Learning and using different task rule updating strategies require extensive practice

Shengjie Xu[a,*], Tanya Wen[b], Tobias Egner[c], Tom Verguts[a] and Senne Braem[a]

[a] Department of Experimental Psychology, Ghent University, Belgium

[b] Meta, U.S.A

[c] Department of Psychology and Neuroscience, Duke University, U.S.A

***Corresponding Author:**

Shengjie Xu

Department of Experimental Psychology

Ghent University

Henri Dunantlaan 2, 9000 Ghent

Belgium

Email: sjx667@gmail.com

ORCID:

Shengjie Xu: 0000-0003-4604-372X

Tanya Wen: 0000-0002-4580-7831

Tobias Egner: 0000-0001-7956-3241

Tom Verguts: 0000-0002-7783-4754

Senne Braem: 0000-0002-2619-8225

Acknowledgements

This work was supported by an ERC Starting grant awarded to S.B. (European Union's Horizon 2020 research and innovation program, Grant agreement 852570). T.V. and S.B. were supported by FWO project G010319N.

Conflicts of Interest: None

Abstract

People adjust how fast they update task rules, depending on the volatility of their environment, by regulating cognitive control accordingly. We investigated whether this adaptation is primarily driven by recently experienced volatility in task demands, or can also be shaped by learned, environmental knowledge indicating expected levels of volatility. To this end, we trained participants on a Wisconsin Card Sorting Test task where different environments required different speeds of task rule updating. We demonstrate that, initially, participants adopted task rule updating strategies depending on the most recent experienced levels of volatility and feedback (Experiment 1). However, after extensive (four days) training (Experiment 2), participants updated task rules using strategies based on developed environmental knowledge. We also show how participants fine-tuned their task rule updating strategies over learning. In sum, our findings provide important insights into how humans update task rules using different strategies and highlight the importance of multi-day training in developing environmental knowledge for the regulation of cognitive control.

Keywords: cognitive control; cognitive flexibility; context; reinforcement learning; multi-day training

Introduction

Responding to stimuli based on a task rule generally comes easy to humans. However, Humans are often faced with environments with changing task rules and the speed at which task rules change may differ, so we also need to learn how quickly we should decide to change tasks. In a more volatile environment, task rules change more frequently, requiring a higher learning rate to update task rules. In a less volatile environment, fast task rule updating can be suboptimal, and it may be better to await more feedback before changing task rules, favoring a lower learning rate. This study of updating task rules in response to environment-specific needs has been the topic of research on cognitive control (Badre, 2025; Egner, 2023; Monsell, 2003; Musslick & Cohen, 2021). However, it remains unclear how much of these behavioral adaptations reflect adaptions to recently experienced volatility, or also relies on learned knowledge about environment-specific levels of task volatility (Braem et al., 2024).

Previous studies robustly showed that humans can quickly adapt to a recently experienced need for control. For example, in probabilistic reversal learning tasks, a robust finding is that humans increase their learning rates when the environment is more volatile (Behrens et al., 2007; Simoens et al., 2024; Wen et al., 2023). To account for these observations, some have proposed that these fast adaptations can be seen as a dynamic process where humans upregulate their updating strategies after encountering a higher prediction error, but downregulate them when prediction errors are lower (Bai et al., 2014; Krugel et al., 2009; Li et al., 2011; Nassar et al., 2010; Pearce & Hall, 1980; Verbeke & Verguts, 2024). This type of adaptations is implemented by rapidly adjusting control settings, such as learning rate, based on recent, short-term experiences.

On the other hand, contemporary theories on cognitive control emphasize that the regulation of cognitive control also requires meta-learning (Abrahamse et al., 2016; Braem &

Egner, 2018; Egner, 2014, 2023; Griffiths et al., 2019; Verguts & Notebaert, 2009; Wang, 2021). In particular, humans can learn how to adjust control settings, sometimes referred to as meta-control (Eppinger et al., 2021), by associating environment-specific features to different control settings (Braem & Egner, 2018; Chiu & Egner, 2017; Xu et al., 2024). Different from control adaptations based on short-term experiences, this type of adaptations refers to an adaptive process where associated control settings, like learning rate, can be evoked when humans later revisit these same environments via long-term knowledge, allowing for faster, adaptive changes in task rule updating without the need to learn from ongoing experience again.

One recent study found that people indeed show such environment-control regulation in a task switching paradigm, but only after four days of one-hour long training (Xu et al., 2024). This suggests that extensive task experience might be needed before people both learn and adopt such environment-specific strategies. However, in traditional task switching paradigms, participants are cued about which task to perform next, leaving little uncertainty about the current need for task rule updating. Such experiment settings might emphasize focusing on external cues instead of learning to use environment-specific task rule updating strategies. In many real-life scenarios, people often need to infer which task rule to perform from feedback in environments without explicit task cues. This requires them to form and use internal knowledge of environment-specific task rule updating strategies via learning, as no reliable external cues are available. Contrasting a high versus low volatility environment in a Wisconsin Card Sorting Test task with probabilistic feedback, Wen and colleagues (2023) recently tested whether people can both learn and transfer the learning rate to a subsequent testing phase with medium levels of volatility, in the same or even different versions of the task. They found that participants indeed showed different learning rates for task rule updating based on high versus low environmental volatility and carried these learning rates

over into the testing phase. However, their original task was a between-subject design, where the testing phase occurred right after the training phase. Consequently, this finding may reflect a direct carry-over effect (Aben et al., 2017; Hubbard et al., 2017; Scherbaum et al., 2012) of control settings evoked by recently experienced volatility, rather than a learning and implementing of environment-specific control settings.

To address this, we developed a customized Wisconsin Card Sorting Test task, based on the design by Wen and colleagues (2023), where participants experienced both high and low volatilities in different environments, and were examined in a testing phase using these environments as cues. To study the role of task experience, we conducted two experiments: one-day training in Experiment 1 and four-day training in Experiment 2. We hypothesized that learned, environment-specific task rule updating strategies can only be observed after extensive training, typically not achievable within a single experimental session. Concretely, we expected participants to rely more on locally experienced differences in demands for task rule changes in a first session, without relying on environment-specific cues, hence showing no difference in task rule updating strategies in the testing phase. However, through learning over successive days, participants can gradually learn (and start relying on) environment-specific task rule updating strategies.

Methods

Most task parameters and design choices were close to the original design of Wen and colleagues (2023), except for a few critical manipulations that were required for a within-subject design of the current study. Anonymized data and analysis scripts can be found: [*an osf link will be provided upon publication*]

Participants

Participants were students from Ghent University who received course credits for

their participation in this study[1]. Similar to our previous work on environment-specific cognitive flexibility (Xu et al., 2024), we recruited 65 participants in Experiment 1, and 62 participants in Experiment 2. Participants with an accuracy below 65% were excluded from both experiments (see also, Wen et al., 2023). In Experiment 1, we excluded nine participants, leaving a final sample size of 55 (43 females, M_{age}= 18.87 years, SD_{age}= 1.29, range = 17-23) for the model fitting and statistical analyses. In Experiment 2, we excluded seven participants, resulting in a final sample size of 55 (49 females, M_{age}= 18.40 years, SD_{age}= 1.57, range = 17-28). Both experiments were approved by the Ethical Committee of the Faculty of Psychology and Educational Sciences of Ghent University.

Wisconsin Card Sorting Test Task

On each trial, participants needed to choose between two choice cards to match the reference card on top (Fig.1A). Each choice card only shared one feature with the top reference card, e.g., blue color on the left choice card and item number three on the right card. There were four feature dimensions on these cards: color (blue, green, red or purple), filling (checkered, dots, wave, or grid), item number (one, two, three or four) and item shape (circle, triangle, plus, or star), resulting in 256 possible combinations. For each participant, only two feature dimensions were randomly selected as relevant for the task, counterbalanced across participants. On each trial, only one feature dimension was relevant to the correct card matching rule. Participants were asked to make as many correct choices as possible throughout the experiment.

Procedure

Each trial began with a 1000 ms fixation period, after which the cards were presented for 2000 ms, or until a choice was made. Next, reward feedback was presented for 500 ms.

[1] Meta was not involved in conducting the participant recruitment in this study.

Reward feedback was probabilistic: On 80% of trials the feedback was correct, but on 20% of trials, participants received positive feedback for incorrect choices or negative feedback for correct choices. When participants failed to respond in 2000 ms, they would receive a "Too slow" message instead.

We studied task choice behavior in response to the different (associated) volatility environments in each block. Similar to Wen and colleagues (2023), volatility was defined as the frequency of rule changes; in the low volatility environment the matching rule changed every 30 trials, and in the high volatility environment it changed every 10 trials. Each volatility environment was associated to two different table backgrounds: wood table or stone table (Fig. 1B) to create environmental cues. In the training phase (Fig. 1C), each table picture was always presented in either the low or high volatility environment, indicating to participants which volatility environment they were in. Table-volatility associations were counterbalanced among participants and were not instructed, so participants needed to learn them through experience. To test whether participants successfully learned and used these environment-volatility associations to influence task rule updating, we inserted probe blocks into the testing phase (Fig. 1C) where the different tables were again presented in the background, but the matching rule changed every 20 trials.

Experiment 1

In Experiment 1 (Fig. 1D), participants began with two practice blocks, where the actual matching rule was presented on each trial to help them get familiar with the task. Next, participants entered the learning phase, where they experienced either high or low volatility environments block by block, in an interleaved manner. Each block consisted of 60 trials and participants underwent a total of 16 blocks. In the second half of the experiment, probe

blocks were inserted. There was at least one learning block between two probe blocks with the same level of volatility. Participants randomly started with a high or low volatility block.

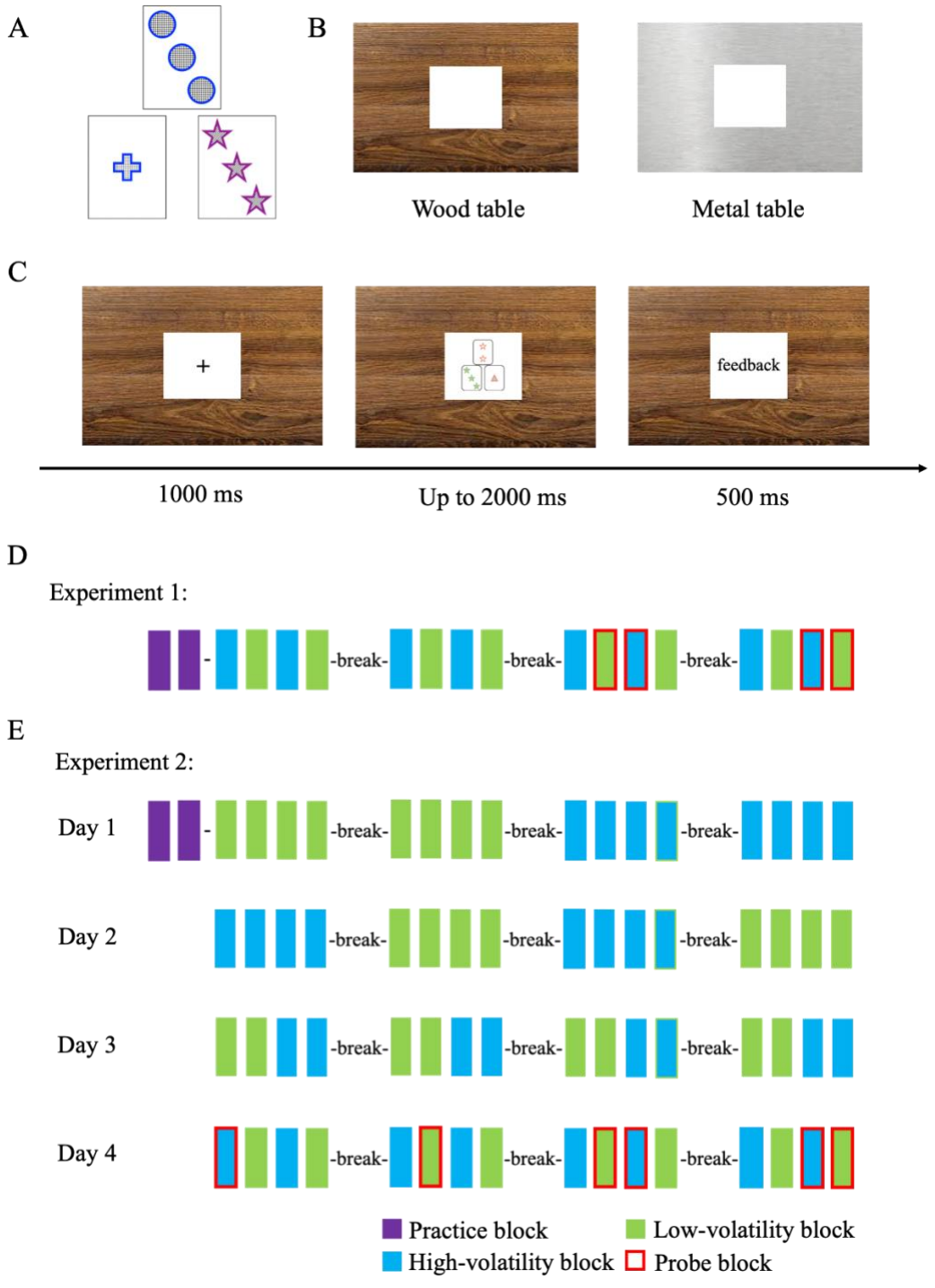


Figure 1: Wisconsin Cart Sorting Test task paradigm. **A**. Example cards presented on each trial. **B**. Two table pictures used as environmental cues. **C**. Trial overview. **D**. Structure overview of Experiment 1. Each box represents one block. The training phase included low-

volatility and high-volatility blocks, while the testing phase consisted of probe blocks in which the level of volatility remained same, but different environment cues (table pictures) were presented. (**E**) Structure overview of Experiment 2.

Experiment 2

In Experiment 2 (Fig. 1E), participants also started with practice blocks. Consistent with Experiment 1, Experiment 2 included 16 blocks per day, with each block consisting of 60 trials. Different from Experiment 1, however, the learning blocks gradually changed from a chunked order to an interleaved order over days (see also, Xu et al., 2024), as humans are better at learning environment-specific knowledge in a blocked manner compared to an interleaved manner (Flesch et al., 2018). The probe blocks only appeared on the last day: one probe block of each environment in the first half and two probe blocks of each environment in the second half. The first volatility environment was randomly chosen and shifted across learning days. Participants were asked to start the experiment each day around the time they started on the first day, allowing for a deviation of up to three hours (24 ± 3 hours).

Behavioral Analysis

To check if our volatility settings induced different performance in each environment, we first analyzed accuracy in the choice data. A correct response was defined as choosing the rule matched the correct rule on a given trial, independent of the feedback, which was invalid on 20% of the trials. In Experiment 1, behavioral accuracy was assessed by a two (volatility: high versus low) × two (phase: learning blocks versus probe blocks) repeated-measures ANOVA (rmANOVA). In Experiment 2, accuracy data from the first three days were examined by a three (day: day 1 versus day 2 versus day 3) × two (phase: learning blocks versus probe blocks) rmANOVA to assess performance improvement across days. Accuracy data from the last days were analyzed by a two (volatility: high versus low) × two (phase:

learning blocks versus probe blocks) repeated-measures ANOVA. We applied a Greenhouse-Geisser sphericity correction to within-subject factors that violated the sphericity assumption and reported corrected statics, and the same procedure was applied for all rmANOVAs.

We then analyzed the switch frequency in both experiments to see whether our volatility settings evoked different task rule updating strategies. Switch frequency was defined as the proportion of trials in which participants chose to switch the matching rule within each block. In Experiment 1, we used a two (volatility: high versus low) × two (phase: learning blocks versus probe blocks) rmANOVA. In Experiment 2, we further checked how switch frequency changed over training, using a two (volatility: high versus low) by three (day: day 1 versus day 2 versus day 3) rmANOVA. We again conducted a two (volatility: high versus low) × two (phase: learning blocks versus probe blocks) rmANOVA on switch frequency data from the last day.

Reinforcement Learning Models

We fitted participants' choice behaviors to two reinforcement learning models: a standard Rescorla-Wagner (RW) model and a dual-rates (DR) model, where we fitted individual learning rates for positive and negative feedback separately. These two models were adapted from Wen et al. (2023) for use in a within-subject design. Similarly, each model was constructed using a hierarchical framework, implementing mixed-effects linear models to account for both fixed and random effects within the data structure for parameter estimates.

For the RW model, at the top level, the conditional learning rate α and inverse temperature β in each experimental condition were derived from their group-level average:

$$\alpha_C[p, v] = \alpha_G + \alpha_R[p, v]$$

$$\beta_C[p, v] = \beta_G + \beta_R[p, v]$$

Here, p stands for the factor of experiment phase (learning versus probe), and v stands for the factor of volatility level (high versus low). $\alpha_C[p, v]$ and $\beta_C[p, v]$ represent the learning rates and inverse temperature parameters in each condition, which are generated by a linear combination of group-level averages across all conditions and participants, α_G and β_G, and conditional random effects, $\alpha_R[p, v]$ and $\beta_R[p, v]$.

Participants' individual parameters were modelled at the hierarchical level below:

$$\alpha[i, p, v] = \alpha_C[p, v] + \alpha'_R[i, p, v]$$

$$\beta[i, p, v] = \beta_C[p, v] + \beta'_R[i, p, v]$$

Similarly, in the equations above, individual parameters in each condition are a linear combination of their fixed effects in each condition, $\alpha_C[p, v]$ and $\beta_C[p, v]$ as well as random effects, $\alpha'_R[i, p, v]$and $\beta'_R[i, p, v]$, at the individual level where i represents each participant.

At the lowest level of the hierarchy, the probability of choosing a card was computed by a SoftMax function, and the values of choosing each card were updated after feedback:

$$P_{i,c,t} = \frac{\exp(\beta[i,p,v] * Q_{i,t,n})}{\exp(\beta[i,p,v] * Q_{i,t,1}) + \exp(\beta[i,p,v] * Q_{i,t,2})}$$

$$Q_{i,t+1,n} = Q_{i,t,n} + \alpha[i, p, v] * (R_t - Q_{i,t,n})$$

Here Q stands for the value of each choice. Learning rate α quantifies the extent of Q value update after receiving feedback, with larger learning rates meaning being more flexible (Behrens et al., 2007). The inverse temperature β reflects how reliably people choose the option with the higher Q value. Larger β values indicate that choices are more deterministic and goal-directed, guided by Q values of options. In contrast, smaller β values indicate more exploratory and stochastic choice behavior. To fit human choice data, the Q values of the two choice cards (n = 1 or 2) were sent to a SoftMax function together with the conditional

inverse temperature parameter to compute the probability of choosing a certain card on trial *t*. Next, the Q value of the chosen card, $Q_{i,t+1,n}$, was updated by the prediction error, which is the difference between the received reward feedback R_t and the Q value of this card on trial *t*, $Q_{i,t,n}$. This update was weighted by the individual conditional learning rate $\alpha[i,p,v]$.

The DR model was constructed in the same hierarchical way, but the Q value update was implemented separately according to the received reward feedback:

$$Q_{i,c,t+1} = \begin{cases} Q_{i,c,t} + \alpha^{+}[p,v] * \left(R_t - Q_{i,c,t}\right) \; if \; R_t = 1 \\ Q_{i,c,t} + \alpha^{-}[p,v] * \left(R_t - Q_{i,c,t}\right) \; if \; R_t = 0 \end{cases}$$

Here $\alpha^{+}[p,v]$ and $\alpha^{-}[p,v]$ represent learning rates for positive reward feedback (positive learning rate), meaning how much Q value is updated when receiving positive feedback R_t = 1 or negative reward feedback (negative learning rate), R_t = 0, respectively.

The same models were also used in Experiment 2. To account for the random effect of different experiment days, we included another fixed factor *day* in the above hierarchical structure. Notably, because there was no testing phase in the first three learning days, the corresponding model parameters were omitted from fitting.

Model Fitting and Model Comparison

Parameters in the two hierarchical reinforcement models were estimated using Hamiltonian Monte Carlo sampling, implemented in Stan (version 2.32.2). The learning rates at the top hierarchy were sampled from a normal distribution $\mathcal{N}(0,1)$, while inverse temperature parameters were sampled from a normal distribution $\mathcal{N}(5,5)$. This prior selection was the same as Wen and colleagues (2023). We also applied non-centered parameterization to reduce parameter dependence across hierarchies for a more reliable estimation. Hierarchical Bayesian modelling estimates the posterior distributions, not point estimates, of parameters in each condition. Therefore, we computed posterior probabilities to

compare parameters in each condition. The model comparison was conducted by using the leave-one-out information criterion (LOOIC) method, with a higher expected log pointwise predictive density (ELPD) value indicating better fitting performance.

Model Validation

We conducted parameter recovery simulations to ensure our models and experimental design allow for reliable model fitting in two volatility environments. As our main interest is the difference between group-level model estimates in the probe blocks, model simulations were implemented with the same volatility level we used in the probe blocks where the matching rule changed every 20 trials for each model, and the feedback validity was set to 80%. We ran 60 simulations which is close to our actual sample size in each experiment. Each volatility condition was simulated with two blocks of 60 trials, hence 240 trials in total in each simulation. Simulated choice behaviors were fitted using the corresponding model. We then ran correlations between the fitted group-level estimates and the true group-level parameters, as well as their difference, to assess our models' performance on reproducing the true group-level parameters and group differences.

For the RW model we ran five simulations with different group-level learning rates in high- and low-volatility conditions. Specifically, each time individual learning rates in a high-volatility condition were sampled from a normal distribution with a group-level mean $\alpha_C \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and a group-level standard deviation $\sigma_\alpha = 0.07$, while individual learning rates in a low-volatility condition were sampled from a normal distribution with a group-level mean $\alpha_C \in \{0.5, 0.4, 0.3, 0.2, 0.1\}$, and a group-level standard deviation $\sigma_\alpha = 0.07$. Inverse temperatures were sampled from a normal distribution with a group-level mean $\beta_C = 3.64$ and a group-level standard deviation $\sigma_\beta = 1.08$ in the high-volatility condition, and from a normal distribution with a group-level mean $\beta_C = 3.77$ and a group-level standard

deviation σ_β = 1.11 in the low-volatility condition. β_C, σ_β and σ_α were determined based on the RW model fitting results on the human dataset, the same procedure was applied to the subsequent DR model simulations. Therefore, we created a group-level learning rate difference of {0, 0.2, 0.4, 0.6, 0.8} across five simulations. Our simulation results (Fig. 1A) revealed that the Pearson correlation coefficient between the true and fitted group-level learning rates in the high-volatility condition was $r > 0.99$, $p < .001$, and $r > 0.99$, $p < .001$, for the low-volatility condition. More importantly, the Pearson correlation coefficient between the true and fitted group-level learning rate differences was $r > 0.99$, $p < .001$, suggesting the RW can correctly capture the true difference between the group-level learning rates in two volatility conditions.

In the DR model simulations, we simulated two types of learning rate for positive and negative feedback respectively. We again created a group-level learning rate difference of {0, 0.2, 0.4, 0.6, 0.8} in the positive/negative learning rates in two volatility conditions. Individual positive/negative learning rates in the high-volatility condition were sampled from a normal distribution with a group-level mean $\alpha_C \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, and a group-level standard deviation σ_α = 0.13, and in the low-volatility condition from a normal distribution with a group-level mean $\alpha_C \in \{0.5, 0.4, 0.3, 0.2, 0.1\}$, and a group-level standard deviation σ_α = 0.12. Inverse temperatures were sampled from a normal distribution with a group-level mean β_C = 3.56 and a group-level standard deviation σ_β = 1.08 in the high-volatility condition, and from a normal distribution with a group-level mean β_C = 3.77 and a group-level standard deviation σ_β = 1.08 in the low-volatility condition. β_C, σ_β and σ_α were determined based on the DR model fitting results on the human dataset. We simulated all possible combinations of the group-level differences in positive and negative learning rates, resulting in 25 (5 * 5) simulations in total. The simulation results (Fig. 1B) revealed that the Pearson correlation coefficient between the true and fitted group-level positive learning rates

in the high-volatility condition was $r = 0.81$, $p < .001$, and $r = 0.85$, $p < .001$ for the low-volatility condition. Similarly, the Pearson correlation coefficient between the true and fitted group-level negative learning rates in the high-volatility condition was $r = 0.90$, $p < .001$, and $r = 0.94$, $p < .001$ for the low-volatility condition. The DR model also correctly capture the group-level difference in positive learning rates, $r = 0.84$, $p < .001$, and negative learning rates, $r = 0.99$, $p < .001$. Across two groups of simulations on RW and DR models, we showed that our models can provide reliable model fits to represent group-level differences in the probe blocks.

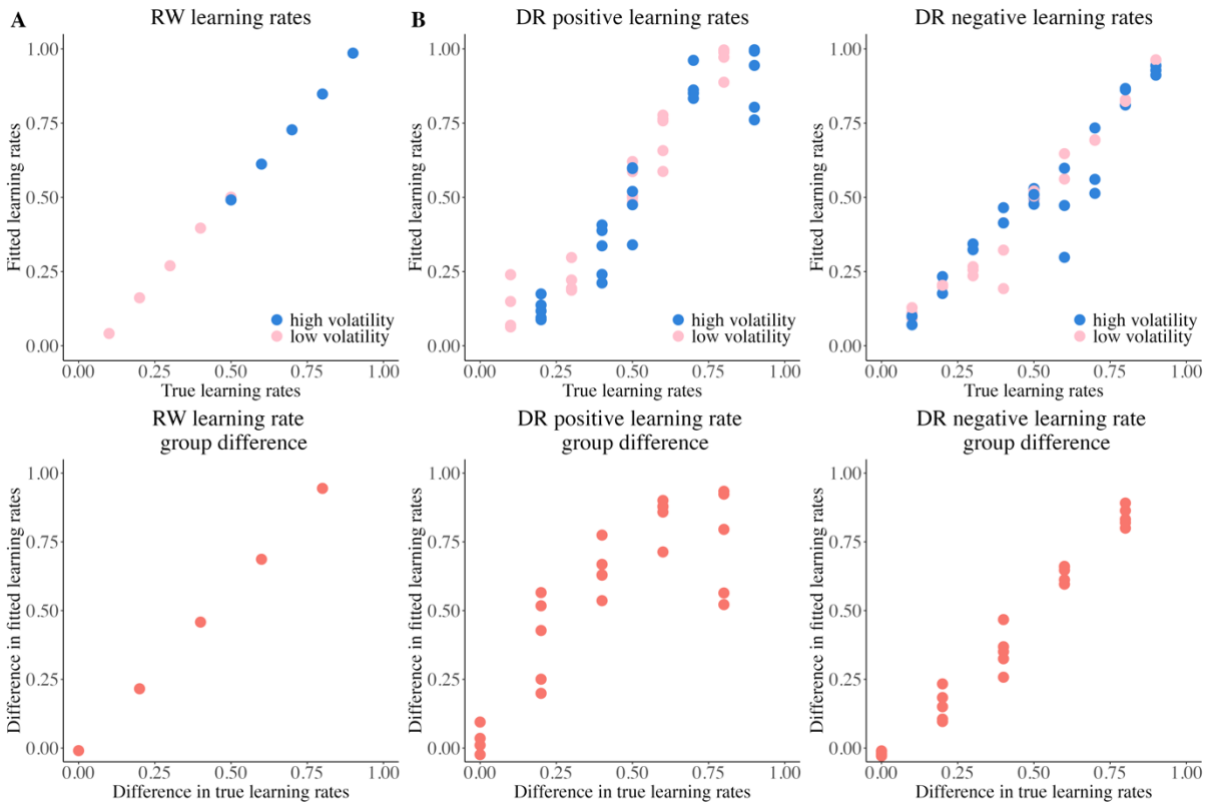


Figure 2: **A**. Correlation results between fitted versus true learning rates (top), and their group-level difference (bottom) in the RW model. **B**. Correlation results between fitted versus true positive learning rates (top left), negative learning rates (top right), and their differences (positive: bottom left; negative: bottom right) in the DR model.

Results

Behavioral Accuracy

Table 1 summarizes participants' accuracy across the two experiments. In Experiment 1, our two (volatility: high versus low) × two (phase: learning blocks versus probe blocks) rmANOVA results demonstrated a significant main effect of volatility, $F(1, 54) = 156.89$, $p < .001$, $\eta_p^2 = 0.74$, showing better performance in the low-volatility environment. However, a significant interaction of phase and volatility, $F(1, 54) = 346.77$, $p < .001$, $\eta_p^2 = 0.87$ suggested that this was driven by the significant accuracy difference in the learning blocks, showing lower accuracy in the high-volatility environment, $t(54) = 20.78$, $p < .001$, Cohen's $d = 2.80$, 95% confidence interval $CI_{\text{high-low}} = [-18.87, -14.54]$, but not the probe blocks, $t(54) = -1.82$, $p = .740$, Cohen's $d = 0.25$, 95% $CI_{\text{high-low}} = [-1.11, 3.86]$.

In Experiment 2, the three (day: day 1 versus day 2 versus day 3) × two (phase: learning blocks versus probe blocks) rmANOVA revealed that participants improved their performance over successive days (Table 1), suggested by a significant main effect of day, $F(1.70, 91.84) = 18.71$, $p < .001$, $\eta_p^2 = 0.26$. They also showed a main effect of volatility, $F(1, 54) = 792.07$, $p < .001$, $\eta_p^2 = 0.94$, which did not further interact with day, $F(1.50, 81.10) = 0.59$, $p = .509$, $\eta_p^2 = 0.01$. The two (volatility: high versus low) × two (phase: learning blocks versus probe blocks) rmANOVA on day 4 revealed a main effect of volatility, indicating that participants still performed better in the low-volatility environment even after extensive training, $F(1, 54) = 389.49$, $p < .001$, $\eta_p^2 = 0.88$. We again observed a significant interaction of phase and volatility, $F(1, 54) = 135.59$, $p < .001$, $\eta_p^2 = 0.72$, indicating worse performance in the high-volatility environment in learning blocks, $t(54) = 18.90$, $p < .001$, Cohen's $d = 2.55$, 95% $CI_{\text{high-low}} = [-20.13, -15.05]$, but not in probe blocks, $t(54) = 0.25$, $p = .807$, Cohen's $d = 0.03$, 95% $CI_{\text{high-low}} = [-2.98, 2.58]$.

Table 1

Behavioral accuracy (%) in two experiments

Experiment 1				Experiment 2									
				Day 1		Day 2		Day 3		Day 4			
Learning		Probe		Learning		Learning		Learning		Learning		Probe	
High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
68.82 (5.46)	85.53 (5.97)	81.65 (6.70)	80.27 (6.44)	67.72 (5.73)	84.37 (6.12)	71.13 (5.37)	87.60 (4.66)	70.41 (7.43)	87.83 (4.95)	70.09 (7.76)	87.68 (5.48)	82.04 (7.15)	82.24 (7.54)

Note. Learning and probe stand for learning blocks and probe blocks. High and low stand for high-volatility and low-volatility. Values given in the table are mean accuracy and values in parentheses are standard deviations.

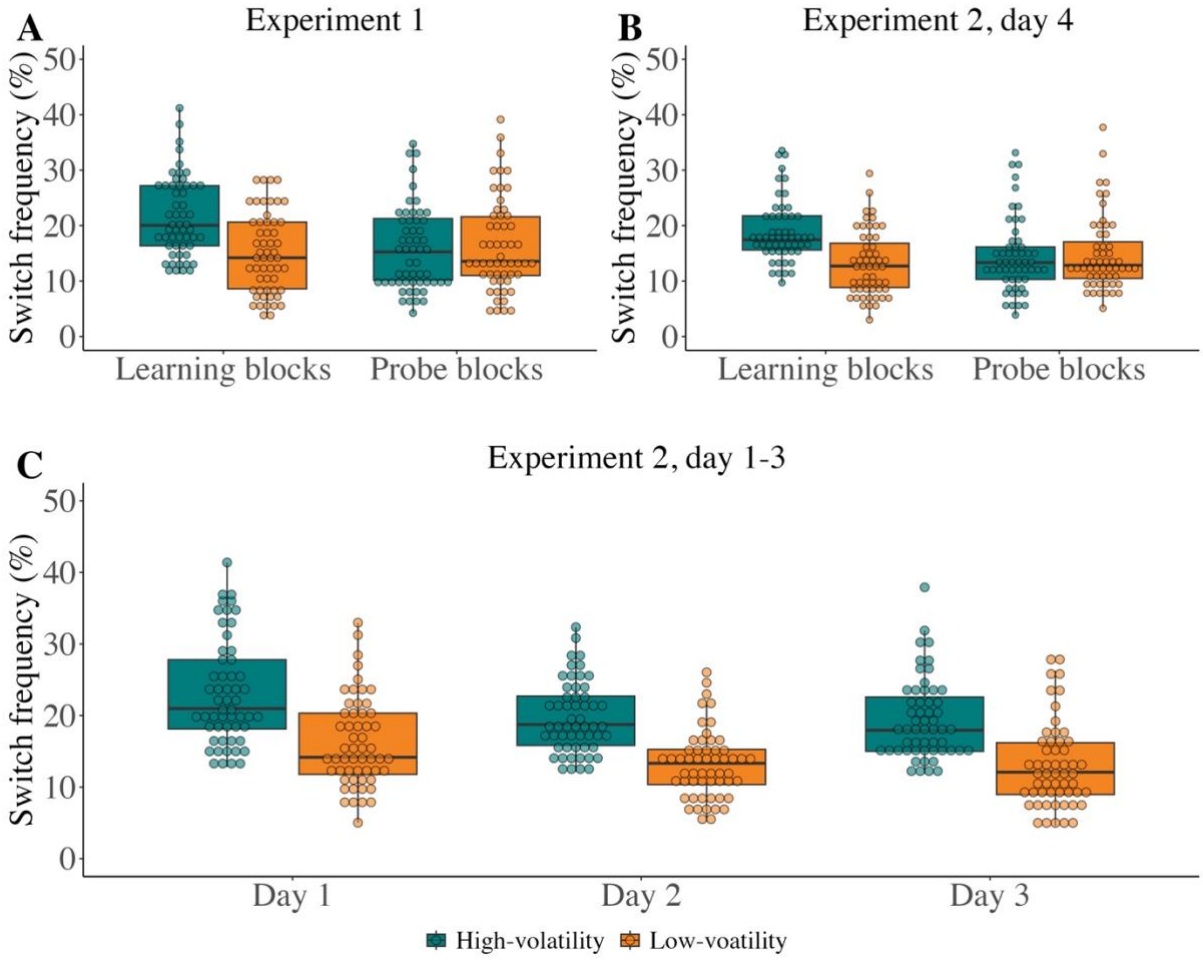


Figure 3: Switch frequency in Experiment 1 (**A**), Experiment 2, day 4 (**B**) and Experiment 2 day 1-3 (**C**). Individual dots represent each participant's values.

Switch Frequency

In Experiment 1, we observed a significant main effect of volatility, $F(1, 54) = 39.99$, $p < .001$, $\eta_p^2 = 0.43$, indicating that, as expected, participants switched more frequently in the high-volatility environment than the low-volatility environment (Fig. 3A), 95% $CI_{\text{high-low}}$ = [0.87, 5.03]. Moreover, the switch frequency was also higher in the learning blocks, $F(1,54) = 17.65$, $p < .001$, $\eta_p^2 = 0.25$, 95% $CI_{\text{learning-probe}}$ = [0.16, 4.35]. However, the significant interaction of volatility and phase, $F(1, 54) = 79.62$, $p < .001$, $\eta_p^2 = 0.60$, indicated that although participants showed a higher switch frequency in the high-volatility environment in learning blocks, $t(54) = 11.32$, $p < .001$, Cohen's $d = 1.53$, 95% $CI_{\text{high-low}}$ = [3.95, 9.37], there was no difference between two volatility environments in the probe blocks, $t(54) = -1.16$, $p = .249$, Cohen's $d = 0.16$, 95% $CI_{\text{high-low}}$ = [-3.76, 2.22].

In Experiment 2, we first observed a significant main effect of training days (Fig. 3C), $F(1.70, 91.88) = 20.85$, $p < .001$, $\eta_p^2 = 0.28$. Post-hoc analyses revealed that, participants switched matching rules less frequently on day 2 compared to day 1, $t(54) = 5.12$, $p < .001$, Cohen's $d = 0.69$, 95% $CI_{\text{day 2-day 1}}$ = [-5.47, -1.39], and their switching frequency remained stable from Day 2 onward, $t(54) = 0.46$, $p = .648$, Cohen's $d = 0.06$, 95% $CI_{\text{day 3-day 2}}$ = [-2.27, 1.82]. In addition, across three training days, participants consistently switched their matching rules more frequently in a high-volatility environment, $F(1, 54) = 262.78$, $p < .001$, $\eta_p^2 = 0.83$, 95% $CI_{\text{high-low}}$ = [4.85, 8.54]. However, the difference in switch frequency between two volatility environments was stable over the course of learning, indicated by an insignificant interaction of training days and volatility, $F(1.55, 83.78) = 0.23$, $p = .740$, $\eta_p^2 < 0.01$.

Similarly, after four days of training in Experiment 2, participants robustly showed a higher switch frequency in the high volatility environment (Fig. 3B), indicated by a main

effect of volatility, $F(1, 54) = 38.94$, $p < .001$, $\eta_p^2 = 0.42$, reflecting a higher switch likelihood in a high-volatile environment, 95% $CI_{\text{high-low}} = [1.00, 4.41]$. We also observed a higher switch frequency in the learning blocks, $F(1, 54) = 16.44$, $p < .001$, $\eta_p^2 = 0.23$, 95% $CI_{\text{learning-probe}} = [-0.33, 3.14]$. There was again a significant interaction between volatility and phase in Experiment 2, $F(1, 54) = 57.00$, $p < .001$, $\eta_p^2 = 0.51$. The post-hoc test results revealed that, while participants kept displaying a higher switch frequency in the learning blocks, $t(54) = 12.30$, $p < .001$, Cohen's $d = 1.65$, 95% $CI_{\text{high-low}} = [3.66, 7.98]$, this effect was again missing in the probe blocks, $t(54) = -0.59$, $p = .557$, Cohen's $d = 0.08$, 95% $CI_{\text{high-low}} = [-2.94, 2.11]$. Taken together, we found that participants switched between task rules more frequently in the high volatility environment, but this effect did not extend to the probe blocks, even after extensive training.

Model Comparison Results

We computed ELPD values between the candidate model and the best model (0 for the best model) and the standard error (SE) of these ELPD difference. The DR model outperformed the RW model in both experiments (Table 2). Therefore, our modelling analyses focused on the DR model parameters in the main text. We report results of the same analyses on the RW model in supplementary materials.

Table 2

Model comparison results

	Model type	Δ ELPD	Δ SE
Exp 1	DR model	0	0
	RW model	-84.0	12.8
Exp 2	DR model	0	0
	RW model	-1509.9	49.6

Environment-Specific Learning Rates

Learning Phase. We first examined whether the different volatility conditions evoked different task rule updating strategies in the learning phase, by comparing model estimates from different conditions. The significance level was defined by the probability that one posterior distribution is greater than the other distribution. In Experiment 1, we observed that positive learning rates were significantly lower in the low-volatility environment (Fig. 4A), posterior probability (p_post) = .016, but not the negative learning rates, p_post = .910. In addition, the inverse temperature parameters were significantly lower in the high-volatility environment, p_post = .003, indicating behavior was slightly less influenced by the learned values and thus more random. Similarly, after four days training in Experiment 2, participants consistently used higher positive learning rates in response to high volatility (Fig. 4B), p_post < .001. Moreover, the inverse temperature parameters were again lower in the high-volatility environment, p_post < .001. However, there was no difference between two volatility environments in negative learning rates, p_post = .278.

Testing Phase. Critical to our hypothesis, we next investigate whether the observed parameter patterns in the learning phase extend to the probe phase. In Experiment 1, there was no difference between two volatility environments in positive learning rates, p_post = .827, negative learning rates, p_post = .708, or inverse temperature parameters, p_post = .348. Consistent with the behavioral findings reported above, these null results suggest that participants did not use any learned environment-specific strategy, but rather showed a more direct, adaptive adjustment to the local needs of the testing phase.

Our key hypothesis was that participants would show environment-specific learning rates in the testing phase after four days training (Experiment 2), as evoked by the environment cue. There was again no difference in positive learning rates, p_post = .215, or

inverse temperature parameters, p_post = .741. However, this time, negative learning rates were higher in the high volatility environment, p_post = .011, suggesting that participants adopted different task rule updating strategies toward negative feedback in the testing phase. These results provide important evidence of the need for extensive training in establishing environment-specific strategies to regulate cognitive flexibility.

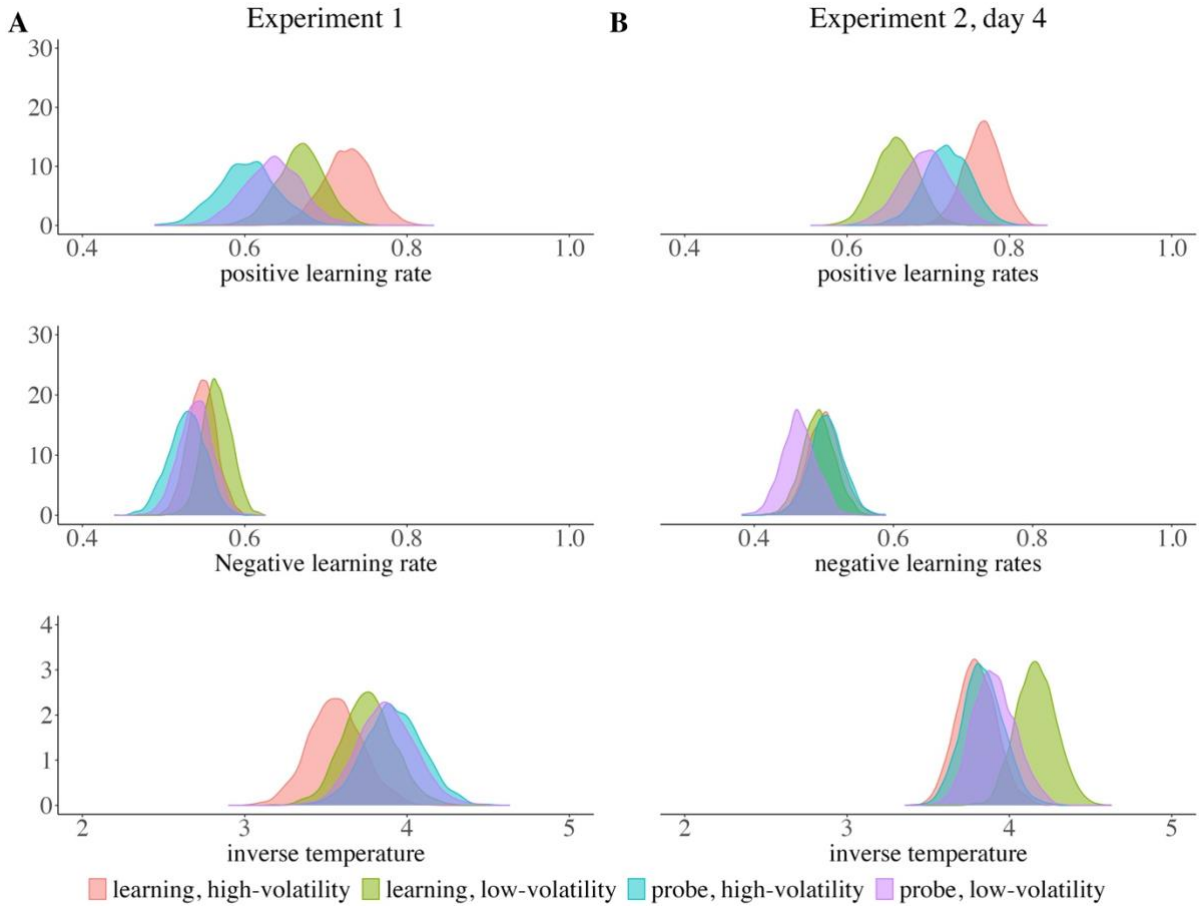


Figure 4: DR model parameter estimates in Experiment 1 (**A**) and Experiment 2 (**B**).

Parameter Evolution over Learning

To study how participants improved their task knowledge and developed this more environment-specific task rule updating strategy, we also analyzed the parameters from the first three learning days in Experiment 2. We first analyzed whether there was an overall trend of parameter estimates across learning days, by comparing parameter estimates between

two successive learning days. The results revealed that there was no difference in positive learning rates (Fig. 5A) between Day 1 and Day 2, p_post = .661 or Day 2 and Day 3, p_post = .144, suggesting positive learning rates, on average, remain stable over time. Interestingly, however, negative learning rates (Fig. 5B) decreased on Day 2, p_post < .001 and became stable as of Day 3 p_post = 0.331. Similarly, the inverse temperature parameters increased on Day 2, p_post < .001. Although the inverse temperature parameters (Fig. 4C) were again larger on Day 3, this difference did not reach significance, p_post = .921.

Next, to compare the parameter difference between two volatility environments over learning, we computed highest density intervals between posterior distributions. Interestingly, the difference between positive learning rates in two volatility environments reduced over learning (95% highest density interval (HDI), Day 1: [0.06, 0.15]; Day 2: [0.01, 0.08]; Day 3: [0.00, 0.07]). Specifically, positive learning rates in the low volatility environment increased on Day 2, p_post = .001, but did not further increase on Day 3, p_post = .225. In contrast, positive learning rates, if anything, decreased in the high volatility environment, albeit not significantly, all p_posts > .063.

In contrast to this evolution in positive learning rates, we observed that the difference between negative learning rates actually increased (95% HDI, Day 1: [-0.01, 0.04]; Day 2: [0.02, 0.07]; Day 3: [0.02, 0.07]) over learning. When studied individually, the negative learning rate in the low volatility environment reduced on Day 2, p_post < .001 and stayed consistent on Day 3, p_post = 0.301. Similarly, in the high volatility environment, the negative learning reduced on Day 2, p_post < 0.001, and kept stable on Day 3, p_post = .362.

Finaly, the difference in inverse temperature parameters decreased over days (95% HDI, Day 1: [0.13, 0.36]; Day 2: [0.04, 0.33]; Day 3: [0.03, 0.30]), meaning participants' choice behavior became more goal-directed through leaning. The inverse temperature

parameters in both volatility environments first increased on Day 2, all p_posts < .001, and remained stable on Day 3, all p_posts > .910.

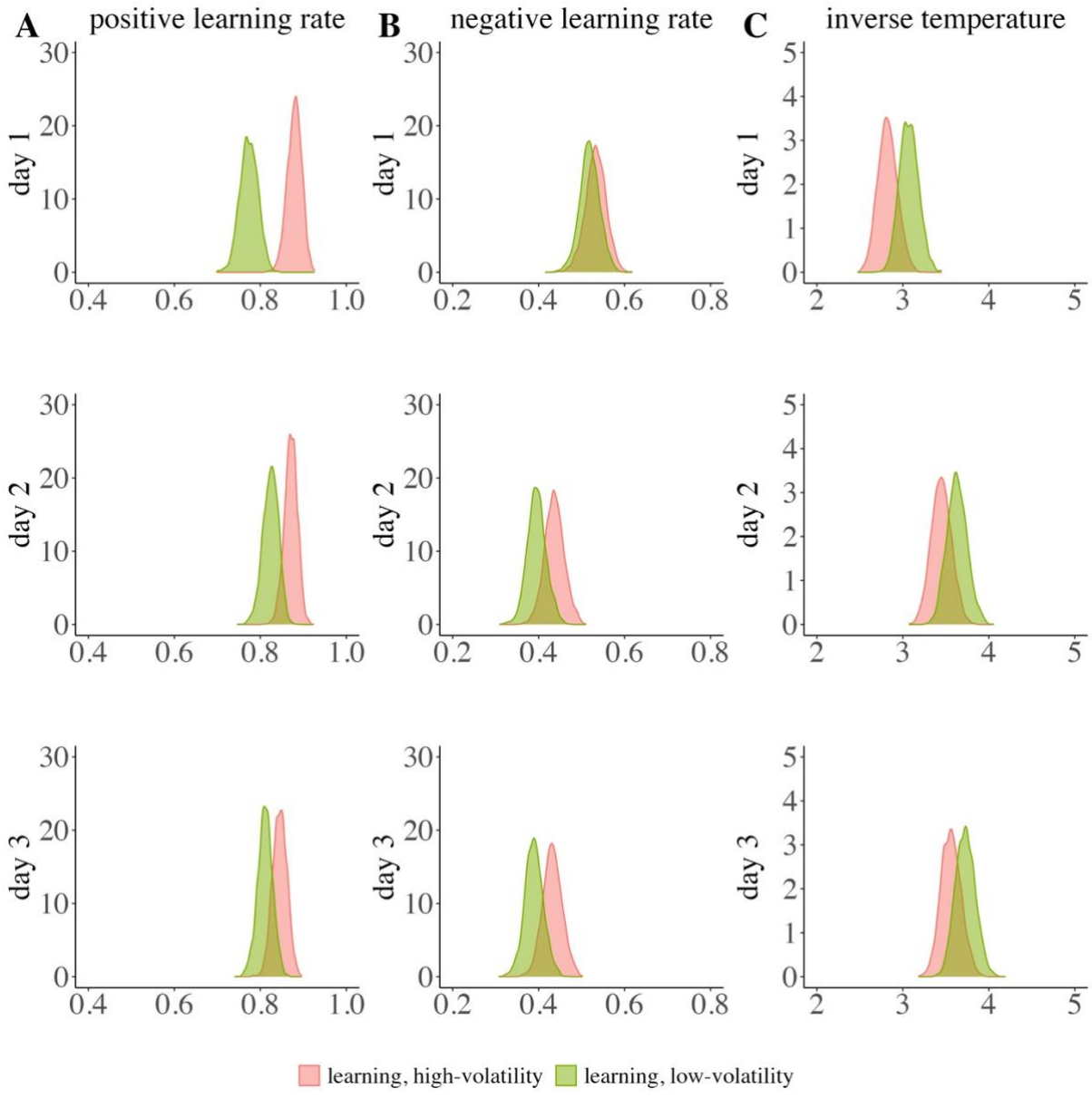


Figure 5: DR model parameter estimates from three learning days in Experiment 2.

Exploratory Analysis: Accuracy across Trials in Probe Blocks

Interestingly, while our modelling analyses suggest learned environment-specific task rule updating strategies after four days of training, our behavioral analyses focusing on average task accuracy or mean switch frequencies were unable to catch these subtle

differences in behavior. Wen and colleagues (2023) conducted a trial-wise analysis based on behavioral accuracy to assess participants' tendency to switch the matching rule before and after switch points. Here, we conducted a similar exploratory analysis to see how participants tracked the correct matching rules on each trial, by examining their behavioral accuracy across trials in the probe blocks. Across both experiments, participants gradually increased accuracy to around 80%, which was expected as we provided valid feedback on 80% of trials. Their accuracy dropped after a matching rule switched and increased again on the following trials. This general pattern indicates that participants were able to efficiently update their task rules according to feedback (Fig. 5), instead of other strategies such as counting the number of trials to predict an upcoming switch. To explore further differences in behavior, we applied uncorrected paired *t* tests to compare participants' accuracy data in both volatility environments on each trial. In Experiment 1 (Fig. 6A), participants only reached a higher accuracy on trial 5, $t(54) = 2.21$, $p = .032$, Cohen's $d = 0.30$, 95% $CI_{\text{high-low}} = [0.01, 18.17]$ and trial 40, $t(54) = 2.42$, $p = .019$, Cohen's $d = 0.33$, 95% $CI_{\text{high-low}} = [1.69, 14.67]$, in probe blocks with a high-volatility picture. There was no significant difference between high-volatility and low-volatility on other trials, all $t(54)$s < 1.96, ps > 0.055. In contrast, in Experiment 2 (Fig. 6B), participants reached better performance in probe blocks with a low-volatility environmental cue on trials 12, $t(54) = 2.67$, $p = .010$, Cohen's $d = 0.36$, 95% $CI_{\text{high-low}} = [2.59, 16.80]$, and trial 14, $t(54) = 2.97$, $p = .004$, Cohen's $d = 0.40$, 95% $CI_{\text{high-low}} = [3.46, 18.36]$, while they still showed higher accuracy with a high-volatility environmental cue on trial 19, $t(54) = 2.11$, $p = .040$, Cohen's $d = 0.28$, 95% $CI_{\text{high-low}} = [0.39, 12.94]$. While explorative and less statistically robust, this analysis suggests that the environment-specific strategies that participants developed after four days of training, were characterized by an increased tendency to switch around trial 10 in the probe blocks.

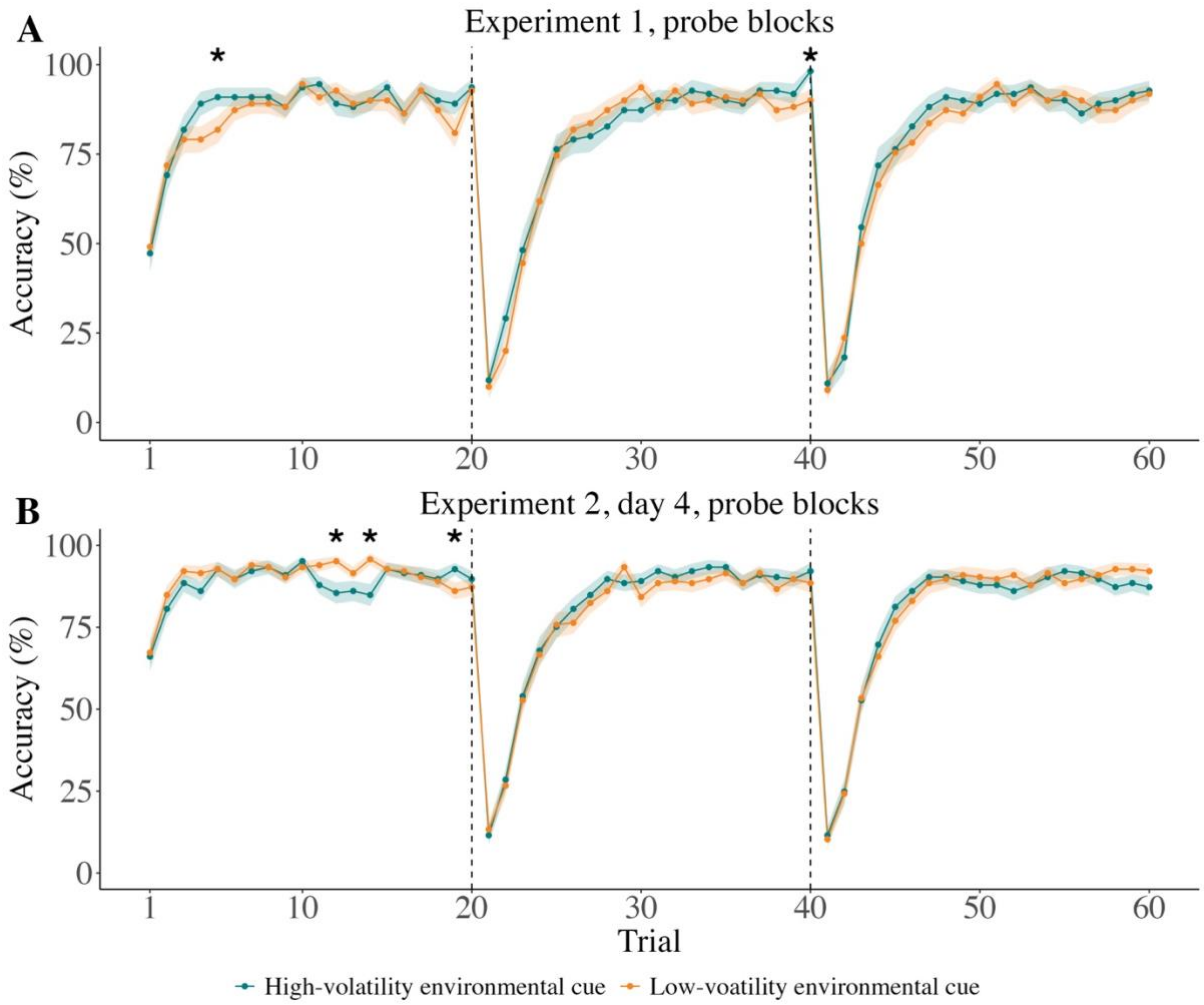


Figure 6: Accuracy over all trials in the probe blocks in Experiment 1 (**A**) and Experiment 2, day 4 (**B**). Dashed lines indicate a matching rule switch on the next trial. Shaded area stands for 95% confidence intervals (CI), * represents $p < .05$.

Discussion

We aimed to examine whether flexibly adjusting task rule updating strategies could occur through both adaptations based on recently experienced levels of volatility, as well as through the learning and use of environment-control knowledge, using a probabilistic Wisconsin Card Sorting Test task. Across two experiments, we observed that participants quickly adapted to experienced volatility showing fast control adaptations based on short-term experiences. Similar to previous studies (Behrens et al., 2007; Nassar et al., 2010; Piray & Daw, 2021; Simoens et al., 2024; Wen et al., 2023), we consistently observed that

participants adopted higher learning rates in a high-volatility environment, and showed higher switch frequencies but worse accuracy, suggesting our volatility settings successfully manipulated participants' task rule updating strategies. Importantly, after four days training, they also learned to associate different negative learning rates to co-occurring environmental cues, suggesting the control adaptations based on long-term knowledge. In addition, participants fine-tuned their task rule updating strategy through learning, as suggested by the distinct evolution of positive versus negative learning rates, highlighting the importance of a multi-day training schema in developing environment-specific task rule updating strategies.

More broadly, our findings indicate that although people rapidly adapt learning rates in response to different levels of volatility according to short-term experiences, they also learn to optimize and fine-tune these learning parameters to form long-term knowledge for future use over extensive practice. While adaptations based on short-term experiences are driven by moment-to-moment demands, adaptations based on long-term knowledge likely require exploring different environmental structures to identify optimal solutions, making it a more time-consuming process (Botvinick et al., 2019; Braem et al., 2024). In recent modelling work, Xu and colleagues (2024) further showed how this distinction between these two types of adaptations is analogous to in-context versus in-weight learning as observed in recurrent neural networks (see also, Binz et al., 2024), with each adaptation contributing to the regulation of task control across different timescales. Adaptations based on short-term experiences provide immediate solutions to control demands at the moment and reflect short-term changes on control settings, but they can be biased by contingencies. In contrast, adaptations based on long-term knowledge reflect more enduring, robust changes on control settings, but developed over learning. In this sense, the regulation of cognitive control can be seen as a meta-control process, where the prediction of the optimal control settings is updated in a manner of a Markov decision process with an integration of both short-term experience

and long-term knowledge (Behrens et al., 2007; Jiang et al., 2014).

Cognitive control has been argued to be organized and implemented hierarchically, analogous to a decision tree process from higher-level abstract features to lower-level concrete features (Badre, 2008; Becker et al., 2024; Eckstein & Collins, 2021; Egner, 2014). Furthermore, the human brain is considered an action-oriented system (Cisek & Kalaska, 2010; Fine & Hayden, 2022). Therefore, it is intuitive that humans tend to start from control adaptations based on short-term experiences as this approach is arguably more straightforward and instantaneous. This is in line with findings that intelligent agents naturally start learning from lower-level concrete features for cheap and fast initial learning (Collins & Frank, 2013; Kool et al., 2016; Morris, 2020; Wen et al., 2023). Despite these advantages, adaptations only based on short-term experiences might not be computationally optimal in the long term. While such adaptations offer fast solutions to control demands, they heavily rely on recent experiences and therefore can be biased by contingencies. In contrast, control adaptations based on long-term knowledge can determine adequate control settings according to learned environment-control associations developed through extended experiences over time. Compared to adaptations based on short-term experiences, this type of adaptation is more robust and efficient but also slower to learn (Decker et al., 2016; Doll et al., 2012; Evans & Stanovich, 2013; Sloman, 1996). Indeed, our results together with previous findings (Xu et al., 2024) suggest the formation and use of these environment-control associations in control adaptations emerged only after extensive training.

Our exploratory trial-wise accuracy results in Experiment 2 suggested a drop in accuracy after trial 10. This could be explained by learned task rule updating strategies to negative feedback in the high-volatility environment, which led to these unnecessary rule switches. Notably, our trial-wise results showed no further differences after trial 20, which also suggests that participants still used recent experiences to regulate task rule updating

strategies even after extensive learning. Indeed, while these two strategies contribute to meeting control demands in different ways, they are not necessarily exclusive to each other (Botvinick et al., 2019; De Neys, 2023).

Interestingly, when focusing on the effects of successive training days we noticed a gradual evolution of learning parameters over training. We observed a distinct evolution of positive versus negative learning rates over days. Specifically, over the course of learning, the difference in positive learning rates between two volatility environments diminished, whereas the difference in negative learning rates increased. Together, these findings suggest that participants might be generally more sensitive to local differences in positive feedback (rewards) at first, when needing to optimize their behavior in response to different volatilities. Given that positive learning rates were also generally higher throughout the experiment, our results could reflect an overall positivity bias (for a review, see Palminteri & Lebreton, 2022), where participants overweighted positive feedback compared to negative feedback. However, although the general size of positive learning rates remained stable over learning, the difference in positive learning rates between high- and low-volatility gradually reduced. This pattern reflects how participants fine-tuned their task rule updating strategies with more training. First, high volatility involves more frequent task rule switching compared to low volatility, resulting in generally higher learning rates. In addition, although positive feedback seems generally motivating in our experiments, it does not indicate a need for task rule switching. Therefore, it is unnecessary to maintain distinct positive learning rates across two volatility environments.

Conversely, negative feedback is a critical signal in our experiments, as it indicates where a shift in task rules occurs, especially in the high-volatility environment. In the low-volatility environment, it more often reflected the noisiness of our feedback. Therefore, the decreasing negative learning rates, together with decreasing switch frequency, across training

days suggest that participants gradually learned to make better use of these noisy signals to avoid overreacting for task optimization. In a similar vein, the increasing inverse temperature parameters further indicated that participants obtained better structured task knowledge (i.e., increased behavioral accuracy) and responded less stochastically over time. More importantly, participants also learned to establish different negative learning rates across the two volatility environments, as indicated by increasing differences in negative learning rates. Consistent with our findings, previous studies using similar designs revealed that adaptive changes in environment-specific learning strategies were best captured by differences in negative learning rates (Simoens et al., 2024; Wen et al., 2023).

We showed that even within uncertain task environments, humans still need a considerable amount of learning before they can acquire and apply environment-control knowledge. However, an open question is which factor(s) make this a longer process. One possible answer is that people might need more time to obtain abstract knowledge, as they may tend to start learning from lower-level intuitive and concrete features first and move to abstract knowledge when necessary (Bugg, 2014; Gershman et al., 2015; Kool et al., 2016). Another answer may be that humans need extensive learning to also enable the translation of learned abstract control strategies to concrete action plans. For example, simply cueing participants with the need for control often does not predict better task performance in response to conflicts or need for task learning (Braem et al., 2017; Brass et al., 2017; Bugg & Smallwood, 2016; Chai et al., 2025; Jiménez et al., 2021). The regulation of cognitive control is often highly task-specific (Siqi-Liu & Egner, 2023; Wen et al., 2023; Zhu et al., 2025; but see Ileri-Tayar et al., 2024), emphasizing the role of concrete knowledge of task processing in the generalization of abstract control strategies. Therefore, the implementation of abstract control strategies may be less favored compared to relying on local adaptations where control settings are shaped by the experience with more concrete task information at the beginning.

However, tracking momentary changes is cognitively demanding and does not always lead to an optimal solution. For example, it may result in convergence to a local minimum. Therefore, humans may choose between different types of adaptations by evaluating their benefits and costs. To account for why extensive practice is required for adaptations based on long-term knowledge, future work should aim to precisely disentangle the formation of abstract control knowledge from its implementation in behavioral adaptation. This would help shed light on the mechanisms underlying different adaptation strategies, particularly how humans evaluate and apply them to optimize behavior.

Taken together, our study investigated behavioral mechanisms underlying the environment-specific regulation of task rule updating strategies. We provide empirical evidence to support a critical contribution of multi-day training schemas to form environment-control associations. Our results provide new insights into the gradual learning of control, highlighting how people eventually learn to adaptively regulate task rule updating strategies with the help of environmental features.

References

Aben, B., Verguts, T., & Van den Bussche, E. (2017). Beyond trial-by-trial adaptation: A quantification of the time scale of cognitive control. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 509–517. http://dx.doi.org/10.1037/xhp0000324

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693–728. https://doi.org/10.1037/bul0000047

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Badre, D. (2025). Cognitive Control. *Annual Review of Psychology*, *76*(Volume 76, 2025), 167–195. https://doi.org/10.1146/annurev-psych-022024-103901

Bai, Y., Katahira, K., & Ohira, H. (2014). Dual learning processes underlying human decision-making in reversal learning tasks: Functional significance and evidence from the model fit to human behavior. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00871

Becker, D., Bijleveld, E., Braem, S., Fröber, K., Götz, F. J., Kleiman, T., Körner, A., Pfister, R., Reiter, A. M. F., Saunders, B., Schneider, I. K., Soutschek, A., Steenbergen, H. van, & Dignath, D. (2024). An integrative framework of conflict and control. *Trends in Cognitive Sciences*, *28*(8), 757–768. https://doi.org/10.1016/j.tics.2024.07.002

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147. https://doi.org/10.1017/S0140525X23003266

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Braem, S., Chai, M., Held, L. K., & Xu, S. (2024). One cannot simply 'be flexible': Regulating control parameters requires learning. *Current Opinion in Behavioral Sciences*, *55*, 101347. https://doi.org/10.1016/j.cobeha.2023.101347

Braem, S., & Egner, T. (2018). Getting a Grip on Cognitive Flexibility. *Current Directions in Psychological Science*, *27*(6), 470–476. https://doi.org/10.1177/0963721418787475

Braem, S., Liefooghe, B., De Houwer, J., Brass, M., & Abrahamse, E. L. (2017). There are limits to the effects of task instructions: Making the automatic effects of task instructions context-specific takes practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 394–403. https://doi.org/10.1037/xlm0000310

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews*, *81*, 16–28. https://doi.org/10.1016/j.neubiorev.2017.02.012

Bugg, J. M. (2014). Conflict-triggered top-down control: Default mode, last resort, or no such thing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 567–587. http://dx.doi.org/10.1037/a0035032

Bugg, J. M., & Smallwood, A. (2016). The next trial will be conflicting! Effects of explicit congruency pre-cues on cognitive control. *Psychological Research*, *80*(1), 16–33. https://doi.org/10.1007/s00426-014-0638-5

Chai, M., Palenciano, A. F., Mill, R., Cole, M. W., & Braem, S. (2025). It’s Hard to Prepare for Task Novelty: Cueing the Novelty of Upcoming Tasks Does Not Facilitate Task Performance. *Journal of Cognition*, *8*(1). https://doi.org/10.5334/joc.423

Chiu, Y.-C., & Egner, T. (2017). Cueing cognitive flexibility: Item-specific learning of switch readiness. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1950–1960. https://doi.org/10.1037/xhp0000420

Cisek, P., & Kalaska, J. F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience*, *33*(Volume 33, 2010), 269–298. https://doi.org/10.1146/annurev.neuro.051508.135409

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111. https://doi.org/10.1017/S0140525X2200142X

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychological Science*, *27*(6), 848–858. https://doi.org/10.1177/0956797616639301

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

Eckstein, M. K., & Collins, A. G. E. (2021). How the Mind Creates Structure: Hierarchical Learning of Action Sequences. *CogSci ... Annual Conference of the Cognitive Science Society. Cognitive Science Society (U.S.). Conference*, *43*, 618–624.

Egner, T. (2014). Creatures of habit (and control): A multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology*, *5*, 1247. https://doi.org/10.3389/fpsyg.2014.01247

Egner, T. (2023). Principles of cognitive control over task focus and task switching. *Nature Reviews Psychology*, 1–13. https://doi.org/10.1038/s44159-023-00234-4

Eppinger, B., Goschke, T., & Musslick, S. (2021). Meta-control: From psychology to computational neuroscience. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(3), 447–452. https://doi.org/10.3758/s13415-021-00919-4

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691612460685

Fine, J. M., & Hayden, B. Y. (2022). The whole prefrontal cortex is premotor cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200524. https://doi.org/10.1098/rstb.2020.0524

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322. https://doi.org/10.1073/pnas.1800755115

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30. https://doi.org/10.1016/j.cobeha.2019.01.005

Hubbard, J., Kuhns, D., Schäfer, T. A. J., & Mayr, U. (2017). Is conflict adaptation due to active regulation or passive carry-over? Evidence from eye movements. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *43*(3), 385–393. https://doi.org/10.1037/xlm0000306

Ileri-Tayar, M., Colvett, J. S., & Bugg, J. M. (2024). Between-task transfer of item-specific control is replicable and extends to novel conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(6), 535–553. https://doi.org/10.1037/xhp0001200

Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews*, *46*, 30–43. https://doi.org/10.1016/j.neubiorev.2014.06.001

Jiménez, L., Méndez, C., Abrahamse, E., & Braem, S. (2021). It is harder than you think: On the boundary conditions of exploiting congruency cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(10), 1686–1704. https://doi.org/10.1037/xlm0000844

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLOS Computational Biology*, *12*(8), e1005090. https://doi.org/10.1371/journal.pcbi.1005090

Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences*, *106*(42), 17951–17956. https://doi.org/10.1073/pnas.0905191106

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*(10), 1250–1252. https://doi.org/10.1038/nn.2904

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Morris, T. H. (2020). Experiential learning – a systematic review and revision of Kolb's model. *Interactive Learning Environments*, *28*(8), 1064–1077. https://doi.org/10.1080/10494820.2019.1570279

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/j.tics.2021.06.001

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621. https://doi.org/10.1016/j.tics.2022.04.005

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), 6587. https://doi.org/10.1038/s41467-021-26731-9

Scherbaum, S., Dshemuchadse, M., Ruge, H., & Goschke, T. (2012). Dynamic goal states: Adjusting cognitive control without conflict monitoring. *NeuroImage*, *63*(1), 126–136. https://doi.org/10.1016/j.neuroimage.2012.06.021

Simoens, J., Verguts, T., & Braem, S. (2024). Learning environment-specific learning rates. *PLOS Computational Biology*, *20*(3), e1011978. https://doi.org/10.1371/journal.pcbi.1011978

Siqi-Liu, A., & Egner, T. (2023). Task sets define boundaries of learned cognitive flexibility in list-wide proportion switch manipulations. *Journal of Experimental Psychology. Human Perception and Performance*, *49*(8), 1111–1122. https://doi.org/10.1037/xhp0001138

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Verbeke, P., & Verguts, T. (2024). Humans adaptively select different computational strategies in different learning environments. *Psychological Review*. https://doi.org/10.1037/rev0000474

Verguts, T., & Notebaert, W. (2009). Adaptation by binding: A learning account of cognitive control. *Trends in Cognitive Sciences*, *13*(6), 252–257. https://doi.org/10.1016/j.tics.2009.02.007

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. https://doi.org/10.1016/j.cobeha.2021.01.002

Wen, T., Geddert, R. M., Madlon-Kay, S., & Egner, T. (2023). Transfer of Learned Cognitive Flexibility to Novel Stimuli and Task Sets. *Psychological Science*, *34*(4), 435–454. https://doi.org/10.1177/09567976221141854

Xu, S., Simoens, J., Verguts, T., & Braem, S. (2024). Learning where to be flexible: Using environmental cues to regulate cognitive control. *Journal of Experimental Psychology: General*, *153*(2), 328–338. https://doi.org/10.1037/xge0001488

Xu, S., Verguts, T., & Braem, S. (2024). *Cognitive flexibility versus stability via activation-based and weight-based adaptations*. OSF. https://doi.org/10.31234/osf.io/w8meh

Zhu, D., Wang, X., Zhao, E., Nozari, N., Notebaert, W., & Braem, S. (2025). Cognitive control is task specific: Further evidence against the idea of domain-general conflict

adaptation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001480

Supplementary Materials

Here we present results from the same modelling analyses reported in the main text for the RW model.

Environment-Specific Learning Rates

Learning Phase. In Experiment 1, we did not observe significant differences in learning rates from the RW model during the learning phase (Fig. S1A; p_post = .312). Although the overall pattern was consistent with our expectation, showing higher learning rates in the high-volatility environment, the difference did not reach significance. However, inverse temperature parameters were still smaller in the high-volatile environment compared to in the low-volatility environment, p_post = .026. In Experiment 2, after four days of training (Fig. S1B), participants showed significantly higher learning rates, p_post < .001, and significantly smaller inverse temperature parameters, p_post < .001, in the high-volatility environment than in the low-volatility environment during the learning phase.

Testing Phase. Key to our main hypotheses, we expected participants to use environmental cues to regulate their task rule updating strategies during the testing phase, showing similar environment-specific model parameter estimates in the learning phase. However, in Experiment 1, we again did not observe any significant differences in neither learning rates, p_post = 0.865, nor inverse temperature parameters, p_post = .411, between two volatility environments. In Experiment 2, consistent with the results from the DR model, we found significantly higher learning rates in the high-volatility environment compared to the low-volatility environment during the testing phase on Day 4, p_post = .016. This again suggested that participants started to use environmental cues to guide their task rule updating strategies in probe blocks after four days of training. However, there was still no difference in inverse temperature parameters between two volatility environments during the testing phase,

p_post = 0.685.

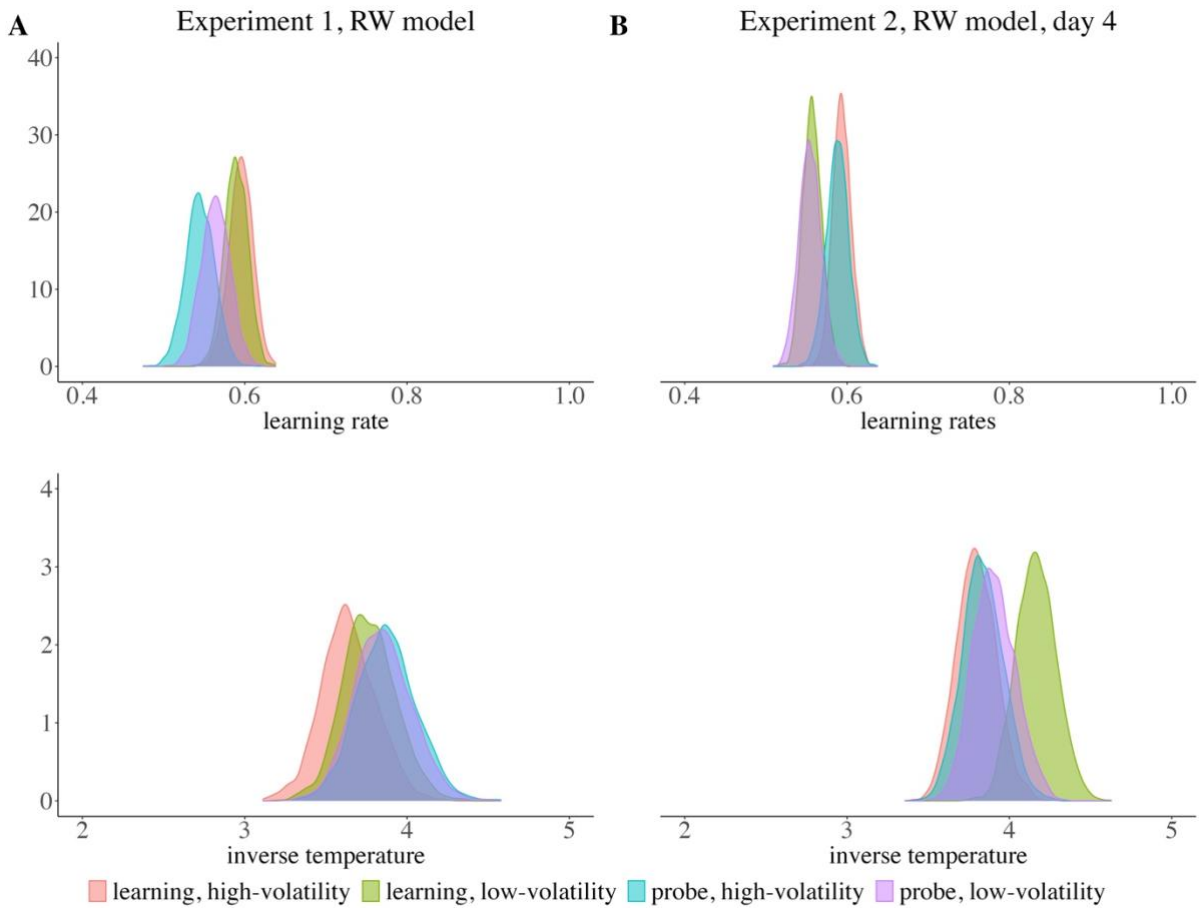


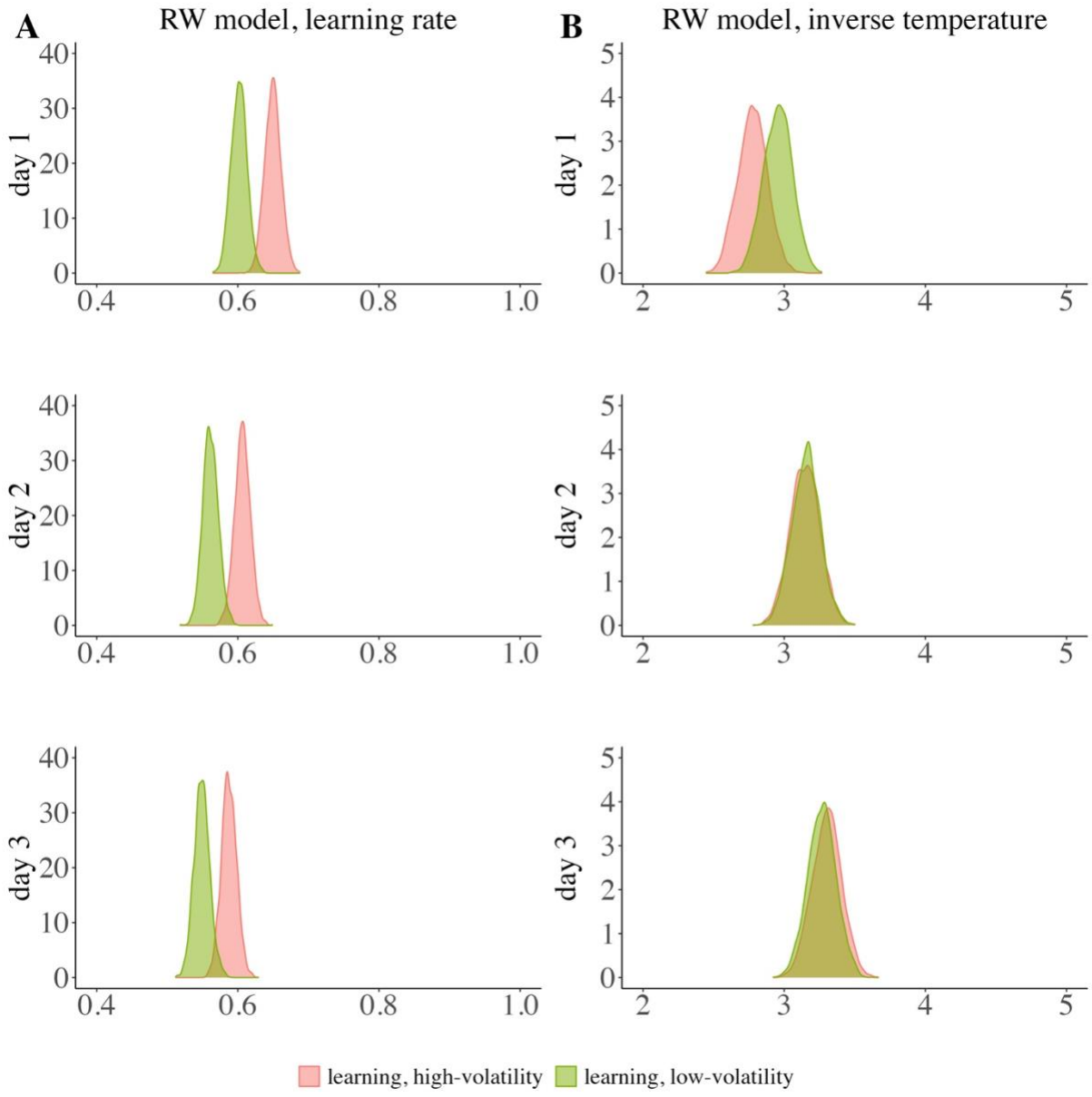
Figure S1: RW model parameter estimates in Experiment 1 (**A**) and Experiment 2 (**B**).

Parameter Evolution over Learning

We first examined changes in the RW model parameter estimates across the first three training days. Similarly, the results (Fig. S2A) revealed that overall learning rates first decreased from Day 1 to Day 2, p_post < .001, and then became stable on Day 3, p_post = .070. More specially, we found the learning rates in the high-volatility environment consistently reduced over learning, all p_posts < .0.032, whereas the learning rates in the low-volatility environment reduced from Day 1 to Day 2, p_post < .001, and stablized on Day 3, p_post < .108. In contrast, overall inverse temperature parameters consistently increased (Fig. S2B) over learning, showing significant differences between Day 1 and Day

2, p_post < .001, as well as significant differences between Day 2 and Day 3, p_post < .001. Similarly, inverse temperature parameters within each volatility environment also kept increasing over learning, all p_posts < .020. These results together again confirmed that participants fine-tuned their task rule updating strategies across training days.

We then measured changes of differences between two volatility environments in learning rates and inverse temperature parameters. Across three training days, learning rates were robustly higher in the high-volatility environment than in the low-volatility environment, all p_posts < .001. However, different from what we observed from the DR model fits, the learning rate differences between two environments from RW model were stable over learning (95% HDI, Day 1: [0.02, 0.07]; Day 2: [0.02, 0.07]; Day 3: [0.02, 0.06]). In contrast, there were significant differences in inverse temperature parameters between two volatility environments on Day 1, p_post < .001, 95% HDI = [0.07, 0.29]. But such differences were not found on Day 2, p_posts = .548, 95% HDI = [-0.11, 0.12] and Day 3, p_posts = .288, 95% HDI = [-0.16, 0.08].



855

856 **Figure S2:** RW model parameter estimates from three learning days in Experiment 2.

857