Title

The Dynamics of Feedback-based Learning is Modulated by Working Memory Capacity

Author names and affiliations

Jil Humann[1*], Adrian G. Fischer[2,3,4*], Markus Ullsperger[1,2,3]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Netherlands;

[2] Otto-von-Guericke-Universität, Magdeburg, Germany

[3] Center for Behavioral Brain Sciences, Magdeburg, Germany

[4] Freie Universität Berlin, Berlin, Germany

[*] Equal contribution.

Corresponding author

Prof. Dr. Markus Ullsperger

Otto-von-Guericke University Magdeburg

Institute of Psychology

Universitätsplatz 2, 39106 Magdeburg, Germany

Telephone: +49 391 6758475

Email: markus.ullsperger@ovgu.de

Running head: Value-based learning and working memory capacity

Abstract

Research suggests that working memory (WM) has an important role in instrumental learning in changeable environments when reinforcement histories of multiple options must be tracked. Working memory capacity (WMC) not only reflects the ability to maintain items, but also to update and shield items against interference in a context-dependent manner; functions conceivably also essential to instrumental learning. To address the relationship of WMC and instrumental learning, we studied choice behavior and EEG of participants performing a probabilistic reversal learning task. Their separately measured WMC positively correlated with reversal learning performance. Computational modeling revealed that low-capacity participants modulated learning rates less dynamically around value reversals. Their choices were more stochastic and less guided by learnt values, resulting in less stable performance and higher susceptibility to misleading probabilistic feedback. Single-trial model-based EEG analysis revealed that prediction errors and learning rates were less strongly represented in cortical activity of low-capacity participants, while the centroparietal positivity, a general correlate of adaptation, was independent of WMC. In conclusion, cognitive functions tackled by WMC tasks are also necessary in instrumental learning. We suggest that noisier representations render items held in WM as well as tracked values in instrumental learning less stable and more susceptible to distractors.

Keywords: feedback-related negativity, EEG, model-based analysis, working memory, reinforcement learning,

Decisions on how to (re-)act on certain stimuli is based on learnt values of stimulus-action contingencies. Reinforcement learning (RL) theory (Rescorla and Wagner 1972; Sutton and Barto 1998) has been successfully applied to the study of neuronal underpinnings of value-based decision making. In real-world scenarios, stimulus-action-outcome contingencies are subject to change which induces uncertainty. This requires continuous tracking of outcomes to previous choices (reinforcement history) and adaptive value updating. The influence of unpredicted outcomes on future decisions, reflected in the learning rate (LR) in RL models, must be adjusted to reinforcement history (Behrens et al. 2007; Jocham et al. 2009; Krugel et al. 2009; McGuire et al. 2014). RL parameters determining value updates, LR and reward prediction error (RPE), have been shown to be encoded in the posterior medial frontal cortex (Behrens *et al.* 2007; Jocham *et al.* 2009; McGuire *et al.* 2014).

Tracking values of multiple and variable choice options involves several cortical and subcortical brain structures and learning mechanisms (Fusi et al. 2007; Boorman et al. 2011; Fischer et al. 2017; Meder et al. 2017). It is conceivable that working memory (WM) may contribute to instrumental learning in complex environments, but this notion has received little attention so far. The number of to-be-learned stimulus-action contingencies influences performance in instrumental learning tasks. This finding has been captured by adding working memory components to a RL model (Collins and Frank 2012). An intimate relation between working memory capacity (WMC) and RL also appears plausible given that cortical areas recruited during instrumental learning show considerable overlap with activations during working memory tasks (Amiez and Petrides 2007; Klein et al. 2007; Constantinidis and Klingberg 2016) and that both functions are influenced by prefrontal and striatal dopamine signaling (Pessiglione et al. 2006; Cools et al. 2008; Krugel *et al.* 2009; Jocham et al. 2011, 2014). Conventional WMC tests yield span measures influenced by limits of simultaneously maintaining multiple items in working memory, updating of its contents, and the ability to

focus attention and to shield maintained content from distraction (Unsworth and Engle 2007; Adam et al. 2015; Unsworth and Robison 2016; Adam et al. 2017). At least descriptively, processes tackled by WMC tasks appear similar to tracking stimulus and action values in instrumental learning, i.e. maintaining them across multiple trials and updating them whenever unexpected outcomes suggest a value change.

Here, we investigate the relationship between WMC and instrumental learning. We applied a reversal learning task with probabilistic outcomes (Figure 1) well-suited to study dynamic LR adjustments around stimulus value reversals (Krugel *et al.* 2009; Li et al. 2011) and recorded EEG. For optimal performance, participants need to find a balance between maintaining learnt stimulus values when they are stable and quickly updating them upon a reversal. Thus, when task rules are stable LR should be low such that probabilistic misleading feedback can be ignored. Upon reversals, LR must be transiently upregulated to enable fast updating of stimulus values. We hypothesized that a low WMC coincides with suboptimal LR regulation. Low-capacity participants might be more prone to being misled by false probabilistic feedback and generally more distractible. This should be reflected in a higher average LR but reduced LR dynamics around contingency reversals in low compared to high WMC participants. Moreover, low-capacity participants' choices might be more stochastic and less guided by learnt stimulus values.

Using single-trial regression on the EEG data, we investigated the dynamics of outcome processing. We focused on EEG correlates of RL parameters: the feedback-related negativity (FRN), the frontocentral P3a, and the centroparietal P3b (Fischer and Ullsperger 2013).The FRN is likely generated in the posterior medial frontal cortex and covaries with signed RPEs (Walsh and Anderson 2012; Fischer and Ullsperger 2013) but also with surprise (Talmi et al. 2013). P3a and P3b are mostly outcome-driven and additionally modulated by the LR. High LR, reflecting strong impact of outcomes on future choices, induces a sustained positive shift

of central feedback-locked EEG activity (Fischer and Ullsperger 2013). The P3b has been suggested to reflect a final pathway of adaptation (Ullsperger et al. 2014), specifically an evidence-accumulation-to-bound process in decision making (O'Connell et al. 2012). We tested whether individual WMC predicts the strength and pattern of representation of RL parameters in feedback-related EEG dynamics.

Materials and Methods

Participants

58 healthy right-handed participants were recruited in Magdeburg, Germany. The data of seven participants were excluded from further analysis for various reasons: six performed below threshold on the math part of the working memory task (less than 85% accuracy or > 11 errors) and there were technical problems during EEG recording in one subject. The final sample thus comprises data from 51 participants (23 female; mean ± SD age, 26.2 ± 4.9 years). None of the participants had a history of psychiatric or neurological conditions and all had normal or corrected-to-normal vision. Oral and written information was provided to the subjects prior to the experiment and informed consent was obtained. All volunteers were naïve to the task and received payment for participation. The study was in accordance with the standards set by the Declaration of Helsinki (Edinburgh amendments) and approved by the institutional review board of the Donders Centre for Cognition.

Working memory capacity task

We used the Automated Operation Span task (AOSPAN; Unsworth et al. 2005) to assess participants' working memory capacity (WMC). The task was translated to German and the code for E-prime 2.0.10 (Psychology Software Tools, Pittsburgh, PA) was adapted accordingly by T. Radüntz (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Berlin, 2012). It requires participants to remember a series of letters (set size 3-7) while solving math problems in between presentation of subsequent letters. AOSPAN scores were calculated as the "All or nothing score" (Conway et al. 2005) with 70 points being the maximally reachable score.

Probabilistic Reversal learning Task

In the probabilistic reversal learning task (Figure 1), participants were presented one of three

different stimuli at a time and decided to either choose or avoid gambling with that stimulus with the goal to maximize the final reward (similar to Fischer and Ullsperger 2013). A gamble resulted in monetary gain or loss, depending on reward contingencies associated with the particular stimulus. Avoiding to gamble removed all financial consequences of that trial (no gains or losses) but participants still received feedback about the counterfactual or fictive outcome of the potential gamble. The three stimuli – white line drawings of a camel, a chicken and a bear on black background - were presented in a pseudo random series that was the same for all participants. Reward contingencies for every stimulus could be 20%, 50% or 80% and stayed constant within one block of 28-42 trials. After every block, reward contingency changed without notice. The experiment consisted of 7 blocks per stimulus, leading to 18 reversals and 714 trials in total. Presentation 10.3 (Neurobehavioral Systems) was used for task presentation. Every trial of the task began with a central fixation cross, presented for a variable time between 300 and 500 ms. After fixation, the stimulus was presented together with the two choice alternatives (a green checkmark for choosing and a red no-go sign for avoiding, sides counterbalanced across subjects) for a maximum of 2000 ms or until a response was given. If participants failed to respond in time, a question mark was shown and the trial was repeated at the end of the block. When a response was made, the stimulus stayed on screen and feedback was given after 500 ms. The outcome was then presented for 750 ms depending on the subject's choice. Choosing to gamble led to either a green smiling smiley and a reward of 10 points or a red frowning smiley and a loss of 10 points according to the reward probability of the stimulus. An avoided gamble had no monetary consequences: the outcome was always 0. Counterfactual/fictive outcomes, indicating what would have happened had the participant chosen to gamble, were shown on screen using the same smileys, but the reward or punishment was crossed out to indicate that the outcome was fictive.

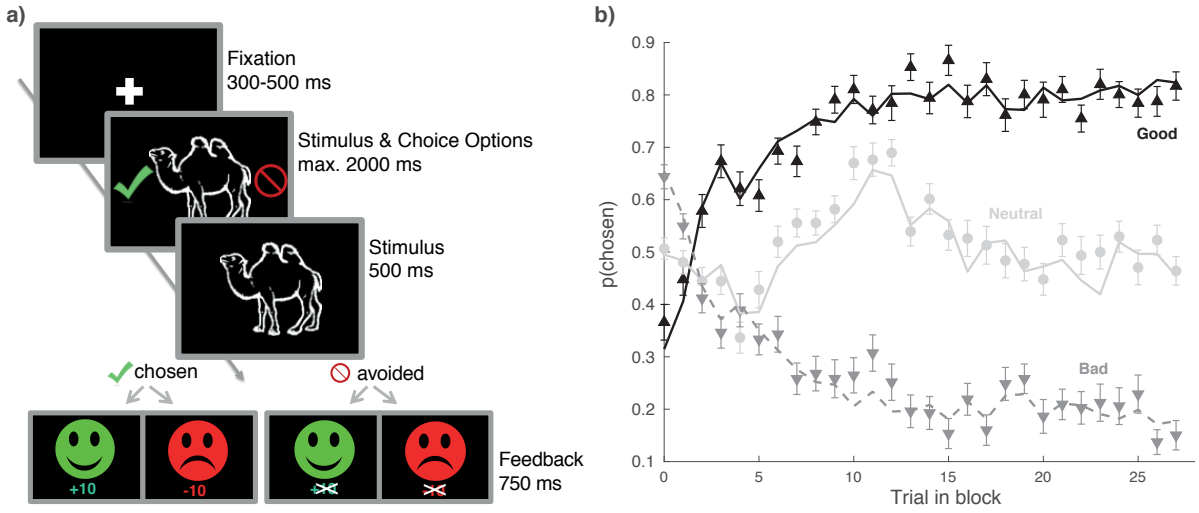

Figure 1. Schematic of the probabilistic reversal learning task and behavioral results. a) Timing of a single task trial. b) Observed (Symbols ± SE) and modeled behavior (lines) during the reversal learning task for different reward contingencies. Learning curves are comparable for good (black upward triangles; black solid line) and bad (dark gray downward triangles, dashed line) stimuli and approach their asymptote at the real reward probability (80 and 20%, respectively). Light gray line and circles: model and behavior for neutral stimuli.

EEG Acquisition and Analysis

EEG was recorded from 64 channels, positioned according to the extended 10/20 system in an elastic cap (Easycap) at a sampling rate of 500 Hz with BrainAmp MR plus amplifiers (Brain Products, Gilching, Germany). Impedances were kept below 5 kΩ. Furthermore, EOG electrodes were positioned above and below the right eye and on the outer canthus of each eye.

Raw data were then analysed offline using custom routines in MATLAB (MathWorks) & EEGLAB (Delorme and Makeig 2004). First, we band-pass filtered the data from 0.5 – 50 Hz and re-referenced offline to common average. Data were then segmented into epochs spanning from -1.5 to 1.5 s around feedback onset and epochs containing deviations greater than 5 SD of the mean probability distribution in every single channel or the whole montage

were automatically rejected. Epoched data were then demeaned and subjected to temporal infomax independent component analysis (ICA) integrated in EEGLAB and independent components reflecting eye blinks or horizontal eye movements were manually removed. Average EEG activity from -250 to -50 ms before feedback presentation was used as the baseline and subtracted from each channel individually.

Computational Model

Reinforcement learning models using variable rather than fixed learning rates have been found to better explain participant's behaviour (Krugel *et al.* 2009). We therefore fitted a hybrid Rescorla-Wagner Pearce-Hall model (Li *et al.* 2011) with a dynamic learning rate to subject's choices to obtain single-trial estimates of reward prediction errors (*RPEs*, δ) and the associability / learning rate (*LR*, α). The RPE is calculated as the difference between actual current reward magnitude R_t and expected value V_t for a trial's stimulus X_t :

$$\delta_t = R_t - V_t(X_t) \tag{1}$$

The *LR* α for a stimulus is updated on the basis of the absolute *RPE* (reflecting surprise) weighted by η and the counterweighted current *LR* of that stimulus:

$$\alpha_{t+1}(X_t) = \eta \times |\delta_t| + (1 - \eta) \times \alpha_t(X_t) \tag{2}$$

The sum of the stimulus' current value and the *RPE* weighted by the *LR* varying from trial to trial and a constant κ determine the expected value of the next trial of a given stimulus:

$$V_{t+1}(X_t) = V_t(X_t) + \kappa \times \alpha_t(X_t) \times \delta_t \tag{3}$$

Thus, in this hybrid model, the weighting factor η determines the extent to which value updates are driven by surprise or rather by the classical Rescorla-Wagner algorithm with a constant *LR*. The model's likelihood to either choose (P_c) or avoid (P_a) the current stimulus was determined according to the softmax rule of its current value V_t :

$$P_{c,t} = \frac{1}{\left(1 + e^{\left(\frac{-V_t + 0.5}{\beta}\right)}\right)}, and\ P_{a,t} = 1 - P_{c,t} \quad (4)$$

The temperature β indicates the degree to which choices are stochastic (high values) or deterministically driven (low values) by the value of V_t. The initial LR (α_1) at the beginning of the task was an additional free parameter, whereas all consecutive updates are governed by the hybrid RL algorithm. The parameters α_1, η and κ were constrained between 0 and 1, and β to be larger than zero and all parameters were non-linearly transformed to allow gaussian estimation of their distributions.

We fit the model using expectation-maximization (Huys et al. 2011) which recursively estimates group level distributions of individual subjects' maximum likelihood estimated parameters. Following every iteration, the group distribution for each parameter is estimated using Laplace approximation and used as a prior on the next step until convergence is reached, i.e., the group posterior log-likelihood does no longer change. A benefit of this approach is that the addition of group distributions into the fit reduces outlier parameter fits. Please note that we fit one distribution across all participants to avoid confounding high and low working memory span groups by the fit procedure.

As a Benchmark, we compared this model to a standard Rescorla-Wagner reinforcement learning algorithm with only two free parameters (α, β). To account for the increased complexity of the hybrid model, we compare both models using Bayesian model comparison via computation of integrated BIC (iBIC) values at group level (Huys *et al.* 2011). iBIC can be interpreted as a measure of how well a model describes the data penalized by how complex it is and smaller numbers indicate better and more parsimonious explanations for the data. Differences above 20 points of iBIC scores indicate very strong evidence in favor of one model (Kass and Raftery 1995). We found that the hybrid RL model clearly outperformed the standard RL algorithm. iBIC scores were 36689 for the hybrid and 36739 for the RL model

(difference score: 50). Combined with the simulation of participants' choices (Figure 1b), we conclude that the hybrid-RL model provides a reasonable explanation for the reversal learning performance data. We then used individual best fitting parameters constrained by the group posterior distribution to derive single-trial estimates for $\delta_t, V_t \; and \; \alpha_t$.

Multiple Single-Trial Robust Regression

We applied multiple single-trial regression of z-scored predictions derived from the computational reinforcement learning model on single-trial EEG activity at each electrode and time point to investigate the effects of the different parameters (1st level) and WMC (2nd level) in a general linear model (GLM). Robust regression that down weights outliers by performing an iteratively reweighted least square method (O'Leary 1990) was employed to determine parameters in the following linear equation:

$$Y = b\_0 \; intercept + \; b\_1 \; RPE + b\_2 \; LR + b\_3 \; OR + \; b\_4 \; RPE \times LR + \; b\_5 \; RPE \times OR$$
$$+ \; b\_6 \; LR \times OR \; + \; b\_7 \; RPE \times LR \times OR + error \qquad (GLM1)$$

where Y is a data vector with amplitude values at each time point for every electrode, $b_{1\text{-}7}$ are the regression coefficients and error reflects residual variance. *LR* and *RPE* are vectors with the z-scored model parameters and *OR* (outcome reality) is a binary regressor coding whether the outcome of a trial was real (1) or fictive (-1). For every participant this procedure results in matrices of regressor weights in the form space (electrode) x timepoint x regressor. To enable further interpretation of interactions, subordinate linear regressions were calculated separately for real and fictive outcomes with all predictors except *OR* and their interactions (GLM2,3). Multiple comparisons were controlled using the Bonferroni-corrected threshold for all time points from 200-550 ms past feedback and all midline electrodes, based on previously reported effects in this latency range and at these electrodes (Fischer and Ullsperger 2013).

To check for effects of response switches, we ran an additional GLM4 with only a binary regressor, coding whether the participant switched to a different response in the next trial when the same stimulus was repeated.

For direct comparisons of regression weights between real and fictive trials, we selected electrodes a priori based on the peaks of RPE effects in the study of Fischer & Ullsperger (2013): Cz for FRN and P3a and Pz for P3b and used 2-tailed t-tests with a Bonferroni-corrected threshold of $p = 10^{-4}$ (.05 / number of sample points in the longest time window) for significance. For analysis of WMC effects on the EEG coding of learning parameters, we correlated WMC scores with the maximum regression weights within the time windows of interest at the preselected electrodes. Furthermore, we split the groups into high and low capacity individuals based on the median AOSPAN score (43), to compare model fits and for plotting regression weights of both groups separately.

Results

Working memory capacity

Overall, participants scored 41.78 ± 16.16 points during the AOSPAN and committed 4.73 ± 2.36 operation errors. There was a significant negative correlation between participants' age and their AOSPAN score ($r = -.346$, $p = .013$).

Reversal learning

On average participants performed the reversal learning task with 76.51 ± 1.69% accuracy, that is choosing the advantageous option. Participants learned avoiding bad (20% rewarded) and choosing good (80% rewarded) stimuli comparably well (Figure 1b): there was no difference in the percentage of correct decisions for good (77.3 ± 11.4 %) compared to bad stimuli (75.7 ± 16.7 %; $t_{50} = .724$, $p=.473$). However, median reaction times (RT) were lower

on good (544.9 ± 123.7 ms) compared to bad stimulus trials (579.3 ± 123.2 ms; t_{50} = -7.475, p = 1.1 x 10^{-9}).

WMC was predictive of reversal learning performance. The higher the AOSPAN score, the faster participants responded (r = -.379, p = .006), the more bonus they gained during the task (r = .452, p = .001) and the more overall advantageous choices they made (r = .413, p = .003). Previous studies suggested that, in part, performance differences in WMC tasks may result from fluctuations in attentional control (Adam *et al.* 2015; Unsworth and Robison 2016). Therefore, we tested for a relationship between individual RT variability (average individual SD of RTs) and WMC, but found no significant correlation (r= -.170, p = .234).

After a reversal from good to bad reward contingency, it took participants on average 4.30 ± 1.94 trials to switch to the now advantageous response. After a reversal from bad to good, they were faster in determining the new correct response (3.51 ± 2.00 trials, t_{50} = 2.354, p = .023). This pattern was independent of WMC (both in median split and correlations of WMC with a positivity bias difference score; all p > .1). High WMC participants, however, were more persistent in their performance after a successful switch than low WMC participants (Figure 2a): There was a positive correlation between AOSPAN scores and the mean persistence, measured as the probability to consistently choose the correct option in the 10 trials after the final reversal error (r = .402, p = .004). This is also reflected in the sensitivity to misleading probabilistic feedback. The higher the WMC, the less likely it was for participants to switch responses after misleading feedback on correct responses (r = -.384, p = .005; Figure 2b).

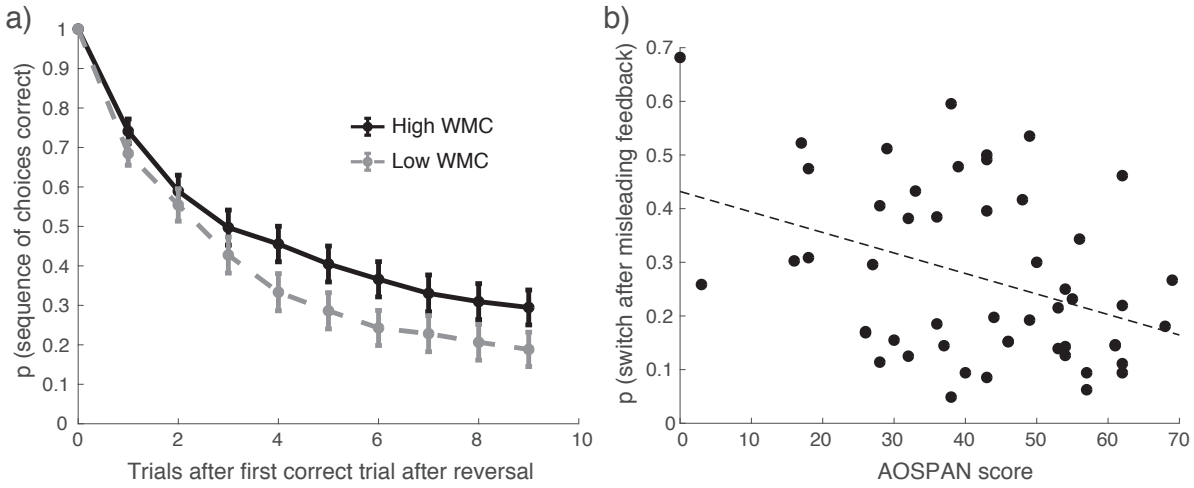

Figure 2. Behavioral effects of working memory capacity (WMC). a) Persistence of behavioral adaptation after a switch for high and low WMC subjects (median split) and d) scatter plot displaying the relationship between sensitivity to misleading feedback and WMC.

Table 1: Model parameter comparison and correlation with working memory capacity

Model Parameter	Mean (SEM)	r
α_{start}	.894 (.025)	-.052
η	.432 (.051)	.438**
κ	.512 (.031)	-.133
β	.275 (.035)	-.426**

Computational Model

Model fits of the hybrid Rescorla-Wagner Pearce-Hall model were comparable for high and low WMC groups (log likelihood [LL] for High WMC 334.72 ± 47.83, Low WMC 355.59 ± 51.76; p = 0.14). Table 1 shows the mean fitted model parameters across all participants and their correlation with WMC. Weighting factor η values were positively correlated with AOSPAN score (r = .438, p = .001). That is, participants with a lower AOSPAN score performed more like a classical Rescorla-Wagner algorithm with a constant LR, while LRs of participants with a higher AOSPAN score had a higher dynamic range. Furthermore, η

predicted reversal learning performance both with regard to earned bonus (r = .705, p = 7.6 x 10^{-9}) and overall percentage correct choices (r= .667, p = 9.1 x 10^{-8}).

Softmax temperature values β were negatively correlated with WMC (r= -.426, p=.002). As temperature values lead to less predictable behavior, this indicates that low capacity subjects chose more randomly than high capacity subjects. Softmax temperature was also negatively correlated with correct choices during reversal learning (r= -.712, p = 4.6 x 10^{-9}). There were no significant WMC correlations with either the constant κ or initial *LR* α_1 parameters (both ps > .3). The modeled dynamic *LR*, averaged across all trials, was negatively correlated with AOSPAN scores (r = -.492, p = .0002). To visualize the dynamics of *LR* changes, we plotted mean trial-wise *LR* values from ten trials before to ten trials after reversals (Figure 3). This shows that, as a result of the higher weighting factor η, the baseline *LR* (reached at the end of a block with constant stimulus values and reflected in the LR at the rule reversal trial, $\alpha_{reversal}$) is indeed lower in high WMC subjects, but after a value reversal it is upregulated quickly. We quantified the dynamic changes of LR after reversals from good to bad or from bad to good as the difference of mean LRs across ten trials after the value reversal and the baseline LR at the reversal trial:

$$\Delta\alpha = mean(\alpha_{reversal+1}, \alpha_{reversal+2}, \dots \alpha_{reversal+10}) - \alpha_{reversal} \qquad (Eq. 5)$$

This dynamic post-reversal increase in LR was correlated with WMC (r = .446, p = .001 and r = -.487, p < .001, respectively).

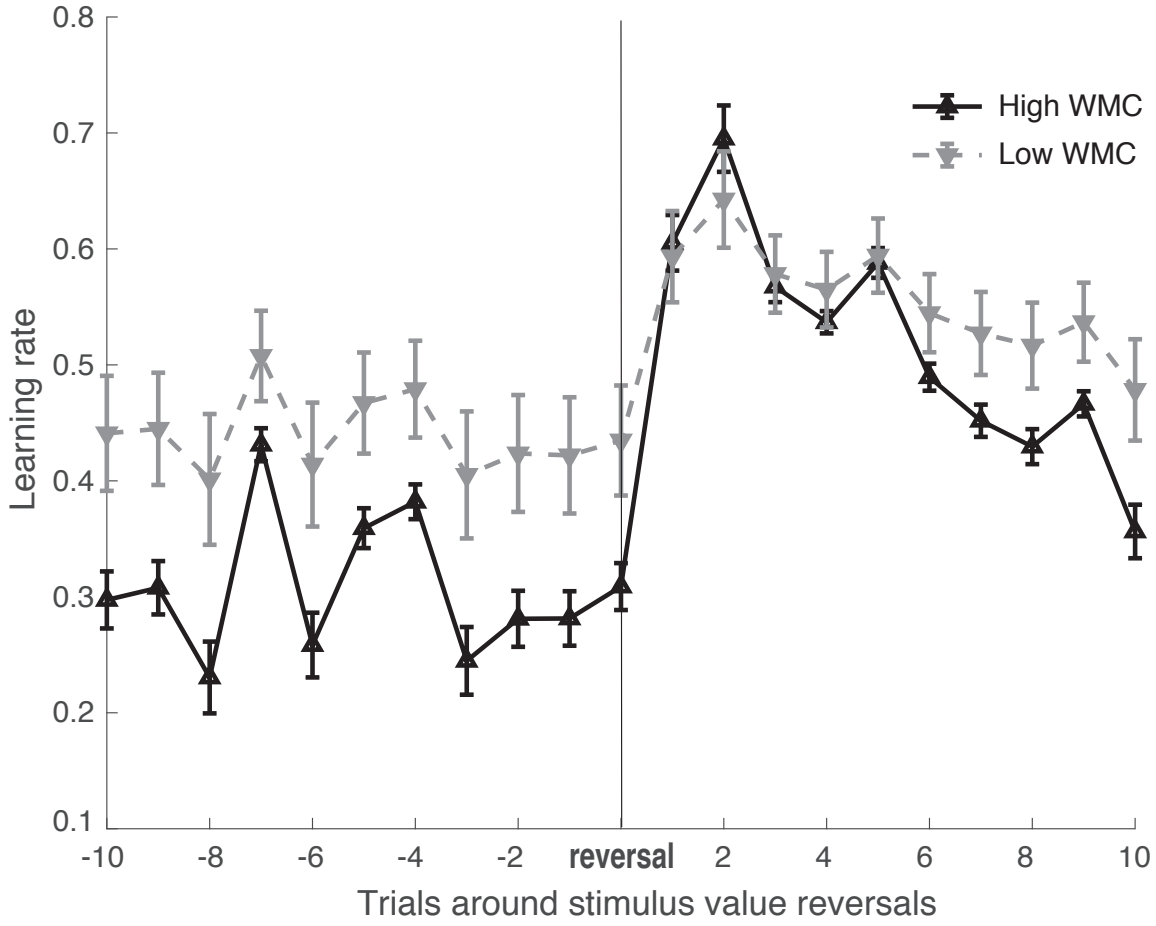

Figure 3. Dynamics of the modeled learning rate around good-to-bad and bad-to-good switches in reward contingencies separately for high (black upward triangles, solid line) and low (gray downward triangles, dashed line) WMC subjects.

EEG dynamics reflects RPEs differently for real and fictive outcomes

The results of the omnibus GLM1 are shown in Figure 4. Significant main effects were found for all predictors and the interactions of *RPE* × *OR* as well as *RPE* × *LR* × *OR*. The complex patterns of main effects and interactions can be interpreted based on the subordinate separate GLMs for real and fictive outcomes, which are displayed in Figures 5a and 5b, respectively. In the first 400 ms, there was a dissociation of RPE correlates during processing of feedback about real and fictive outcomes, replicating the results reported previously in a probabilistic learning task without reversals (Fischer and Ullsperger 2013). Real outcomes were associated

with a positive covariation of the RPE with the EEG amplitude between 200 and 300 ms (contributing to the FRN) and a negative covariation between 300 and 450 ms (contributing to the P3a) at frontocentral electrodes. In contrast, only for fictive outcomes, an early negative occipital effect around 190 to 240 ms after feedback was found (significant difference between real and fictive conditions, at electrode Oz; peak $t_{50} = 5.565$, $p < 10^{-5}$ at 214 ms). No further covariation with RPE was found for fictive outcomes for latencies up to 400 ms. The direct contrast of both feedback conditions revealed a significant difference at electrode Cz in both the FRN time windows (peak $t_{50} = 6.063$, $p < 10^{-6}$ at 248 ms) and the P3a time window (peak $t_{50} = -7.075$, $p < 10^{-8}$ at 408 ms); reflecting the absence of an RPE effect on these deflections in the event-related potential elicited by fictive outcomes.

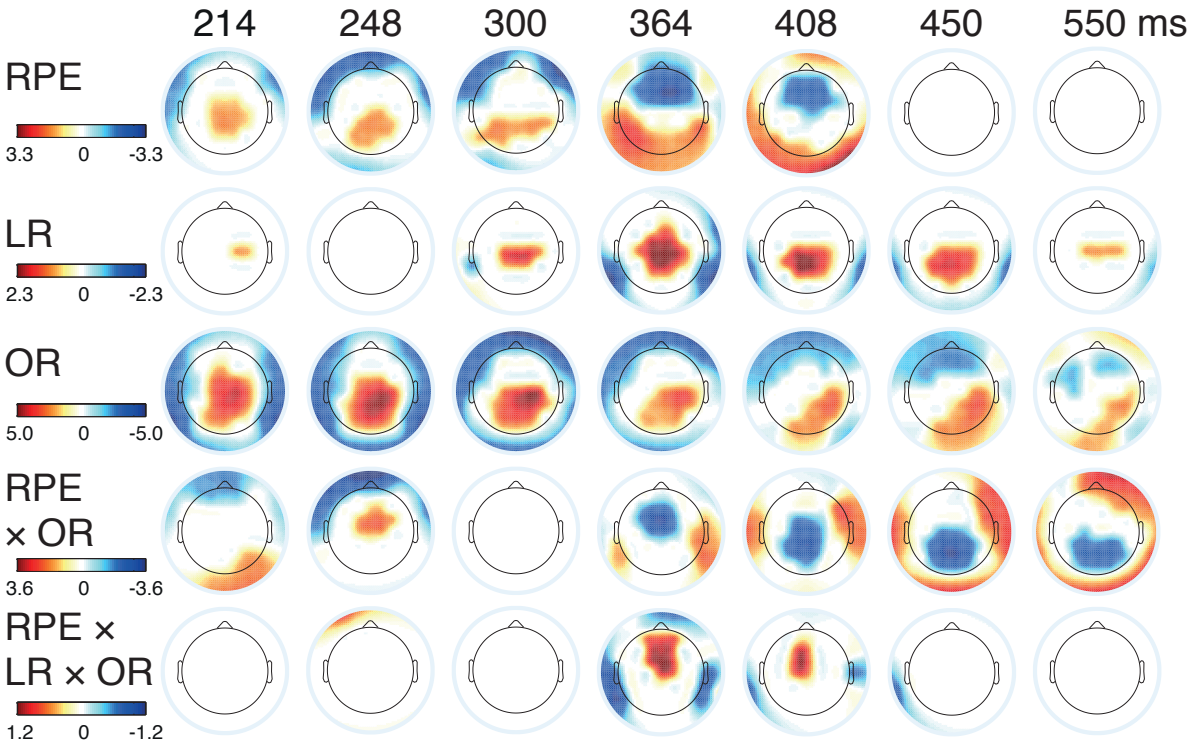

Figure 4. Results of the omnibus general linear model analysis of feedback-locked EEG data. Topographical distributions of group mean regression weights significantly different from zero (corrected) are displayed for latencies between 214 and 550 ms after feedback onset for main effects of reward prediction error (*RPE*), learning rate (*LR*), outcome reality (*OR*,

real/fictive) and their interactions. Interactions *RPE* × *LR* and *LR* × *OR* were not significant and are not displayed. Color bars show range of regression weights (arbitrary units).

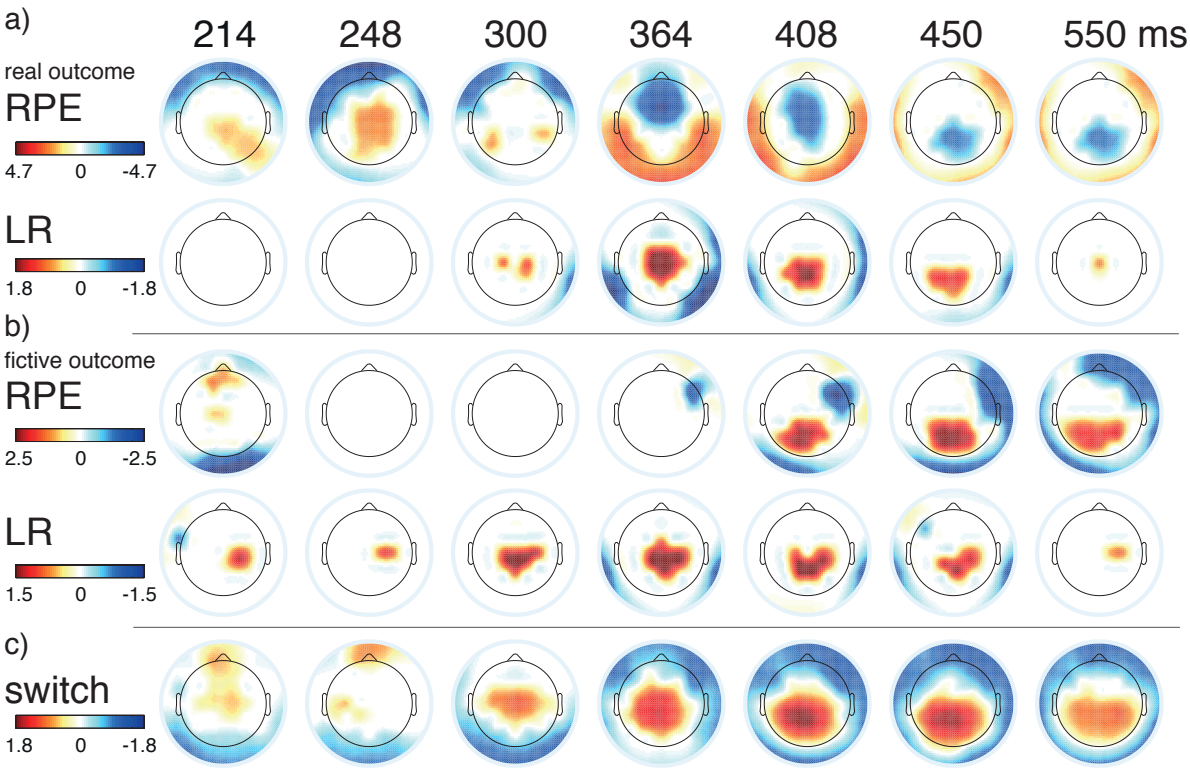

Figure 5. Separate general linear model analyses of feedback-locked EEG data. Topographical distributions of group mean regression weights significantly different from zero (corrected) are displayed for latencies between 214 and 550 ms for main effects of RPE and LR separately (a) for real outcomes and (b) for fictive outcomes. (c) shows the results for a separate analysis on all outcomes with a single binary regressor coding whether the choice at the next encounter with the just-seen stimulus will be different from the current response (switch) or not. Color bars show range of regression weights (arbitrary units).

During the P3b time window (430-600 ms) processing of both real and fictive outcomes seems to converge: *RPE* correlated with centroparietal EEG activity negatively for real outcomes and positively for fictive outcomes. This reversal of polarities reflects the fact that in the real feedback conditions, negative RPEs are unfavorable outcomes and thus signal the

need to adapt, whereas in the fictive condition a positive RPE is actually the unfavorable outcome. Notably, in the P3b time window the overall GLM1 shows no general RPE effect but a significant interaction of *RPE* and *OR* at centroparietal electrodes, which suggests that RPE modulation of the P3b amplitude is equal in size in both conditions and just differs in sign as a result of quantifying RPEs with respect to stimulus values and not subjective values.
Feedback-locked EEG also reflected dynamically changing LR (Figures 4 and 5a,b). We found a significant main effect of *LR* reflected in a sustained positive covariation with the EEG amplitude from 300 to 550 ms: The higher the LR, the more positive-going is the EEG at centroparietal electrodes thereby contributing to the P3a and P3b (Cz: peak $t_{50} = 7.454$, $p < 10^{-8}$ at 364 ms ; Pz: peak $t_{50} = 7.184$, $p < 10^{-8}$ at 418 ms). There was no significant interaction between outcome condition (real vs. fictive) and *LR*, suggesting that the LR effect did not differ, although the LR effect appeared to start earlier (at around 200 ms) for fictive outcomes (Figure 5b).

WMC affects EEG coding of learning parameters

Subjects with high WMC scores show a stronger covariation of *RPE* and P3a and P3b in real (correlation at electrode Cz between peak negative t-value in real trials between 300 and 450 ms and AOSPAN score: r = -.450, p = .001; between 450 and 600 ms: r = -0.36, p=.01), but not in fictive trials (P3a, r = -.073, p = .612; P3b, r = -0.15, p=.279) than subjects with low AOSPAN scores (Figure 6a,b). There also seems to be a stronger covariation during the FRN time-window, but this effect was not significant (correlation between peak t-value in real trials between 200 and 300 ms and AOSPAN score: r = .191, p = .18).
WMC also influenced the effect of *LR* on feedback-related EEG dynamics (Figure 6c). There was a positive correlation between peak t-values in the P3a time window (300 - 450 ms) at

electrode Cz and AOSPAN score (r = .291, p = .038; both conditions collapsed). A similar but weaker correlation was found for the P3b time window (450-550 ms) at Pz (r=.141, p= .322). These results suggest that the reinforcement learning parameters *RPE* and *LR* are represented in feedback-locked EEG activity more in participants with higher WMC. However, in order to test whether the representation of adaptive mechanisms in the EEG activity generally varies as a function of WMC, we performed an additional analysis in which we regressed EEG amplitudes against a binary switch regressor indicating whether participants adapt, i.e. switch, their choice behavior at the next encounter with the same stimulus or not. In other words, this analysis shows any feedback-related EEG activity predictive of subsequent behavioral adaptation. Figure 5c displays the result, a sustained P3b-like centroparietal positivity which is maximal in the latency range 360-460 ms. Importantly, we found no correlation of this EEG switch effect with WMC (Cz, r = .156, p = .276). This demonstrates that the WMC effects on cortical representations of learning parameters reported above do not reflect generally weaker brain-behavior interactions in low capacity-participants.

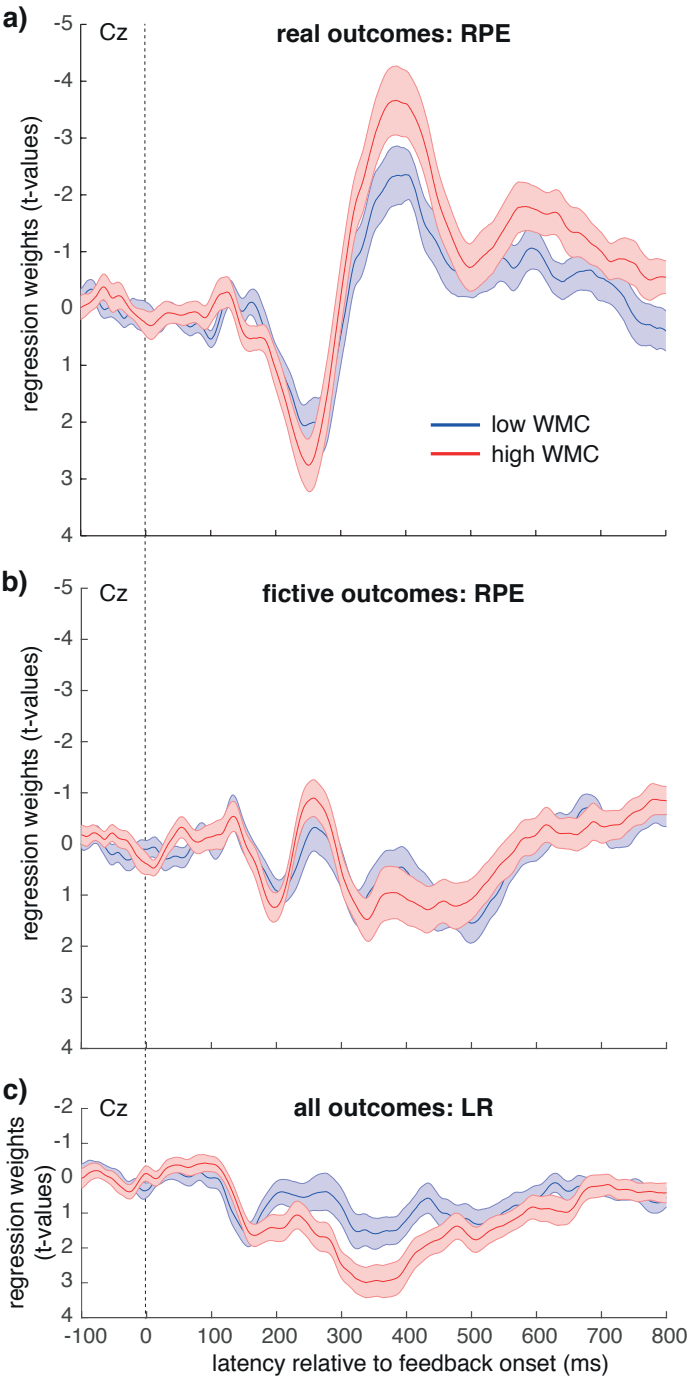

Figure 6. WMC effects on feedback-related EEG correlates of learning parameters. (a,b) Time course of regression weights for the RPE regressor at electrode Cz for real (a) and fictive (b) outcomes separately for participants with high WMC (dashed gray) and low WMC (solid black). High and low WMC reflect scores above and below the group median WMC score, respectively. (c) Time course of regression weights for the LR regressor at electrode Cz for all outcomes separately for participants with high WMC (dashed gray) and low WMC (solid black). Shades depict standard errors of the mean. The vertical dotted line indicates feedback onset.

Discussion

In the present study we replicated previous findings regarding the EEG correlates of cortical activity during learning and decision making (Fischer and Ullsperger 2013). By implementing a reversal learning task in which reward contingencies of the different choice options changed independently and unbeknownst to the participants, we decorrelated LR dynamics from

learning progress and time on task – substantially extending our previous findings. Importantly, we found that dynamically changing LR was reflected in a sustained centroparietal shift of the EEG modulating the P3b. Furthermore, we confirmed a dissociation of cortical RPE correlates of real and fictive outcome processing: whereas real feedback elicits a frontocentral covariation of the EEG with RPE from 200 to 400 ms contributing to the FRN and P3a, counterfactual feedback was associated with an early occipitotemporal RPE effect but neither FRN- nor P3a-like activity. For both conditions, EEG activity after 400 ms converged on a centroparietal P3b, which was driven by unfavorable outcomes and predicted subsequent shifts in choice behavior. This supports the notion that the feedback-related P3b reflects activity of a common pathway of adaptation resulting in value update and guiding future decisions.

The main goal of the present study, however, was to investigate whether and how WMC influences reinforcement learning, in particular the ability to flexibly adjust the weight assigned to current action outcomes in guiding future decisions. We found that higher WMC was associated with better performance in the reversal learning task, reflected in faster and overall more advantageous choices. Lower WMC has been suggested to result, in part, from general fluctuations in task engagement or lapses in attention (Adam *et al.* 2015; Unsworth and Robison 2016). However, general fluctuations in attentional control seem not to be the main cause of reduced reversal learning performance in low WMC participants, as we found no general increase in response time variability. Instead, we found learning-specific differences between high- and low-WMC participants. Low WMC was associated with a greater sensitivity to misleading probabilistic feedback. In other words, in low-span participants infrequent incorrect feedback (negative on good stimuli or positive on bad stimuli) resulted more often in maladaptive shifts towards avoiding good stimuli and choosing

bad stimuli instead of persevering with the previously learned response. Additionally, after having successfully shifted the response after a rule reversal, low-span participants were more prone to shift away from the correct choice again. These behavioral findings suggest that higher WMC supports the ability to integrate outcomes over a longer reinforcement history. Computational modeling supports this conclusion. The hybrid Rescorla-Wagner-Pearce-Hall model allowed us to dissociate general learning, as captured by the averaged modeled LR, and the speed of adaptation to new situations, as captured by the changes in LR around reversals. In line with our hypothesis, we found that low WMC individuals indeed showed less optimal regulation of the LR, reflected in the peri-switch dynamics. At the end of blocks of trials with stable response rules, LR should reach its minimum reflecting maximal reduction of uncertainty about the task rules (O'Reilly 2013; McGuire *et al.* 2014). However, in low-WMC individuals the gradual decrease of LR across blocks is minimal, rendering them vulnerable to misleading probabilistic feedback at any time during the task. After rule reversals, participants persevered with their previous choice behavior for three to five trials. Unfavorable feedback on these trials results in the accumulation of large RPEs (in contrast to isolated peaks in RPEs on probabilistic misleading feedback). In Pearce-Hall learning algorithms, surprise –which can be expressed as high absolute prediction errors– increases the associability or LR (Pearce and Hall 1980; Roesch et al. 2012). The hybrid model we applied to the behavioral data combines features from both Rescorla-Wagner and Pearce-Hall algorithms (Li *et al.* 2011). The weighting factor η describes the extent to which learning dynamics is driven by Pearce-Hall or Rescorla-Wagner algorithms. High η values indicate that the LR is strongly influenced by surprise, i.e., the accumulation of large absolute RPEs, which is the case after unexpected reversals of action-reward contingencies. This enables dynamic adjustments of the LR around rule shifts as seen in high-WMC participants. Indeed, the weighting factor η correlated with WMC. Low η values in low-WMC individuals resulted in a more constant

LR, which on average was higher than in high WMC participants. This suggests that low WMC is associated with problems assigning weight appropriately to salient events: whereas misleading feedback is overweighted, informative feedback after rule reversals is used less efficiently for value updates. In addition, the softmax temperature was higher in low-WMC participants suggesting that their choices were more stochastic and less strongly driven by the learnt values. Both model parameters thus explain the reduced reversal learning performance with low WMC.

Interestingly, the EEG correlates of learning parameters varied as a function of WMC. While the qualitative patterns of feedback-locked EEG activity did not depend on WMC, the amplitudes of the RPE effect in the time range of the P3a and P3b as well as the sustained central positive effect of LR were significantly reduced in low WMC participants. This may suggest that cortical representation of learning parameters is weaker or noisier in low WMC individuals. The frontocentral localization of the WMC effect on EEG correlates of RL is compatible with findings from neuroimaging that updating of action values and of complex beliefs as well as LR and RPE are encoded in the posterior medial frontal cortex (Behrens *et al.* 2007; Jocham *et al.* 2009; O'Reilly et al. 2013; McGuire *et al.* 2014; Fischer *et al.* 2017). Consistent with this, anterior cingulate sulcus lesions impair the ability to track reinforcement history and to adjust the LR accordingly (Kennerley et al. 2006). Thus, our current results suggest that interindividual differences in WMC are manifested in learning-related activity of the posterior medial frontal cortex. However, it remains to be investigated whether these individual differences are related to differential functioning of this region or to differential inputs to it via functional connectivity (McGuire *et al.* 2014). Interestingly, previous studies have shown that higher WMC is also associated with larger amplitudes of the error-related negativity (Miller et al. 2012; Coleman et al. 2018), an event-related potential elicited by

action slips in speeded reaction time tasks, which is also generated in the pMFC (Debener et al. 2005).

Importantly, in contrast to the representations of LR and RPE in the EEG, the effect of subsequent switches of choice behavior on the centroparietal EEG in the latency range of the P3b did not vary with WMC. This suggests that the common final pathway of adaptation does not differ between high- and low-span participants. Thus, the mechanisms guiding future choices appear to be general and independent of WMC, but the weighting of inputs to the decision-making process and its efficiency appears to vary as a function of WMC.

How do these findings on instrumental learning relate to WMC? It appears that the cognitive functions needed to score high in WMC span tasks are necessary in instrumental learning in uncertain environments as well. These cognitive functions encompass the ability to appropriately weight salient external inputs, be it to-be-remembered items or distractors in span tasks or informative vs. misleading feedback in probabilistic reversal learning. The current results may suggest that representations of items held in working memory but also of values learnt in instrumental learning tasks and learning parameters such as RPE and LR are represented with lower signal-to-noise ratio in low capacity participants compared to high-capacity participants. Future research needs to elucidate whether working memory training, which can improve WMC (Constantinidis and Klingberg 2016), has transfer effects onto instrumental learning and appropriately adjusting dynamic learning rates as well.

To conclude, the present study demonstrated that working memory is essential for efficient instrumental learning in changeable and uncertain environments which require tracking of multiple values in parallel. Low WMC is associated with a reduced ability to appropriately weight salient events and to regulate value updating accordingly, and leads to increased stochasticity of choice behavior. While EEG correlates of value adaptations are independent

of WMC, the representation of learning parameters is weaker in low-capacity participants, suggesting that choice behavior is more influenced by other, idiosyncratic factors.

Conflict of Interest

The authors declare no competing financial interests.

Acknowledgements

The authors would like to thank Julia Berghäuser for help with data collection. This work was supported by the Deutsche Forschungsgemeinschaft (Collaborative Research Center SFB 779, Neurobiology of Motivated Behavior, TP A12).

References

Adam KC, Mance I, Fukuda K, Vogel EK. 2015. The contribution of attentional lapses to individual differences in visual working memory capacity. J Cogn Neurosci. 27:1601-1616.

Adam KCS, Vogel EK, Awh E. 2017. Clear evidence for item limits in visual working memory. Cognitive psychology. 97:79-97.

Amiez C, Petrides M. 2007. Selective involvement of the mid-dorsolateral prefrontal cortex in the coding of the serial order of visual stimuli in working memory. Proc Natl Acad Sci U S A. 104:13786-13791.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF. 2007. Learning the value of information in an uncertain world. Nat Neurosci. 10:1214-1221.

Boorman ED, Behrens TE, Rushworth MF. 2011. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biol. 9:e1001093.

Coleman JR, Watson JM, Strayer DL. 2018. Working memory capacity and task goals modulate error-related ERPs. Psychophysiology. 55.

Collins AG, Frank MJ. 2012. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 35:1024-1035.

Constantinidis C, Klingberg T. 2016. The neuroscience of working memory capacity and training. Nat Rev Neurosci. 17:438-449.

Conway AR, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. 2005. Working memory span tasks: A methodological review and user's guide. Psychon Bull Rev. 12:769-786.

Cools R, Gibbs SE, Miyakawa A, Jagust W, D'Esposito M. 2008. Working memory capacity predicts dopamine synthesis capacity in the human striatum. J Neurosci. 28:1208-1212.

Debener S, Ullsperger M, Siegel M, Fiehler K, von Cramon DY, Engel AK. 2005. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. J Neurosci. 25:11730-11737.

Delorme A, Makeig S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 134:9-21.

Fischer AG, Bourgeois-Gironde S, Ullsperger M. 2017. Short-term reward experience biases inference despite dissociable neural correlates. Nature communications. 8:1690.

Fischer AG, Ullsperger M. 2013. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. Neuron. 79:1243-1255.

Fusi S, Asaad WF, Miller EK, Wang XJ. 2007. A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. Neuron. 54:319-333.
Huys QJ, Cools R, Golzer M, Friedel E, Heinz A, Dolan RJ, Dayan P. 2011. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS computational biology. 7:e1002028.
Jocham G, Klein TA, Ullsperger M. 2011. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci. 31:1606-1613.
Jocham G, Klein TA, Ullsperger M. 2014. Differential modulation of reinforcement learning by D2 dopamine and NMDA glutamate receptor antagonism. J Neurosci. 34:13151-13162.
Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M. 2009. Adaptive coding of action values in the human rostral cingulate zone. J Neurosci. 29:7489-7496.
Kass RE, Raftery AE. 1995. Bayes Factors. Journal of the American Statistical Association. 90:773-795.
Kennerley SW, Walton ME, Behrens TE, Buckley MJ, Rushworth MF. 2006. Optimal decision making and the anterior cingulate cortex. Nat Neurosci. 9:940-947.
Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M. 2007. Genetically determined differences in learning from errors. Science. 318:1642-1645.
Krugel LK, Biele G, Mohr PN, Li SC, Heekeren HR. 2009. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc Natl Acad Sci U S A. 106:17951-17956.
Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND. 2011. Differential roles of human striatum and amygdala in associative learning. Nat Neurosci. 14:1250-1252.
McGuire JT, Nassar MR, Gold JI, Kable JW. 2014. Functionally dissociable influences on learning rate in a dynamic environment. Neuron. 84:870-881.
Meder D, Kolling N, Verhagen L, Wittmann MK, Scholl J, Madsen KH, Hulme OJ, Behrens TEJ, Rushworth MFS. 2017. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. Nature communications. 8:1942.
Miller AE, Watson JM, Strayer DL. 2012. Individual differences in working memory capacity predict action monitoring and the error-related negativity. J Exp Psychol Learn Mem Cogn. 38:757-763.
O'Connell RG, Dockree PM, Kelly SP. 2012. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nat Neurosci. 15:1729-1735.

O'Reilly JX. 2013. Making predictions in a changing world-inference, uncertainty, and learning. Front Neurosci. 7:105.
O'Reilly JX, Schuffelgen U, Cuell SF, Behrens TE, Mars RB, Rushworth MF. 2013. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc Natl Acad Sci U S A. 110:E3660-3669.
O'Leary DP. 1990. Robust Regression Computation Using Iteratively Reweighted Least Squares. SIAM Journal on Matrix Analysis and Applications. 11:466-480.
Pearce JM, Hall G. 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev. 87:532-552.
Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 442:1042-1045.
Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokrasy WF, editors. Classical conditioning II: Current research and theory. New York: Appleton-Century-Crofts p 64-99.
Roesch MR, Esber GR, Li J, Daw ND, Schoenbaum G. 2012. Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. Eur J Neurosci. 35:1190-1200.
Sutton RS, Barto AG. 1998. Reinforcement Learning: An Introduction. Cambridge, MA: The MIT Press.
Talmi D, Atkinson R, El-Deredy W. 2013. The feedback-related negativity signals salience prediction errors, not reward prediction errors. J Neurosci. 33:8264-8269.
Ullsperger M, Fischer AG, Nigbur R, Endrass T. 2014. Neural mechanisms and temporal dynamics of performance monitoring. Trends Cogn Sci. 18:259-267.
Unsworth N, Engle RW. 2007. On the division of short-term and working memory: an examination of simple and complex span and their relation to higher order abilities. Psychol Bull. 133:1038-1066.
Unsworth N, Heitz RP, Schrock JC, Engle RW. 2005. An automated version of the operation span task. Behav Res Methods. 37:498-505.
Unsworth N, Robison MK. 2016. The influence of lapses of attention on working memory capacity. Mem Cognit. 44:188-196.
Walsh MM, Anderson JR. 2012. Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. Neurosci Biobehav Rev. 36:1870-1884.