Transfer of learned cognitive flexibility to novel stimuli and task sets

Tanya Wen[1*], Raphael M. Geddert[1], Seth Madlon-Kay[1], Tobias Egner[1,2]

[1]Center for Cognitive Neuroscience, Duke University, Durham, NC, USA

[2]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

[*]tanya.wen@duke.edu

Abstract

Adaptive behavior requires learning about the structure of the environment to derive optimal action policies, and previous studies have documented transfer of such structural knowledge to bias choices in new environments. Here, we asked whether people could also acquire and transfer more abstract knowledge across different task environments, in particular, expectations about demands on cognitive control. Over three experiments, participants performed a probabilistic card-sorting task in environments of either a low or high volatility of task rule changes (requiring low or high cognitive flexibility) before transitioning to a medium-volatility environment. Using reinforcement learning modeling, we consistently found that previous exposure to high task rule volatility led to faster adaptation to rule changes in the subsequent transfer phase. This transfer of expectations about demands on cognitive flexibility was both task- (Experiment 2) and stimulus- (Experiment 3) independent, thus demonstrating the formation and generalization of environmental structure knowledge to guide cognitive control.

Keywords: cognitive flexibility, reinforcement learning, generalization, task switching, meta-flexibility

Introduction

Adaptive behavior requires us to identify and keep in mind the currently relevant "rules of the game" – that is, which responses to which stimuli likely lead to desirable outcomes (also known as task sets; Monsell, 2003). Moreover, given that the world is ever-changing, optimal regulation of task sets involves resolving a tradeoff between needing to implement the current task set and shielding it from distraction (cognitive stability) versus being ready to update (or switch) task sets in response to changing environmental contingencies (cognitive flexibility; Goschke, 2003; Nassar & Troiani, 2020). Importantly, neither stability nor flexibility are inherently beneficial; rather, it is the ability to dynamically adapt one's flexibility level to suit varying environmental demands, referred to as meta-flexibility, that facilitates optimal cognition (Goschke, 2013).

To adjust cognitive flexibility in an optimal manner, one must infer which task sets to use at a given time by observing environmental statistics (Behrens et al., 2007; Yu et al., 2020), such as associations between stimuli, responses, and outcomes. Accordingly, humans continuously monitor and integrate new information, creating representations of the set of states that exist

within a task, the transitions between them, and which actions likely lead to desirable outcomes in each state (Niv, 2019). This learning process can be characterized by reinforcement learning (RL) models, where people update their value estimations of actions in various states through trial and error (Barraclough et al., 2004; Lee et al., 2012; Sutton & Barto, 1998). Crucially, classic RL models involve a learning rate parameter, which estimates the degree to which newly encountered information is used to update one's beliefs. Previous studies have shown that people's learning rates are typically low during periods of environmental stability and high during periods of volatility (Behrens et al., 2007; Browning et al., 2015; Jiang et al., 2014, 2015; Massi et al., 2018).

We posit that successfully matching cognitive flexibility levels to varying demand contexts could be mediated by learning and transferring knowledge about the demand structure of one's environment. For example, while learning a novel task, it may be beneficial to exploit relevant information that was acquired in the past (Kemp et al., 2010; Mark et al., 2020; Yu et al., 2020). Previous studies have demonstrated that structural knowledge of an environment in the form of cognitive maps of stimulus associations (Mark et al., 2020) and correlated bandit arms (Baram et al., 2020; Schulz et al., 2020) can foster transferrable expectations about the structure of new environments. However, to the best of our knowledge, it has not been tested whether learning parameters driving cognitive control processes, such as task set updating, can be transferred to different temporal contexts, task sets, or stimuli. We here combined these two prior insights – volatility learning and structure transfer – to create a novel test of the acquisition and transfer of cognitive control policies, specifically, one's level of cognitive flexibility or switch-readiness.

To this end, the current study investigated whether participants learning to update task sets more or less frequently (i.e., at different learning rates) in one context would transfer their expectations to another context. Specifically, we conducted three experiments employing a probabilistic version of the Wisconsin Card Sorting Task (Berg, 1948; Van Eylen et al., 2011) wherein two groups of participants were initially exposed to either a low- or high-volatility learning environment, with seldom vs. frequent rule changes, respectively. Next, participants from both groups switched to the same medium-volatility transfer environment, which had an intermediate rate of rule changes. RL models were fit to participants' rule-choice behavior to quantify the rates of rule-updating (i.e., the learning rates) of the two groups in both task phases. We predicted that participants who were initially exposed to the more stable, low-volatility condition would have a lower rule learning rate compared to participants who experienced the high-volatility condition, and that the learning rates would generalize to the transfer phase. Across the three experiments, we systematically decreased the task and stimulus overlap to investigate whether the similarity between the learning and transfer phases influenced learning rate transfer.

General Methods

In three experiments, we examined whether participants could acquire and transfer knowledge about cognitive flexibility demands across different contexts. In all experiments, participants were split into two groups (low- and high-volatility) that completed a learning phase where the

task sets switched less or more frequently, respectively. Next, we tested them in a medium-volatility environment transfer phase, where the switch rate of task rules was the same for both groups. Our main question was whether expectations about the frequency of task-set updating acquired in the learning phase would generalize to the subsequent transfer phase.

Procedure

Figure 1 illustrates two example trials of the task paradigm in the learning phase of all three experiments. On each trial, three cards arranged in a pyramid were simultaneously presented on the screen, randomly chosen with the following constraints. The card on the top served as the reference card, and the cards at the bottom were choice cards. One of the two choice cards shared the same value in one dimension (e.g., shape) as the reference card, but had different values in all the other dimensions (e.g., color, filling, number). The other choice card shared the same value as the reference card in a second dimension (e.g., color), while being different in the three other dimensions (e.g., shape, filling, number). Additionally, there were no shared values on any dimensions between the two choice cards. Only two of the four dimensions, randomly assigned for each participant, were relevant as possible matching rules during the experiment. The two relevant dimensions were explicitly instructed to the participant (and practiced, see below) prior to the experiment. Only one of the two dimensions was the valid matching rule at any one time, and the valid rule changed over time. It was the participants' goal to figure out, via trial-and-error learning, which matching rule was currently valid on a given trial.

Each trial began with a 1 s fixation period. Then, participants were asked to match the reference card to the correct choice card based on the dimension they believed to be the currently valid matching rule, using the "z" or "m" button to indicate the left or right choice card, respectively. The cards remained on the screen for up to 2 s or until participants made a response. If participants did not respond in time, they would receive a "Too slow!" feedback and would be asked to press the spacebar to begin the next trial. Otherwise, they would be given either an "Oh no!" or "Bonus + $0.01" feedback for 500 ms. The feedback validity was 80%; that is, participants had an 80% chance of receiving positive feedback (and a 20% chance of negative feedback) on correct responses, and vice versa for incorrect responses. Participants were informed of the 80% feedback validity before the experiment. The correct sorting dimension stayed the same for a fixed number of trials before changing to the other dimension, but participants were not explicitly informed about the frequency of rule changes.

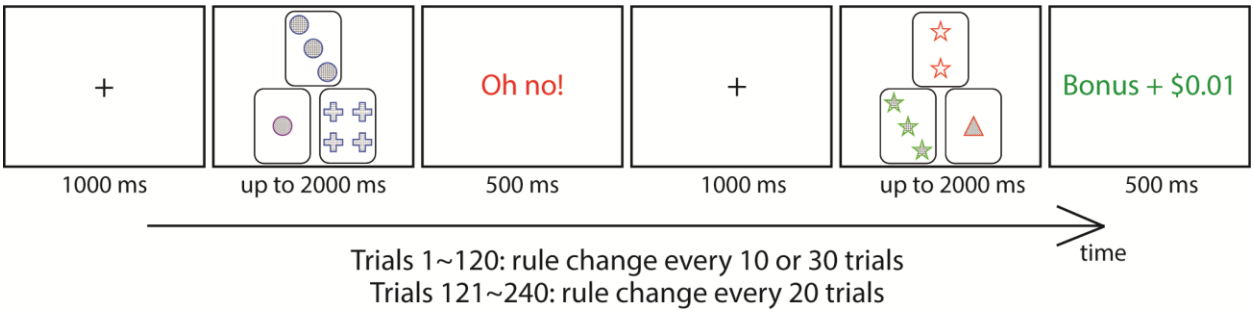

Figure 1. Illustration of the task paradigm. Each trial began with a fixation period, followed by a display of the reference card (top) and two choice cards (bottom) that required a participant response, followed by feedback. Participants were asked to match the correct choice card with the reference card according to the dimension (i.e., color, shape, filling, or number) they believed to be the currently relevant matching rule. In the example above, participants had to sort cards according to color or shape. In the first half of the experiment (learning phase), the sorting rule changed every 30 trials for participants in the low-volatility group, and every 10 trials for participants in the high-volatility group. In the second half of the experiment (transfer phase), the sorting rule changed every 20 trials for both groups.

Before starting the main experiment, participants were asked to perform a practice task consisting of 40 trials, with the sorting rule changing after 20 trials. The practice task was similar to the main experiment, except that the sorting rule was explicitly displayed on the screen. Participants had to achieve at least 90% accuracy on the practice task to move on to the main experiment. In the main experiment, both the low- and high-volatility groups completed a total of 240 trials. In the first half of the experiment (the learning phase), the sorting rule changed every 30 trials for participants in the low-volatility group, and every 10 trials for participants in the high-volatility group. In the second half of the experiment (the transfer phase), the sorting rule changed every 20 trials for both groups. In Experiment 1, the stimuli and task sets remained the same during the transfer phase as in the learning phase; in Experiment 2, the stimuli remained the same, but task sets were novel; and in Experiment 3, both stimuli and task sets were novel. There were no explicit instructions informing participants about the currently relevant rule, such that participants always had to rely on the feedback to figure out the currently relevant sorting rule to maximize their earnings.

Behavioral analyses

For each experiment, we compared the accuracies of the low- and high-volatility groups, with an accurate trial defined as responding according to the correct sorting rule, regardless of feedback. We then split the data into learning and transfer phases, and calculated the accuracies for each phase, and entered participants' mean accuracy values into a phase (learning vs. transfer) × volatility (low vs. high) ANOVA. This was to ensure that any group differences in RL model parameters were not confounded by differences in overall accuracy.

We hypothesized that, to perform optimally, participants would need to learn the volatility (rate of change) of the current phase, and use this understanding of periodic changes in the currently valid rule to inform their evaluation of the feedback. Specifically, they should be more likely to infer from negative feedback that a rule change had occurred if that feedback was encountered around the time of the periodic rule change point than at other times. To test this, we divided the trials into time-bins; one time-bin ranged from -3 to +3 trials around the change point (boundary trials), while all remaining trials were assigned to the other time-bin (non-boundary trials). This was done for both learning and transfer phases for both the low- and high-volatility groups in each experiment. Switch likelihood was calculated as the proportion of rule switch responses that

participants made among all trials following negative feedback. A phase (learning vs. transfer) × boundary (boundary vs. non-boundary trials) × volatility (low vs. high) ANOVA was conducted.

Reinforcement-learning modeling

We fit RL models (Sutton & Barto, 1998) to the choice behavior to estimate learning rates for the low- and high-volatility groups in the learning and transfer phases of the three experiments, using Markov Chain Monte Carlo (MCMC) sampling via the "stan" function from the RStan package (Stan Development Team, 2020) in R (R Core Team, 2020). We ran four MCMC chains for 1000 samples, discarding the first 150 as warm-up.

Our first model was a standard hierarchical RL model (RW-RL; Rescorla & Wagner, 1972), fit to the rule choice behavior. A hierarchical model was used as it takes into account the within-subject error of each subject's parameter estimate, unlike in the classic approach of comparing the mean value of each parameter for each condition after estimating point estimates for each subject (Daw, 2009). The model consisted first of a Q-learning model (Watkins & Dayan, 1992), whereby value estimates for each rule are updated over time based on feedback. Specifically, after an individual (i) on trial (t) chooses a matching rule, $C_{i,t} \in \{1,2\}$ (e.g., color or shape), and feedback is received for that choice, $R_{i,t} \in \{0,1\}$ (0 if negative feedback and 1 if positive feedback), the value estimate of that rule, $V(C)$, is updated according to the following:

$$V_{i,t+1}(C_{i,t}) = V_{i,t}(C_{i,t}) + \alpha_i \left(R_{i,t} - V_{i,t}(C_{i,t})\right) \quad (1)$$

where α_i is each individual's learning rate. The first trial of each experiment, $V_{i,1}$, as well as the first trial of the transfer phase in Experiments 2 and 3, $V_{i,121}$, were initialized with a separate starting utility, with the prior distribution of $N(0.5,0.5)$. This was not done for the first trial of the transfer phase of Experiment 1 because in that experiment there was no change in stimuli or task between learning and transfer phases.

In order to estimate the distribution of learning rates across experiments, conditions, and individuals, we estimated a multi-level model with three levels of hierarchy. The top level of the hierarchy described how the average learning rate varied across different conditions, while the middle level described how learning rates varied among individuals within a condition. Finally, the bottom level, described by equations (1), (4), and (7), modeled how individuals learned from feedback over the course of the task and predicted their choices. The advantage of using a single hierarchical model across all experiments and conditions is to pool information across conditions, resulting in less noisy estimates and reducing overfitting to individual conditions (Gelman, Hill, & Yajima, 2012).

At the top level of the hierarchy, we assumed the average learning rates for each condition were generated by a mixed-effect general linear model:

$$\begin{gathered} \mu_\eta(p,g,e) = \phi_\eta + \lambda_\eta(p,g,e) \\ \lambda_\eta(p,g,e) \sim N\left(0,\tau_\eta^2\right) \\ \phi_\eta \sim N(0,1) \end{gathered} \quad (2)$$

$$\tau_\eta \sim N^+(0,1)$$
$$\mu_\alpha(p,g,e) = logit^{-1}\left(\mu_\eta(p,g,e)\right)$$

Each condition was defined by a combination of a phase p (learning or transfer), a group g (high or low volatility), and an experiment e (experiment 1, 2, or 3). The hyperparameter ϕ_η is the population average learning rate for all subjects across all conditions. The condition-level random effects $\lambda_\eta(p,g,e)$ determine how far the average of each condition, i.e., phase (p) for each group (g) in each experiment (e), is from the average of the population mean, with the variance τ_η^2, governing the overall variability across conditions.

Next, for the middle level of the hierarchy, we modeled individual learning rates as arising from a mixed-effect general linear model:

$$\begin{gathered}\eta_{i,p} = \mu_\eta(p,g,e) + \gamma_{i,p} \\ \gamma_{i,p} \sim N\left(0,\sigma_\eta^2\right) \\ \sigma_\eta \sim N^+(0,1) \\ \alpha_{i,p} = logit^{-1}\left(\eta_{i,p}\right)\end{gathered} \tag{3}$$

where the random effects $\gamma_{i,p}$ determine how far each individual i is from the average for their condition, $\mu_\eta(p,g,e)$, with σ_η^2 governing the overall variability across subjects within each condition.

Having thus fit $\alpha_{i,p}$ estimates for all subjects, we examined the contrasts between conditions of interest. In particular, for each experiment and phase, we compared whether there were any differences in learning rate between the low- and high-volatility groups. We further examined whether this differed across experiments.

We also sought to examine whether the effects of learning rates were differentially driven by positive feedback (rewarded) versus negative feedback (unrewarded) trials. Intuitively, the negative feedback trials would be expected to drive rule switches, as participants would presumably recognize that their currently applied rule was incorrect. We therefore fit a second model, which we call the two-rates (2R) RL model, in which learning rate was fit separately for positive (+) and negative (−) reward (r) feedback trials (Donahue and Lee, 2015). The value for each rule, $V(C)$, is here updated according to:

$$V_{i,t+1}\left(C_{i,t}\right) = \begin{cases} V_{i,t}\left(C_{i,t}\right) + \alpha_{+,i}\left(R_{i,t} - V_{i,t}\left(C_{i,t}\right)\right) if R_{i,t} = 1 \\ V_{i,t}\left(C_{i,t}\right) + \alpha_{-,i}\left(R_{i,t} - V_{i,t}\left(C_{i,t}\right)\right) if R_{i,t} = 0 \end{cases} \tag{4}$$

where $\alpha_{+,i}$ is each individual's learning rate for positive feedback trials, and $\alpha_{-,i}$ is each individual's learning rate for negative feedback trials. Similar to the RW-RL model, a hierarchical general linear model is used to estimate mean effects of each condition:

$$\mu_{r,\eta}(p,g,e) = \phi_{r,\eta} + \lambda_{r,\eta}(p,g,e) \tag{5}$$

$$\lambda_{r,\eta}(p,g,e) \sim N\left(0, \tau_{r,\eta}^2 \middle| 2\right)$$
$$\sigma_{r,\eta} \sim N^{+}(0,1)$$
$$\mu_{r,\alpha}(p,g,e) = logit^{-1}\left(\mu_{r,\eta}(p,g,e)\right)$$

where $\phi_{r,\eta}$ are hyperparameters representing the population mean learning rates for the feedback level r for all subjects across the two phases and three experiments. The random effects $\lambda_{r,\eta}(p,g,e)$ represent the deviations of each condition from the population means, and $\tau_{r,\eta}^2$ govern the overall variability in each condition.

Another hierarchical general linear model was used to model individual learning rates:

$$\eta_{r,i,p} = \mu_{r,\eta}(p,g,e) + \gamma_{r,i,p} \tag{6}$$
$$\gamma_{r,i,p} \sim N\left(0, \sigma_{r,\eta}^2\right)$$
$$\sigma_{r,\eta}^2 \sim N^{+}(0,1)$$
$$\alpha_{R_{i,t},i,p} = logit^{-1}\left(\eta_{R_{i,t},i,p}\right)$$

where $\mu_{r,\eta}(p,g,e)$ are hyperparameters representing the mean learning rates for positive and negative feedback trials for each condition, and $\gamma_{r,i,p}$ is the deviation of each individual from the condition means, with $\sigma_{r,\eta}^2$ variability.

Finally, we also examined how the two volatility groups differed in terms of their action policies. For both the RW-RL and the 2R-RL model, we assumed that subjects chose rules probabilistically based on the value estimates according to a softmax distribution (Daw 2009). Thus, choice probabilities of selecting each rule (e.g., color or shape) for each trial were computed as follows:

$$p_{i,t}\left(chooseC_{i,t}\right) = \frac{e^{\beta_i V_{i,t}(C_{i,t})}}{\sum_{j=1}^{2} e^{\beta_i V_{i,t}(C_{i,j})}} \tag{7}$$

Here, β_i is a hyperparameter known as the inverse temperature, which represents how sensitive choice probabilities are to differences in choice value (Katahira 2015). β_i values were calculated for each subject similar to η_i, with a hyperparameter representing the populations's mean ϕ_β, and how many standard deviations, $\lambda_\beta(p,g,e)$, each condition deviated from their group's mean, and then another hyperparameter $\mu_\beta(p,g,e)$ representing each condition's mean, and the deviations of each individual, $\gamma_{i,p}$, from the condition mean.

$$\mu_\beta(p,g,e) = \phi_\beta + \lambda_\beta(p,g,e) \tag{8}$$
$$\gamma_{\mu_\beta}(p,g,e) \sim N\left(0, \tau_\beta^2\right)$$
$$\tau_\beta \sim N^{+}(5,5)$$

$$\beta_{i,p} = \mu_\beta(p,g,e) + \gamma_{i,p} \tag{9}$$
$$\gamma_{i,p} \sim N\left(0, \sigma_\beta^2\right)$$

$$\sigma_\beta \sim N^+(0,5)$$

While we report results from both the RW-RL and 2R-RL models, using them to characterize a general overall learning rate ($\alpha_{i,p}$) as well as separate learning rates for positive ($\alpha_{+,i,p}$) and negative feedback ($\alpha_{-,i,p}$) trials, we compared model fits between the two models using the leave-one-out information criterion (Vehtari, Gelman, & Gabry, 2017), and found that the 2R-RL model fit the data better. The expected log pointwise predictive density (ELPD) difference between the two models was -70.1, and its standard error (SE) difference was 30.6. This suggests that in the current experiments, participants did indeed have different learning rates for positive and negative feedback trials.

In the following analyses, we compared the posterior distribution of parameter estimates for learning rates and inverse temperature, in the learning phase and transfer phase, across the three experiments. We report $\hat{\delta}$, representing the mean difference between conditions in the model. In analyses with multiple factors, the results are reported in the format of an ANOVA. That is, we report the main effect of a factor, by comparing the means of each level of that factor, disregarding all other factors. In the case of interactions, we examine the difference of differences between levels in each factor. We also report credible interval (*CI*), which is the Bayesian equivalent of a confidence interval (with a slightly different technical interpretation). All credible intervals reported are central 95% intervals of the posterior differences. Additionally, given our a priori expectation that the learning rate in the high volatility group could only be the same or higher than the low volatility group in the transfer phase, for tests comparing the two volatility groups, we also report $p(\hat{\delta} < 0)$, which is the proportion of the posterior difference which falls below zero (corresponding to the logic of a one-tailed *p*-value).

Parameter estimates for the RW-RL model are summarized in Table 1, and for the 2R-RL model in Table 2. Our main analyses focused on learning rates, however. We report the results of inverse temperatures in the Supplementary Results.

Data and code sharing

All data and code of experiments and analyses have been deposited on GitHub, https://github.com/tanya-wen/Meta-flexibility.

Experiment 1

Experiment 1 examined whether prior exposure to low- vs. high-volatility rule switching environments biased people's propensity to infer rule changes in response to negative feedback in subsequent medium-volatility environments. More specifically, it explored the transfer of learning rates between initial and subsequent environments that differed solely in terms of rule change volatility, with task stimuli and categorization rules held constant. Later experiments then addressed the question of whether transfer occurs when task rules and stimuli also changed.

Method

Participants

Due to a lack of comparable prior studies, we could not base our target sample size on an empirical effect size. We therefore opted for a relatively large target sample size (N=~80). Eighty-eight participants were recruited from Amazon Mechanical Turk (MTurk) and randomly assigned to one of two experimental groups. Participants were compensated at a base pay rate of $2.50 plus any additional bonuses (mean = $1.86, SD = $0.10) earned during the experiment. Thirteen participants were excluded from the analysis due to an overall accuracy lower than 65%, leaving a final sample size of 75. The low-volatility group had 39 participants (22 male, 15 female, 2 did not wish to reply; age range: 26-56, mean = 36.69, SD = 8.81) and the high-volatility group had 36 participants (24 male, 11 female, 1 did not wish to reply; age range: 22-60, mean = 39.44, SD = 10.40).

Stimuli

Task stimuli consisted of 256 unique "cards" with one to four display items consisting of a specific shape (circle, triangle, plus, or star) in a particular color (blue, green, red, or purple) and a particular filling (checkered, dots, wave, or grid). We refer to these card properties as "dimensions" (number, shape, color, filling) that can take particular "values" (e.g., 1, 2, 3, or 4 for the number dimension). Each trial involved a display of three such cards. See Figure 1 for example stimuli.

Procedure

In the first half of the experiment, participants had to sort cards according to two of the four dimensions (e.g., color and shape; randomly assigned across participants). Sorting rules alternated every 30 trials for the low-volatility group and every 10 trials for the high-volatility group. In the second half (transfer phase) of the experiment, sorting rules alternated every 20 trials in the transfer phase. There was no explicit separation between the first and second half of the experiment.

Results

Behavior

Figure 2 illustrates the rule sequence and choice data from a representative participant from each group. In spite of the 80%-validity probabilistic feedback, participants were able to track the correct rule most of the time (low-volatility group: mean = 79.28%, SD = 5.76%; high-volatility group: mean = 73.19%, SD = 4.13%). A phase (learning vs. transfer) × volatility (low vs. high) ANOVA showed a significant main effect of phase ($F(1,73) = 39.82$, $p < 0.001$), group ($F(1,73) = 99.74$, $p < 0.001$), as well as phase × volatility interaction ($F(1,73) = 27.27$ $p < 0.001$). The main effect of phase was driven by participants having a higher accuracy for the transfer compared to the learning phase ($t = 3.56$, $p < 0.001$), presumably due to a practice effect. The main effect of volatility was driven by the low-volatility group having an overall higher accuracy than the high-volatility group ($t = 5.22$, $p < 0.001$), and this difference was significant in the learning phase ($t = 11.03$, $p < 0.001$), but not in the transfer phase ($t = -0.99$, $p = 0.33$). This

effect was expected, given the greater number of (error-inducing) rule reversals in the high-volatility group's learning phase.

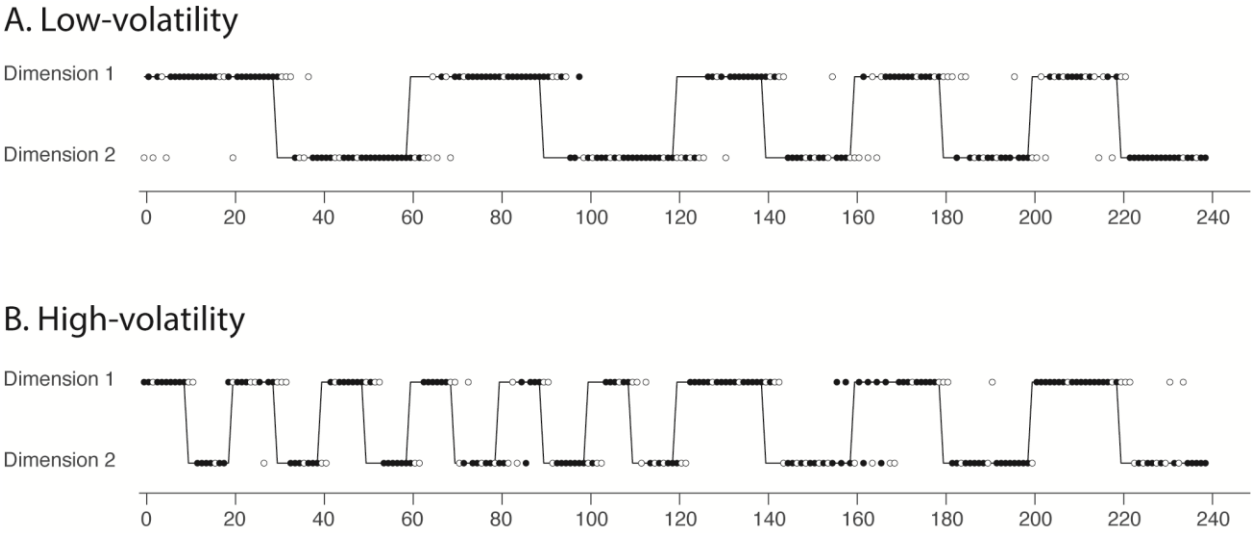

Figure 2. Dimension rule sequences and a representative participant from (A) the low-volatility group and (B) the high-volatility group. On each trial, participants chose a card based on their belief of the currently valid dimensional matching rule, here called Dimension 1 or 2 (circles). They received positive feedback (filled circles) when sorting according to the correct dimension (black line) 80% of the time and for incorrect choices 20% of the time; and they received negative feedback (open circles) for correct choices 20% of the time and for incorrect choices 80% of the time.

We examined whether participants were more likely to switch task rules after receiving negative feedback around the time of a true rule change than at other times (see Methods). A phase (learning vs. transfer) × boundary (boundary vs. non-boundary trials) × volatility (low vs. high) ANOVA showed significant main effects for phase ($F(1,73) = 46.50$, $p < 0.001$) and boundary ($F(1,73) = 22.59$, $p < 0.001$), which were driven by higher switch likelihood in the transfer phase, and around the time of a true rule change. There was no main effect of volatility ($F(1,73) = 0.73$, $p = 0.39$), and none of the interactions were significant (all Fs(1,73) < 2.98, all ps > 0.08). Thus, participants in both volatility groups acquired knowledge about the structure of their task environment, characterized by the periodicity of rule changes.

Reinforcement-learning models

As a confirmatory analysis, we first tested whether learning rates in the high-volatility group were higher than in the low-volatility group during the learning phase, where the matching rule switched every 10 compared to 30 trials. As expected, the RW-RL model showed that learning rates for participants in the high-volatility group were significantly larger than the low-volatility group ($\hat{\delta} = 0.08$, $CI = [0.01, 0.16]$, $p(\hat{\delta} < 0) = 0.01$). Similarly, in the 2R-RL model, we also found a main effect of volatility, with the high-volatility group exhibiting higher learning rates than the low-volatility group ($\hat{\delta} = 0.10$, $CI = [0.004, 0.19]$, $p(\hat{\delta} < 0) = 0.02$). There was no main

effect of feedback ($\hat{\delta}$ = -0.04, CI = [-0.14, 0.07]), and no feedback × volatility interaction ($\hat{\delta}$ = -0.01, CI = [-0.11, 0.09]).

Our main interest centered on the learning rates during the transfer phase. We found that according to the RW-RL model, the high-volatility group continued to show a higher learning rate than the low-volatility group during this phase ($\hat{\delta}$ = 0.14, CI = [0.06, 0.21], $p(\hat{\delta} < 0) < 0.001$). The 2R-RL model showed main effects of volatility ($\hat{\delta}$ = 0.15, CI = [0.07, 0.23], $p(\hat{\delta} < 0) < 0.001$), driven by higher learning rates for the high-volatility group; there was also a main effect of feedback ($\hat{\delta}$ = -0.15, CI = [-0.24, -0.07]), driven by the learning rates being higher for positive feedback trials; and no feedback × volatility interaction ($\hat{\delta}$ = -0.03, CI = [-0.11, 0.05]). These results document that volatility-driven learning rates acquired during the first half of the task generalized to the second half transfer phase where volatility was equated between groups.

Table 1. Mean parameter estimates (standard deviations) from the RW-RL model

	Model fitted parameters						
Parameters	ϕ_η	τ_η		ϕ_β	τ_β		
	0.47	0.26		4.17	0.65		
	(0.09)	(0.08)		(0.21)	(0.19)		
Experiment and parameters	Transformed parameters						Generated quantities
Experiment 1	μ_η	σ_η^2	λ_η	μ_β	σ_β^2	λ_β	μ_α
		0.61			1.42		
		(0.03)			(0.07)		
Learning phase							
Low-volatility	0.23		-0.96	4.73		0.91	0.56
	(0.11)		(0.55)	(0.25)		(0.52)	(0.03)
High-volatility	0.59		0.47	4.20		0.05	0.64
	(0.12)		(0.56)	(0.25)		(0.48)	(0.03)
Transfer phase							
Low-volatility	0.16		-1.24	4.73		0.91	0.54
	(0.11)		(0.55)	(0.26)		(0.53)	(0.03)
High-volatility	0.74		1.06	4.32		0.25	0.68
	(0.12)		(0.57)	(0.26)		(0.50)	(0.03)
Experiment 2							
Learning phase							
Low-volatility	0.23		-0.97	4.44		0.44	0.56
	(0.12)		(0.55)	(0.24)		(0.47)	(0.03)
High-volatility	0.71		0.93	3.83		-0.55	0.67
	(0.13)		(0.57)	(0.25)		(0.49)	(0.03)
Transfer phase							

Low-volatility	0.52	0.19	3.71	-0.74	0.63
	(0.11)	(0.53)	(0.23)	(0.50)	(0.03)
High-volatility	0.69	0.88	3.96	-0.33	0.67
	(0.11)	(0.55)	(0.23)	(0.46)	(0.03)
Experiment 3					
Learning phase					
Low-volatility	0.36	-0.45	4.79	1.01	0.59
	(0.11)	(0.53)	(0.26)	(0.54)	(0.03)
High-volatility	0.54	0.29	4.55	0.63	0.63
	(0.11)	(0.52)	(0.25)	(0.49)	(0.03)
Transfer phase					
Low-volatility	0.33	-0.58	3.37	-1.29	0.58
	(0.11)	(0.54)	(0.24)	(0.51)	(0.03)
High-volatility	0.61	0.54	3.26	-1.47	0.65
	(0.12)	(0.54)	(0.23)	(0.55)	(0.03)

Table 2. Mean parameter estimates (standard deviations) from the 2R-RL model

	Model fitted parameters					
Parameters	ϕ_{η_+}	τ_{η_+}	ϕ_{η_-}	τ_{η_-}	ϕ_β	τ_β
	0.83	0.69	0.40	0.32	4.39	0.80
	(0.24)	(0.17)	(0.11)	(0.10)	(0.27)	(0.21)

Experiment and parameters	Transformed parameters								Generated quantities	
Experiment 1	$\phi_{+,\eta}$	$\sigma^2_{+,\eta}$	$\lambda_{+,\eta}$	$\phi_{-,\eta}$	$\sigma^2_{-,\eta}$	$\lambda_{-,\eta}$	μ_β	σ^2_β	γ_{μ_β}	$\mu_{+,\alpha}$	$\mu_{+,\alpha}$
		1.46			0.59			1.28			
		(0.12)			(0.05)			(0.08)			
Learning phase											
Low-volatility	0.40		-0.63	0.28		-0.40	4.97		0.77	0.60	0.57
	(0.27)		(0.47)	(0.12)		(0.46)	(0.26)		(0.45)	(0.06)	(0.03)
High-volatility	0.90		0.10	0.65		0.81	4.26		-0.18	0.71	0.66
	(0.33)		(0.53)	(0.14)		(0.52)	(0.28)		(0.46)	(0.07)	(0.03)
Transfer phase											
Low-volatility	0.64		-0.28	0.10		-1.01	4.78		0.50	0.65	0.53
	(0.26)		(0.45)	(0.12)		(0.51)	(0.25)		(0.43)	(0.06)	(0.03)
High-volatility	1.59		1.14	0.60		0.62	4.30		-0.12	0.83	0.64
	(0.32)		(0.54)	(0.13)		(0.49)	(0.25)		(0.43)	(0.05)	(0.03)
Experiment 2											
Learning phase											
Low-volatility	-0.30		-1.72	0.47		0.23	5.38		1.30	0.43	0.62
	(0.24)		(0.56)	(0.11)		(0.48)	(0.29)		(0.53)	(0.06)	(0.03)
High-volatility	0.46		-0.56	0.91		1.67	4.15		-0.31	0.61	0.71
	(0.30)		(0.51)	(0.15)		(0.56)	(0.27)		(0.44)	(0.07)	(0.03)
Transfer phase											

Low-volatility	1.61	1.19	0.11	-0.99	3.80	-0.79	0.83	0.53
	(0.26)	(0.54)	(0.12)	(0.53)	(0.22)	(0.46)	(0.04)	(0.03)
High-volatility	1.36	0.80	0.56	0.49	4.07	-0.42	0.79	0.64
	(0.29)	(0.51)	(0.12)	(0.48)	(0.24)	(0.41)	(0.05)	(0.03)
Experiment 3								
Learning phase								
Low-volatility	0.23	-0.90	0.46	0.21	5.31	1.22	0.56	0.61
	(0.27)	(0.49)	(0.12)	(0.48)	(0.29)	(0.52)	(0.06)	(0.03)
High-volatility	1.07	0.36	0.36	-0.15	4.69	0.41	0.74	0.59
	(0.29)	(0.46)	(0.12)	(0.48)	(0.25)	(0.45)	(0.06)	(0.03)
Transfer phase								
Low-volatility	1.05	0.34	0.00	-1.35	3.50	-1.19	0.74	0.50
	(0.27)	(0.47)	(0.13)	(0.55)	(0.23)	(0.49)	(0.05)	(0.03)
High-volatility	1.29	0.69	0.42	0.06	3.32	-1.42	0.78	0.60
	(0.28)	(0.47)	(0.13)	(0.49)	(0.22)	(0.49)	(0.05)	(0.03)

Discussion

Our results showed that participants adapted their rule switching strategies to the volatility of the task environment. Participants in both volatility groups learned the periodicity of rule changes, as reflected in more frequent rule switches around task boundaries, as well as in a higher learning rate in participants in the high- compared to the low-volatility environment. Importantly pre-exposure to high- compared to low-volatility environments led to a higher learning rate in a subsequent medium-volatility environment. In other words, learned expectations about the level of cognitive flexibility required in the environment endured over time.

Experiment 2

The results of Experiment 1 suggest that people transfer expectations about the volatility of task rules across adjacent episodes in time (even when the underlying volatility changes) under conditions of identical stimuli and task rules. This represents a "near transfer" of rule learning rate. It is possible that the transfer effect observed in Experiment 1 was driven by participants forming associations between periodic rule switches and the specific rule representations they were learning (see Siqi-Liu & Egner, 2020) rather than reflecting learning at a more abstract level, of making inferences about the general rate at which rules seem to change. Experiment 2 therefore tested whether learning rates would still transfer when the sets of rules that participants learned about changed between the learning and transfer phases (while the stimuli remained the same). Specifically, we probed whether exposure to low- or high-volatility learning environments involving two of four possible task rules (e.g., shape and color matching) would bias the learning rate in subsequent medium-volatility environments with the other two possible task rules (i.e., number and filling matching). Obtaining transfer under these conditions would indicate that the expectations about the rate of rule changes that are being transferred are independent of the specific task rules, thus representing a form of meta-learning.

Method

Participants

Ninety-four participants were recruited from MTurk and randomly assigned to one of the two volatility groups. Participants were compensated at a base pay rate of $2.50 plus any additional bonuses (mean = $1.86, SD = $0.10) earned during the experiment. Twelve participants were excluded from the analysis due to overall accuracy lower than 65%, leaving a final sample size of 82. The low- volatility group had 42 participants (20 male, 22 female; age range: 23-70, mean = 39.64, SD = 11.80) and the high-volatility group had 40 participants (22 male, 18 female; age range: 24-76, mean = 38.80, SD = 11.36).

Stimuli

The stimuli were the same as in Experiment 1.

Procedure

As in Experiment 1, in the first half of the experiment, participants had to sort cards according to two of the four dimensions (e.g., color and shape; randomly assigned across participants). However, unlike Experiment 1, before the start of the second half (transfer phase) of the experiment, participants were taken to another instruction screen and informed that they would now be sorting cards according to the other two dimensions that were previously irrelevant in the first half (e.g., filling and number). There was no practice for the transfer phase, and it started as soon as participants indicated they were ready.

Results

Behavior

Participants were able to perform the task reasonably well (low-volatility group: mean = 79.22%, SD = 5.66%; high-volatility group: mean = 73.58%, SD = 4.45%). The phase (learning vs. transfer) × volatility (low vs. high) ANOVA showed a significant main effect of phase ($F(1,80) = 22.80$, $p < 0.001$), a main effect of volatility ($F(1,80) = 24.89$, $p < 0.001$), and a phase × volatility interaction ($F(1,80) = 40.93$, $p < 0.001$). As in Experiment 1, the main effect of phase was driven by participants having a higher accuracy for the transfer compared to the learning phase ($t = 4.99$, $p < 0.001$), presumably due to generic task practice effects. The main effect of volatility was driven by the low-volatility group having a higher accuracy than the high-volatility group ($t = 3.71$, $p < 0.001$). The interaction was again driven by the low-volatility group having higher accuracy compared to the high-volatility group in the learning ($t = 8.39$, $p < 0.001$), but not in the transfer phase ($t = -0.54$, $p = 0.59$).

The phase (learning vs. transfer) × boundary (boundary vs. non-boundary trials) × volatility (low vs. high) ANOVA showed significant main effects for phase ($F(1,80) = 39.74$, $p < 0.001$) and boundary ($F(1,80) = 49.44$, $p < 0.001$), which were driven by higher switch likelihood in the transfer phase, and around the time of a true rule change. There was no main effect of volatility ($F(1,80) = 0.003$, $p = 0.96$), and none of the other interactions were significant (all Fs$(1,80) <$

3.08, all *p*s > 0.08). In sum, participants in both volatility groups showed evidence of learning the periodicity of rule changes in their respective learning environments.

Reinforcement-learning models

In the learning phase, the high-volatility group had a higher learning rate than the low-volatility group, as estimated by the RW-RL model ($\hat{\delta} = 0.11$, $CI = [0.03, 0.19]$, $p(\hat{\delta} < 0) < 0.01$). Learning rates from the 2R-RL model also showed a main effect of volatility ($\hat{\delta} = 0.14$, $CI = [0.05, 0.23]$, $p(\hat{\delta} < 0) = 0.001$), which was again driven by the high-volatility group having a higher learning rate. We additionally found a main effect of feedback ($\hat{\delta} = 0.15$, $CI = [0.04, 0.25]$), due to higher learning rates for the negative feedback trials. There was no feedback × volatility interaction ($\hat{\delta} = -0.04$, $CI = [-0.14, 0.05]$).

In the transfer phase, the RW-RL model showed no significant difference of learning rates between groups ($\hat{\delta} = 0.04$, $CI = [-0.03, 0.11]$, $p(\hat{\delta} < 0) = 0.13$), though the high-volatility group showed a numerically higher learning rate. The 2R-RL model showed no main effect of volatility ($\hat{\delta} = 0.04$, $CI = [-0.04, 0.10]$, $p(\hat{\delta} < 0) = 0.15$), but there was a main effect of feedback ($\hat{\delta} = -0.23$, $CI = [-0.30, -0.15]$), which was driven by higher learning rates for positive feedback trials. Critically, results also showed a significant feedback × volatility interaction ($\hat{\delta} = 0.07$, $CI = [0.01, 0.14]$). The interaction was driven by the high-volatility group showing a higher learning rate compared to the low-volatility group for negative feedback ($\hat{\delta} = 0.11$, $CI = [0.03, 0.19]$, $p(\hat{\delta} < 0) < 0.01$), but not for positive feedback trials ($\hat{\delta} = -0.04$, $CI = [-0.15, 0.07]$, $p(\hat{\delta} < 0) = 0.74$). Thus, learning rates acquired during the learning phase generalized to rule switching performance in the transfer phase despite a change in specific task rules between phases, but only during negative feedback.

Discussion

Even though in Experiment 2 the task rules that participants were switching between changed from the learning to the transfer phase, we observed robust evidence for transfer of rule learning rates. These results suggest that participants do not transfer a specific association between particular task rules and change point estimates in the present paradigm, but rather that they form and transfer a more abstract expectation of the volatility of the rules governing the environment, as reflected in the learning rate. This transfer was expressed primarily in response to negative feedback rather than to positive feedback, in line with the assumption that negative feedback trials in particular cause participants to switch rules.

Experiment 3

Experiments 1 and 2 showed that task rule learning rates can generalize in time and across task rules. The cross-rule transfer effect observed in Experiment 2 clearly indicates that participants formed a rule-independent expectation of rule volatility, though transfer of learning rates here may have been aided by the fact that the new rules were still applied to the same stimuli. To further probe rule learning rate transfer to a more dissimilar environment, in Experiment 3, we

provided a test of "far transfer" by testing whether prior experiences of low- or high-volatility environments would bias the tendency to shift sets in subsequent medium-volatility environments with both novel rules *and* novel stimuli.

Method

Participants

One-hundred-and-one participants were recruited from MTurk and randomly assigned to one of two volatility groups. Participants were compensated at a base pay rate of $2.50 plus any additional bonuses (mean = $1.82, SD = $0.09) earned during the experiment. Twenty participants were excluded from the analysis due to overall accuracy lower than 65%, leaving a final sample size of 81. The low- volatility group had 39 participants (26 male, 12 female, 1 did not wish to reply; age range: 27-67, mean = 41.41, SD = 10.94) and the high-volatility group had 42 participants (23 male, 18 female, 1 did not wish to reply; age range: 21-56, mean = 36.29, SD = 7.88).

Stimuli

We used the same stimuli as in Experiments 1 and 2 for the learning phase of Experiment 3. To test whether the rule learning rates could transfer to other tasks with novel stimuli, for the Experiment 3 transfer phase we used face stimuli taken from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). A total of 64 emotion-neutral faces (16 Asian males, 16 Asian females, 16 Caucasian males, and 16 Caucasian females) were used.

Procedure

In the first half of the experiment (the learning phase), participants had to sort cards according to two of the four dimensions (e.g., color and shape; randomly assigned across participants). Before the start of the second half (transfer phase) of the experiment, participants were taken to another instructions screen and informed that they would now be sorting face images according to either gender (male vs. female) or race (Asian vs. Caucasian). As in the card-matching task, on each trial three faces were displayed arranged in a pyramid, with the face on the top serving as the reference face, and the faces at the bottom as choice faces. Each of the two choice faces shared only one matching domain (gender or race) as the reference face. There was no practice for the transfer phase, and it started as soon as participants indicated they were ready.

Results

Behavior

Both groups were able to perform the task reasonably well (low-volatility group: mean = 74.75%, SD = 5.66%; high-volatility group: mean = 71.96%, SD = 4.81%). The phase (learning vs. transfer) × volatility (low vs. high) ANOVA showed a significant main effect of phase ($F(1,79) = 11.38$, $p = 0.001$), a main effect of volatility ($F(1,79) = 5.74$, $p = 0.02$), and a phase × volatility interaction ($F(1,79) = 16.28$, $p < 0.001$). As in the prior two experiments, the main effect of phase was driven by higher accuracy for the transfer phase ($t = 3.36$, $p < 0.001$). In line

with previous results, mean accuracy was higher for the low-volatility group in the learning phase (t = 4.45, p < 0.001), but not in the transfer phase (t = -1.24, p = 0.22).

The phase (learning vs. transfer) × boundary (boundary vs. non-boundary trials) × volatility (low vs. high) ANOVA showed significant main effects for phase (F(1,79) = 127.42, p < 0.001) and boundary (F(1,79) = 4.15, p < 0.05), which were driven by higher switch likelihood in the transfer phase, and around the time of a true rule change. There was no main effect of volatility (F(1,79) = 0.10, p = 0.75), and none of the other interactions were significant (all Fs(1,78) < 2.11, all ps > 0.15). Thus, participants in both groups acquired knowledge about the structure of their task environments.

Reinforcement-learning model

In the learning phase, the RW-RL model found no significant difference between learning rates in the two volatility groups ($\hat{\delta}$ = 0.04, CI = [-0.03, 0.12], $p(\hat{\delta} < 0)$ = 0.12), though the high-volatility group had a numerically higher learning rate compared to the low-volatility. The 2R-RL model showed the high-volatility group had higher learning rates than the low-volatility group ($\hat{\delta}$ = 0.08, CI = [-0.01, 0.17], $p(\hat{\delta} < 0)$ = 0.03). There was no main effect of feedback ($\hat{\delta}$ = -0.05, CI = [-0.15, 0.05]). However, there was a significant feedback × volatility interaction ($\hat{\delta}$ = -0.10, CI = [-0.19, -0.01]). Post hoc analysis suggest that the interaction was driven by the high-volatility group showing higher learning rates than the low-volatility group for positive ($\hat{\delta}$ = 0.18, CI = [0.02, 0.34], $p(\hat{\delta} < 0)$ = 0.01) but not negative feedback ($\hat{\delta}$ = -0.02, CI = [-0.10, 0.05], $p(\hat{\delta} < 0)$ = 0.74).

In the transfer phase, the RW-RL model had higher learning rates in the high-volatility group compared to the low-volatility group ($\hat{\delta}$ = 0.07, CI = [-0.01, 0.15], $p(\hat{\delta} < 0)$ = 0.04). Similarly, in the 2R-RL model, the high-volatility group had higher learning rates than the low-volatility group ($\hat{\delta}$ = 0.07, CI = [-0.003, 0.15], $p(\hat{\delta} < 0)$ = 0.03). We found a main effect for feedback ($\hat{\delta}$ = -0.21, CI = [-0.29, -0.12]), driven by higher learning rates during positive feedback trials. We found no feedback × volatility interaction ($\hat{\delta}$ = 0.03, CI = [-0.05, 0.11]). Although this interaction was not significant, based on our previous findings we further examined the volatility effect in positive and negative feedback separately, and found that the volatility effect was significant in the negative ($\hat{\delta}$ = 0.10, CI = [0.02, 0.19], $p(\hat{\delta} < 0) < 0.01$), but not positive feedback trials ($\hat{\delta}$ = 0.04, CI = [-0.09, 0.18], $p(\hat{\delta} < 0)$ = 0.25). Thus, learning rates acquired during the learning phase generalized to rule switching performance in the transfer phase, in spite of a change in both the stimulus materials and task rules between phases.

In sum, we observed rule- and stimulus-independent "far transfer" of rule learning rates, in particular for negative feedback trials. In a final analysis, we sought to directly compare the degree of learning rate transfer between experiments, testing whether transfer differed quantitatively as a function of whether rules and stimuli remained the same (Experiment 1), whether rules changed (Experiment 2), or whether both rules and stimuli changed between learning and transfer phases.

Cross-experiment transfer effect comparison

We compared the transfer phase learning rates across the three experiments. Results from the RW-RL model showed a main effect of volatility ($\hat{\delta} = 0.08$, $CI = [0.03, 0.12]$, $p(\hat{\delta} < 0) < 0.001$), with the high-volatility groups having higher learning rates than the low-volatility groups. There were no pairwise differences in overall learning rates across the three experiments (max $|\hat{\delta}| = 0.04$, $CI = [-0.09, 0.01]$). There were no interactions between volatility and experiments (max $\hat{\delta} = 0.05$, $CI = [-0.001, 0.10]$).

With learning rates from the 2R-RL model, we found a main effect of volatility ($\hat{\delta} = 0.09$, $CI = [0.04, 0.13]$, $p(\hat{\delta} < 0) < 0.001$), with the high-volatility groups having higher learning rates than the low-volatility groups. There was also a main effect of feedback ($\hat{\delta} = 0.20$, $CI = [0.14, 0.25]$), driven by higher learning rates during negative feedback trials. There were no differences in mean learning rates between pairwise comparisons across experiments (max $\hat{\delta} = 0.4$, $CI = [-0.01, 0.09]$). There were no volatility × feedback × experiment interactions (max $\hat{\delta} = 0.05$, $CI = [-0.0004, 0.10]$).

Discussion

These results echo the previous experiments in providing evidence for a transfer of task rule learning rates, even though in Experiment 3 this involved applying new rules to new stimuli ("far" transfer). Our results again also suggest that this transfer occurs primarily for negative rather than positive feedback trials. A comparison of the three experiments showed that the degree of this transfer did not differ between experiments, and was therefore unaffected by the similarity between learning and transfer tasks.

General Discussion

The goal of the current study was to examine whether participants acquire and transfer expectations about cognitive flexibility demands across different contexts. In particular, we tested whether learning to change task sets more or less frequently in one context would affect learning rates in subsequent contexts. We replicated previous findings showing that participants adjusted their learning rates according to environment volatility (Behrens et al., 2007; Massi et al., 2018), with high volatility environments leading to higher learning rates in task rule updating. Crucially, we further found that the inductive biases acquired during the learning phase affected learning rates in a subsequent transfer phase (Experiment 1) and generalized to novel task rules (Experiment 2) and novel rules and stimuli (Experiment 3). This was reflected by an overall higher learning rate in the transfer phase for participants previously exposed to a high-volatility environment compared to those previously exposed to a low-volatility environment, which was mainly driven by learning from negative feedback (unrewarded) trials. Taken together, this demonstrates that people form and transfer an abstract, stimulus- and task-independent, expectation of the volatility of the rules governing their environment, expressed in a more or less cognitively flexible rule updating strategy. To the best of our knowledge, this is the first demonstration of people's ability to extract and re-use cognitive control learning parameters that transcend specific stimuli and tasks.

Previous behavioral studies have shown that participants can strategically adapt their readiness to switch tasks in line with changes in contextual switch-likelihood (reviewed in Braem & Egner, 2018). For instance, when manipulating the frequency of cued task switches over different blocks of trials, participants exhibit smaller switch costs (slower and more error prone responses for switching than repeating tasks) in blocks where switches are frequent compared to when they are rare (e.g., Chiu & Egner, 2017; e.g., Dreisbach & Haider, 2006; Leboe et al., 2008; Monsell & Mizon, 2006; Siqi-Liu & Egner, 2020). However, unlike in the present experiments, this change in switch-readiness seems to be limited to "biased" task sets that are associated with either more frequent switch or repeat trials, and does not generalize to intermingled "unbiased" task sets where the likelihood of switching versus repeating a task is equal (Siqi-Liu & Egner, 2020). This suggests that meta-flexibility in cued task switching is task-set or stimulus specific, rather than being due to participants developing a more global flexible cognitive strategy or processing mode that would promote switching in general (i.e., to *any* other task). A possible explanation underlying this dearth of transfer in prior studies could be that frequent forced (cued) switching motivates participants to try and keep the multiple relevant task sets in working memory, which would ease the switching between the respective tasks, but may not necessarily transfer to new tasks (Dreisbach & Fröber, 2019).

How was transfer of switch-readiness achieved in the present study then? We posit that this is likely due to reliance on a different mechanism for modulating meta-flexibility. In particular, one's set point on the cognitive stability-flexibility continuum has also been conceptualized in terms of an "updating threshold" – the ease with which new task rule information is allowed to enter working memory (Dreisbach & Fröber, 2019; Goschke, 2003, 2013). Thus, another way to increase flexibility, besides attempting to keep relevant task sets in working memory, may be to lower this updating threshold, which in turn could increase flexibility in a more generalizable fashion (Dreisbach & Fröber, 2019). One reason for not observing such generalization of flexibility in past studies might be that they typically relied on cued or forced-choice switching protocols (Koch et al., 2018), while task-switching in everyday life often lack explicit instructions or cues as to when to switch; instead, people have to discover the underlying rules based on environmental feedback amidst uncertainty (Behrens et al., 2007; Niv et al., 2015; Van Eylen et al., 2011). Self-initiated switches without an explicit cue are also thought to require a higher degree of disengagement from the previous task to perform the switch (Manly et al., 2002; Van Eylen et al., 2011). A couple of studies that examined the influence of forced-choice task switching on voluntary task switches, found that increasing the proportion of forced choices, in particular in combination with high switch rates, increases voluntary task switching rates (Chiu et al., 2020; Fröber & Dreisbach, 2017). Taken together, this suggests that in conditions where experience is used to learn task models (Niv, 2019), meta-flexibility may be achieved by altering one's updating threshold or the rate at which this threshold is reached (both would be observed as a change in learning rate in RL modeling), which in turn may promote transferable effects (Goschke, 2003; Marković et al., 2019), as observed in the present study.

While we used RL models to characterize learning rates, we use them here only as a method for quantifying the speed of behavioral adaptation. We do not propose that these models are representative of the true generative process underlying behavior. While the simple 'model-free'

learning models we used here do not capture the use of structured knowledge (Hampton et al., 2006; Radulescu et al., 2019), we hypothesized that participants in our task would infer a higher-order task structure, in the form of knowledge about the periodicity of rule changes, to guide their behavior. We found evidence for this prediction, as participants were more likely to switch task sets after a negative feedback during trials surrounding the true rule change during the learning and transfer phases (Costa et al., 2015; Vilà-Balló et al., 2017). Furthermore, this effect was stronger in the transfer compared to the learning phase, suggesting participants acquired and utilized knowledge about the structure of their task-learning environment (Schulz et al., 2020).

Previous studies have shown that inferring hidden underlying structural forms such as the relationships between stimuli, periodicities, or cognitive maps can enable rapid generalization of behavior to new environments (Behrens et al., 2018; Halford et al., 1998; Kemp et al., 2010; Mark et al., 2020). For instance, in Mark et al. (2020), two groups of participants learned either hexagonal graphs or community structure graphs on the first day, and on the second day, they learned a new graph with either the same or the alternate structure. The authors found that the experience on the first day shaped prior expectations over the underlying structure on the second day, shown by improved task performance in the group that had the correct prior structural knowledge of the graph. By analogy, it is likely that in the current study, the structure of the card sorting tasks was generalized over task rules and stimuli using similar mechanisms of applying previously learned abstracted knowledge, in this case, the updating threshold or learning rate (Baram et al., 2020). Consequently, the more or less frequent switches encountered during the learning phase would drive expectations and switch readiness during the transfer phase. Note that in the present study, unlike in the above studies of structure learning, the transfer phase was designed to detect a differential bias in task set updating between participants undergoing different learning regimes, rather than to produce superior performance in one group than the other. However, the inferences drawn are ultimately identical: task structure knowledge acquired in the learning phase creates an inductive bias that is reflected in how participants make inferences about their environment in the transfer phase.

In conclusion, we here present a novel paradigm to show that participants transfer volatility-conditioned rule learning rates to new temporal, task, and stimulus contexts. This transfer of a task- and stimulus-independent rule learning parameter represents the formation and generalization of structural task knowledge for guiding cognitive control strategies. Given that impairments in the ability to adopt a contextually appropriate level of cognitive flexibility are thought to be central to various clinical conditions (e.g., Browning et al., 2015; Manly et al., 2002; Nassar & Troiani, 2020; Van Eylen et al., 2011), this new task protocol holds promise for developing a model-based assessment of individual differences in this ability, and for delineating the underlying neural mechanisms in future studies.

References

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2020). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement

learning problems. *Neuron*. https://doi.org/10.1016/j.neuron.2020.11.024

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410. https://doi.org/10.1038/nn1209

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. In *Neuron* (Vol. 100, Issue 2, pp. 490–509). Cell Press. https://doi.org/10.1016/j.neuron.2018.10.002

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). https://doi.org/10.1038/nn1954

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology*, *39*(1), 15–22. https://doi.org/10.1080/00221309.1948.9918159

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

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

Chiu, Y. C., Fröber, K., & Egner, T. (2020). Item-specific priming of voluntary task switches. *Journal of Experimental Psychology. Human Perception and Performance*, *46*(4), 434–441. https://doi.org/10.1037/xhp0000725

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2015). Reversal learning and dopamine: A Bayesian perspective. *Journal of Neuroscience*, *35*(6), 2407–2416. https://doi.org/10.1523/JNEUROSCI.1989-14.2015

Dreisbach, G., & Fröber, K. (2019). On How to Be Flexible (or Not): Modulation of the Stability-Flexibility Balance. *Current Directions in Psychological Science*, *28*(1), 3–9. https://doi.org/10.1177/0963721418800030

Dreisbach, G., & Haider, H. (2006). Preparatory adjustment of cognitive control in the task switching paradigm. *Psychonomic Bulletin and Review*, *13*(2), 334–338. https://doi.org/10.3758/BF03193853

Fröber, K., & Dreisbach, G. (2017). Keep flexible – Keep switching! The influence of forced task switching on voluntary task switching. *Cognition*, *162*, 48–53. https://doi.org/10.1016/j.cognition.2017.01.024

Goschke, T. (2003). Voluntary action and cognitive control from a cognitive neuroscience perspective. In *Voluntary action: Brains, minds, and sociality*. (pp. 49–85). https://psycnet.apa.org/record/2003-06267-003

Goschke, T. (2013). Volition in Action: Intentions, Control Dilemmas, and the Dynamic Regulation of Cognitive Control. In *Action Science: Foundations of an Emerging Discipline* (pp. 409–434). The MIT Press. https://doi.org/10.7551/mitpress/9780262018555.003.0016

Halford, G. S., Bain, J. D., Maybery, M. T., & Andrews, G. (1998). Induction of Relational Schemas: Common Processes in Reasoning and Complex Learning. *Cognitive Psychology*, *35*(3), 201–245. https://doi.org/10.1006/cogp.1998.0679

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006

Jiang, J., Beck, J., Heller, K., & Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nature Communications*, *6*(1), 1–11. https://doi.org/10.1038/ncomms9165

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

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to Learn Causal Models. *Cognitive Science*, *34*(7), 1185–1243. https://doi.org/10.1111/j.1551-6709.2010.01128.x

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking-an integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557–583. https://doi.org/10.1037/bul0000144

Leboe, J. P., Wong, J., Crump, M., & Stobbe, K. (2008). Probe-specific proportion task repetition effects on switching costs. *Perception and Psychophysics*, *70*(6), 935–945. https://doi.org/10.3758/PP.70.6.935

Lee, D., Seo, H., & Jung, M. W. (2012). *Neural Basis of Reinforcement Learning and Decision Making*. https://doi.org/10.1146/annurev-neuro-062111-150512

Manly, T., Hawkins, K., Evans, J., Woldt, K., & Robertson, I. H. (2002). Rehabilitation of executive function: Facilitation of effective goal management on complex tasks using periodic auditory alerts. *Neuropsychologia*, *40*(3), 271–281. https://doi.org/10.1016/S0028-3932(01)00094-X

Mark, S., Moran, R., Parr, T., Kennerley, S. W., & Behrens, T. E. J. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18254-6

Marković, D., Goschke, T., & Kiebel, S. J. (2019). Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. In *bioRxiv*. https://doi.org/10.1101/847566

Massi, B., Donahue, C. H., & Lee, D. (2018). Volatility Facilitates Value Updating in the Prefrontal Cortex. *Neuron*, *99*(3), 598-608.e4. https://doi.org/10.1016/j.neuron.2018.06.033

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

Monsell, S., & Mizon, G. A. (2006). Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 493–516. https://doi.org/10.1037/0096-1523.32.3.493

Nassar, M. R., & Troiani, V. (2020). The stability flexibility tradeoff and the dark side of detail. *Cognitive, Affective and Behavioral Neuroscience*, 1–17. https://doi.org/10.3758/s13415-020-00848-8

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic Reinforcement Learning: The Role of Structure and Attention. In *Trends in Cognitive Sciences* (Vol. 23, Issue 4, pp. 278–292). Elsevier Ltd. https://doi.org/10.1016/j.tics.2019.01.010

Rescorla, R. ., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory* (pp. 64–99).

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, 101261. https://doi.org/10.1016/j.cogpsych.2019.101261

Siqi-Liu, A., & Egner, T. (2020). Contextual Adaptation of Cognitive Flexibility is driven by Task- and Item-Level Learning. *Cognitive, Affective and Behavioral Neuroscience*, *20*(4), 757–782. https://doi.org/10.3758/s13415-020-00801-9

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*.

Van Eylen, L., Boets, B., Steyaert, J., Evers, K., Wagemans, J., & Noens, I. (2011). Cognitive flexibility in autism spectrum disorder: Explaining the inconsistencies? *Research in Autism Spectrum Disorders*, *5*(4), 1390–1401. https://doi.org/10.1016/j.rasd.2011.01.025

Vilà-Balló, A., Mas-Herrero, E., Ripollés, P., Simó, M., Miró, J., Cucurell, D., López-Barroso, D., Juncadella, M., Marco-Pallarés, J., Falip, M., & Rodríguez-Fornells, A. (2017). Unraveling the role of the hippocampus in reversal learning. *Journal of Neuroscience*, *37*(28), 6686–6697. https://doi.org/10.1523/JNEUROSCI.3212-16.2017

Watkins, C. J. C. H., & Dayan, P. (1992). *Q-Learning* (Vol. 8).

Yu, L. Q., Wilson, R. C., & Nassar, M. R. (2020). Adaptive learning is structure learning in time. *Psyarxiv*, 1–27. https://doi.org/10.31234/OSF.IO/R637C