### Title
Neural Index of Reinforcement Learning Predicts Improved Stimulus-Response Retention under High Working Memory Load.




### Journal
Journal of Neuroscience, 43(17)

### ISSN
0270-6474



### Authors
Rac-Lubashevsky, Rachel
Cremer, Anna
Collins, Anne GE
et al.



### Publication Date
2023-04-26

### DOI
10.1523/jneurosci.1274-22.2023

JNeurosci
THE JOURNAL OF NEUROSCIENCE

***Research Articles: Behavioral/Cognitive***

# Neural index of reinforcement learning predicts improved stimulus-response retention under high working memory load

https://doi.org/10.1523/JNEUROSCI.1274-22.2023

**Cite as:** J. Neurosci 2023; 10.1523/JNEUROSCI.1274-22.2023

Received: 29 June 2022
Revised: 19 January 2023
Accepted: 20 February 2023

# Neural index of reinforcement learning predicts improved stimulus-response retention under high working memory load

**Abbreviated title**: Neural learning indices predict policy retention

Rachel Rac-Lubashevsky[1,2], Anna Cremer[3], Anne Collins[4,5], Michael J Frank[1,2*] and Lars Schwabe[3*]

**1** Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, Rhode Island, United States of America, **2** Carney Institute for Brain Science, Brown University, Providence, Rhode Island, United States of America, **3** Department of Cognitive Psychology, Universitat Hamburg 20146, Germany, **4** Department of Psychology, University of California, Berkeley, United States **5** Helen Wills Neuroscience Institute, University of California, Berkeley, United States

* These authors contributed equally

**Address for Correspondence**

Rachel Rac-Lubashevsky

Department of Cognitive, Linguistic & Psychological Sciences

Brown University, Providence, Rhode Island, United States of America,

rac.hunrachel@gmail.com

- The number of figures is 7. The number of tables is 1.
- The number of words for Abstract is 233; For introduction is 650; For Discussion is 1499.

**Acknowledgment**

This research was funded by a grant of the Landesforschungsfoerdung Hamburg, Germany (LFF FV 38) to L.S and by R01 MH084840-08A1 to MJF.

**Abstract**

Human learning and decision making is supported by multiple systems operating in parallel. Recent studies isolating the contributions of reinforcement learning (RL) and working memory (WM) have revealed a trade-off between the two. An interactive WM-RL computational model predicts that while high WM load slows behavioral acquisition, it also induces larger prediction errors in the RL system that enhance robustness and retention of learned behaviors. Here we tested this account by parametrically manipulating WM load during RL in conjunction with EEG, in both male and female participants, and administered two surprise memory tests. We further leveraged single trial decoding of EEG signatures of RL and WM to determine whether their interaction predicted robust retention. Consistent with the model, behavioral learning was slower for associations acquired under higher load but showed parametrically improved future retention. This paradoxical result was mirrored by EEG indices of RL, which were strengthened under higher WM loads and predictive of more robust future behavioral retention of learned stimulus-response contingencies. We further tested whether stress alters the ability to shift between the two systems strategically to maximize immediate learning versus retention of information and found that induced stress had only a limited effect on this trade-off. The present results offer a deeper understanding of the cooperative interaction between WM and RL and show that relying on WM can benefit the rapid acquisition of choice behavior during learning but impairs retention.

**Significance statement**

Successful learning is achieved by the joint contribution of the dopaminergic reinforcement learning (RL) system and working memory (WM). The cooperative WMRL model was productive in improving our understanding of the interplay between the two systems during learning, demonstrating that reliance on RL computations is modulated by WM load. However, the role of WM/RL systems in the retention of learned stimulus-response associations remained unestablished. Our results show that increased neural signatures of learning, indicative of greater RL computation, under high WM load also predicted better stimulus-response retention. This result supports a trade-off between the two systems, where degraded WM increases RL processing which improves retention. Notably, we show that this cooperative interplay remains largely unaffected by acute stress.

## Introduction

Everyday behavior, like selecting what to wear and what to eat, involves reinforcement learning (RL). Canonical RL models incrementally accumulate expected values of stimulus-action pairings over the course of multiple experiences. While this RL system learns rather slowly and incrementally, it can be augmented by the joint support of working memory (WM), especially when learning new arbitrary contingencies (Yoo & Collins, 2021). WM enables fast learning by robustly maintaining, in an accessible form, the representations of relevant stimulus-action associations to support ongoing processing such as value-based learning and decision-making. However, when WM capacity is exceeded, it suffers from interference, causing relevant representations to be lost or corrupted (Oberauer et al., 2016). Indeed, while the WM system is beneficial for supporting early learning, its contribution to successful learning is constrained by limited capacity (Collins & Frank, 2012). On the other hand, the incremental RL system has a much broader capacity and is more robust as long as the reward contingencies remain stable. Previous studies have thus shown a transition from capacity- and delay-sensitive WM to RL over the course of learning (Collins & Frank, 2012; 2018).

Moreover, recent studies examining the joint contributions of WM and RL to learning have suggested that these systems are not modular, but rather interactive (Collins, 2018; Collins & Frank, 2018; Collins et al, 2017a,b). fMRI and EEG studies provided support for a cooperative interaction: when stimulus-reward information is stored in WM, neural indices of reward prediction errors (RPEs) are reduced (Collins et al., 2017a; Collins & Frank, 2018). Conversely, RPEs were larger under high load, leading to accelerated "neural learning curves" putatively indicative of more robust RL (despite slowed behavioral learning due to degraded WM). This dissociation suggested that while a high WM load slows learning, it might also improve retention, due to accumulative RPEs that reinforce the RL system. Supporting this prediction, in the surprise test phase, participants showed better retention performance for stimulus-response contingencies and their reward values when they had been learned under higher compared to lower WM demands (Collins et al., 2017b; Collins, 2018; Wimmer & Poldrack, 2020). However, two major limitations remained from this prior work.

First, the previous study showing enhanced retention of stimulus-response associations had only tested low and high WM conditions (Collins, 2018), with only subtle albeit significant differences in performance (around 5% difference between set size 3 vs. 6). We thus parametrically manipulated WM demands (Collins et al., 2017b) to test the prediction that retention performance of stimulus-response associations would scale monotonically as a function of increased WM demand, despite monotonically slowed learning in these conditions. Second, while the neural and behavioral findings have been documented on their own, it has not yet been established whether cooperative neural interactions within WM/RL systems during learning are predictive of future retention. Moreover, it is unclear whether neural RL learning curves reflect reward expectations, or whether they reflect learned policies (as predicted by Q learning vs. actor-critic; Jaskir & Frank 2022; Li & Daw 2011). We thus sought to test these relationships directly by recording EEG during learning and then administering two retention tests. The EEG measures of RL were used to assess whether the neural RL measure is predictive of participants' ability to retrieve learned reward expectations and/or the retention of stimulus-response contingencies.

As a secondary aim, we also examined the impacts of acute stress on RL and WM processes. There is accumulating evidence, across various domains of learning, that acute stress reduces goal-directed decision making and alters prefrontal cortex functioning (see review by Arnsten, 2009), thereby promoting a shift from cognitively demanding but flexible systems towards simpler but more rigid systems (e.g., Wirz, et al., 2018; Kim et al., 2001; Schwabe & Wolf, 2009; Vogel, Fernández, Joëls, & Schwabe, 2016; Meier, Staresina, & Schwabe, 2022). We thus tested whether stress could reduce WM's ability to effectively guide learning and instead enhance the relative contribution of RL processing.

## Methods

### *Participants*

Eighty-six healthy volunteers (43 women, age 18-34, mean = 24.56, SD = 3.84) participated in this experiment. All participants were right-handed, had normal or corrected-to-normal vision, and were screened for possible EEG contraindications. Individuals with a current medical condition, medication intake, or lifetime history of any neurological or psychiatric disorders were excluded from participation. All participants provided written informed consent before the beginning of testing and received moderate monetary compensation. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Sciences at the University of Hamburg.

*Experimental procedure*

*Learning task*

Interactions of RL and WM were tested using the RLWM task (Collins 2018, Collins & Frank, 2012; 2018), programmed in MATLAB using the Psychophysics Toolbox. In this task (see Fig. 1A), each trial started with a presentation of a stimulus in the center of the screen, on a black background and participants had to learn which of the three actions (key presses A1, A2, A3) to select based on trial-by-trial reward feedback. Stimulus presentation and response time were limited to 1.4 sec. Incorrect choices led to feedback 0, while correct choices led to reward, (reward was 1 or 2 points fixed with the probability of 0.2, 0.5, or 0.8). Stimulus probability assignment was counterbalanced within participants to ensure equal overall value of different set sizes (see below) and motor actions. The key press was followed by audio-visual feedback (the word “Win!” with an ascending tone or the word “Loss!” with a descending tone). If participants did not respond within 1.4 sec, the message “Too slow!” appeared. Feedback was presented for 0.4 – 0.8 sec and followed by a fixation cross for 0.4 – 0.8 before the next trial started.

To manipulate WM demands, the number of stimulus-action contingencies to be learned varied by block between 1 to 5 (denoted as ns), with new stimuli set presented at each new block (e.g., colors, fruits, or animals). There were four blocks in which set size=2, two blocks in which set size=4, and three block in which set size=1, 3, 5 for a total of 15 blocks and 645 trials. Within a block, each stimulus was presented 15 times. 108 stimuli were pseudo-randomized and 43 stimuli were presented for each participant. Stimulus category assignment to block set size was counterbalanced across subjects. Block order was also counterbalanced with the exception of set size=1 which served as control (block numbers 8 and 14 were saved for set size=1).

The following instructions were given to participants: “In this experiment, you will see an image on the screen. You need to respond to each image by pressing one of the three buttons on the Gamepad: 1, 2, or 3 with your right hand. Your goal is to figure out which button makes you win for each image. You will have a few seconds to respond. Please respond to every image as quickly and accurately as possible. If you do not respond, the trial will be counted as a loss. If you select the correct button, you will gain points. You can gain either 1 or 2 points designated as "$" or "$$". Some images will give you more points for correct answers on average than other images. You can only gain points when you select the correct button for each image. At the beginning of each block, you will be shown the set of images for that block. Take some time to identify them correctly. Note the following important rules: There is ONLY ONE correct response for each image. One response button MAY be correct for multiple images, or not be correct for any image. Within each block, the correct response for each image will not change”.

*Test phase*

After the learning phase, participants completed two surprise test phases (Fig 1 B, C). The first was a reward retention test that has been used in earlier studies (e.g., Collins et al., 2017b). *The reward retention test* was designed to test whether expected values are learned by default since several previous studies showed that participants can select actions based on their relative expected values at the transfer phase even when they only had to learn which item was best (e.g., Frank et al, 2007; Palminteri et al, 2015). In this phase, on each trial participants were requested to select the more rewarding stimulus from a pair of stimuli that had each been encountered during the learning phase. All stimuli that were used in the learning phase were presented in the test phase at least once. The two stimuli were pseudo-randomly selected to sample across all possible combinations of set sizes, blocks and probabilities. To ensure no new learning at this phase, participants did not receive any feedback on their responses. Note that in this test, participants could not leverage information they had learned about which response to select (the ‘policy'); instead they had to use novel response mappings to simply indicate which stimulus had been more rewarded. Participants’ ability to select the more rewarding stimulus therefore required successful integration of the probabilistic reward magnitude history over learning for each stimulus.

The second test was *the stimulus-response retention test* which assesses whether participants remember the correct response for each stimulus that they had encountered previously during learning. Each of the stimuli used in the learning phase (except stimuli from block 1 and block 15 to limit primacy and recency effects) was presented four times individually, and participants were requested to press the key that was associated with the respective stimulus. Stimulus order was pseudo-randomized to make sure that each stimulus was presented in each quarter of the test phase. No feedback was presented to rule out new learning during this test phase. Note that because this phase was preceded by the reward test phase, and because it followed many serial blocks of learning, it is not plausible that participants could hold information for previously encountered stimuli in WM, and thus retention depends on the memory for stimulus-action associations (the policy) as formalized by the RL system (Collins 2018; Jaskir & Frank 2022).

*---- Figure 1 here -------*

*Behavioral data analysis*

Statistical analyses were performed using R (R Core Team, 2020; https://www.r-project.org/) and the lme4 package (v1.1-26; Bates et al., 2015). Data were fitted using generalized mixed-effect models (glmer) with the Binomial family function. To avoid the Type I error rate without sacrificing statistical power, we followed the parsimonious mixed model approach (Matuschek et al., 2017). We selected the random-effects structure that contained only variance components that were supported by the data by running singular value decomposition (Bates et al., 2015; Matuschek et al., 2017).

*Behavioural analysis of learning task*

To quantify the effect of RL versus WM, we analyzed learning performance (the proportion of correct responses) with general mixed effect regression on trial-by-trial data from 86 participants, as a function of both WM and RL variables and their interactions. The WM variables include the number of stimulus-response associations to be learned (denoted as *setSize*), and the number of intervening trials since the last time the stimulus was presented and a correct response was made (denoted as *delay*) reflecting WM interference or maintenance time in WM. The RL variable is the total number of previous correct responses for a stimulus (denoted as *Pcor)*. Participants and all the predictors were selected as random variables.

*Behavioral analysis of the reward retention test*

To quantify the possible effect of expected value learning under different WM loads, we analyzed test performance (the proportion of selecting the right vs left stimulus) with general mixed effect regression on trial-by-trial data from 86 participants, as a function of six variables: value difference (denoted as *delta_Q*; is positive when the right stimulus had higher value and negative when the left stimulus had higher value), mean Q value of the stimulus pair (denoted as *mean value (Q)*), mean set size of the stimulus pair (denoted as *mean_setSize)*, the difference in set size (denoted as *delta_ setSize*; is positive when the right stimulus was learned in higher set size), *block* (the block number in which they were learned, indicating how recently it was learned), and *perseveration* (binary coding of repetitions in response, repeat/switch). Participants, the effect of value difference (delta_Q), and the effect of set size difference (delta_setSize) were entered as random variables.

*Behavioral analysis of the reward retention test together with EEG RL index*

We ran a new regression model on the reward retention test data (including only the 77 participants that had EEG data), adding the difference in the EEG RL index between the pair of stimuli at choice. Because the neural RL index (see a detailed description of this measure below) could have both positive and negative values all the predictors that were calculated as difference scores were taken as absolute scores and the model predicted performance accuracy (proportion of choosing the higher value stimulus). Test performance accuracy was analyzed as a function of: The absolute model estimated value difference between the right and left stimulus (*abs_delta_Q* ); the absolute difference in the EEG RL index between the right and left stimulus (*abs_delta_EEG_RL*); the mean value (estimated from the model) of the stimulus pair (*mean Q value*); the mean set size of the stimulus pair (*mean set size*); the absolute difference in the block number where the right and left stimulus were learned (*abs_delta_block*); response bias towards the previously selected response (*perseveration;* binary coding of repetitions in response). Participants, the effect of value difference (*abs_delta_Q*), and the effect of EEG RL index difference (*abs_delta_EEG_RL*) were entered as random variables.

*Behavioral analysis of the stimulus-response retention test*

In a general mixed-effect regression analysis we tested accuracy for correctly recalling the response associated with a presented stimulus learned during the training phase as a function of *set size* (the set size block in which they were learned), *block* (the block number in which they were learned, indicating how recently it was learned) and *model Q* (the model estimated Q value of each stimulus calculated as the average Q value of the final 6 iterations during learning) and *perseveration* (the tendency to repeat the response selected in the previous trial at test coded as 1 for repeat and 0 for switch). The interactions between set size and model Q value, set size and block, and between set size and perseveration were also added as predictors. Participants and the interaction between model Q and set size were entered as random variables.

*Behavioral analysis of the stimulus-response retention test together with EEG RL index*

We ran the same regression model on the stimulus-response retention test data as before (including only the 77 participants that had EEG data), adding two new predictors: the average EEG RL index for each stimulus-response association (see a detailed description of this measure below) and the interaction between EEG RL index and set size. Participants, the interaction between model Q and set size, and the interaction between EEG RL index and set size were entered as random variables.

*Electroencephalogram (EEG) recording and processing*

During the learning task, participants were seated approximately 80 cm from the monitor in an electrically shielded and sound attenuated cabin. EEG was recorded using a 64-channel BioSemi ActiveTwo system (BioSemi, Amsterdam, The Netherlands) with sintered Ag-AgCl electrodes organized according to the 10-20 system. The sampling rate was 2048 Hz. The signal was digitized using a 24-bit A/D converter. Additional electrodes were placed at the left and right mastoids, approximately 1 cm above and below the orbital ridge of each eye and at the outer canthi of the eyes for measurement of eye movements. The EEG data were re-referenced offline to a common average. Electrode impedances were kept below 30 kΩ. EEG and EOG were amplified with a low cut-off frequency of 0.53 Hz (=0.3 s time constant).

The EEG data were processed using EEGLAB (Delorme and Makeig, 2004) and ERPLAB (Lopez-Calderon and Luck, 2014). The continuous EEG was bandpass-filtered offline between 0.5–20 Hz and down-sampled to 125 Hz, then it was segmented into epochs ranging from 500 ms pre-stimulus up to 3000 ms post-stimulus. The epoched data were visually inspected and those containing large artifacts due to facial electromyographic (EMG) activity or other artifacts, except for eye blinks were manually removed (e.g., large fluctuations in voltage across several electrodes that were in an order magnitude above neighbouring activity). Independent components analysis (ICA) was next conducted only on the 64 scalp electrodes using EEGLAB's runica algorithm. Components containing blink or oculomotor artifacts, were subtracted from the data resulting in an average of 1.6 components removed per participant (ranging between 0 to 3 components). Finally, the epoched data was subjected to automatic bad-electrodes and artifact detection algorithm (100µV voltage threshold with a moving window width of 200ms and a 100ms window step) which was followed by manual verification. Bed-electrodes were interpolated and trials containing large artifacts were removed. Nine participants were removed from all the reported EEG analyses due to a high EEG artifact rate (>40% in one or more of the conditions) resulting in 77 participants that were used in the EEG analysis.

*Data processing for behavior and EEG regression analysis*

Omission trials, trials with very fast RT (<200ms), and trials before the first correct response was made were excluded from all analyses. Setting the delay and Pcor variables to have 1 as their lowest level was done to insure an interpretable analysis of these variables (Collins & Frank, 2012). The delay predictor (the number of trials since the stimulus was presented and a correct response was made) used in the regression analyses was inverse transformed (-1/delay) to avoid the disproportion effect of very large but rare delays (when a correct response was given early in the block but was then followed by several error responses for that stimulus).

*Modeling*

RL and WM contributions to participants ‘choices were estimated with the previously developed RLWM computational model (the model described below is identical to that used in Collins and Frank, 2018; see more details described in the original paper). The RLWM is a mixture of a standard RL module with a delta rule and a WM module that has perfect memory for information that is within its limited capacity and is sensitive to delay (reflecting memory decay and interference from other intervening stimuli). For each stimulus-action association, the RL module estimates the expected value (“Q”) and updates those values incrementally on every trial as a function of the reinforcement history. This computation is complemented by the WM module where information in the capacity-limited WM feeds into RL expectations, thereby affecting RL prediction errors and learning (see Fig. 2).

**Basic RL module:** To maintain consistency with prior studies with this task and model, and to keep the model as simple as possible, we use Q learning for the model-free algorithm, but an actor critic could also have been used (there are multiple options to capture incremental model-free RL, including methods that learn expected values for each choice and select on that basis (a canonical instance is Q learning and is often used in human studies) as well as methods that learn to directly optimize the policy (a canonical variant is an actor-critic model). Both classes of models similarly predict behavioral adjustment in RL tasks and specific designs are needed to distinguish between them (e.g., Gold et al, 2012; Geana et al 2021). The main goal here is to simply summarize the incremental RL process as distinct from the WM process.

Reward values were coded as 0 or 1 for correct or incorrect (model fits are not improved if using 1 vs 2 points in the Q learning system, and behavioral learning curves are similar for stimuli that yield higher or lower probability of 2 points; Collins et al., 2017b). For each stimulus *s* and action *a* association, the RL module estimates the expected reward value $Q$ and updates those values incrementally on every trial:

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

The Q value was updated as a function of the learning rate $\alpha$ (reflecting how fast reward expectations are updated) and the reward prediction error $\delta$, calculated as the difference between the observe reward, $R_t$ and the expected reward, $Q_t$ at each trial: $\delta_t = R_t - Q_t$.

Choices were probabilistically determined using a softmax choice policy:

$$p(a|s) = exp\left(\beta Q(s,a)\right) / \sum \left(exp\left(\beta Q(s,a_i)\right)\right)$$

Here, $\beta$ is the inverse temperature determining the degree to which differences in $Q$ values are translated into more deterministic choices, and the sum is over the three possible actions. Q-values were initialized to $1/n_A$, where $n_A = 3$ is the number of actions (i.e., the prior that any action is correct is 1/3).

**WM module**: This module updates stimulus-action-outcome associations in a single trial. It assumes that stimulus-action-outcome information, when encoded and maintained in WM, could serve to update reward expectation rapidly and accurately (i.e., perfect retention of the previous trial's information). When not limited by capacity and decay (see below), the WM module is therefore represented by a Q learning system with a learning rate of 1 ($\alpha = 1$).

*Decay*: To account for potential forgetting on each trial due to delay or WM interference, we included a decay parameter $\phi$ ($0 < \phi < 1$) which pulls the estimates of Q values toward their initial value, [$Q_0 = 1/n_A$, number of actions $n_A = 3$].

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

Only the WM module was subject to forgetting (decay parameter $\varphi_{WM}$), to capture WM's well documented short-term stability, in contrast to RL's robustness.

*WM contributes to choice:* Because WM is capacity limited, only *K* stimulus and action associations can be remembered. A constraint factor reflects the *a priori* probability that the item was stored in WM: $w_{WM}(0) = P_0\,(WM) = K/n_s$ (i.e., the set size in the current block relative to capacity *K*) and implies that the maximal use of WM policy relative to RL policy depends on the probability that an item is stored in WM. This probability is then scaled by $\rho$ ($0 < \rho < 1$), the participant's overall reliance of WM vs RL (where higher values reflect greater confidence in WM).

$$w_{WM}(0) = \rho * min(1, K/n_s)$$

**Cooperative model***:* While the original model (Collins & Frank, 2012) assumed independent RL and WM modules that compete to guide behavior, our more recent work suggests that WM expectations influence RL updating (Collins & Frank, 2018). Thus, WM contributes part of the reward expectation for the RL model, according to the equation: $\delta_t = R_t - [w_{WM} \times Q_{WM} + (1 - w_{WM}) \times Q_{RL}]$, where $w_{WM}$ is the weighting parameter (the degree to which WM is weighted relative to RL, which is stronger in low set sizes), and $Q_{WM}$ is the expected reward from the WM module. This RPE is then used to update the RL Q value: $Q_{t+1} = Q_t + \alpha \times \delta_t$

This interactive computation of RL forms the basis of the simulated predictions shown in Figure 2. Nevertheless, as explained in Collins and Frank (2018), we test these predictions by fitting models in which RL and WM modules are independent (independence is assumed in the original models, which still provide good fits to the data, because when information is within WM, WM dominates updating and contributes to rapid learning curves, and hence the interactive models' smaller RPEs and RL Q values for small set sizes are not influential on behavioral accuracy during learning; however, this model makes differential predictions for neural learning curves and future retention). We then assess systematic deviations from independence informed by these simulations (e.g, neural Q learning curves should grow more rapidly in high than low set sizes; Fig. 2).

----- *figure 2 is here* -----

*Data processing for univariate EEG analysis*

To extract the neural correlates in the EEG signal of conditions of interest we employed a mass univariate approach (Collins & Frank, 2018). A multiple regression analysis was conducted for each participant, in which the EEG amplitude at each electrode site and time point was predicted by the conditions of interest: set-size (number of stimulus-response-outcome associations given in a block), model-derived RL expected value (denoted as Q), delay (number of trials since this stimulus was presented and a correct response was given) and the interaction of these three regressors, while controlling for other factors like reaction time (log-transformed) and trial number within block. Furthermore, the EEG signal was reduced to a selected window of -100 to +700 ms around stimulus onset, and was baseline corrected from –100 to 0 ms before the onset of the stimulus. To account for remaining noise in the EEG data, the EEG signal (at each time point and electrode) was z-scored across all trials and so were all the predictors before they were entered to the robust multilinear regression analysis (Collins & Frank, 2018).

*Corrected ERPs*

To plot corrected ERPs, we computed the predicted voltage using the multiple-regression model described above while setting a single regressor to 0 (set size, delay, expected Q value, or reaction time); we subtracted this predicted voltage from the true voltage (for every electrode and time point within each trial), leaving only the fixed effect, the variance explained by that regressor, and the residual noise of the regression model. ERPs were computed as the average corrected voltage from all trials that belong to the same level of condition. Note that the array of expected Q values was divided to 4 quartiles and trials within each quartile were averaged for plotting ERPs.

*Trial-by-trial similarity index of WM and RL*

As explained above, a multiple regression analysis was conducted for each participant, in which the EEG amplitude at each electrode site and time point was predicted by the conditions of interest (set size, delay, RL expected value, and their interactions). We used the previously identified analysis method (Collins & Frank, 2018; Rac-Lubashevsky & Frank 2021) to identify spatiotemporal clusters (masks) of the three main predictors in the GLM (set-size, delay, and model-derived RL expected value). Specifically, we tested the significance of each time point at each electrode across participants against 0 using only trials with correct responses.

We then used cluster-mass correction by permutation testing with custom written Matlab scripts. Cluster-based test statistics were calculated by taking the sum of the t-values within a spatiotemporal cluster of points that exceeded the P = 0.001 threshold for a t-test significance level. This was repeated 1000 times, generating a distribution of maximum cluster-mass statistics under the null hypothesis. Only clusters with greater t-value sum than the maximum cluster-mass obtained with 95% chance permutations were considered significant. We then assessed each trial's neural similarity to the spatiotemporal mask by computing the dot product between the activity in the individual trial (voltage maps of electrode × time) and the identified masks (t-value maps of electrode × time). This computation produced a trial-level similarity measure intended to assess the trial-wise experienced WM load and delay effects, as well as trial-wise RL contributions.

The EEG RL index predictor used in the general mixed-effect regression analyses of both test phases was calculated by averaging the EEG RL index in the final 6 iterations of each stimulus. This was done for each stimulus-response association within each participant.

*Stress manipulation*

All testing took place in the morning between 8am and noon. Upon their arrival in the lab, participants' baseline measures of blood pressure and salivary cortisol were taken. Afterwards, participants were prepared for the EEG and completed the mood questionnaire MDBF (Steyer, et al., 1994) that measures subjective mood on the scales negative vs. elevated mood, calmness vs. restlessness, and wakefulness vs. tiredness, before and after the treatment as well as after the learning task. 42 participants underwent the Socially-evaluated Cold Pressor Test (SECPT; Schwabe et al., 2008) and 44 participants were assigned the warm water control condition. The SECPT is a standardized stress protocol in experimental stress research that combines physiological and psychosocial stress elements and has been shown to result in robust stress responses (Schwabe & Schächinger, 2018). During the SECPT, participants in the stress group immersed their right hand for three minutes in ice water (0-2°C), while being videotaped and evaluated by a non-reinforcing, cold experimenter. In the control condition, participants immersed their hands in warm water (35-37°C), without being videotaped or evaluated by an experimenter. About 25 minutes after the treatment, participants received the learning task instructions and completed a brief training session after which they completed the learning task and the test phases 1 and 2. In total, the experiment lasted about 130 minutes.

**Results**

In line with previous findings in this task (e.g., Collins et al. 2017b), our data demonstrated separable contributions of RL and WM systems to performance. The contribution of incremental RL was observed as the proportion of correct responses increased with the progress in the block (Fig. 3A) and with the increase in reward history (*pcor:* $\beta$=.67, SE=.05, z(46926)=13.17, $p$<.001). WM contributions were observed as learning was strongly affected by set size with a greater proportion of correct responses in low set sizes than in high set sizes (*set size:* $\beta$=-.28, SE=.05, z(46926)= -5.39, $p$<.001). Learning curves were more gradual in higher set sizes than in low set sizes (Fig. 3A; and slower Fig. 3B). Moreover, performance decreased with increasing delay in larger set sizes (*delay* × *ns,* $\beta$=-.09, SE=.05, z(46926)= -2.59, $p$=.009; Fig. 3C). These relative contributions of WM decreased with learning as the detrimental effect of delay attenuated with the increase of accumulated rewards (*ns* × *Pcor:* $\beta$=.13, SE=.04, z(46926)=3.35, $p$<.001; *delay* × *Pcor:* $\beta$=.34, SE=.04, z(46926)=9.17, $p$<.001; *ns* × *Delay* × *Pcor*, $\beta$=.20, SE=.03, z(46926)= 6.37, $p$<.001; Fig. 3D-3E), reflecting a transition from WM to RL. Together these results confirm the cooperative interaction of early WM contributions that diminish as RL becomes more dominant.

*--- Figure 3 is here---*

*Behavioral Performance: Reward Retention Test*

Results replicated previous findings in this phase (Collins et al, 2017b). Participants were more likely to select the stimulus for which they had been rewarded more often during learning as a function of the difference between the number of rewards experienced for these stimuli (*delta_Q:*

$\beta$=.41, *SE*=.04, z(19796)=9.76, $p$<.001). Moreover, also replicating previous findings, this value discrimination effect was enhanced when stimulus values were learned under higher set sizes rather than under lower set sizes (*mean_setSize × delta_Q:* $\beta$=.11, *SE*=.02, z(19796)=6.04, $p$<.001). For display purposes, the median split in the absolute delta_Q score is shown as high and low-value differences (see Fig. 4A). Furthermore, participants were generally less likely to select the stimulus learned under a higher set size than under a low set size (*delta_setSize,* $\beta$=-.69, *SE*=.09, z(19796)=-7.61, $p$<.001), an effect previously attributed to participants learning a cost of mental effort in a high set size (Collins et al 2017b). There was no effect for the difference in the block in which the item values were learned nor was the set size effect modulated by block number ($p$ >.82). We also controlled for response perseveration; no significant tendency was observed for repeating the same response used in the previous trial ($p$ >.69).

*--- Figure 4 is here---*

*Behavioral Performance: Stimulus-response retention test*

Supporting the key model prediction that retention of stimulus-response associations should improve as load increases, we observed better recall performance for associations learned under high rather than low set sizes (*set size:* $\beta$=.84, *SE*=.05, z(11894)=15.83, $p$<.001). And, indeed this effect was parametric, with substantially better performance as set size increased (see Fig. 4B-C). This effect is particularly striking given that performance is parametrically worse for the higher set size items during learning (compare Fig. 3A and Fig. 4C). Not surprisingly, recall accuracy in the test phase was positively predicted by the estimated Q value of the probed stimulus-response association (*model Q:* $\beta$=.27, *SE*=.04, z(11894)=6.97, $p$<.001), that is, associations that were learned better were also better remembered. Importantly, this effect grew when the set size was high (*model Q × set size:* $\beta$=.15, *SE*=.04, z(11894)=3.64, $p$<.001; see Fig. 4B). Recall accuracy was also subject to the influence of recency as associations learned during more recent than early blocks were also recalled more accurately (*block:* $\beta$=.22, *SE*=.03, z(11894)=8.61, $p$<.001). This recency effect increased for associations learned under higher set sizes (*set size × block:* $\beta$=.09, *SE*=.02, z(11894)=4.13, *p*<.001). No effect of perseveration in responses was observed (*p*>.11).

*EEG correlates of WM and RL during learning*

The model-based EEG analysis indicated significant effects for all three variables of interest: set size, delay, and RL. Consistent with previous EEG results in this task (Collins & Frank, 2018) and with the prediction that separable systems contribute to learning, the neural signals of RL exhibited an early frontal activity (around 300ms post-stimulus onset; see Fig.5) that preceded the parietal neural signal of set-size (peaked around 540 ms; see Fig.5), supporting the engagement of the RL system early in the trial followed by the cognitively effortful WM process. The neural signals of RL exhibited an additional late temporal activity (around 600ms post-stimulus onset) that overlapped in time with the set size effect. Finally, a significant frontal and parietal effect of delay was also observed to initiate early at 300ms.

*--- Figure 5 is here ---*

To quantify how the neural measure of RL is modulated by WM and RL processes, we analyzed the trial-by-trial level EEG RL index (reflecting how strong is the RL computation at a given trial) with linear effects regression from 77 participants, as a function of set size (*setSize* =1,2,3,4,5), the number of previous correct (*pcor*=1:15), and the interactions between them (see Methods). As expected due to incremental learning, neural indices of RL increased parametrically as a function of reward history (*pcor:* $\beta$=.17, *t*(38377)=34.77, *p*<.001). Importantly, confirming model predictions, neural RL signals increased to a larger extent as the set size grew (*pcor × setSize:* $\beta$=.04, *t*(38377)=7.53, *p*<.001; Fig. 4F). This finding corroborates previous reports that RL computations are larger in high set sizes due to diminishing WM contributions and thus increasing the accumulation of reward prediction errors (Collins et al., 2017b; Collins & Frank, 2018).

*--- Figure 6 is here ---*

We next assessed the core prediction that the neural RL index is related to future retention, and more specifically, the cooperative model prediction that the speeded neural RL curves in high set sizes are related to better retention of learned contingencies. Notably, while this prediction did not hold for the reward retention phase (*abs_delta_EEG_RL:* $p$=.65; *mean_setSize* × *abs_delta_EEG_RL:* $p$=.61; Fig 4D), it was clearly borne out for the stimulus-response retention phase (*EEG RL:* $\beta$=.23, z(10613)=4.51, $p$<.001; Fig 4E). Stimuli that had been associated with a larger EEG RL index during learning were associated with better recall of the associated response at test; this effect held even when controlling for the non-neural predictors (which replicated the prior analysis). Figure 4E shows that a high EEG RL index (by median split) was predictive of better retention performance at test. The finding that the neural index of RL is related to policy retention but not reward retention is relevant for models that dissociate whether model-free RL in the brain encodes expected values or policies (see model method section and Discussion). Note that a slightly different regression model was used for testing the neural RL index effect on the reward retention test performance than the behaviour model used previously (see Method section for more detail). Nevertheless, the key behavior results were replicated in this analysis as performance increased with the increase in the absolute value differences (*abs_delta_Q:* $\beta$=.31, *SE*=.03, z(17743)=8.82, $p$<.001) and while this effect was not further modulated by set size (*mean_setSize* × *abs_delta_Q,* $p$=.63), performance accuracy did improve with set size (*mean_setSize:* $\beta$=.07, *SE*=.02, z(17743)=3.23, $p$=.001; see Fig 4D).

*Acute stress modulation of RL and WM interaction*

*Manipulation check*

Subjective, autonomic and endocrine data indicated that the stress induction by the SECPT was successful. The SECPT was rated as significantly more unpleasant, stressful, and painful than the warm water control procedure: [more difficult, t(84) = 9.941, $p$ <.001, $d$ = 2.14; more unpleasant, t(84) = 9.088, $p$ < .001, $d$ = 1.96; more stressful, t(84) = 7.72, $p$ <.001, $d$ = 1.66; and more painful t(84) = 11.42, $p$ < .001, $d$ = 2.46; see rating reports in Table 1]. Furthermore, we observed significant

Treatment-by-Time interactions for subjective stress ratings [negative mood: $F_{2,164} = 10.53$, $p < .001$, $\eta_g^2 = .02$; restlessness: $F_{2,164} = 9.47$, $p < .001$, $\eta_g^2 = .02$] and autonomic arousal measures [systolic blood pressure (SBP): $F_{4,336} = 26.22$, $p < .001$, $\eta_g^2 = .06$; diastolic blood pressure (DBP): $F_{4,336} = 26.99$, $p < .001$, $\eta_g^2 = .09$; and heart rate: $F_{3,252} = 10.70$, $p < .001$, $\eta_g^2 = .02$]. As expected, these autonomic responses returned relatively quickly to baseline after the treatment (see Fig.6). The stress and no-stress control groups did not differ in any of the autonomic arousal measures pre-treatment (all p-values>.07).

*--- Figure 6 is here ---*

*--- Table 1 is here ---*

Salivary cortisol (sCORT) responses were assessed by running ANOVA with Time (T1, T2, T3, T4) as the within-subject factor and Treatment (SECPT vs. warm water control group) as the between-subject factor. We observed a significant effect for Time ($F_{3,234} = 28.53$, $p < .001$, $\eta_p^2 = .27$) but not for Treatment ($F_{1,78} = 3.03$, $p = .08$, $\eta_p^2 = .04$). An expected Treatment × Time interaction was observed ($F_{3,234} = 6.97$, $p < .001$, $\eta_p^2 = .08$), with the stress group displaying greater sCORT levels immediately before the learning task (23 min post-treatment) [$t(78) = 2.80$, $p = .006$, $d = 0.63$] but only marginal difference was observed at half time during learning task (50 min post-treatment) [$t(78) = 1.90$, $p = .06$, $d = 0.43$]. No difference in sCORT levels was observed at baseline [$t(78) = 0.61$, $p = .54$] nor at the end of the learning task (80 min post-treatment) [$t(78) = 0.11$, $p = .91$], suggesting that stress-induced cortisol elevations gradually decreased during the learning task (Fig. 10). Note that 6 participants were excluded from the cortisol analysis because they did not provide sufficient saliva for analysis.

*Learning Phase performance by stress group*

To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning thereby weakening the relative contribution of WM in the training phase in the stress group compared to the control group, we ran the same general mixed-effect regression model on trial-by-trial training data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group). This analysis revealed that learning by set size interaction was modulated by stress (*pcor × set size × stress_group:* $\beta$=-.20, *SE*=.08, z(46926)=-2.60, *p*=.009) and so was the learning by delay interaction (*pcor × delay × stress_group:* $\beta$=.22, *SE*=.07, z(46926)=3.04, *p*=.002). To understand the nature of these interactions we ran two follow-up analyses using the same general mixed-effect regression model on trial-by-trial training data, separately in the control (N=44) and the stress group (N=42). These analyses showed that learning curves were additive to the set size effect in the stress group (*pcor × set size: p*=.74) but not in the control group (*pcor × set size:* $\beta$=.22, *SE*=.05, z(24031)=4.30, *p*<.001) which showed a greater drop in performance during high set sizes (see Fig. 7A-B). The attenuated delay effect with learning was significant for both the stress group (*pcor × delay:* $\beta$=.47, *SE*=.05, z(22895)=8.41, *p*<.001) and the control group (*pcor × delay:* $\beta$=.23, *SE*=.05, z(24031)=4.74, *p*<.001; see Fig. 7C-D).

*--- Figure 7 is here ---*

*Reward Retention Test performance by stress group*

To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning thereby strengthening RL conurbations during the training phase and leading to better retention of learned information in the stress group compared to the control group, we ran the same general mixed-effect regression model on trial-by-trial reward retention test data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group) and analyzed test performance (the proportion of selecting the right vs left stimulus). This analysis replicated the results of the behavior analysis without the group factor. No effect of stress was observed (*p*>.15; Fig 7E).

*Stimulus-response retention test performance by stress group*

To test the hypothesis that acute stress may reduce WM's ability to effectively guide learning thereby strengthening RL conurbations during the training phase and leading to better retention of learned information in the stress group compared to the control group, we ran the same general mixed-effect regression model on trial-by-trial stimulus-response retention test data from 86 participants but added stress group as a factor (42 participants in the stress group and 44 participants in the control group) and analyzed test performance. This analysis revealed that the effect of set size on recall accuracy of stimulus-response associations interacted with stress (*set size × stress_group:* $\beta$=.22, *SE*=.10, z(11894)=2.30, *p*=.02; Fig. 7F) but follow up analysis on each group separately showed significant effect of set size on recall accuracy in both the control group ($\beta$=.72, *SE*=.07, z(6129)=10.72, *p*<.001) and the stress group ($\beta$=.95, *SE*=.08, z(5765)=11.76, *p*<.001).

**Discussion**

Taken together, our findings provide insight into the intricate interplay between WM and RL during learning, and its opposing influences on acquisition vs. retention of stimulus-response associations. A recent study proposed a cooperative WMRL model, whereby RPEs in the RL system are not only computed relative to RL expected values but are also modulated by expectations held in WM (Collins & Frank, 2018). This model accounted for fMRI and EEG findings in which neural RPEs were diminished for smaller WM loads (Collins et al., 2017; Collins & Frank, 2018). Moreover, this model accounted for findings that on a given trial, larger neural indices of WM expectations were predictive of subsequent RPEs during the outcome, even within a given set size (Collins & Frank, 2018). This model led to a key prediction that enhanced RL processes under high WM load would support more robust retention of learned association, despite the substantially slower acquisition. Preliminary behavioral evidence for such a behavioral prediction had been reported by Collins (2018), who showed enhanced retention of items learned in set size 6 compared to set size 3. However, that study did not employ neural recordings and thus did not test whether the neural WMRL interaction was the underlying mechanism for these effects. Here we provide several lines of evidence in support of this claim.

First, our behavioral and EEG results replicated key findings in the RLWM task and in the subsequent memory tests. In the learning task, we observed worse acquisition with increasing set size and with delays between successive stimulus presentations, but as learning progressed (with the increase in reward history) the negative effect of delay in high set sizes diminished considerably. This observation further supports the model prediction that RL dominates over WM with the accumulation of rewards over time. Second, at the neural level, we also replicated findings in which neural RL indices preceded the cognitively costly WM process during stimulus processing (Collins and Frank, 2018). Moreover, we found robust evidence that EEG signals of RL increased more rapidly across trials under high than low load (Fig. 4F), a key prediction of the cooperative model (Fig. 2), even though behavioral learning was slower in these conditions.

Importantly, we observed that associations learned under higher WM load had increasingly higher recall accuracy in the stimulus-response retention test (Fig. 4C). This result extends the previously reported retention benefit of associations learned under high compared to low set sizes (Collins, 2018). We showed that this effect is parametric across five levels of WM load, and moreover that the greatest retention deficits occurred for the very lowest set sizes in which participants could easily learn the task purely via WM. Furthermore, we replicated previous results in the reward retention test (Collins et al., 2017) and demonstrated that participants have differential sensitivity to the proportion of trials in which they were rewarded for either of the stimuli and this effect grew with set size.

Finally, to gain a better understanding of the mechanism responsible for the benefits in both retention tests, we leveraged a within-trial neural indexing approach of EEG dynamics. We showed that neural indices of RL during acquisition were predictive of subsequent retention in the stimulus-response retention, even after controlling for set size. This result supports the key model prediction that RL processes during learning, which are stronger under high WM load, are responsible for increasing policy retention, when WM is no longer available. In contrast, neural indices of RL were not predictive of performance in the reward retention test.

This result supports theoretical and empirical studies suggesting that model-free learning in the brain (especially the corticostriatal system) directly learns a stimulus-response policy using prediction errors from another system (“actor-critic”; Collins & Frank 2014; Jaskir & Frank 2022; Klein et al 2017). By this account, the “actor” selecting policies would have no direct access to experienced reward values, but only the propensity for a specific response for each of them. Participants could plausibly access their “critic” values for each stimulus and compare them in the reward retention phase, but they would not have had to do so during learning. Indeed, participants show above chance performance in such discriminations, but only subtly (accuracy rises up to 60% at best); in contrast, accuracy in the stimulus-response retention test, which directly assesses what the actor would have learned, is far superior (roughly 80% for the higher set sizes), despite being tested with further delays since learning.

For most simple RL tasks, these two classes of model-free RL algorithms (those that focus on learning expected values and the actor-critic), are largely indistinguishable as they both predict that an agent progressively chooses those actions that maximize reward. However, several theoretical and empirical studies suggest that the basic RL system in humans satisfies predictions of an actor-critic in behavior, imaging, and in theoretical models of corticostriatal contributions to RL (Collins & Frank 2014; Jaskir & Frank 2022; Li & Daw 2011; Klein et al 2017; Gold et al 2012; Geana et al., 2021). Moreover, the model fits here did not improve if we allowed the Q learning agent to learn the difference between 2 vs 1 point, and instead suggested that participants learned to simply maximize task performance, which effectively makes Q learning equivalent to an actor-critic at the level of task performance. Nevertheless, a Q learner would, at minimum, learn the reward value of a stimulus in terms of the percentage of times they were correct (i.e., whether they got 1 or 2 points vs 0). Yet, the EEG marker of RL is still not related to performance in reward retention test even when correct performance there would be counted as simply choosing the stimulus that had yielded higher proportion of correct responses. While our neural RL index cannot distinguish between an EEG metric of “Q values”, or “actor weights”, the findings that it only predicts performance in the stimulus-response test provides initial evidence supporting the actor interpretation where the neural RL index reflects the policy rather than its reward value.

While we focussed mainly on how the RLWM mechanism informs retention, we also tested whether the interaction between RL and WM can be modulated by acute stress. Stress is known to have a major impact on learning and decision-making processes (Cremer et al., 2021; Raio, et al., 2017; Starcke & Brand, 2012). Previous work had shown that acute stress alters prefrontal cortex functioning thus impairing executive control over cognition (e.g., cognitive inhibition, task switching, working memory maintenance; Bogdanov & Schwabe, 2016; Brown et al., 2020; Hamilton &

Brigman, 2015; Goldfarb et al., 2017; Plessow et al., 2012; Schwabe & Wolf, 2011; Schwabe, et al., 2011; Vogel et al., 2016). On the other hand, acute stress was also shown to increase striatal dopamine activity (Vaessen et al., 2015) leading to better working-memory updating (Goldfarb et al., 2017) and improving executive control over motor actions (i.e., response inhibition; Leong and Packard, 2014; Schwabe & Wolf, 2012). We, therefore, predicted that stress would affect the WM vs. RL trade-off such that it will impede WM's contribution to learning and will instead enhance the relative contribution of RL computations. Current results did not confirm this hypothesis as only subtle differences were observed between the stress and control groups during the learning task and at the tests.

It is possible that the 25 minutes' delay between the stressor and the beginning of the learning task hindered the stress response on behavior as it was previously suggested that both noradrenaline and cortisol levels need to be elevated in order for stress to affect WM performance (Roozendaal, et al., 2006; Barsegyan et al., 2010; Elzinga & Roelofs, 2005). Another intriguing possibility is that individuals with higher WM capacity were more resilient against cognitive impairments induced by stress and were also less biased toward habitual decision-making (Cremer et al., 2021; Otto et al., 2013; Quaedflieg et al., 2019). Future work should test directly the specific effect of stress on WM and RL interactions while taking into account participants' WM capacity as a factor.

To conclude, our results contribute to a better understanding of the coupled mechanism of WM and RL that can dynamically shift between relying more on the effortful but fast and reliable WM system or the slow, more error-prone RL system that has retention benefits. We reported trial-by-trial evidence in the neural signal for this trade-off during learning and showed that greater reliance on the RL system when WM is degraded (i.e., when WM load is high) predicted better memory retention of learned stimulus-response associations. An intriguing possibility that remains to be tested is that the shift between the two systems is strategic and can be modulated by one's preference or ability to maximize immediate learning vs retention. However, it remains to be seen if clinical populations with impairments in one or both systems of WM and RL, might alter the flexible shifting between the two systems, possibly biasing the use of one system more than the other even when it is less advantageous.

**References**

1. Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature reviews neuroscience, 10*, 410-422.
2. Barsegyan, A., Mackenzie, S. M., Kurose, B. D., McGaugh, J. L., & Roozendaal, B. (2010). Glucocorticoids in the prefrontal cortex enhance memory consolidation and impair working memory by a common neural mechanism. *Proceedings of the National Academy of Sciences*, *107*, 16655-16660.
3. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823. doi: https://doi.org/10.48550/arXiv.1406.5823
4. Bogdanov, M., & Schwabe, L. (2016). Transcranial stimulation of the dorsolateral prefrontal cortex prevents stress-induced working memory deficits. *Journal of Neuroscience, 36*, 1429-1437.
5. Brown, T. I., Gagnon, S. A., & Wagner, A. D. (2020). Stress disrupts human hippocampal-prefrontal function during prospective spatial navigation and hinders flexible behavior. *Current Biology, 30*, 1-13.
6. Carvalheiro, J., Conceição, V. A., Mesquita, A., & Seara-Cardoso, A. (2021). Acute stress impairs reward learning in men. *Brain and Cognition*, *147*, 105657.
7. Collins, C. J., Yi, F., Dayuha, R., Duong, P., Horslen, S., Camarata, M., ... and Hahn, S. H. (2021). Direct measurement of ATP7B peptides is highly effective in the diagnosis of Wilson disease. *Gastroenterology*, *160*, 2367-2382.
8. Collins, A. G. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of cognitive neuroscience*, *30*, 1422-1432.
9. Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024-1035.
10. Collins, A. G., & Frank, M. J. (2018). Within-and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*, 2502-2507.
11. Collins, A. G., Ciullo, B., Frank, M. J., & Badre, D. (2017a). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*, 4332-4342.
12. Collins, A. G., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017b). Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biological psychiatry*, *82*, 431-439.
13. Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*, 13747-13756.
14. Cremer, A., Kalbe, F., Gläscher, J., & Schwabe, L. (2021). Stress reduces both model-based and model-free neural computations during flexible learning. *NeuroImage, 229,* 117747.
15. Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods, 134*, 9-21.
16. Elzinga, B. M., & Roelofs, K. (2005). Cortisol-induced impairments of working memory require acute sympathetic activation. *Behavioral neuroscience, 119,* 98.
17. Frank, M. J., Samanta, J., Moustafa, A. A., & Sherman, S. J. (2007). Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science, 318*, 1309-1312.
18. Hamilton, D. A., & Brigman, J. L. (2015). Behavioral flexibility in rats and mice: contributions of distinct frontocortical regions. *Genes, Brain and Behavior, 14*, 4-21.

19. Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald III, A. W., Ragland, J. D., Silverstein S. M., & Frank, M. J. (2022). Using Computational Modeling to Capture Schizophrenia-Specific Reinforcement Learning Differences and Their Implications on Patient Classification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 7*, 1035-1046.
20. Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S., Collins A.G.E., & Frank, M. J. (2012). Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. *Archives of general psychiatry, 69*, 129-138.
21. Goldfarb EV, Froböse, MI, Cools R, & Phelps EA (2017). Stress and cognitive flexibility: cortisol increases are associated with enhanced updating but impaired switching. *Journal of Cognitive Neuroscience, 29*,14-24
22. Jafarpour, A., Buffalo, E. A., Knight, R. T., & Collins, A. G. (2022). Event segmentation reveals working memory forgetting rate. *Iscience*, *25*, 103902.
23. Jaskir, A., & Frank, M. J. (2022). On the normative advantages of dopamine and striatal opponency for learning and choice. bioRxiv 483879. https://doi.org/10.1101/2022.03.10.483879.
24. Kim, J., Lee, H., Han, J., & Packard, M. (2001). Amygdala is critical for stress-induced modulation of hippocampal long-term potentiation and learning. *Journal of neuroscience, 21,* 5222-5228.
25. Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature communications, 8*, 1-12.
26. Leong, K. C., & Packard, M. G. (2014). Exposure to predator odor influences the relative use of multiple memory systems: role of basolateral amygdala. *Neurobiology of Learning and Memory*, *109*, 56-61.
27. Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience, 31*, 5504-5511.
28. Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in human neuroscience*, *8*, 213.
29. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305-315.
30. Meier, J. K., Staresina, B. P., & Schwabe, L. (2022). Stress diminishes outcome but enhances response representations during instrumental learning. *Elife, 11*, e67517.
31. Oberauer, K., Farrell, S., Jarrold, C., and Lewandowsky, S. (2016). What limits working memory capacity?. *Psychological bulletin, 142*, 758.
32. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences, 110,* 20941-20946.
33. Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature communications, 6*, 1-14.
34. Plessow, F., Kiesel, A., & Kirschbaum, C. (2012). The stressed prefrontal cortex and goal-directed behaviour: acute psychosocial stress impairs the flexible implementation of task goals. *Experimental brain research, 216,* 397-408.
35. Quaedflieg, C. W. E. M., Stoffregen, H., Sebalo, I., & Smeets, T. (2019). Stress-induced impairment in goal-directed instrumental behaviour is moderated by baseline working memory. *Neurobiology of learning and memory, 158,* 42-49

36. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

37. Rac-Lubashevsky, R., & Frank, M. J. (2021). Analogous computations in working memory input, output and motor gating: Electrophysiological and computational modeling evidence. *PLoS computational biology, 17*, e1008971.

38. Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J., & Phelps, E. A. (2017). Stress attenuates the flexible updating of aversive value. *Proceedings of the National Academy of Sciences, 114*, 11241-11246.

39. Roozendaal, B., Okuda, S., De Quervain, D. F., & McGaugh, J. L. (2006). Glucocorticoids interact with emotion-induced noradrenergic activation in influencing different memory functions. *Neuroscience, 138,* 901-910.

40. Schwabe, L., & Schächinger, H. (2018). Ten years of research with the Socially Evaluated Cold Pressor Test: Data from the past and guidelines for the future. *Psychoneuroendocrinology, 92,* 155-161.

41. Schwabe, L., & Wolf, O. T. (2009). Stress prompts habit behavior in humans. *The Journal of Neuroscience, 29,* 7191-7198.

42. Schwabe, L., & Wolf, O. T. (2011). Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behavioural brain research, 219*, 321-328.

43. Schwabe, L., & Wolf, O. T. (2012). Stress modulates the engagement of multiple memory systems in classification learning. *The Journal of Neuroscience, 32,* 11042-11049.

44. Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*, 890-895.

45. Schwabe, L., Höffken, O., Tegenthoff, M., & Wolf, O. T. (2011). Preventing the stress-induced shift from goal-directed to habit action with a β-adrenergic antagonist. *Journal of Neuroscience, 31,* 17317-17325.

46. Simon-Kutscher, K., Wanke, N., Hiller, C., & Schwabe, L. (2019). Fear without context: acute stress modulates the balance of cue-dependent and contextual fear learning. *Psychological Science, 30,* 1123-1135.

47. Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and Biobehavioral Reviews, 36*, 1228-1248.

48. Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1994). Testtheoretische Analysen der Mehrdimensionalen Befindlichkeitsfragebogens (MDBF) [Test-theoretical analyses of the Multidimensional Mood State Questionnaire]. *Diagnostica, 40,* 320-328.

49. Vaessen, T., Hernaus, D., Myin-Germeys, I., & van Amelsvoort, T. (2015). The dopaminergic response to acute stress in health and psychopathology: a systematic review. *Neuroscience and Biobehavioral Reviews*, *56*, 241-251.

50. Vogel, S., Fernández, G., Joëls, M., & Schwabe, L. (2016). Cognitive adaptation under stress: a case for the mineralocorticoid receptor. *Trends in cognitive sciences, 20*, 192-203.

51. Wimmer, G. E., & Poldrack, R. A. (2022). Reward learning and working memory: Effects of massed versus spaced training and post-learning delay period. *Memory & cognition, 50,* 312-324.

52. Wirz, L., Bogdanov, M., & Schwabe, L. (2018). Habits under stress: mechanistic insights across different types of learning. *Current Opinion in Behavioral Sciences, 20,* 9-16.


**Figure 1.** Experimental protocol of the learning task and the two test phases. (A) In the learning phase, in each block participants use deterministic reward feedback to learn which of three actions to select

for each stimulus image. The set size (or the number of stimuli; ns) varies from one to five across blocks. After each response feedback was presented audio-visually (see text for more detail). (B) The surprise reward-retention test protocol. In this task, participants are asked to recall the reward value of stimuli learned during the learning phase by choosing the stimulus they perceive to have been more rewarded within a pair of stimuli presented on every trial. (C) The surprise stimulus-response retention test protocol is a test of the learned stimulus-response "policy". Here, participants are asked to recall the correct action for the probed stimulus. No feedback was given at either test phase.

**Figure 2.** Cooperative interaction between the RL and WM systems (adapted from Collins and Frank, 2018): A. Both WM and RL inform expected Q values and thus inform reward prediction errors (RPEs). When the number of stimuli to learn (ssz or "set size") is within WM capacity (e.g., ssz=2 on the left) the expected Q value of each contingency can be held in WM, thereby reducing RPE's during early learning compared to those that would occur from RL alone. When set size exceeds WM capacity (e.g., ssz=5 on the right), degraded WM results in larger RPEs. B. Computational model simulations (recreated from Collins and Frank, 2018) capture the RL and WM interaction, showing that larger RPEs persist for longer when WM load is taxed (high ssz), thereby accumulating expected Q values in the RL system. C. Note that Q learning curves in panel B evolve more rapidly in high ssz, despite the opposite pattern in simulated behavioral learning curves (whereby WM contributes to rapid learning in low ssz.

**Figure 3.** Behavioral results from the learning phase. (A-B) Performance learning curves and reaction times (RT) for each set size as a function of the number of iterations of a stimulus (stim). (C) Performance as a function of WM load, the detrimental effect of delay is greater in high set sizes. (D-E) Reduced effects of both delay and set size as learning progresses from early (up to two previous correct choices) to late (the last two trials of each stimulus) trials in a block, suggestive of a transition from WM to RL.

**Figure 4.** Behavior performance at the test phase. (A) Effect of value difference and set size on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of differences in the number of experienced rewards (Q value diff) and the set size in which they were learned. The median split of absolute value differences is shown (high-Q value difference trials depicted in red and low-Q value difference trials in blue). (B-C) Effect of set size on the stimulus-response retention test performance. The proportion of correct recall in the test phase increases as a function of the estimated Q values of the probed association and as a function of the set size in which it was learned. The median split of the estimated stimulus-response Q values is shown (high Q value associations in red and low Q value associations in blue). (D) Effect of EEG RL index on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of the set size in which they were learned but was not further modulated by the magnitude of the EEG RL index of the stimuli. The median split of absolute differences in EEG RL indices is shown (high-EEG RL index difference in red and low-EEG RL index difference in blue). (E) Effect of the neural RL index on recall accuracy in the stimulus-response retention test. The neural RL index is shown as the median split across all the RL indices. Stimuli with high RL index are depicted in red and stimuli with low RL index are depicted in blue. (F) The EEG RL index increases parametrically with the increase in accumulated rewards. These neural learning curves parametrically increase with set size. Error bars represent standard errors.

**Figure 5.** EEG decoding of RL and WM effects during choice. Corrected event-related potentials (ERPs) exhibiting the effect of three main predictors (set-size in green, delay in blue, RL value quartiles in red; from top to bottom row) on the voltage of significant electrodes (FCz, CPz, and Poz for set size and delay, and FCz, CPz, and C3 for RL). The black line reflects the significant time points after permutation correction. On the right, the effect of each predictor in the row is exhibited with a scalp map topography at an early (300ms) and late (540ms) time points. The color in the scalp map represents significant thresholded t-values.

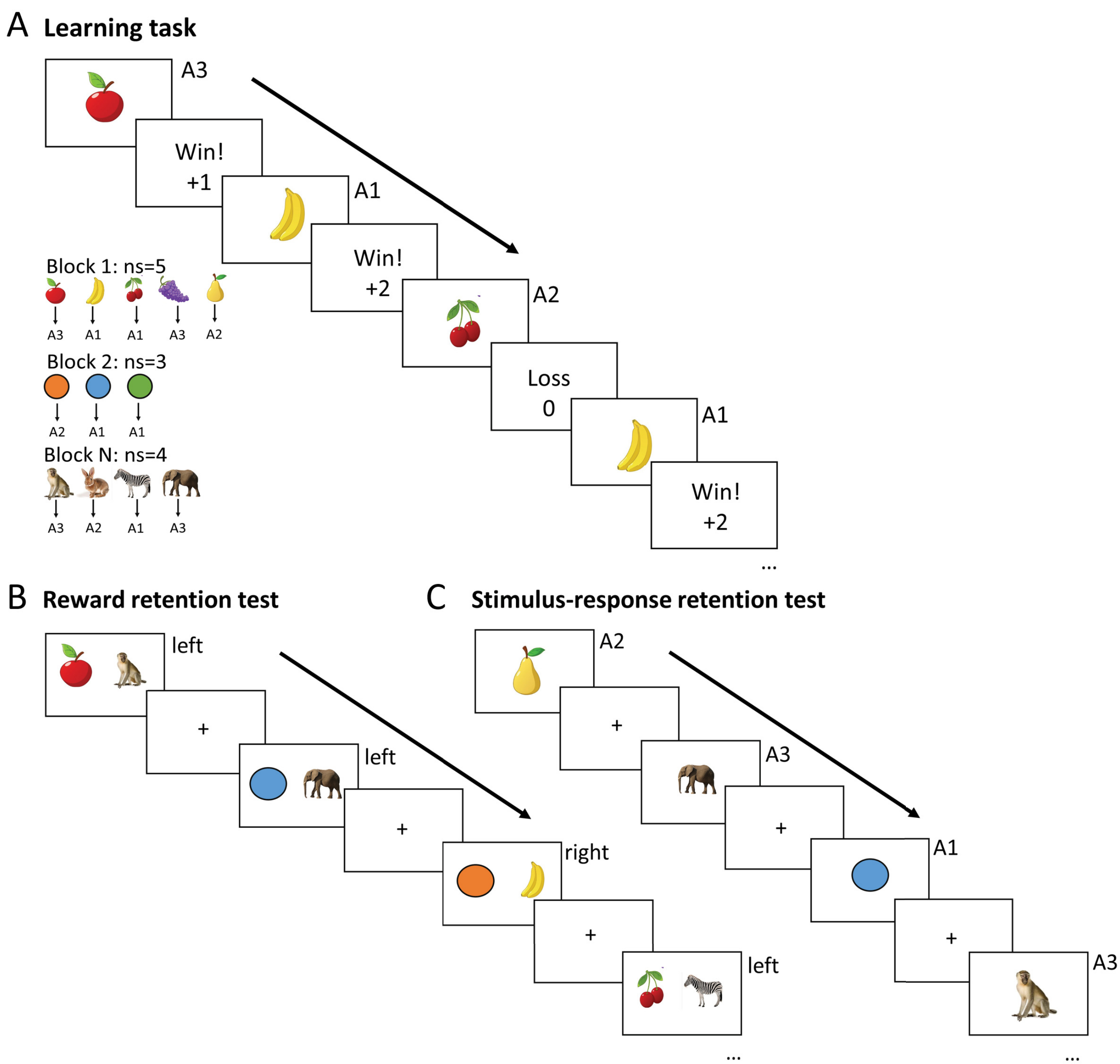
**Figure 6.** Successful stress induction. The exposure to the stressor led to significant increases in (A) systolic blood pressure, (B) diastolic blood pressure, (C) heart rate, and (D) salivary cortisol levels; error bars represent standard errors. The control group is depicted in dark blue and the stress group in red. $^{**}p < 0.01$, $^{***}p < 0.001$ for the comparison between the stress group and the control group.

**Figure 7.** Stress effects during the learning and test phases. (A) Learning curves across iterations as a function of set size in the control group (B) and stress group. (C) Learning curves across the number of previous correct as a function of delay (1 to 5 where 5 reflects delay of five and above) in the control group (D) and stress group. (E) Effect of stress on the reward retention test performance. The proportion of correct selection of the more rewarding stimulus from a pair of the probed stimuli increases as a function of the set size in both the control group (depicted in black) and in the stress group (depicted in red). (F) Effect of stress on recall accuracy in the stimulus-response retention test. The proportion of correct recall in the stimulus-response test increases as a function of the set size in both the control group (depicted in black) and the stress group (depicted in red). Error bars represent standard errors.

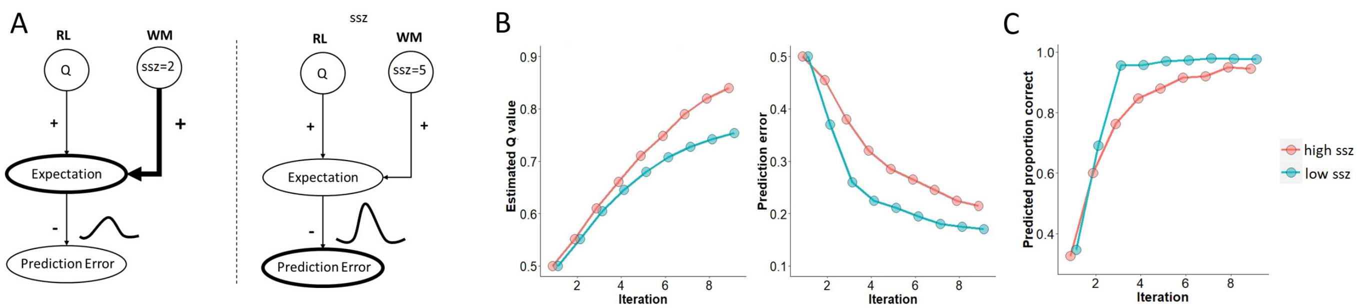
**Table 1.** Subjective mood and procedure ratings across the experiment in both control and stress groups. The mean and standard deviation of the ratings before and after the procedures are reported for the control group (upper part) and for the stress group (bottom part).

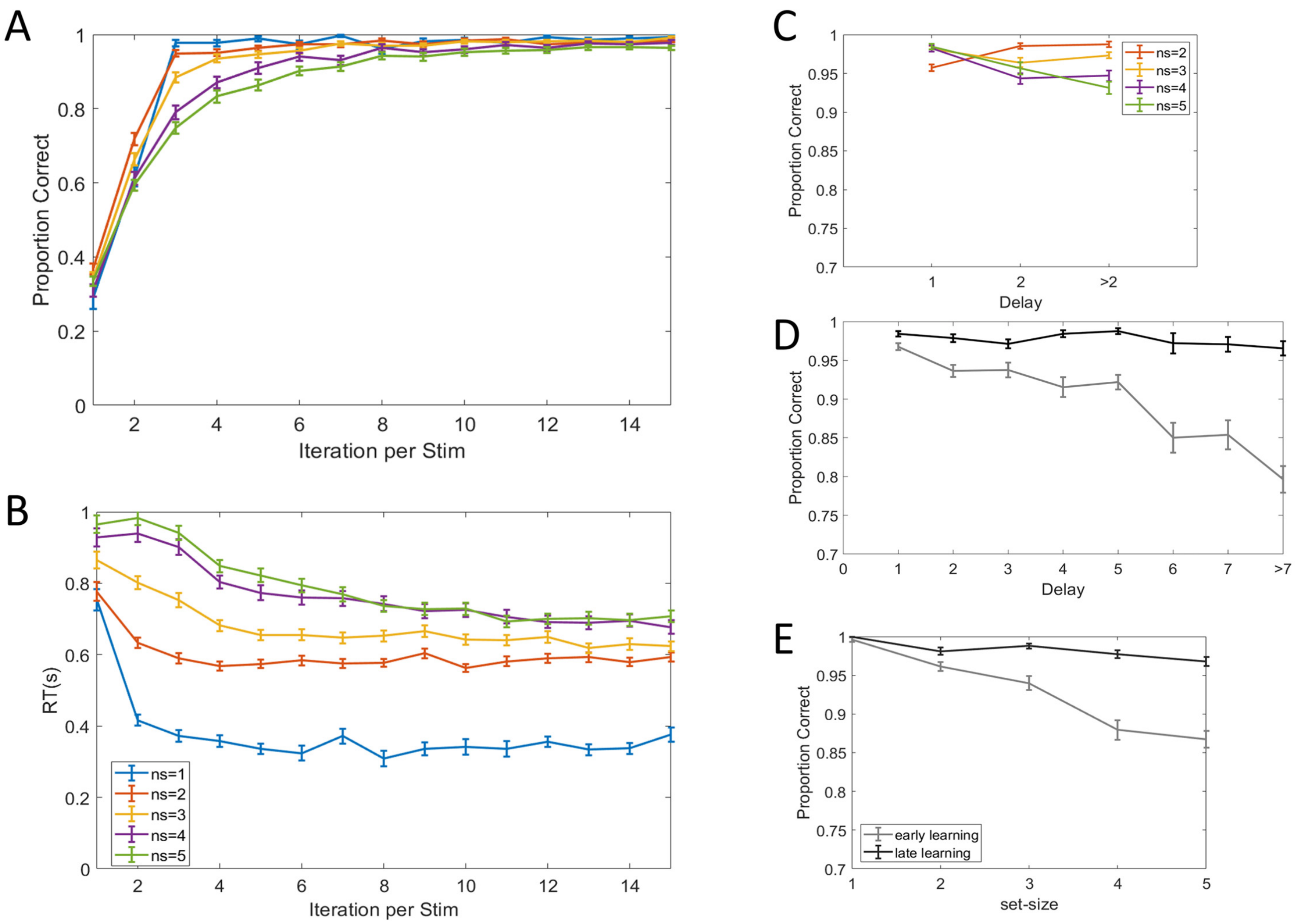
| | **Control group** | | |
|---|---|---|---|
| | Before | After | End of testing day |
| **Subjective mood** | | | |
| Depressed mood vs. elevated mood | 33.69 (4.99) | 34.26 (4.72) | 33.86 (4.66) |
| Restlessness vs. calmness | 32.476 (6.08) | 33.83 (5.14) | 33.24 (4.61) |
| Sleepiness vs. wakefulness | 28.571 (6.48) | 28.31 (6.88) | 26.64 (6.78) |
| **Rating of control procedure** | | | |
| difficult | - | 4.09 (13.21) | - |
| unpleasant | - | 9.52 (21.88) | - |
| stressful | - | 4.20 (15.23) | - |
| painful | - | 3.79 (14.62) | - |

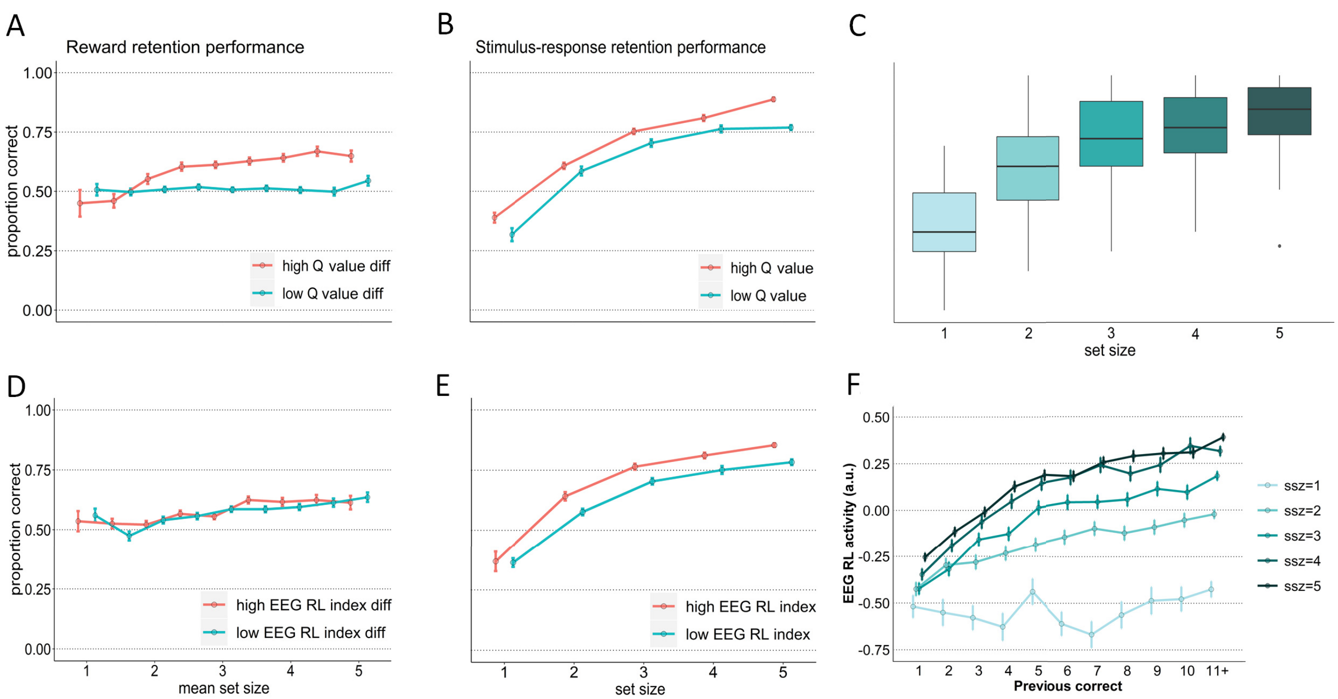
| | **Stress group** | | |
|---|---|---|---|
| | Before | After | End of testing day |
| **Subjective mood** | | | |
| Depressed mood vs. elevated mood | 33.76 (3.51) | 31.57 (5.32) | 33.43 (3.99) |
| Restlessness vs. calmness | 32.99 (4.24) | 30.45 (6.14) | 32.43 (4.72) |
| Sleepiness vs. wakefulness | 28.98 (5.71) | 29.86 (6.16) | 26.45 (6.12) |
| **Rating of stressor** | | | |
| difficult | - | 50.69 (28.01) | - |
| unpleasant | - | 58.73 (28.09) | - |
| stressful | - | 40.17 (26.70) | - |
| painful | - | 55.40 (25.97) | - |

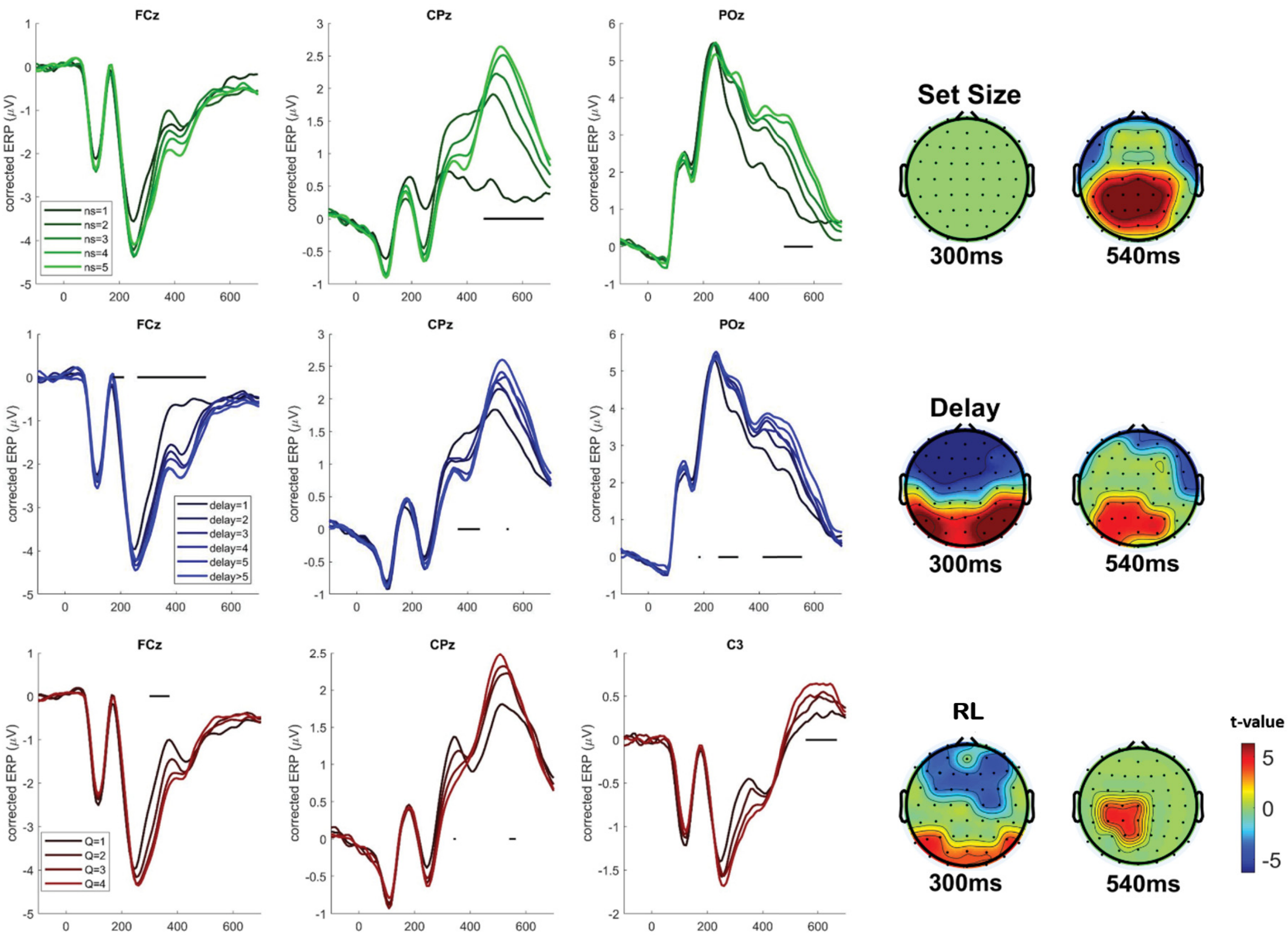


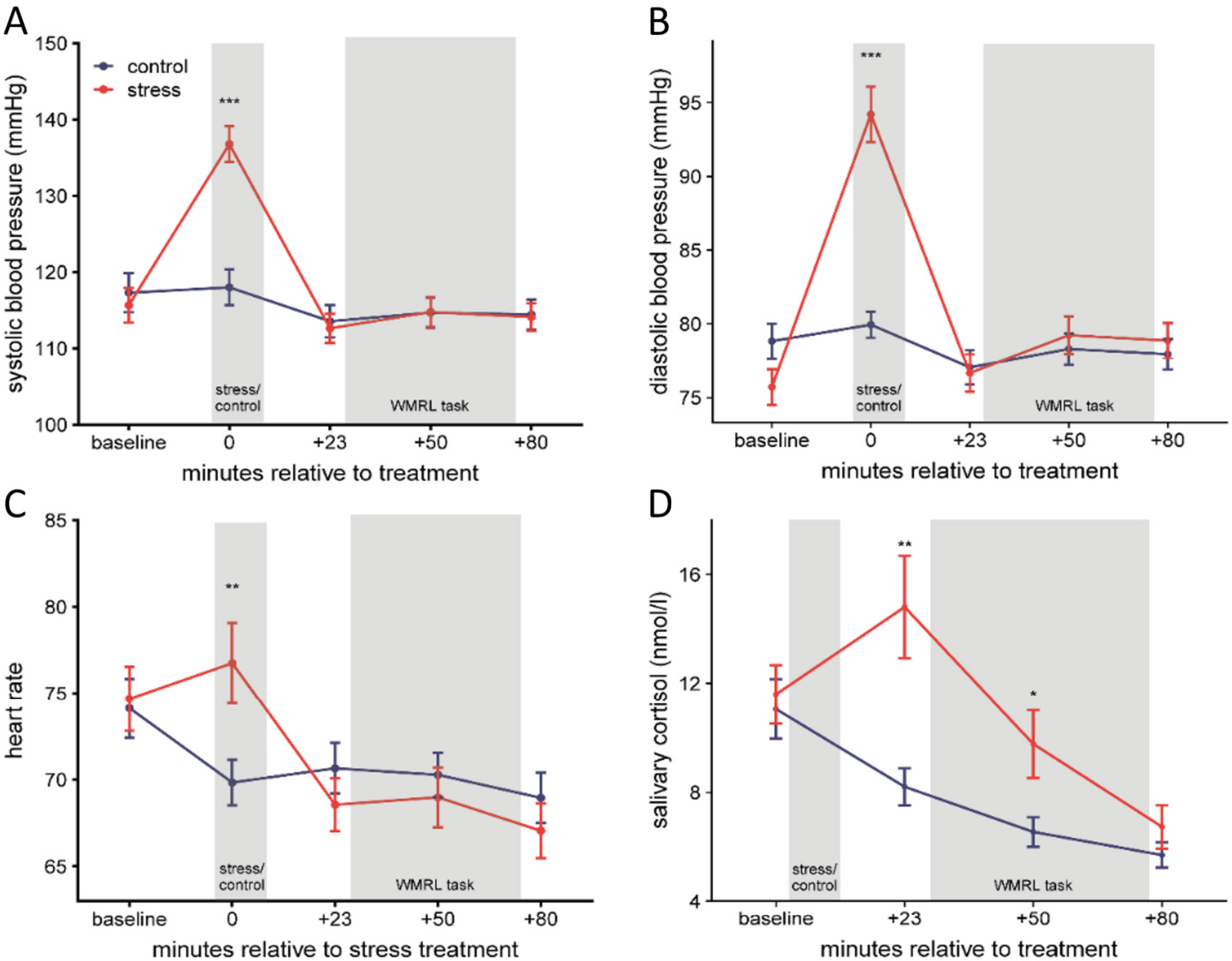
A Learning task
A3
Win!
+1
A1
Win!
+2
A2
Loss
0
A1
Win!
+2
...
Block 1: ns=5
A3 A1 A1 A3 A2
Block 2: ns=3
A2 A1 A1
Block N: ns=4
A3 A2 A1 A3
B Reward retention test
left
+
left
+
right
+
left
...
C Stimulus-response retention test
A2
+
A3
+
A1
+
A3
...

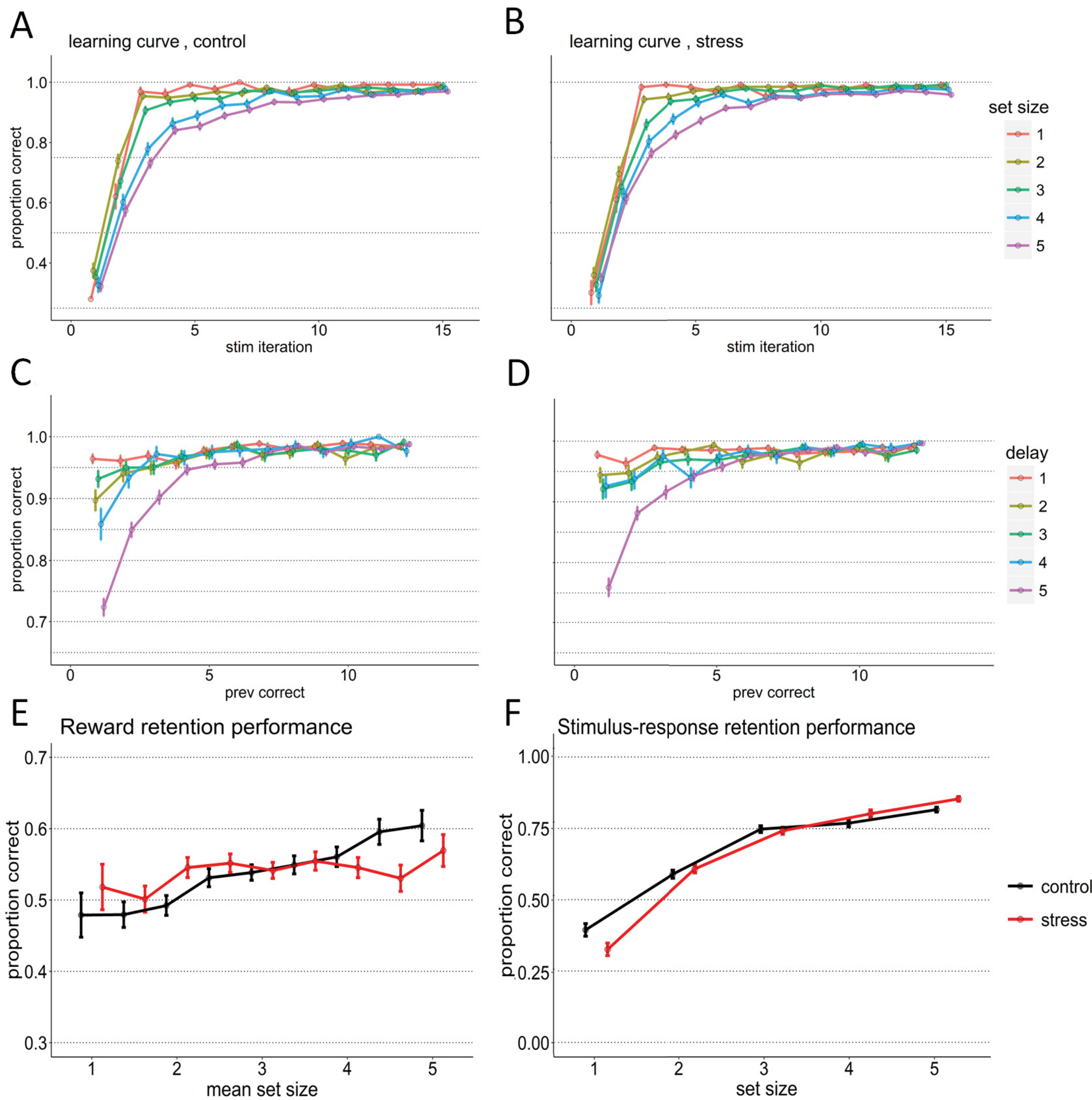

A
RL
WM
Q
ssz=2
+
+
Expectation
-
Prediction Error
ssz
RL
WM
Q
ssz=5
+
+
Expectation
-
Prediction Error
B
Estimated Q value
Iteration
Prediction error
Iteration
C
Predicted proportion correct
Iteration
high ssz
low ssz


A
Proportion Correct
Iteration per Stim
B
RT(s)
Iteration per Stim
ns=1
ns=2
ns=3
ns=4
ns=5
C
Proportion Correct
Delay
ns=2
ns=3
ns=4
ns=5
D
Proportion Correct
Delay
E
Proportion Correct
set-size
early learning
late learning


A
Reward retention performance
B
Stimulus-response retention performance
C
proportion correct
high Q value diff
low Q value diff
high Q value
low Q value
set size
D
E
F
high EEG RL index diff
low EEG RL index diff
high EEG RL index
low EEG RL index
mean set size
set size
EEG RL activity (a.u.)
Previous correct
ssz=1
ssz=2
ssz=3
ssz=4
ssz=5


FCz
CPz
POz
C3
corrected ERP (μV)
ns=1
ns=2
ns=3
ns=4
ns=5
delay=1
delay=2
delay=3
delay=4
delay=5
delay>5
Q=1
Q=2
Q=3
Q=4
Set Size
Delay
RL
300ms
540ms
t-value
5
0
-5


A
control
stress
***
systolic blood pressure (mmHg)
150
140
130
120
110
100
stress/
control
WMRL task
baseline
0
+23
+50
+80
minutes relative to treatment
B
***
diastolic blood pressure (mmHg)
95
90
85
80
75
stress/
control
WMRL task
baseline
0
+23
+50
+80
minutes relative to treatment
C
**
heart rate
85
80
75
70
65
stress/
control
WMRL task
baseline
0
+23
+50
+80
minutes relative to stress treatment
D
**
*
salivary cortisol (nmol/l)
16
12
8
4
stress/
control
WMRL task
baseline
+23
+50
+80
minutes relative to treatment


A
learning curve , control
proportion correct
stim iteration
B
learning curve , stress
stim iteration
set size
1
2
3
4
5
C
proportion correct
prev correct
D
prev correct
delay
1
2
3
4
5
E
Reward retention performance
proportion correct
mean set size
F
Stimulus-response retention performance
proportion correct
set size
control
stress