ERPs, we computed the predicted voltage by the multiple-regression model when setting a single regressor to 0 (set size, delay, RPE, or reaction time); we subtracted this predicted voltage from the true voltage, leaving only the fixed effect, the variance explained by that regressor, and the residual noise of the regression model.

Trial-by-trial markers of WM and RL. We used sensitivity to set size as a marker of WM-dependent processing and sensitivity to model-inferred RL-Q value or RL-RPE as a marker of RL processing. To compute trial-by-trial markers of each process, we first defined a spatiotemporal mask as a result of the analysis of GLM weights: The stimulus-locked WM mask was defined at each time point and electrode by 0 if the effect of set size was not significant, and the t value of the effect if it was significant. A similar process defined the stimulus-locked Q mask and the FB-locked RPE mask. Index of WM activation SWM in a given trial was then computed by how well its activity pattern matched the mask, which we computed as a cross-product of the mask by the activity pattern. The same process was applied to stimulus-locked Q mask, yielding a trial-by-trial index stimulus-locked RL activity SQ, and to the FB-locked RPE-mask, yielding a trial-by-trial index of FB-locked RL activity FPE. EEG learning curves plotted in Fig. 6 show averaged indices as a function of trial iteration number. We tested the model predictions by computing the difference in index from first to second correct trial and comparing this value between low and high set sizes.

Link between stim-locked and FB-locked. We used the previously defined indices to ask whether WM- and RL-related activity at stimulus presentation predicted RL-related activity at FB. To do so, we tried to explain FPE index in a multiple regression, including the behavioral RPE, set size, and the EEG SWM and SQ indices.

Regressor correlations. Regressors used for the within-trial multiple regression analysis defined in *Link Between Stim-Locked and FB-Locked* are significantly correlated with each other (Fig. S5). Based on ref. 15, we chose to not orthogonalize regressors against each other in our main analysis. However, to clarify that SQ and SWM still accounted for variance in FPE in addition to set size and behavioral RPE, we performed successive regressions on residuals, accounting first for set size and RPE, then SQ

and SWM (in either order). We found identical results to those described in main text and Fig. 5*C* (Fig. S5, *Inset*).

SI Discussion

A popular framework for investigating multiple systems that contribute to reinforcement learning is the model-based vs. model-free RL framework (16). As mentioned in the introduction, the WM system is a primitive for the model-based RL system (which requires multiple components in addition to WM, such as planning and learning the environments' model). A more elaborate model-based RL is only relevant in learning tasks with sequential environments—where a choice determines the next state and not just the immediate reward—and thus where model-based planning is potential. Our protocol does not have this dynamic and is thus the type that is often considered fully within the purview of model-free RL. Our experimental protocol (*i*) demonstrates that executive functions (in the form of WM) are used even in the simplest, apparently model-free RL tasks; and (*ii*) provides parametric modulation that allows us to carefully investigate the role of WM (and hence primitives that support MB) in RL, such as its interactions, beyond the competition for choice that has been studied in the model-based vs. model-free setting.

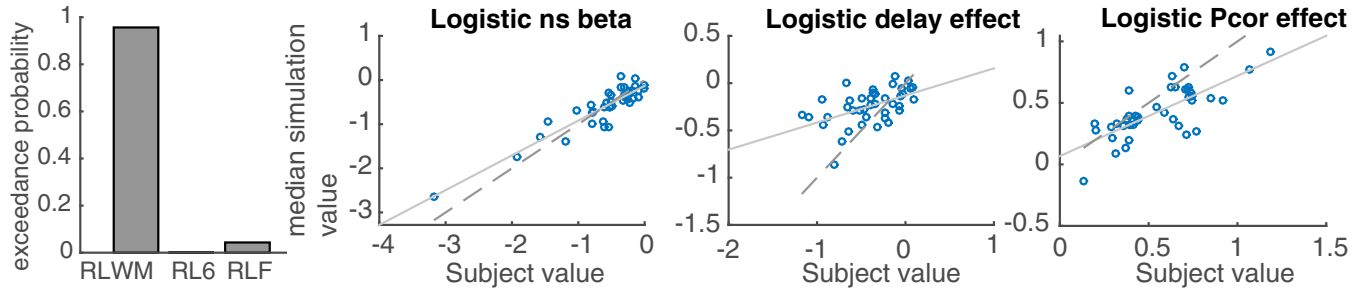

Fig. S1. Model validation. (*Left*) RLWM model provides superior fit to trial-by-trial data compared with competing single-actor models, assessed by exceedance probability. RL6 is a standard RL model allowing separate learning rates per set size; RLF is a standard RL model with forgetting. (*Center* and *Right*) Logistic regression analysis of model behavior simulated with model parameters fit on participants' behavior (100 simulations per subject). Median regression weights per subject for set size (*ns*), delay, and iterations are strongly related to actual participants' regression weights.

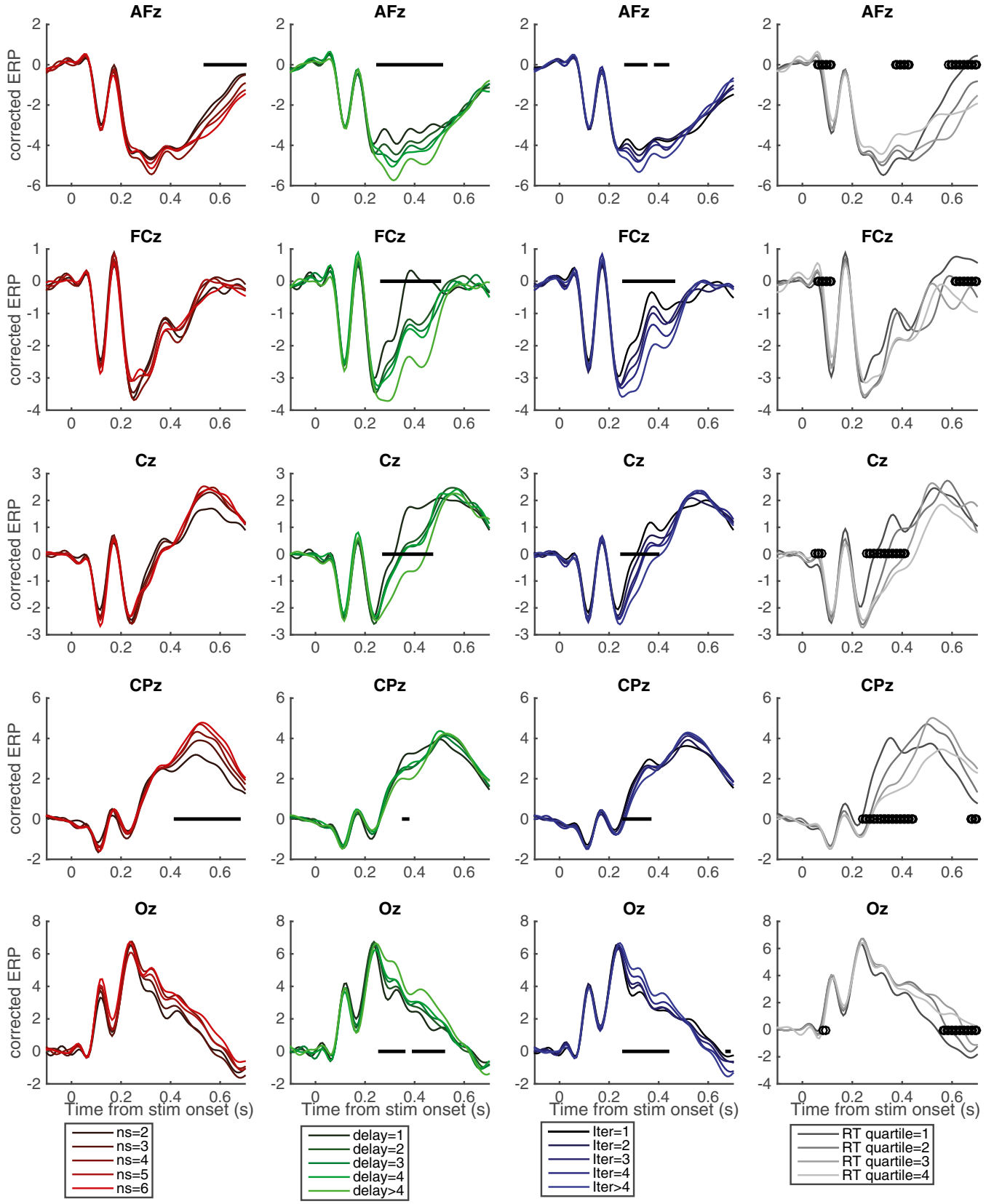

Fig. S2. See Fig. 3*B*. Rightmost column shows corrected ERPs for different reaction time quartiles. Iter, iterations.

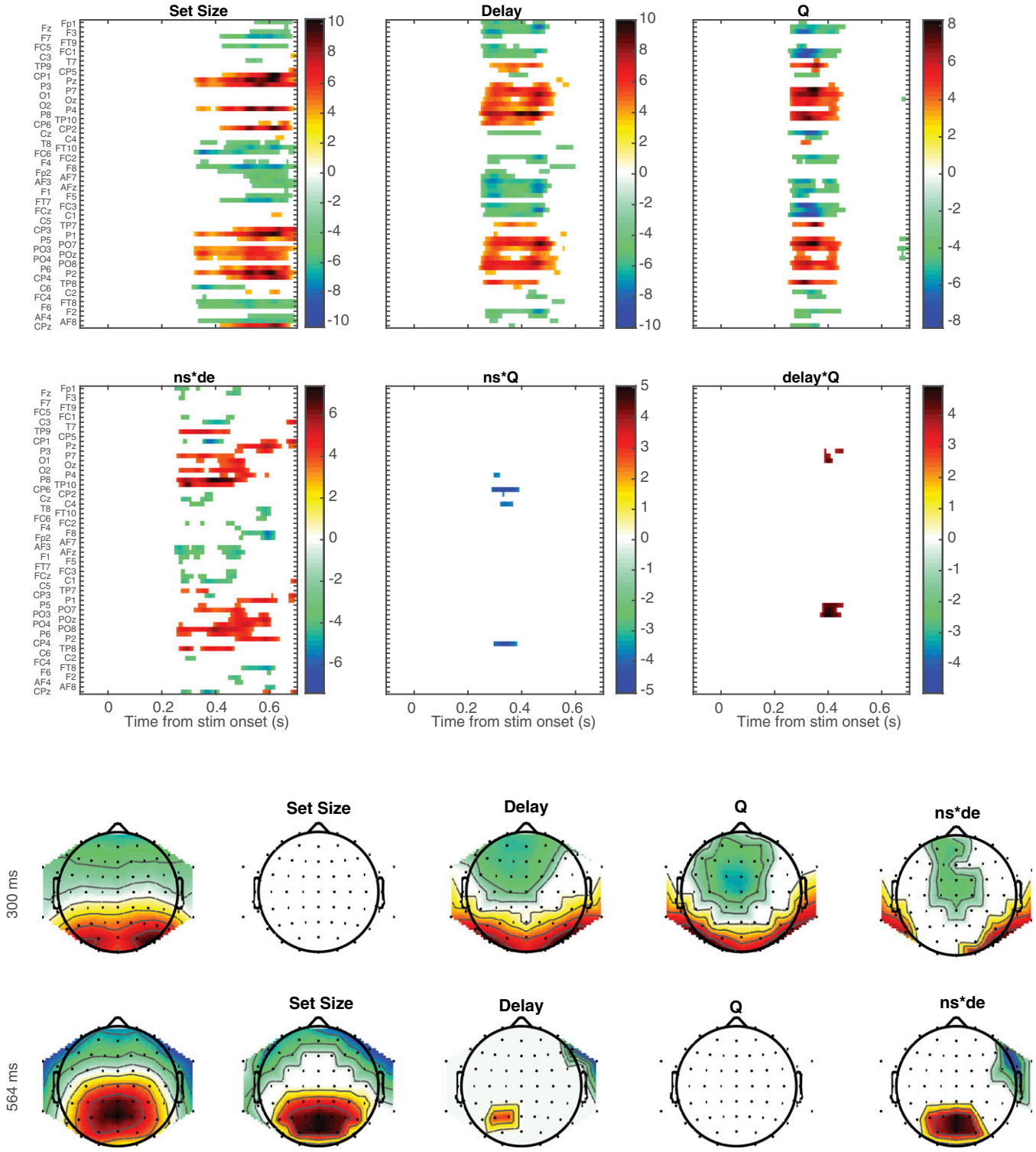

Fig. S3. See Fig. 3C. (*Upper*) Heat maps show stimulus-locked multiple-regression analysis results for all time points and electrodes, as well as for the interactions of the three main factors (set size *ns*, delay, and *Q* values). (*Lower*) Scalp topography for two time points.

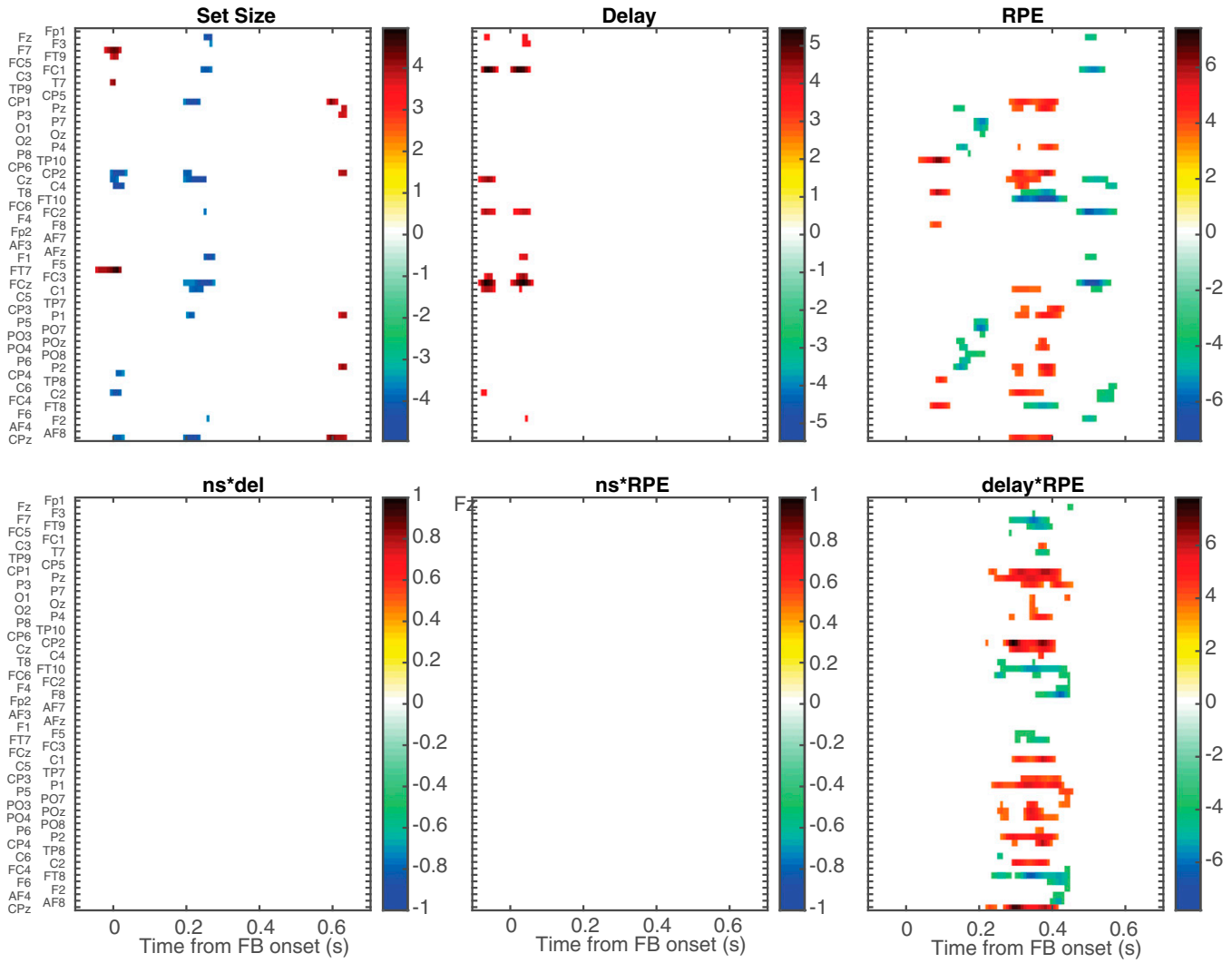

Fig. S4. See Fig. 4. Heat maps show FB-locked multiple-regression analysis results for all time points and electrodes, as well as for the interactions of the three main factors (set size *ns*, delay, and RPE). Results for all electrodes and time points. (*Upper*) Main effects. (*Lower*) Interaction effects.

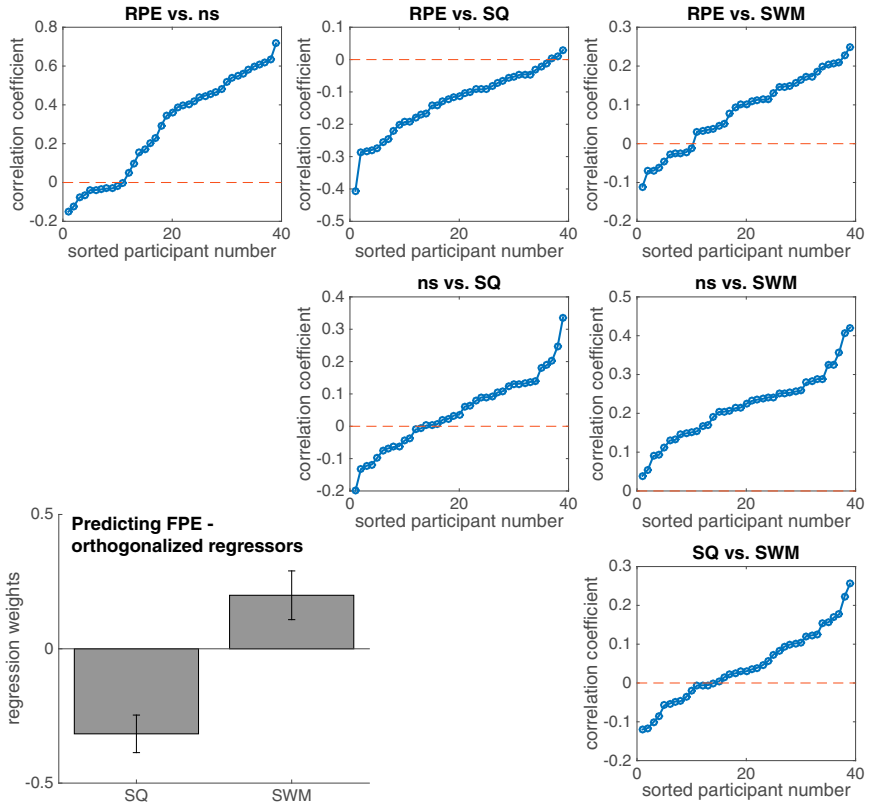

Fig. S5. (*Top*, *Middle*, and *Bottom Right*) Correlation between trial by trial regressors used in Fig. 5 analysis. Each point is a single participant's correlation coefficient, and participants are sorted based on increasing coefficients for each plot. (*Bottom Left*) See Fig. 5C. Analysis with serially orthogonalized regressors (ns, RPE, and {SQ,SWM}) shows similar results to the main text analysis.

Table S1. Summary statistics for fit parameters for models RLWM, RLF, and RL6

Model summary statistic	K	ρ	ϕ_{WM}	α	β	α(1)	α(2)	α(3)	α(4)	α(5)	α(6)	ϕ_{RL}	ε	1-pers
RLWM														
Mean	3.58	0.92	0.28	0.04								0.04	0.04	0.3
Median	3	0.98	0.27	0.02								0.04	0.04	0.19
SD	0.98	0.13	0.19	0.08								0.04	0.03	0.27
Min	2	0.36	0.02	0								0	0	0.05
Max	5	1	1	0.48								0.23	0.15	1
RL6														
Mean					38	0.73	0.52	0.49	0.29	0.1	0.06	0.08	0.06	0.3
Median					44.2	0.8	0.47	0.41	0.05	0.02	0.01	0.07	0.05	0.22
SD					13.2	0.28	0.36	0.42	0.39	0.18	0.18	0.04	0.03	0.27
Min					6.9	0.03	0.01	0.02	0.01	0	0	0	0	0.03
Max					50	1	1	1	1	0.85	0.84	0.2	0.17	1
RLF														
Mean				0.14	39.8							0.12	0.06	0.52
Median				0.02	46.5							0.11	0.05	0.49
SD				0.29	14.6							0.07	0.04	0.32
Min				0.01	4.2							0.01	0	0.02
Max				1	50							0.39	0.17	1

Max, maximum; min, minimum.