Lowered inter-stimulus discriminability hurts incremental contributions to learning

Aspen H. Yoo[1,2], Haley Keglovits[3], Anne G.E. Collins[1,2]

April 10, 2023

University of California, Berkeley. Department of Psychology[1]
University of California, Berkeley. Helen Wills Neuroscience Institute [2]
Brown University. Department of Cognitive, Linguistic and Psychological Sciences[3]

Abstract

How does the similarity between stimuli affect our ability to learn appropriate response associations for them? In typical laboratory experiments learning is investigated under somewhat ideal circumstances, where stimuli are easily discriminable. This is not representative of most real-life learning, where overlapping "stimuli" can result in different "rewards" and may be learned simultaneously (e.g., you may learn over repeated interactions that a specific dog is friendly, but that a very similar looking one isn't). With two experiments, we test how humans learn in three stimulus conditions: one "best case" condition in which stimuli have idealized and highly discriminable visual and semantic representations, and two in which stimuli have overlapping representations, making them less discriminable. We find that, unsurprisingly, decreasing stimuli discriminability decreases performance. We develop computational models to test different hypotheses about how reinforcement learning (RL) and working memory (WM) processes are affected by different stimulus conditions. Our results replicate earlier studies demonstrating the importance of both processes to capture behavior. However, our results extend previous studies by demonstrating that RL, and not WM, is affected by stimulus distinctness: people learn slower and have higher across-stimulus value confusion at decision when stimuli are more similar to each other. These results illustrate strong effects of stimulus type on learning and demonstrate the importance of considering parallel contributions of different cognitive processes when studying behavior.

1 Introduction

Humans are efficient learners, but how fast we learn, depends heavily on what we learn about. For example, a teacher learning the name of two new transfer students may only need to be told their names once, but they may need much more trial and error for each student if they're learning the name of the entire class at the same time. Furthermore, if the students look alike, learning may require even more effort. Here, we formally explore how stimulus discriminability (in a semantic and visual domain) impacts learning, and whether the multiple processes involved in learning are affected differently.

Specifically, we investigate stimulus discriminability in a stimulus-action association task in which both reinforcement learning (RL) and working memory (WM) processes are utilized (e.g., Collins & Frank, 2012). Reinforcement learning (RL) broadly refers to the process that characterizes how people learn incrementally through valenced feedback (Sutton & Barto, 1998). Working memory (WM) is a flexible, but capacity-limited process involved in actively maintaining perceptually unavailable information over a short period of time (Cowan, 2017). While there has been an increase in investigating the interplay between these two essential processes (for a review, see Yoo & Collins, 2022), there still is much to be learned about how the two interact in different settings.

For example, researchers in both RL and WM fields consider stimulus carefully when designing experiments, but each field tends to focus on different aspects of stimuli. RL studies tend to use a variety of stimuli across tasks. Sometimes they use stimuli with low semantic information, such as gabor patches, fractals, and foreign alphabet characters (e.g., Farashahi, Rowe, Aslami, Lee, & Soltani, 2017; Niv et al., 2015; Oemisch et al., 2019; Wilson & Niv, 2012; Wunderlich, Beierholm, Bossaerts, & O'Doherty, 2011; Radulescu, Niv, & Ballard, 2019; Daw, Gershman, Seymour, Dayan, & Dolan, 2011), under the assumption that relying on stimuli that are easy to name and have high semantic discriminability (i.e., have different names), such as different common objects, shapes, and colors (e.g., Collins & Frank, 2012; Collins, 2018; Farashahi, Xu, Wu, & Soltani, 2020), may affect behavior (perhaps by employing more explicit processes like WM). WM studies' choice of stimuli is much more explicit, due to traditional WM being formalized as being modality specific (i.e., containing separate visual and verbal storage units; Baddeley & Hitch, 1974). Stimuli that are nameable (e.g., spoken words, digits, or words) are considered to relate to verbal WM (e.g., Conrad, 1964), while less easily nameable stimuli (e.g., orientations, spatial frequencies) correspond to visual WM (e.g., Luck & Vogel, 1997; Wilken & Ma, 2004).

From previous research, it is apparent that there is some consideration of how different stimuli may affect behavior. However, it is still unclear how stimulus discriminability affects RL, WM, or their interplay. How do different types of stimuli affect RL and WM processes during an associative learning task? Specifically, are RL and WM differently affected by how distinct stimuli are? To address our question, we designed and collected data on two stimulus-response association learning experiments, manipulating stimulus discriminability. Learning was measured in three stimulus conditions. There is evidence that human learning differs for abstract and naturalistic stimuli

(Farashahi et al., 2020), so one of our primary criteria when choosing stimulus sets was for them to be similarly "naturalistic" and similarly familiar (vs. novel). Our first condition, the "Standard" condition, we used a standard stimulus set, in which the stimuli images that were discriminable visually and semantically. Second, the "Text" condition had stimuli which were simply text printed of different nouns, designed to limit visual information while maintaining semantic information. Finally, in our "Variants" condition, stimulus sets contained different example images of the same noun, designed to decrease semantic discriminability across stimuli without simplifying the stimuli themselves (i.e., images alone had full semantic information, but as a group caused interference by all being associated with the same name). We investigated the effect of these conditions through behavioral comparisons of learning behavior across the three conditions and two load conditions, as well as computational modeling to try to understand changes in the underlying RL and WM processes across conditions.

Generally, we predicted that both RL and WM would be necessary to capture behavior in all conditions, but that the processes would behave differently across the three stimulus conditions. However, due to 1) the fact that both Text and Variants conditions likely had lowered discriminability in both visual and semantic dimensions and 2) the potentially competing effects between RL and WM, it was difficult to predict exactly how changes in RL, WM, and their interplay would affect the ultimate behavioral performance across conditions. Take, for example, the Variants condition vs. the Standard condition. An assumption in the RL literature is that learning associations from stimuli with semantic information (e.g., Standard condition) may recruit "more explicit" processes like WM, and thus that a Variants condition could avoid contamination from explicit processes and better access to implicit learning ones. However, the assumption that decreasing semantic discriminability would lower the contribution of WM in learning is untested. In fact, the visual WM literature consistently demonstrates that WM representations need not be verbalizable at all. Additionally, people are able to reliably discriminate between WM representations of naturalistic stimuli with the same label (Brady, Störmer, & Alvarez, 2016). Similarly, if RL is indeed an implicit process, as often hinted in the literature, then stimulus condition should not impact it much. However, if RL instead relies heavily on distinct semantic information across stimuli, performance should suffer in the Variants condition. Thus, while we had a strong prediction that stimulus type would impact learning, and could impact the different processes supporting learning in different ways, we did not have a strong prediction as to the exact nature of this impact. We designed the study with an eye to behavioral modeling to help understand the intertwined processes.

Our results confirmed that stimulus type impacted learning; we observed lower performance in the Variants and Text conditions relative to the Standard condition, demonstrating that overall discriminability is important in learning. The behavioral deficit was particularly pronounced in the Variants condition. Through computational modeling, we found that stimulus conditions seemed to specifically affect RL, and not WM.

2 Experiment 1

In Experiment 1, participants completed a Conditional Associative Learning paradigm, learning correct stimulus-action associations through feedback.

2.1 Experimental Methods

2.1.1 Participants

88 participants were recruited through Amazon Mechanical Turk (MTurk), provided informed and written consent, and verified they were adults. The study was in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of University of California, Berkeley (IRB 2016-01-0820). Participants received $0.50 base payment for participating, and earned bonus payments for the time they spent on the task and their accuracy. Participants were informed that each correct response would increase their payment, and were reminded of this when starting each block. On average, participants made $3.30 and spent 42 minutes on the task. Participants who were performing below chance after the fourth or eighth block were discontinued from completing the task, but were compensated for their time. Participants who performed under 40% accuracy overall were additionally excluded from further analyses. 19 participants did not complete the task and 10 participants did not meet the accuracy threshold, leaving 59 participants in the final online sample.

2.1.2 Experimental design

Participants completed a Conditional Associative Learning paradigm (Petrides, 1985), adapted to investigate the contributions of RL and WM in learning (Collins & Frank, 2012; Collins, Brown, Gold, Waltz, & Frank, 2014). At the beginning of each block, participants viewed a screen that displayed the set of stimuli that would be used on that block. They were instructed that each stimulus had a single correct button press associated with it, and that their goal was to learn the correct association using trial-and-error. On each trial in the block, participants viewed a centrally-presented stimulus from this set and had up to 1500 milliseconds to press one of three buttons on a keyboard to respond (Figure 1A). Participants received binary, deterministic reward feedback after each response indicating whether the response was correct for this stimulus. If participants failed to respond within 1500ms, the screen indicated "response too slow," and were coded as nonresponses for subsequent analyses. Each stimulus was presented approximately 13 times within a block (stimuli were presented as few as 11 and as many as 14 times). Participants learned sets of either 3 or 6 images (stimuli) at a time, resulting in two set sizes for analysis. The larger set size (6 stimuli) resulted in greater WM load as well as longer delay times between repetitions of the same stimulus, and thus were more difficult. Because all stimuli were presented approximately the same number of times, the total number of trials per block was either 39 or 78. All blocks had the same number of keypress options (3), and the information about any stimulus-key pairing was not informative of any others within or across blocks (i.e., it was not the case in the 3 stimuli blocks that each stimulus mapped to a different key). Thus, chance performance was 33%.

In addition to the set size condition, each block also belonged to one of the three following stimulus conditions (Figure 1B):

- Standard: stimuli are images of different subcategory members belonging to the same category (e.g., vegetables: broccoli, celery, potato), and easily discriminable both semantically and visually.
- Text: stimuli are words printed in black letters on a white background, corresponding to subcategory name (e.g., the words "broccoli", "celery", "potato"). This condition is designed to provide full semantic information as Standard, but lowered visual discriminability within stimulus set.
- Variants: stimuli are different images of the same subcategory (e.g., different images of broccoli). This condition is designed to provide rich visual information, but limited distinct semantic information relative to the Standard condition – each image within a set was designed to call to mind the same word to limit the ability to have unique verbal labels for each image.

One of our primary criteria for choosing the stimuli across conditions was for them to be similarly naturalistic and familiar/recognizable to the participants. There is evidence that humans learn differently between abstract and naturalistic stimuli (Farashahi et al., 2020). Furthermore, differences in familiarity could also impact learning. Stimuli in the Standard condition were based on prior studies using the RLWM design (Collins & Frank, 2012), and were taken from ImageNet, a crowdsourced dataset commonly used to train the computer vision networks on image classification.

Variants condition images were also acquired from ImageNet, but chosen to call to mind the same word. Based on reported verbal strategies from prior studies using RLWM tasks, we predicted that allowing for extraneous visual variance could lead to alternative labeling strategies (for example, labeling a broccoli on a farm "farm" and a broccoli on a kitchen table as "table"), so we additionally minimized the possibility of additional distinguishing features (e.g., all images of broccoli on a plain background). While there is less visual discriminability in the Variants condition than the Standard one, the images are certainly not perceptually confusable, for they vary along lower-level visual dimensions (e.g., broccoli in different orientations, of different size, shades of green). Ultimately to keep stimuli naturalistic, we opted to use images that alone, had full semantic information (i.e., were individually nameable), but as a group caused interference (i.e., were all associated with the same name).

With similar motivation, we chose to use Text for a condition that had full semantic information while limiting visual information. While it would have been ideal to use images that looked alike but depicted different things, we could not think of such visual stimuli while satisfying the naturalistic and familiar constraints we imposed on our stimulus conditions. We thus compromised by simply writing the words out (i.e., showing a picture of black letters on a white screen), lowering visual information overall without sacrificing semantic information.

Each block had a unique category (e.g., vegetables, farm animals, clothing items), so a participant would not see, for example, stimuli corresponding to "farm animals" in both the Standard and Variants conditions. Which category was assigned to each stimulus condition, and what order they were presented in, was counterbalanced across participants, so participants saw different subsets of the entire stimulus set. The block order of the set size and stimulus conditions were also pseudorandomized across participants. Participants completed two blocks per set size x stimulus condition as well as one practice and one final block, completing a total of 780 trials over 14 blocks. We did not consider the first and last block in any analyses to remove potential effects of practice or fatigue, leaving 702 trials for analysis.

2.2 Experimental Results

Learning was successful in all conditions, indicated by an increasing proportion of correct responses as a function of stimulus iteration (Figure 1C). As in prior studies using the RLWM design, participants responded slower in the set size 6 blocks than in the set size 3 blocks. However, a two-way repeated measures ANOVA with stimulus condition, set size, and their interaction showed that while the difference between the set sizes was significant ($p < .001$), there was no effect of stimulus condition ($p = .62$) on reaction time, nor an interaction between condition and set size ($p = .57$). Reaction times are not analyzed further, but are shown in Supplementary Figure 6. To describe experimental effects on accuracy, we conducted a two-way repeated-measures ANOVA with stimulus condition, set size, and their interaction as independent variables, as well as separate intercept terms for each participant. There was a significant effect of set size, such that set size 3 blocks had overall better mean performance ($M = .79$, $SEM = .02$) than set size 6 blocks ($M = .66$, $SEM = .02, F(1, 58) = 106.2, p < .001$, Figure 1C), supporting the involvement of WM in learning and replicating prior work using this paradigm (e.g., Collins, 2018). There was a significant main effect of condition ($F(2, 116) = 43.95, p < .001$), such that performance in the Variants condition ($M = .66, SEM = .02$) was significantly lower than both Standard ($M = .78, SEM = .02, p < .001$) and Text conditions ($M = .74, SEM = .02, p < .001$). Standard and Text conditions were not significantly different ($p = .18$). The p-values for posthoc tests are Bonferroni corrected. Finally, there was a significant interaction between condition and set size ($F(2, 116) = 6.803, p = .002$); this was due to a stronger effect of condition in set size 6 ($F(2, 116) = 38.8, p < .001$) than set size 3 blocks ($F(2, 116) = 8.71, p < .001$). This suggests that stimuli differences are more critical for learning when learning more stimulus-action associations simultaneously.

While the ANOVA reveals gross overall effects, it neglects the progress of learning across set sizes and conditions; to better qualify this experimental effect we conducted a logistic regression. For each participant and condition, we investigated whether we can predict trial-by-trial accuracy based on the previous number of correct outcomes for that stimulus, the set size, and the delay since last correct. We found results consistent with previously reported studies (e.g., Collins & Frank, 2012; Collins et al., 2014), such that the probability of a correct response on the current trial was positively related to previous number of correct (as expected from incremental RL-like learning), and negatively related to set size and delay in all conditions (as expected from WM contributions to learning; predictors are illustrated in Figure 1D).

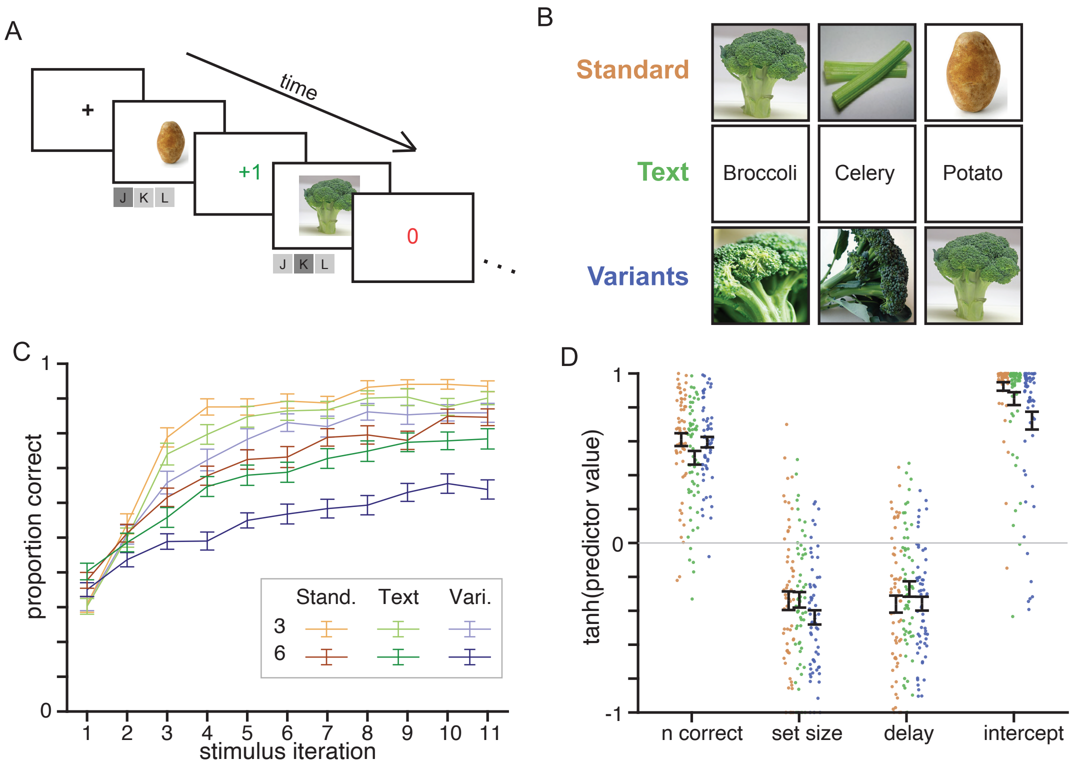

Figure 1: **Experiment 1 task and learning curves.** A. Behavioral task. Participants learn through trial and error, with veridical, deterministic feedback, the correct response to each stimulus. B. Example "vegetable" stimuli, for the three different stimulus conditions: Standard, Text, Variants. Stimulus categories were different for each block, so participants would never see (for example) a broccoli in multiple learning blocks. C. Learning curves ($M \pm SEM$ over participants) show the proportion of correct choices as a function of the number of times a stimulus has been encountered within a block (stimulus iteration), for each stimulus condition (color) and set size (value/saturation). While 11 stimulus iterations are illustrated, some stimuli were presented more times. D. Logistic regression weights (hyperbolic tangent transformed) for each condition (colors) and participant (dots; error bars indicate $M \pm SEM$ across participants).

2.3 Modeling methods

While descriptive statistics allow us to qualify the effects of set size and learning for each condition, these tests do not allow us to understand how the underlying processes, RL and WM, produce these behavioral differences across conditions. For this, we turn to behavioral modeling. Like previous publications using similar tasks and models (e.g., Collins & Frank, 2012; Viejo, Khamassi, Brovelli, & Girard, 2015; Jafarpour, Buffalo, Knight, & Collins, 2022), we assume participants' responses depend on both RL and WM processes. We describe the general "RLWM" framework, then consider different models that make different condition-specific predictions.

2.3.1 General model formulation

In this section, we describe the building blocks of the models we will be testing. We describe the basic learning rules for the RL and WM processes and how a policy is derived from each process's representation of stimulus-action associations.

Learning rules In this section, we discuss the learning rules for the RL and WM processes. We refer to the stimulus (s) action (a) value pairs as Q-value for RL process, $Q(s, a)$, as is standard in the model free reinforcement learning literature, and the corresponding stimulus-action association pairs for WM process as WM, $\mathrm{WM}(s, a)$. When we refer to operations that apply to both functions interchangeably, we generalize using the term "value function," which we denote $V(s, a)$.

RL learning rule. This is the classic Rescorla-Wagner model, in which the observer iteratively learns the value of each stimulus-action response through trial-and-error feedback. After observing reward r_t, the participant updates the Q-value as follows:

$$\forall s, a\, Q_0(s, a) = \frac{1}{N_a}$$

$$Q_{t+1}(s, a) \leftarrow Q_t(s, a) + \alpha(r_{t+1} - Q_t(s, a)),$$

where N_a is the number of possible actions (3 in our experiment) and α is the learning parameter. The larger α, the more informative the current trial is in the Q-value. To allow for learning asymmetry (e.g., Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Niv, Edlund, Dayan, & O'Doherty, 2012; Gershman, 2015; Sugawara & Katahira, 2021), we use two different learning rates for positive (correct) and negative (incorrect) rewards. We fit models in which both α and α_- are free parameters, as well models in which α_- is fixed to 0 (e.g., Xia et al., 2021; Eckstein et al., 2022). In the main manuscript, we report only the models in which $\alpha -_= 0$, for relaxing this assumption did not improve model fit and did not change the main results or conclusions (Supplementary 6.7.2).

WM learning rule. The WM observer updates the association value of stimulus-action pairs immediately to the observed reward, but this "perfect" information is subject to memory decay. The value association update is as follows:

$$\forall s, a\, \mathrm{WM}_0(s, a) = \frac{1}{N_a}$$

$$\mathrm{WM}_{t+1}(s, a) \leftarrow r_{t+1},$$

for $r = 1$, which can be thought of as a Rescorla-Wagner update rule with an $\alpha = 1$ and $\alpha_- = 0$. The WM decay is implemented by, on every trial, having all stimulus-action associations decay towards their starting value:

$$\forall s, a\, \mathrm{WM}_{t+1}(s, a) \leftarrow (1 - \lambda)\mathrm{WM}_{t+1}(s, a) + \lambda \mathrm{WM}_0(s, a),$$

where λ is the decay rate. With this formulation, WM's stored values regress to uninformative values, $\mathrm{WM}_0(s, a)$, for items that have been seen longer ago.

Calculating response probability. We assume that the observer chooses action a_i with probability based on a softmax function:

$$p_V(a_i|s) = \frac{e^{\beta V_t(s,a_i)}}{\sum_{i=1}^{3} e^{\beta V_t(s,a_i)}},$$

where β is the inverse temperature parameter and controls the stochasticity in choice, with higher values leading to a more deterministic choice of the best value action. Here, we fix β to an arbitrarily high number, 100. Fixing β to a high number enforces behavior we find to be a necessary theoretical baseline: it simulates behavior that is true to the way WM is theorized (it enforces a close to perfect one-back WM policy under low load) whilst still being consistent with the general formulation of RL models. Additionally, it is common practice in "RLWM" models (e.g., Jafarpour et al., 2022; McDougle & Collins, 2020), and improves interpretability of parameters (i.e., parameter recovery is only successful when β is fixed). $V_t(s, a_i)$ depends on the given state s, action a_i, and process (RL vs. WM).

Perseveration. Models with perservation incorporate the tendency of agents to respond based on previous actions, irrespective of the current stimulus and reward (e.g., Sugawara & Katahira, 2021).

$$V_t(s, a_i) = V_t(s, a_i) + \phi C_t(a_i),$$

where ϕ denotes how strongly a participant perseverates in their responses, and $C_t(a_i)$ is the choice trace vector of action a_i. The models in the main text set $C_t(a_i) = 1$ if the choice on trial $t-1$ was a_i, and 0 otherwise. (We fit all models without perseveration, and fits were significantly worse across models. We additionally allow perseveration choice to be affected by trials more than one trial back, with decay parameter τ; this addition does not approve the fits. Details can be found in Supplementary 6.7.3).

Response policy. The probability of responding action a_i given state s, $p(a_i|s)$ is a weighted sum of the contribution from the RL and WM process.

$$p(a_i, s) = \omega_n p_{\mathrm{WM}}(a_i|s) + (1-\omega_n) p_{\mathrm{RL}}(a_i|s),$$

where the mixture weight ω_n is a value between 0 and 1, corresponding to the WM contribution for blocks with set size n. In a fully RL-driven model, $\omega_n = 0$; in a fully WM-driven model, $\omega_n = 1$. We predict that $\omega_6 < \omega_3$ because there is lower WM contribution in higher set size conditions, but we do not impose this constraint during model fitting.

Random responses. We additionally assume that, with proportion ϵ, participants randomly choose an action. We are agnostic to whether this behavior reflects a response lapse, a random guess, or greedy exploration. The final response policy at time t, π_t is thus

$$\pi_t(a_i|s) = (1-\epsilon) p(a_i|s) + \frac{\epsilon}{N_a}.$$

2.3.2 Models

In this section, we describe the six models we considered. All models assume that both RL and WM are involved in the learning process, but make different assumptions about whether and how each of the two processes are affected by stimulus conditions. We did not consider models in which only RL or only WM are involved, for neither would be able to capture data across set sizes, let alone across conditions (Supplementary Figure 22). First, we test three models in which RL process is affected specifically. We test one model in which condition-differences in learning are assumed to be a result of different learning rates (RL learning rate). We test alternative models that assume confusion *within* a stimulus set results in noisier learning: either that updating the current stimulus accidentally updates other stimuli in the same block (RL credit assignment), or that retrieving the values of the current stimulus is confused with other stimuli (RL decision confusion). Second, we consider two models in which the WM process is affected specifically, either through differing decay (WM decay) or decision confusion (WM decision confusion) across conditions. Finally, we consider a model that assumes that the RL and WM processes aren't changed in isolation based on stimulus condition, but the interaction between the two (RL WM weight). This model hypothesizes that the observer relies on RL and WM to different degrees, depending on stimulus condition. Alternative assumptions, different specifications for perseveration or nonzero negative learning rate α_- are presented in Supplementary Materials 6.7, but these did not better explain our data than the models presented here.

Condition-specific RL learning rate. Motivated by the observation that stimulus condition influences accuracy, we first consider a model which assumes that stimulus condition impacts how quickly RL updates Q-values. We implement this assumption by fitting three separate α parameters, one for each stimulus condition. We denote the learning parameter for Standard, Text, and Variants stimuli as α_s, α_t, and α_v, respectively.

Condition-specific RL credit assignment. In the "RL credit assignment" observer, we test the assumption that the lowered performance in different conditions is not due to lowered learning rates, but increased difficulty to distinguish the stimuli which leads to credit assignment confusion. Credit assignment confusion occurs when updating Q values not only for the current trial's stimulus, but also for other stimuli, leading to potential future interference between stimuli. For example, when a reward is obtained for a given choice and stimulus, the rewarded choice would also be credited to other stimuli, although those stimuli may require a different correct action.

With standard RL and WM learning rules, the observer only updates state-action values for the current stimulus, s_i. With credit assignment confusion, all other stimuli in the current block (which are not relevant to the current trial) are also updated to a lesser degree, parameterized by weight $0 \leq \eta \leq 1$:

$$\forall s_j \neq s_i : V_{t+1}(s_j, a) \leftarrow V_t(s_j, a) + \alpha\eta(r_{t+1} - V_t(s_i, a)).$$

We fit credit assignment confusion parameters to Text and Variants conditions only, denoted η_t and η_v, respectively. We did attempt to fit a model with credit assignment confusion in the

Standard condition, η_s, and did not include in the main manuscript because parameter recovery was not successful for that model; this is likely because a combination of other parameters (e.g., α, β, λ, ϵ) can characterize noise in a way that is behaviorally difficult to distinguish from credit assignment alone. In this sense, we assume that any credit assignment confusion in the Standard condition would be generally captured by noise parameters, and that the **additional** confusion in the Text and Variants conditions would be captured by the condition-specific parameters. This additional confusion is our primary interest, for we are interested in the difference in performance across conditions.

Condition-specific RL decision confusion. In the "RL decision confusion" observer, we test the assumption that the lowered performance in different conditions is due to across-stimulus decision confusion when the observer is calculating their response policy. In other words, the confusion is not in the encoding of the state-action values (like the RL credit assignment model), but the retrieval of values when making a decision. Decision confusion is implemented during the decision stage, such that all stimuli in the current block that are not relevant to the current trial are also used to calculate the response policy for the RL process:

$$V_t'(s, a_i) = (1 - \zeta)V_t(s, a_i) + \zeta \frac{1}{N_s - 1} \left(\sum_{\neg s} V_t(\neg s, a_i) \right), \tag{1}$$

where N_s is number of stimuli, parameter ζ is a scalar between 0 and 1, and indicates how much across-stimulus decision confusion there is. A value of 0 indicates no decision confusion, and a value of 1 would indicate full confusion. We fit decision confusion parameters for the Text and Variants conditions, denoted ζ_t and ζ_v, respectively. Like in the RL credit assignment model, we implicitly assume there is no RL decision confusion in the Standard condition, $\zeta_s = 0$, for modeling parsimony and recoverability, or that RL decision confusion is absorbed by other noise in that condition. In that sense, again, this model assumes additional processes in the Text and Variants conditions, to attempt to capture observed performance drops.

Condition-specific WM decay In this model, we test the assumption that WM decay is solely responsible for performance differences across conditions. Rather than learning the values faster in certain conditions, we just remember the associations better. We denote the WM decay for Standard, Text, and Variants stimuli as λ_s, λ_t, and λ_v, respectively.

Condition-specific WM decision confusion This model is the WM analog to the RL decision confusion model. In this model, we test the assumption that participants have across-stimulus decision confusion when calculating the response policy for the WM process, according to Equation 1.

Condition-specific weight In this model, we test the assumption that different weights between the RL and WM processes results in different behavior, rather than condition differences resulting from changes in either process. So, when encountering different stimuli, either system could be modulated to have a larger or smaller effect. In this model, the weights ωs differ across condition and set size, and are denoted with subscript. For example, ω_{6s} corresponds to the RLWM weight of a set size 6 Standard stimulus condition. We include the simplifying assumption that the differences across conditions in set size 3 blocks are minimal, and use ω_3 for all set size 3 stimulus conditions. Thus, the Condition-specific weight model has four ω parameters, $\omega_3, \omega_{6s}, \omega_{6t}$, and ω_{6v}.

2.3.3 Parameters and estimation

The parameters for each model, $\boldsymbol{\theta}$ are displayed in Table 1. All models we consider contain the following fitted base parameters: RL learning rules with positive learning rate α, WM with forgetting rate λ, perseveration with proportion ϕ, response policies which are a weighted combination of RL and WM components with a weighted sum (determined by weight ω_3 and ω_6 for set size 3 and 6, respectively), and random responses with proportion ϵ. Model-specific parameters are presented in the, aptly named, "Model-specific parameters" column.

For each participant and each model, we maximized the logarithm of the likelihood (LL) of the data given the parameters and model $\log(p(\text{data}|\boldsymbol{\theta}))$, using fmincon in MATLAB with 20 random starting points. The largest LL, LL^*, and the associated parameter $\boldsymbol{\theta}$ are assumed to be the global maximum-likelihood parameter estimates.

Model	**Base parameters**	**Model-specific parameters**
RL learning rate	$\alpha_s, \lambda, \phi, \omega_3, \omega_6, \epsilon$	α_t, α_v
RL credit assignment	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	η_t, η_v
RL decision confusion	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	ζ_t, ζ_v
WM decay	$\alpha, \lambda_s, \phi, \omega_3, \omega_6, \epsilon$	λ_t, λ_v
WM decision confusion	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	ζ_t, ζ_v
RL WM weight	$\alpha, \lambda, \phi, \omega_3, \omega_{6s}, \epsilon$	ω_{6t}, ω_{6v}

Table 1: **Model parameters**. Free parameters for each model. Base parameters are loosely comparable across all models; model-specific parameters are additional ones fit to capture condition-specific effects.

2.3.4 Model and parameter recovery

A crucial, but often overlooked, step in interpreting model parameters and in quantitative model comparison is making sure parameter values are meaningful and that models are identifiable (Nilsson, Rieskamp, Wagenmakers, & Nilsson, 2011; Palminteri, Wyart, & Koechlin, 2017; Wilson & Collins, 2019). In order to establish the interpretability of model parameters, one should test that the same parameters that generate a data set are the ones estimated through the model parameter estimation method. Successful parameter recovery exists when one is able to "recover" the same (or similar) parameter values that generated the data.

Successful model recovery is an important step for making conclusions from quantitative model comparisons. Successful model recovery occurs when the same model that generates a data set is the model that best fits it (according to your chosen model comparison metrics), when compared to all other models in the comparison set. We obtained reasonable parameter recovery and model recovery; details and figures for both analyses are in Supplementary sections 6.4 and 6.5).

2.3.5 Model comparison

Because all of our models have 8 parameters, we report model goodness-of-fit by simply comparing LL^*, the maximum LL across all runs for a participant and model. In addition to LL^*, we compared fits across participants with group Bayesian Model Selection (BMS; Stephan, Penny, Daunizeau, Moran, & Friston, 2009; Rigoux, Stephan, Friston, & Daunizeau, 2014). While summed LL^* assumes all participants are generated by the same model, BMS explicitly assumes that participants can be best fit by different models. BMS assumes that the distribution of models is fixed but unknown across the population, and uses the log marginal likelihoods for each model and participant to infer the probability of each model across the group. This method is sensitive to both the distribution and magnitude of the differences in log-evidence. From this, we can compute the protected exceedance probability (pxp), which is how likely a given model is to be more frequent than the other models in the comparison set, above and beyond chance. A lower summed LL^* and higher pxp indicate better model fit to data.

2.4 Modeling Results

Both metrics gave similar results, favoring the RL learning rate model over the RL credit assignment, WM decay, WM decision confusion, and RL WM weight models. The RL decision confusion model performed similarly well to the RL learning rate model. We illustrate individual-participant, median ΔLL^*s, summed ΔLL^*s, and pxps in Figure 2B.

Second, we qualitatively compared the models' ability to generate data similar to that of the real data. For example, posterior predictive checks are an important step in assessing model fits, particularly for data with sequential trial dependencies (Palminteri et al., 2017); a simple model of the weather that predicts today's weather is the same as yesterday's may result in high likelihoods without being able to actually predict weather patterns. For each participant, we simulated data using the MLE parameters for each participant, and find that the qualitative fits to the data (Figure 2A) reflect the quantitative model comparison; the models that feature either condition-specific RL learning rates or condition-specific RL decision confusion provide a better fit to the true data than other models. These results suggests that different stimulus conditions affect exclusively the RL process, by how efficiently it learns from or uses reward information.

2.5 Interim conclusions

In Experiment 1, we asked how limiting discriminability in semantic or visual information across stimuli changes people's ability to learn stimulus-response associations in a load-dependent RL task. First, we replicated the set size effect, showing that for all task conditions a load of 6 stimuli produced worse performance than blocks with only 3 stimuli, indicating WM's role in task

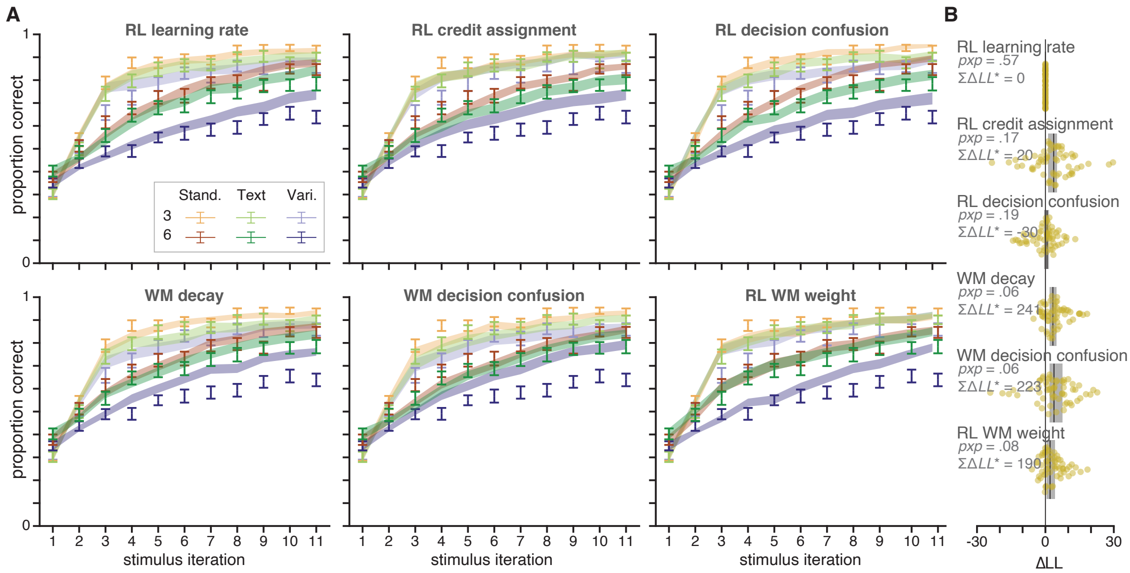

Figure 2: **Experiment 1 Modeling Results.** A. Learning curves for each condition (color) and set size (value/saturation) across participants for data (errorbars, $M \pm SEM$) and model predictions (fills, $\pm SEM$). Only the first 11 stimulus iterations are illustrated, but all iterations were used in modeling. B. Difference in LL scores for each model, relative to the RL learning rate model. Dots indicate individual participants, black line indicates median, and grey box indicates 95% bootstrapped confidence interval of the median. Difference of summed ΔLL^*s across participants and protected exceedance probability displayed for each model. Lower LL^*s and higher pxps indicate better model fit.

performance. Second, and to our main question, we found that limiting either discriminable visual or semantic information across stimuli detrimented performance. This condition effect interacted with load such that it had a larger effect in higher load conditions, suggesting that the condition may tax the RL system that is more responsible for behavior in the larger load conditions.

We used computational modeling to investigate if we could explain the process by which this performance detriment occurs, and found that a model that either assumes that people have lower RL learning rates or have higher confusion across stimuli when calculating the RL response policy was able to capture the data reasonably well qualitatively, and quantitatively better than other models. However, all models predict slightly higher performance in the Variants condition set size 6 relative to human performance (Figure 2). In Experiment 2, we designed an experiment to more directly test the contribution of RL in learning, by adding a surprise memory test.

3 Experiment 2

Our second experiment was designed to replicate and extend the behavioral and modeling results of the first experiment. First, participants completed the same stimulus-response paradigm as in Experiment 1. Participants then completed a "Test phase," after a WM distractor task, designed to clear WM. During the Test phase, all stimuli from all Learning phase blocks were presented again in random order, and participants responded which of the three response keys they believed to be the correct response. No feedback on correctness was given. This phase probed how well stimulus-response pairs were learned by a RL process, presumably without the aid of WM.

3.1 Experimental Methods

3.1.1 Participants

Thirty-seven participants (22 female, mean age 21) were recruited through a UC Berkeley online site and received course credit for experimental participation. Participants in this experiment did not receive any bonus compensation based on performance. We obtained informed, written consent from all participants. The study was in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board of University of California, Berkeley (IRB 2016-01-0820). Seven participants were excluded for psychiatric diagnosis disqualifications, withdrawing early, not being fluent in English, or monitor malfunctions in the testing rooms, leaving 30 (19 female, mean age 21) participants in the final online sample.

3.1.2 Experimental design

Participants completed the same stimulus-response learning paradigm, with the same numbers of trials and blocks, as in Experiment 1. In addition to this "Learning Phase," participants additionally completed a WM distractor task and a "Test Phase," which they were not told about ahead of time.

In the distractor task, participants completed 5 blocks of a N-back task. This task was designed to tax the WM system, clearing any working memory information about stimulus-response mappings from the Learning Phase, and is not analyzed in main manuscript. More details about this task can be found in the Supplementary Materials Section 6.2. It took approximately 10 minutes to complete.

Lastly, participants completed a surprise Test Phase, in which all stimuli from the Learning phase blocks were presented again in random order. Because the Test phase was beyond both WM capacity (54 associations tested) and maintenance period for most stimuli, this phase probed how well stimulus-response pairs were learned by a RL process alone. For each trial, a stimulus was presented, participants responded which of the three response keys they believed to be the correct response, and no feedback on correctness was given. Each of the 54 unique stimuli from the learning block was presented four times, for a total of 216 trials. Only stimuli from the middle 12 blocks (i.e., excluding stimuli from the first and last block) were included in this test phase to limit primacy or recency effects of memory (Murdock Jr., 1962). Because each Learning phase block corresponded to a unique category (i.e., a participant would see stimuli corresponding to "vegetables" in only one stimulus condition), there should not be any category-specific interference between blocks. All trials were completed in a single block.

3.2 Experimental Results

Here, we analyze the behavioral results from the Learning phase and Test phase. First, we analyze learning phase data as done in Experiment 1 (Fig. 3A, middle). We conducted the repeated measures ANOVA, with proportion correct as the dependent variable and set size and stimulus condition as independent variables. There was a significant effect of set size ($F(1, 29) = 185.1$, $p < .001$), condition ($F(2, 58) = 24.66$, $p < .001$), and interaction between set size and condition ($F(2, 58) = 11.90$, $p < .001$). For condition, performance in the Variants condition ($M = .69, SEM = .03$) was significantly lower than that of the Standard ($M = .79, SEM = .02, p < .001$) and Text ($M = .76, SEM = .02, p = .02$) conditions. Performance was not significantly different for Standard and Text conditions $p = .53$). The interaction was driven by a nonsignificant condition effect in set size 3 blocks ($F(2, 58) = 2.44$, $p = .10$) but a strong condition effect in set size 6 blocks ($F(2, 58) = 27.07$, $p < .001$). We then conducted the logistic regression to test whether the likelihood of responding correctly on the current trial could be predicted from the previous number correct for that stimulus, the set size, and the delay since last correct. We found results consistent to Experiment 1 such that the probability of getting a correct response on the current trial was positively related to previous number of correct, and negatively related to set size and delay (Fig. 3A, right). Reaction time analyses revealed the same pattern of results as in Experiment 1: participants responded slower in the set size 6 blocks than in the set size 3 blocks, but an ANOVA showed that while the difference between the set sizes was significant ($p < .001$), there was no effect of stimulus condition ($p = .11$) or an interaction between condition and set size ($p = .80$; Supplementary Figure 6).

Second, we analyzed the participants' performance on the Test phase. Collins and others

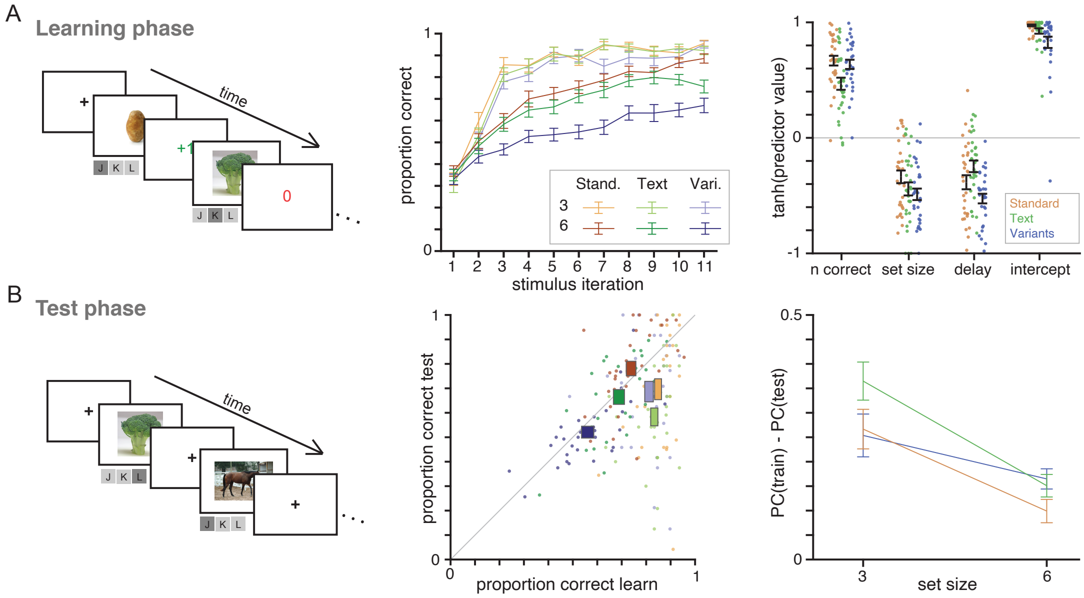

Figure 3: **Experiment 2 task and results.** A. Learning phase. *Left*: Task design. *Middle*: Proportion of correct choices increases as a function of stimulus iteration for all stimulus and set size conditions but slower for set size 6, especially in the Variants condition. *Right*: Logistic regression. For all three conditions, participants are more likely to select the correct response when it is a lower set size block, shorter delay, and when they have gotten more correct responses on that stimulus previously. B. Test phase. *Left*: task design. Participants viewed all stimuli previously learned and reported their believed correct response. No correctness feedback was given. *Middle*: Proportion correct in training (x-axis) and testing (y-axis) phase for condition (color), showing individual participants (dots) or $M \pm sem$ across participants (boxes). *Right*: Tortoise and hare effect: there is a larger deficit in long-term retention (difference in proportion correct (PC) from train to test) with stimuli learned in set size 3 blocks than set size 6 blocks. This deficit was not significantly different across conditions.

(2018) demonstrated an interaction between RL and WM processes for long-term retention of the correct stimulus-action pair. Items in lower set size blocks had better performance during the Learning phase compared to higher set size blocks, but interestingly, a larger detriment in performance in the Test phase. This "tortoise and hare" effect demonstrated a trade off between RL and WM process; while WM assists performance during learning, it detriments long-term retention of the stimulus-action pairs. For all conditions and set sizes, performance was above chance ($t(29) > 6.35, p < .001$), suggesting long-term retention of stimulus-response associations even without explicit instruction to do so. Second, there was a significant positive correlation across participants between the proportion correct in the Learning and Test phases ($r = .40, p = .03$). Finally, the difference between performance in Learning phase and Test phase was much larger in trials corresponding to stimuli learned in set size 3 blocks than ones learned in set size 6 blocks ($t(29) = 6.41, p < .001$), replicating the tortoise and hare effect, showing interference of WM with RL learning. We conducted a one-way repeated measures ANOVA and found no statistical difference in the magnitude of this "tortoise and hare" effect across conditions ($F(2, 58) = 2.207, p = .12$). This nonsignificance of magnitude of deficit suggests that the difference in WM used between set size 3 and 6 in each condition is nonsignificantly different.

3.3 Modeling methods

3.3.1 Replication of Experiment 1

We first analyzed the Learning phase of Experiment 2 identically to that of Experiment 1. Details on the six models, fitting procedure, and model comparison can be found in Section 2.3.2-2.3.5.

3.3.2 Investigating Test phase

We additionally investigate model fit by jointly fitting Learning and Test phase data. In other words, all data are used to calculate the likelihood of parameter given model parameters and data. The likelihood of learning phase data are computed identically to the previous procedure. For test phase data, we assume that participants only have access to RL values, not WM association weights; thus the likelihood of test phase trials relies only on the Q-values learned during the learning phase, which are frozen through the test phase in absence of feedback (Collins, 2018). LLs are optimized in the same way as Experiment 1, and model are compared in the same way as Experiment 1. We fit the two best fitting models: the condition-specific RL learning rate and condition-specific RL decision confusion models.

We additionally test, for the RL learning rate and RL decision confusion models, the assumption that RL and WM processes are not independently updating value in during the learning phase, but actually interact during learning. As in Collins (2018), we implement this assumption such that WM contributes cooperatively during learning when calculating the RPE used by the RL process:

$$\delta_t = r_t - (\omega_n WM_t(s, a) + (1 - \omega_n) Q_t(s, a)). \quad (2)$$

We refer to this set of model as models "with interaction" (e.g., RL learning rate model with this modification is the "RL learning rate + interaction" model).

For all models, we additionally fit a softmax inverse temperature parameter, β, for the Test phase, under the assumption that response noise in using RL Q-values will likely differ for each participant between Training and Test phase due to failures in long-term retention of stimulus-response associations.

3.4 Modeling Results

We modeled the data in Experiment 2 in two ways. First, we fit only the Learning phase data, as in Experiment 1, to see if we could replicate those results. Second, we jointly fitted parameters on Learning and Test phase data, to see if modeling results differed from results when only fitting Training phase data.

Replication of Experiment 1 Modeling results were remarkably consistent with Experiment 1; the condition-specific RL learning rate model fit the substantially better than most models across participants, and similarly as well as the RL decision confusion model. These two models were best able to produce model predictions that looked qualitatively similar to that of the actual data (Fig. 4A). They were additionally able to capture the data quantitatively the best (Fig. 4B).

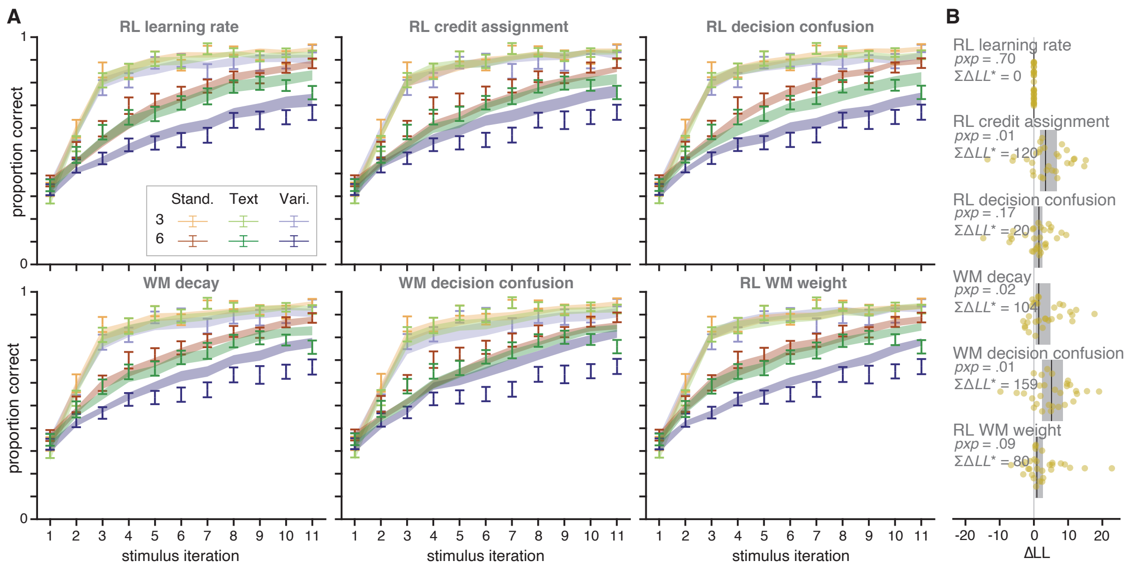

Figure 4: **Experiment 2 modeling results: replication of experiment 1** A. Learning curves for each condition (legend at top) across participants for data (errorbars, $M \pm SEM$) and model predictions (fills, $M \pm SEM$). Only the first 11 stimulus iterations are illustrated, but all iterations were used in modeling. B. Difference in LL^* for each model relative to the RL learning rate model. Dots indicate individual participants, black line indicates median, and grey box indicates 95% bootstrapped confidence interval of the median. Difference of summed LL^*s across participants and protected exceedance probability displayed for each model. Lower LL^*s and higher pxps indicate better model fit.

Investigating Test Phase Model validation plots are illustrated in Figure 5. Quantitatively, model performance was very similar (lower summed ΔLL^* and higher pxp indicates better model fits to data). RL learning rate summed $\Delta LL^* = 0$, $pxp = .25$; RL decision confusion summed $\Delta LL^* = 49$, $pxp = .23$; RL learning rate + interaction summed $\Delta LL^* = -44$, $pxp = .27$; RL decision confusion + interaction summed $\Delta LL^* = -8$, $pxp = .25$).

Qualitatively, the models that assume an interaction between RL and WM during learning were able to capture Test phase data better for the Standard and Text condition (orange and green), but models that assume no interaction were able to capture Test phase data better in the Variants condition (blue). As a follow up, we considered models that had condition-specific interaction strengths, but they were not able to fit the data substantially better than those reported here (Supplementary 6.7.5).

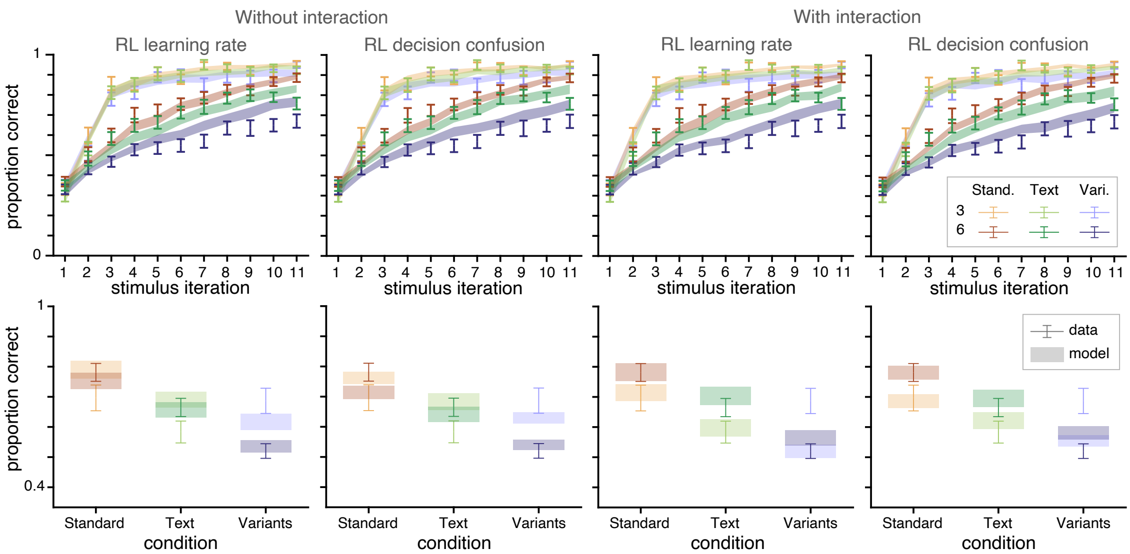

Figure 5: **Exp 2 learning and test phase model validation**. Model validation for RL learning rate and RL decision confusion models without (left two plots) and with (right two plots) an interaction between RL and WM processes during learning. Model predictions (fill) and data (error bars) for models jointly fitted on Training (top) and Test phase (bottom) data.

4 Further model investigations

4.1 Interpreting model parameters

We investigated the parameter values for the two best-fitting models: the condition-specific RL learning rate and the condition-specific RL decision confusion models (individual and group parameter values for models fit on Learning phase displayed in Supplementary 6.6).

We first investigated whether it was reasonable to combine participants across the two experiments, for the models that were fitted to only Learning phase data. For each model, we conducted Welch's t-tests for each parameter with a Bonferroni correction across parameters. We found for both winning models, no parameters were significantly different across experiments ($p > .41$). For all following analyses, we combine participant parameters across experiments.

To investigate the differences between condition-specific parameters for each the model, we conducted Wilcoxon signed-rank test with a Bonferroni correction across the number of pairwise tests. First, we investigated whether the learning rates, αs, across conditions differ in the condition-specific RL learning rate model. The learning rate for Variants condition (α_v: $M = .01$, $SEM = .003$) was significantly lower than that of Text condition (α_t: $M = .03$, $SEM = .006$, $z = -7.40$, $p < .001$) and Standard condition (α_s: $M = .04$, $SEM = .008$, $z = -6.37$, $p < .001$). The difference in learning rates for Standard and Text condition were not statistically significant ($z = 2.25$, $p = .07$). For the models fit to both Learning and Test phase data in Experiment 2, the results are largely consistent, finding that learning rate for the Variants (no interaction model: $M = .01, SEM = .001$, interaction model: $M = .008, SEM = .0008$) condition is lower than that of Standard (no interaction: $M = .04, SEM = .03, z = -4.37, p < .001$; interaction: $M = .04, SEM = .02, z = 4.41, p < .001$) and Text (no interaction: $M = .01, SEM = .003, z = -2.99, p = .008$; interaction: $M = .02, SEM = .004$,$z = 3.38, p = .002$) conditions. However, models that were fitted on both phases also found a statistically significant difference between Text and Standard conditions (no interaction: $z = 2.77, p = .02$; interaction: $z = 2.79, p = .02$).

For the RL decision confusion model, we found that the decision confusion for the Variants condition (ζ_v: $M = .44$, $SEM = .02$) was significantly higher than that of the Text condition (ζ_t: $M = .22$, $SEM = .03$, $z = 6.02$, $p < .001$). This effect is also true for the models fitted on Learning and Test phase of Experiment 2; decision confusion is greater in the Variants condition than the Text condition in both the models that assume no interaction between RL and WM (Variants: $M = .36, SEM = .04$, Text: $M = .18, SEM = .04$, $z = 2.95, p = .003$) and those that do (Variants: $M = .40, SEM = .04$, Text: $M = .20, SEM = .04$, $z = 3.38, p = .001$).

4.2 Alternative models

As in all modeling papers, we cannot possibly sample all possible models of this data. In our final analysis, we test two additional models that embody more complex hypotheses, as a control. We fit just the Learning phase data, and do not assume any interaction between RL and WM during learning.

Condition-specific RL learning rate and WM decay Our previous models assumed that only one process was affected by stimulus condition. In this model, we test the assumption that both processes are affected. To minimize additional complexity, we consider the model that lets the two most likely parameters from each process be condition dependent; specifically, this model assumes that RL learning rate and WM decay both depend on stimulus condition. Theoretically, this model allows us to test the assumption that both processes may differently but jointly contribute to differences in behavior. This model has the following 10 parameters $\alpha_s, \alpha_v, \alpha_t, \lambda_s, \lambda_v, \lambda_t, \phi, \omega_3, \omega_6, \epsilon$.

Superfree The "Superfree" model fits each condition entirely separately. Thus, it is extremely unconstrained, overparameterized, and lacks theoretical justification on its own. However, it provides a *qualitative* upper bound for the explainability of all models considered in this paper. We consider this model an important metric to use when considering the goodness-of-fit of models during model validation. This model has a total of 21 parameters, consisting of 7 parameters for each condition: $\alpha, \lambda, \phi, \zeta, \omega_3, \omega_6, \epsilon$.

4.2.1 Model comparison and results

For model comparison with the new additions, we focus on the previous winning models, as well as the previous best candidate model where WM parameters were condition dependent. Specifically, we select 1) RL learning rate and 2) RL decision confusion, and 3) the WM decay model. Because the models considered in this section have different numbers of parameters, we use corrected Akaike Information Criterion (AICc; Hurvich & Tsai, 1987) to quantitatively compare model goodness-of-fit. Like AIC (Akaike, 1972), AICc penalizes models with more parameters, using parameters as a proxy for model flexibility (and additionally corrects for potentially low trial numbers):

$$\text{AICc} = -2LL^* + 2k + \frac{2k(k+1)}{N_{\text{trials}} - k - 1}$$

where k is the number of parameter and N_{trials} is the number of trials. We chose to use AICc verses other model comparison metrics, because it provided us the best model recoverability, although it penalizes parameters less strictly than Bayesian Information Criterion (BIC). We report the median and mean of the difference between the AICc of one model and the RL learning rate model (ΔAICc); larger values provide larger support in favor of the RL learning rate model. In addition to reporting the protected exceedance probability of each model pxp, we report the expected posterior probability of each model, denoted $\exp_r$. These two metrics provide us a more heterogeneous interpretation of model goodness-of-fit, such that different models may be superior for different subsets of participants. All quantitative results for Experiment 1 and 2 are reported in Table 2 and Table 3, respectively.

Our results in this section are consistent with our other modeling results, for both experiments and for all model comparison metrics. First, as shown previously, both RL-only models individually fit better than the WM-only models in both experiments. Second, they individually fit better than the new model that assumed both RL and WM were affected by stimulus condition, suggesting that assuming condition-dependent WM changes does not provide any additional explanatory power to assuming only RL is affected (though, results of model recovery may weaken the interpretation of this result; Fig 18, 19) . Third, the model that assumed both RL and WM were both affected fit better than the WM-only model, suggesting that condition-specific RL modulation is key to fitting human behavioral data.

Interestingly, the RL-only models are not favored over the Superfree model in either experiment. These quantitative results do not reflect a simple overfitting; the Superfree model is not the best fitting model for data simulated by other models (i.e., model recovery is successful for our chosen model comparison metrics. Figure 18), and is qualitatively superior at capturing behavior in the set size 6, Variants condition (Figure 25). While the Superfree model seems to be capturing *some* *aspects* of behavior that others model are not, the overparameterization of the model (indicated by poor parameter recovery, Figure 16) makes it difficult to understand, in a meaningful way, why. On the other hand, the RL learning rate model still provides a superior fit for a nontrivial proportion of participants (Experiment 1 / 2: $\exp_r = .31$ / $.33$), suggesting that it is a competitive model, whilst still being interpretable.

	RL learning rate	**RL decision confusion**	**WM decay**	**RL learning rate + WM decay**	**Superfree**
pxp	0.21	0.01	0.00	0.00	0.77
$\exp_r$	0.31	0.18	0.04	0.08	0.39
mean(ΔAICc)	0	-1	8	0	-4
med(ΔAICc)	0	1	7	1	2

Table 2: **Experiment 1 quantitative model comparison**. Protected exceedance probability (*pxp*), expected posterior probabilities ($\exp_r$), mean AICc differences relative to RL learning rate (mean(ΔAICc)), and median AICc difference (med(ΔAICc)). Positive AICc values indicate that RL learning rate provides a better fit to the data.

	RL learning rate	**RL decision confusion**	**WM decay**	**RL learning rate + WM decay**	**Superfree**
pxp	0.30	0.04	0.04	0.05	0.56
$\exp_r$	0.33	0.09	0.04	0.14	0.39
mean(ΔAICc)	0	1	7	1	-1
med(ΔAICc)	0	3	3	2	0

Table 3: **Experiment 2 quantitative model comparison**. Protected exceedance probability (*pxp*), expected posterior probabilities ($\exp_r$), mean AICc differences relative to RL learning rate (mean(ΔAICc)), and median AICc difference (med(ΔAICc)). Positive AICc values indicate that RL learning rate provides a better fit to the data.

5 Discussion

In this study, we investigated how the type of information across a stimulus set affected learning. Participants learned the correct response to stimuli that had different levels of discriminability relative to other stimuli in the same block. In behavior across two experiments, we show that,when there are more items to learn about concurrently, performance suffers minimally in the Text condition relative to the Standard condition, but substantially in the Variants condition.

Through computational modeling, we found that the differences in learning behavior across stimulus conditions were driven by deficits in specifically the RL process. The models that best predicted behavior was the one that either assumed that, across conditions, the RL learning rate changed or that there was confusion in the RL system at the decision stage. These models fit better than those that assumed stimulus condition affected credit assignment in RL, WM decay, decision confusion in WM, or the weight between RL and WM. Additionally, models that assumed the RL was alone affected fit better than a model that assumed both RL and WM were affected by stimulus condition.

What could be causing the differences in learning across the two lowered-discriminability stimulus conditions? Perhaps there is a preference for the modality of stimulus. Perhaps the deficit in the Variants condition was driven by a lack of semantic distinctness. Many RL studies actively select non-nameable stimuli with the (often implicit) goals of targeting putatively implicit processes (Frank, Seeberger, & O'Reilly, 2004; Daw et al., 2011) and limiting the contributions of other, more explicit cognitive processes. Consequently, they rely on the hypothesis that stimulus information in the semantic domain may impact learning, and in particular the balance of RL processes and higher level processes such as inference or memory. In contrast to that interpretation, our results suggest that the semantic distinguishability of the stimuli affects RL itself, not a different process and not its interaction with another process. Our results are consistent with that of Radulescu and others (2022), who more directly tested nameability of stimuli on learning. Like us, they found that more nameable stimuli were associated with higher RL learning rates, and that the effect of nameability on performance was more apparent in larger set size conditions. This interpretation is consistent with the results in the Text condition as well. Because stimuli were still semantically discriminable, performance on the Text stimulus condition was not significantly worse than that of the Standard stimulus condition.

In contrast to the RL process, our computational results suggest a lack of impact of stimulus condition on the WM process. Perhaps this is due to sufficient information being available to WM regardless of stimulus condition. Let's consider the Variants condition, in which a lack of semantically distinct information across stimuli does not hurt learning behavior. In other words, there was sufficient visual information between stimuli that WM processing was not affected. This explanation seems feasible given the research on WM for visual stimuli. The visual WM literature has demonstrated that, despite WM being information-constrained, people are able to learn and prioritize information in WM that is most relevant to performance (e.g., Yoo, Klyszejko, Curtis, & Ma, 2018; Bays, 2014; Klyszejko, Rahmati, & Curtis, 2014; Emrich, Lockhart, & Al-Aidroos, 2017; Sims, 2015), even when stimuli are extremely simple and non-verbalizable (e.g., oriented lines, dots in space). Perhaps prioritization of relevant information would be easier with naturalistic stimuli; WM performance for naturalistic stimuli demonstrated to be better than with simple stimuli (Brady et al., 2016), and even more so for objects familiar to participants (Starr, Srinivasan, & Bunge, 2020, even when doing a simultaneous verbal task, to ensure verbal WM is not assisting). Our results and this literature together suggest that, unlike RL, WM can learn actions associated with a stimulus set with low semantic discriminability, as long as there is high visual discriminability (and vice versa). In other words, WM is able to discriminate stimuli and maintain stimulus-response associations equally well with only visual or semantic information. It is important to note, though, that while we designed these stimulus sets with visual and semantic modalities in mind, we did not quantify the difference between discriminability across conditions. Thus, it is possible that our interpretation of how visual vs. semantic information affects processing may be overly simplified.

What other processes could be causing the differences in learning in the RL process across stimulus conditions, beyond a simple modality preference? It is known that learning a category structure becomes more difficult with increased similarity of exemplars between categories (Love, Medin, & Gureckis, 2004; Nosofsky, 1986) and increasing number of dimensions required to distinguish categories (Nosofsky, Palmeri, & McKinley, 1994; Shepard, Hovland, & Jenkins, 1961). This difficulty is apparent in the Variants condition, in which participants had to distinguish between stimuli based on relatively low-level visual differences that are not often of ecological importance. This is in contrast to the Text condition, in which stimuli are so easily discriminable due to the association of the word with its meaning – a relatively automatic association, as seen in the well-replicated Stroop task (1935) – despite having relatively similar low-level visual characteristics across stimuli. In the Variants condition, unlike the Text condition, what features were important to pay attention to itself became something that needed to be learned (e.g., Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017), and likely affected behavior. For example, "learning traps" can occur in behavior (Rich & Gureckis, 2018), due to selective attention, simplification, or dimensionality reduction (Nosofsky et al., 1994; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). The poor performance in the Variants condition could have been because the relevant discriminating features in the Variants condition (e.g., luminosity, absolute size, orientation of object) are, in the other two experimental conditions and often in real life, trivial compared to object identity – your value assessment for an apple doesn't depend on how bright the room is. The combination of interference (due to interleaved condition blocks) and a learning trap (previous experience within and beyond the experiment indicating these low-level features are unimportant) could have resulted in difficulty successfully using these features to discriminate between stimuli for RL. Other studies corroborate this conclusion, finding stimulus type (e.g., naturalistic stimuli learned better than abstract stimuli; Farashahi et al., 2020) and response "state" (e.g., motor responses learned better than stimulus responses; Rmus & Collins, 2020) affect learning. Regardless of exact cognitive mechanism at play, these results demonstrate the importance of considering how a learning state is defined.

Our results have strong implications for understanding the neural circuits that support flexible learning. Previous research has focused on clarifying how the brain integrates past choice and reward history to make a choice given a stimulus, with little consideration to the inputs of this computation - such as the stimuli. Past findings have shown that multiple distinct neural systems contribute to learning. Reinforcement learning computations appear to be implemented in cortico-basal ganglia loops (Alexander, DeLong, & Strick, 1986; Haber, 2011; Collins & Frank, 2014), with striatum playing a crucial role in supporting iterative, reward-dependent learning (e.g., McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Frank et al., 2004; Frank & O'Reilly, 2006). Prefrontal cortex activity also reflects reward prediction errors in feedback-based learning tasks (e.g., Barto, 1995; Schultz, Dayan, & Montague, 1997; Shohamy et al., 2004; Daw et al., 2011), but is typically thought to be more related to flexible goal-directed behavior (e.g., Hampton, Bossaerts, & O'Doherty, 2006; Valentin, Dickinson, & O'Doherty, 2007). Specifically, there has been evidence that PFC function supports WM in the context of learning, in parallel to subcortical RL (Collins & Frank, 2012; Collins, Ciullo, Frank, & Badre, 2017). While there is a growing understanding of the multiple neural mechanisms that support learning, and in particular the RL circuits in the brain, the inputs to this network are not often carefully considered - RL computations assume known stimuli, actions, and rewards as inputs to learn a policy (Rmus, McDougle, & Collins, 2021). Here, our work shows that the inputs, in particular the state space, matter: the nature of the stimuli impacted RL computations, slowing learning and potentially increasing choice confusion. It would be interesting in future research to do network-level modeling to understand how this behavior may arise from more diffuse/overlapping input representations.

Neuroscientific research in RL contrasts with that of WM, which has spent a considerable amount of effort investigating how stimulus information affects WM representations in the brain. Namely, neuroscientific research has demonstrated that WM in the brain is highly distributed, and that the brain areas involved vary depending on the type of information being maintained (for review, see Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017). For example, in addition to the prefrontal cortex, retinotopic maps in occipital and parietal cortices are related to the WM maintenance of visual information (e.g., Harrison & Tong, 2009; Riggall & Postle, 2012). However, despite neural WM representations being represented through sensory cortices, WM still behaves similarly in the context of learning and decision making, where the conjunction of stimuli and correct choices is the most important information to be maintained. Perhaps this associative, higher-level information is successfully represented in the PFC, regardless of specific stimulus information. Future research with brain imaging could shed more light on this.

There are, of course, limitations to our results. First, while our model fits are reasonable, there are still some qualitative deviations in our model validation and the data we collected. In particular, learning performance in the Variants condition in set size 6 was lower than the RL learning rate model predictions. Perhaps learning detriments in the Variants condition is a combination of other, unconsidered processes interacting with either RL or WM. There has been ample research that computationally, behaviorally, and neurologically demonstrate that other processes interact with RL and/or WM. For example, episodic memory interacts with memoranda maintained in WM (e.g., Hoskin, Bornstein, Norman, & Cohen, 2019)and choice in RL tasks (e.g., Bornstein & Norman, 2017). Attention also affects both WM (e.g., Chun, Golomb, & Turk-Browne, 2011; Souza, Thalmann, & Oberauer, 2018) and RL (e.g., Farashahi et al., 2017; Leong et al., 2017; Niv et al., 2015). While it would be lovely to be able to study all these processes in tandem, it is simply out of the scope of this project; the design of our experiment would likely not allow different processes to be distinguished behaviorally or computationally.

Second, and more critically, we were not able to conclusively distinguish whether it was lower learning rate or increased across-stimulus confusion during the RL response policy calculation. Perhaps the experimental design is too simple to distinguish the choice noise that occur from both cases. However, these "RL learning rate" and "RL decision confusion" models are distinguishable according to model recovery (Supplementary 6.5), so it is not simply that they make similar predictions. Additionally, these results do not suggest just a simple increase in noise, since other models that also result in increased behavioral noise (i.e., RL credit assignment, WM decay, and WM decision confusion models) do not fit the data quantitatively or qualitatively as well. Thus, our results do strongly suggest an impact on *specifically* the RL process. Understanding the exact nature of that impact will require additional study, likely with different paradigms.

Our two experiments were conducted in fairly different demographics and experimental environments: Experiment 1 was conducted online on MTurk and Experiment 2 was conducted in person in an undergraduate population. Despite subtle differences in behavior across the two experiments (namely, the difference in statistical significance of condition differences in set size 3 blocks), we find remarkable consistency in behavior, model rankings, qualitative goodness of fits of winning models, and estimated parameters across experiments. Thus, we see the two experiments as a broad replication of results as a sign of robustness of the findings.

Overall, this study replicates results demonstrating the importance of both RL and WM in the study of learning. This study provides evidence that stimulus matters in learning, potentially pointing to the importance of semantic information in learning. We find an interesting result that condition differences only affected the RL process, while the WM process was largely spared. This paper strongly demonstrates the importance of considering how a learning state is defined. Future research should continue to investigate how different stimuli/states affect learning and, at the very least, consider how the experimental choice of stimuli affects learning behavior.

Data and code availability. Participant and simulated data are available at https://osf.io/f4hst/. Plotting and analysis code are available at https://github.com/aspenyoo/RLWM_stim_discrim. None of the experiments were preregistered.

References

Akaike, H. (1972). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381. doi: 10.1146/annurev.ne.09.030186.002041

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. doi: 10.1016/S0079-7421(08)60452-1

Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. In J.C. Houk & D. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(10), 3632–3645. doi: 10.1523/JNEUROSCI.3204-13.2014

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. doi: 10.1038/nn.4573

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464. doi: 10.1073/pnas.1520027113

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. doi: 10.1016/j.tics.2016.12.007

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101. doi: 10.1146/annurev.psych.093008.100427

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10). doi: 10.1162/jocn$_{a0}$1238

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756. doi: 10.1523/JNEUROSCI.0989-14.2014

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working Memory Load Strengthens Reward Prediction Errors. *Journal of Neuroscience*, *37*(16), 4332–4342. doi: 10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. doi: 10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(3), 337–366. doi: 10.1037/a0037015

Conrad, R. (1964). Acoustic Confusions in Immediate Memory. *British Journal of Psychology*, *55*(1), 75–84. doi: 10.1111/j.2044-8295.1964.tb00899.x

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. doi: 10.3758/s13423-016-1191-6

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *eLife*, *11*, e75474. Retrieved from `https://doi.org/10.7554/eLife.75474` doi: 10.7554/eLife.75474

Emrich, S. M., Lockhart, H. A., & Al-Aidroos, N. (2017). Attention mediates the flexible allocation of visual working memory resources. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(7), 1454–1465. doi: 10.1037/xhp0000398

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*(1), 1768. doi: 10.1038/s41467-017-01874-w

Farashahi, S., Xu, J., Wu, S.-W., & Soltani, A. (2020). Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. *Cognition*, *205*, 104425. doi: 10.1016/j.cognition.2020.104425

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. doi: 10.1073/pnas.0706111104

Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral Neuroscience*, *120*(3), 497–517. doi: 10.1037/0735-7044.120.3.497

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943. doi: 10.1126/science.1102941

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*(5), 1320–1327. doi: 10.3758/s13423-014-0790-3

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154. doi: 10.1080/03640210701802071

Haber, S. N. (2011). Neural Circuits of Reward and Decision Making: Integrative Networks across Corticobasal Ganglia Loops. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural Basis of Motivational and Cognitive Control.* MIT Press.

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. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(32), 8360–8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. doi: 10.1038/nature07832

Hoskin, A. N., Bornstein, A. M., Norman, K. A., & Cohen, J. D. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 338–354. doi: 10.3758/s13415-018-00674-z

Hurvich, M. C., & Tsai, L. C. (1987). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Jafarpour, A., Buffalo, E. A., Knight, R. T., & Collins, A. G. E. (2022). Event segmentation reveals working memory forgetting rate. *iScience*, 103902. doi: 10.1016/j.isci.2022.103902

Klyszejko, Z., Rahmati, M., & Curtis, C. E. (2014). Attentional priority determines working memory precision. *Vision Research*, *105*, 70–76. doi: 10.1016/j.visres.2014.09.002

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. doi: 10.1016/j.neuron.2016.12.040

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. doi: http://dx.doi.org.libproxy.berkeley.edu/10.1037/0033-295X.111.2.309

Luck, J. S., & Vogel, K. E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346. doi: 10.1016/s0896-6273(03)00154-5

McDougle, S. D., & Collins, A. G. E. (2020). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39. doi: 10.3758/s13423-020-01774-z

Murdock Jr., B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488. doi: 10.1037/h0045106

Nilsson, H., Rieskamp, J., Wagenmakers, E.-J., & Nilsson, H. (2011). *Hierarchical Bayesian parameter estimation for cumulative prospect theory* (Vol. 55). doi: 10.1016/j.jmp.2010.08.006

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. doi: 10.1523/JNEUROSCI.2978-14.2015

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuro-*

science, *32*(2), 551–562. doi: 10.1523/JNEUROSCI.5498-10.2012

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi: http://dx.doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. doi: 10.1037/0033-295x.101.1.53.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. doi: 10.1016/s0896-6273(03)00169-7

Oemisch, M., Westendorff, S., Azimi, M., Hassani, S. A., Ardid, S., Tiesinga, P., & Womelsdorf, T. (2019). Feature-specific prediction errors and surprise across macaque fronto-striatal circuits. *Nature Communications*, *10*(1), 176. doi: 10.1038/s41467-018-08184-9

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. doi: 10.1016/j.tics.2017.03.011

Petrides, M. (1985). Deficits on conditional associative-learning tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *23*(5), 601–614. doi: 10.1016/0028-3932(85)90062-4

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

Radulescu, A., Vong, W. K., & Gureckis, T. M. (2022). Name that state: How language affects human reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 4).

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology. General*, *147*(11), 1553–1570. doi: 10.1037/xge0000466

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. doi: 10.1523/JNEUROSCI.1892-12.2012

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *NeuroImage*, *84*, 971–985. doi: 10.1016/j.neuroimage.2013.08.065

Rmus, M., & Collins, A. G. E. (2020). What is a Choice in Reinforcement Learning? In *Proceedings of the The Annual Meeting of the Cognitive Science Society.*

Rmus, M., McDougle, S. D., & Collins, A. G. E. (2021). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73. doi: 10.1016/j.cobeha.2020.10.003

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*, 1593–1599. doi: 10.1126/science.275.5306.1593

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. doi: http://dx.doi.org.libproxy.berkeley.edu/10.1037/h0093825

Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico[U+2010]striatal contributions to feedback[U+2010]based learning: converging data from neuroimaging and neuropsychology. *Brain*, *127*(4), 851–859. doi: 10.1093/brain/awh100

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, *15*(3), 2. doi: 10.1167/15.3.2

Souza, A. S., Thalmann, M., & Oberauer, K. (2018). The precision of spatial selection into the focus of attention in working memory. *Psychonomic Bulletin & Review*, *25*(6), 2281–2288. doi: 10.3758/s13423-018-1471-4

Starr, A., Srinivasan, M., & Bunge, S. A. (2020). Semantic knowledge influences visual working memory in adults and children. *PLOS ONE*, *15*(11), e0241110. doi: 10.1371/journal.pone.0241110

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian Model Selection for Group Studies. *NeuroImage*, *46*(4), 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *18*(6), 643–662. doi: 10.1037/h0054651

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific Reports*, *11*(1), 3574. doi: 10.1038/s41598-020-80593-7

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction.* Cambridge, Mass: MIT Press.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(15), 4019–4026. doi: 10.1523/JNEUROSCI.0564-07.2007

Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, *9*. doi: 10.3389/fnbeh.2015.00225

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. doi: 10.1167/4.12.11

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. doi: 10.7554/eLife.49547

Wilson, R. C., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, *5*, 189. doi: 10.3389/fnhum.2011.00189

Wunderlich, K., Beierholm, U. R., Bossaerts, P., & O'Doherty, J. P. (2011). The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of Neurophysiology*, *106*(3), 1558–1569. doi: 10.1152/jn.01051.2010

Xia, L., Master, S. L., Eckstein, M. K., Baribault, B., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2021). Modeling changes in probabilistic reinforcement learning during adolescence. *PLOS Computational Biology*, *17*(7), e1008524. doi: 10.1371/journal.pcbi.1008524

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, 1–18. doi: 10.1162/jocn$_{a0}$1808

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports*, *8*, 16162. doi: 10.1038/s41598-018-34282-1

6 Appendix

In the Supplementary Materials, we include additional analyses that broadly support the main text. We include details on participant reaction times on the Learning phase, N-back distractor task, qualitative differences in error types between the two winning models, parameter recovery, model recovery, and alternative models that were tested. In the alternative models, we included analyses of RL, WM, and RLWM models; whether model goodness-of-fit changes with a fixed or fitted perseveration rate and negative learning rate; and whether perseveration choice trace is greater than one trial back.

6.1 Reaction times

Plotted below are the individual subject (dots) and group mean (bars) reaction times in seconds, split by stimulus condition and set size.

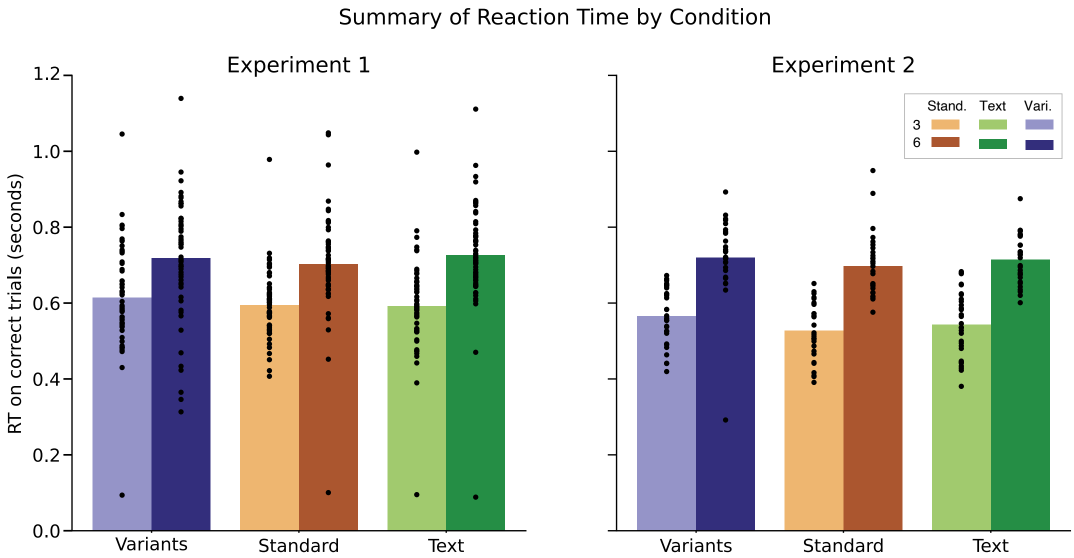

Figure 6: **Subject Reaction Times by Experiment.** Mean (bar) and individual participant (dots) reaction times for each condition, for the learning phase of Experiment 1 (left) and Experiment 2 (right). Reaction times were not used as a means of exclusion for either experiment.

6.2 N-back distractor task

The first block was a practice block with N=2, then the following four blocks incrementally increased from N=2 to N=5. Each block had on average 40 trials, and the stimulus shown on each trial was a colored rectangle; potential rectangle colors were common and distinct from one another (e.g., blue, yellow, pink, black, green). Code for the N-back task can be found at https://github.com/AlexanderFengler/ExperimentDesign_NBackTask.

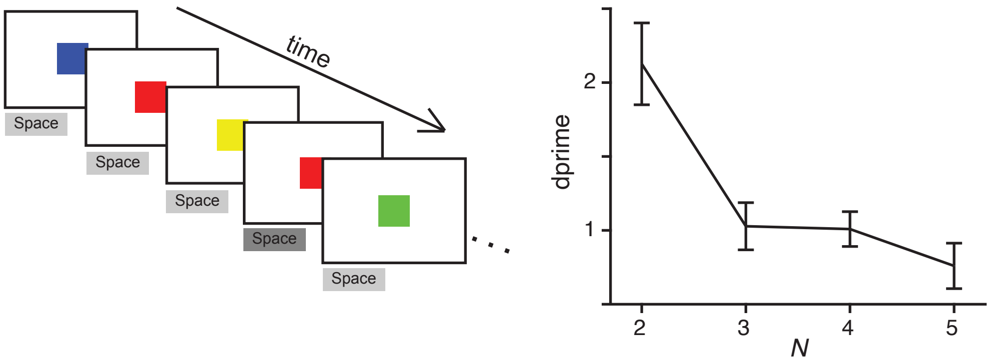

Figure 7: N-back task. *Left*: task design. Participants viewed a series of colors and made a key press every time the color N trials ago was the same as the color of the current block. This illustration demonstrates all correct responses on a $N = 2$ back task. *Right*: d' decreases a function of N, indicating worse performance with increasing set size.

6.3 Qualitative difference between models: error types

We found that the models that assumed that either there was a condition-specific effect on RL learning rate or a condition-specific effect on RL decision confusion were able to fit the data best. While the goal of our paper is not to find one model that explains all datasets we collected, it is still an interesting question to ask what the differences are between participants best fit by each of the models. In this section, we highlight one qualitative difference between the two winning models.

To investigate qualitative differences between models, we analyzed the key press errors. Unlike learning curves, the two models *should* generate different predictions on error types. For the RL learning rate model, errors are primarily driven by a lower rate of learning, so errors should be randomly distributed across incorrect keys. On the other hand, if people are confusing stimuli at the decision stage, errors should not be random. Specifically, the RL decision confusion model should predict that errors would be skewed toward the key presses that are rewarded in other stimuli.

For all set size 3 blocks, there was an imposed structure such that there was a key for which two images were correct, a key for which one image was correct, and a key for which no images were correct. (The correct keys were counterbalanced across blocks.) Because the correct answers were not evenly distributed across key presses, we were able to investigate if errors are random or reflect the distribution of correct keys across all trials (i.e., independent of current stimulus). We cannot do this analysis on set size 6 blocks, since each key had 2 images each associated with it.

For each participant, we split up errors by whether the correct answer was the key that was correct for two stimuli (which we will refer as the "2" key) or if the correct answer was correct for only one stimulus (the "1" key). We then calculated the proportion of the incorrect key presses that were correct for a different stimulus (incorrectly pushing the "1" or "2" key), versus a key that was never rewarded (the "0" key). If errors are random, as predicted by the RL learning rate model, this proportion would be around 0.5. If errors result from decision confusion, participants' error should be biased toward stimuli rewarded in other trials. However, there are other reasons that decisions would be biased toward stimuli rewarded in other trials (e.g., a general avoidance of never-rewarded key). If errors are truly a result of decision confusion, there should be higher confusion in trials in which 1 is correct but 2 is pushed, than trials in which 2 is correct but 1 is pushed.

For visualization, we grouped the participants by whether they were better fit by the RL learning rate or RL decision confusion model (i.e., which model had a higher LL^*). In Experiment 1, 35 participants were best fit by the RL learning rate model, and 24 best fit by the RL decision confusion model. In Experiment 2, 19 participants were best fit by the RL learning rate model, and 11 best fit by the RL decision confusion model. Proportion of error types for both Learning and Test phase are illustrated in Figure 6.

For both phases, we conducted a two-way ANOVA for each group of participants, to investigate whether the error types were different according to condition (Standard, Text, Variants), correct key (2 or 1), and interaction between the condition and correct key. For the RL learning rate group, in both Learning and Test phase, we found no significant main effect of condition, correct key press, and no significant interaction. Preference for key rewarded in other trials in Learning ($t(53) = 7.30, p < .001. M = .60, SEM = .01$) and Test ($M = .64, SEM = .02, t(18) = 6.59, p < .001$) phase was significantly different than chance.

For participants best fit by the RL decision confusion model, there was a significant main effect of correct key press in both Learning ($F(1, 34) = 25.01 p < .001$) and Test phase ($F(1, 34) = 15.05, p < .001$). There was no main effect of condition or interaction between condition and correct key press. In the Learning phase, there was a greater bias toward other rewarded keys in trials when the correct answer was 1 ($M = .74, SEM = .03$) than 2 ($M = .60, SEM = .01$), and both were significantly different than chance ($t(34) > 7.11, p < .001$). In the Test phase, both were significant prefer rewarded keys in other trials, but greater bias toward rewarded keys when correct answer was 1 ($M = .78, SEM = .04, t(10) = 7.44, p < .001$) than 2 ($M = .56, SEM = .02, t(10) = 3.03, p = .01$).

Model predictions do not successfully capture qualitative data patterns. Neither of the models are able to capture the avoidance of the unrewarded key in both phases, suggesting there is another process at work we did not include in the model. The RL decision confusion model is able to capture the qualitative effect of greater bias in "1" trials over "2" trials in Learning phase, but not in Test phase. Perhaps the RL decision confusion is able to capture greater bias in early learning, but stimulus confusion is lessened by late learning Q-values (which the test phase is based on).

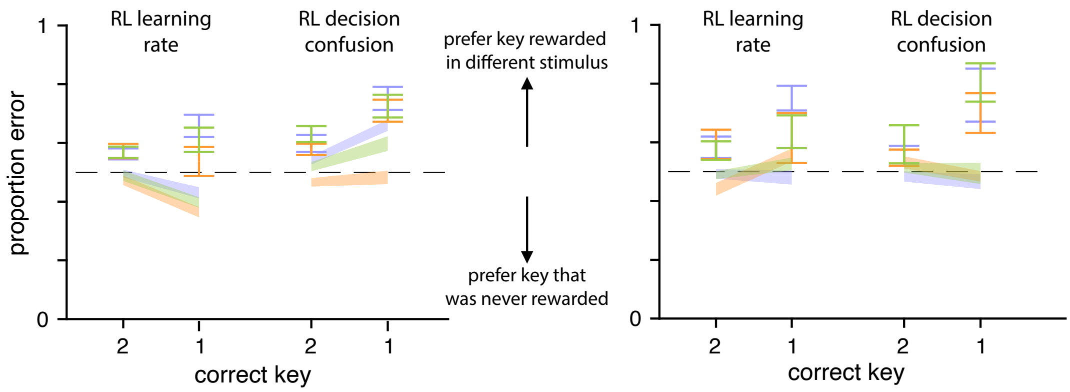

Figure 8: **Error types by winning model.** The proportion of incorrect key presses that were rewarded for other stimuli, based on how many stimuli shared the same key press (x-axis). Randomly responding between the two incorrect keys is shown with the dashed black line; above chance means a preference toward the key rewarded for a different stimulus. $M \pm SEM$ data (error bars) and model predictions (fills) for Learning (left; both experiments) and Test (right; Exp 2) phase.

6.4 Parameter recovery

In order to establish the interpretability of model parameters, one should test that the same parameters that generate a data set are the ones recovered through the model parameter estimation method (Wilson & Collins, 2019). Successful parameter recovery exists when the parameter values that maximize the likelihood of the data given the model parameters are close to the parameter values that generated the data. Successful parameter recovery is necessary to interpret estimated parameter values.

For each model, we generated parameters by sampling the fitted parameter vectors from participants across both experiments. We sampled 50 participants without replacement. Our goal here was to use parameter values that best reflect the regime of the parameter space that matches data we are interested in. We also completed parameter recovery by sampling parameters from a nonparametric distribution informed by the fitted parameter values, rather than using the exact values. Because there are arbitrary decisions required to define this distribution, we did not include the results here. However, the results are qualitatively the same.

For each model and simulated participant, we simulated data with the sampled parameters, then estimated parameters using the same model fitting methods described in the main text. Finally, we plot the true and estimated parameters against one another. For each plot, values clustered along the diagonal indicate successful parameter recovery.

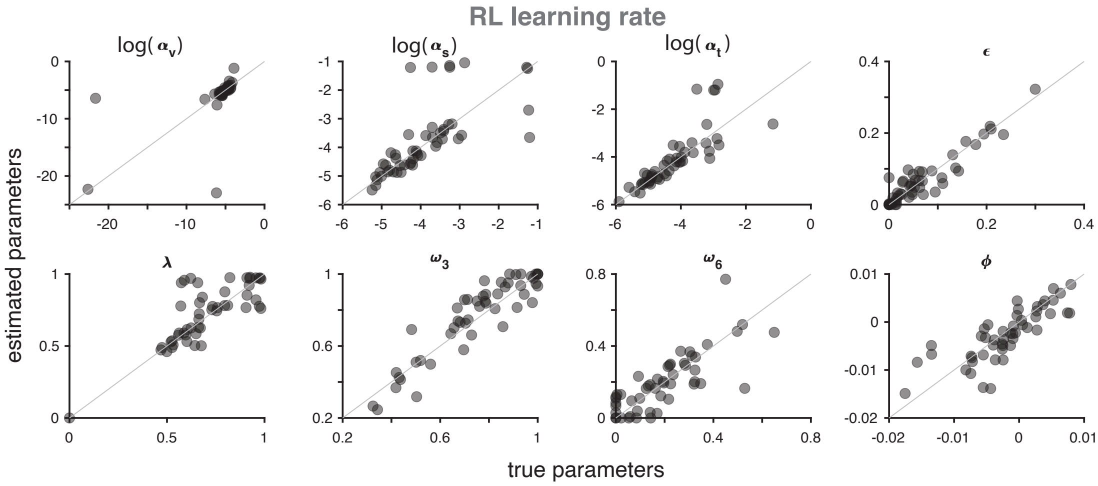

Figure 9: Parameter recovery plots for condition-specific RL learning rate model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

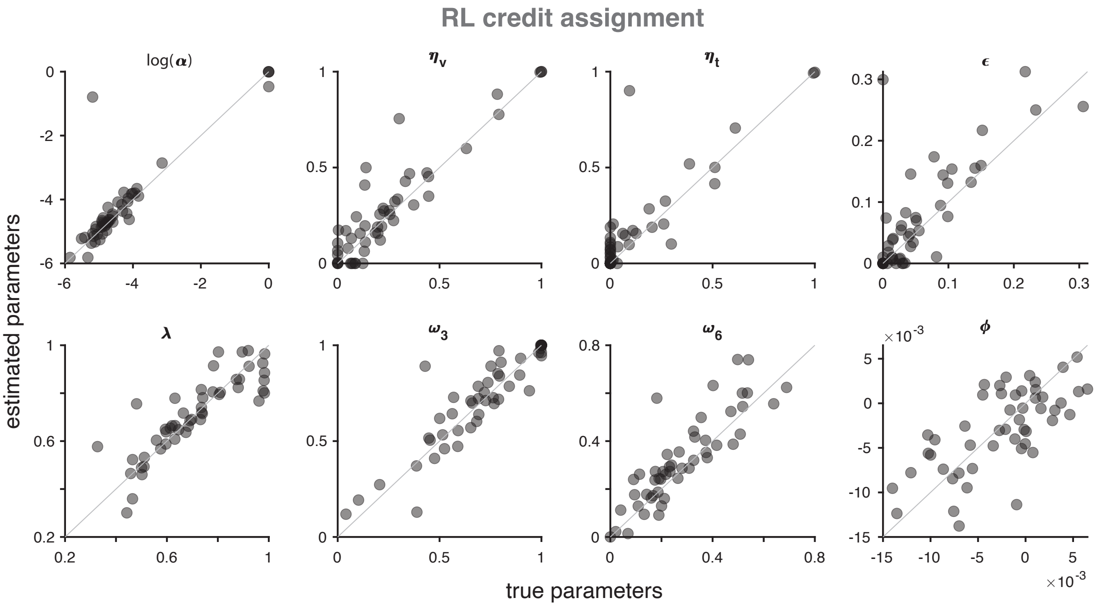

Figure 10: Parameter recovery plots for condition-specific RL credit assignment model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

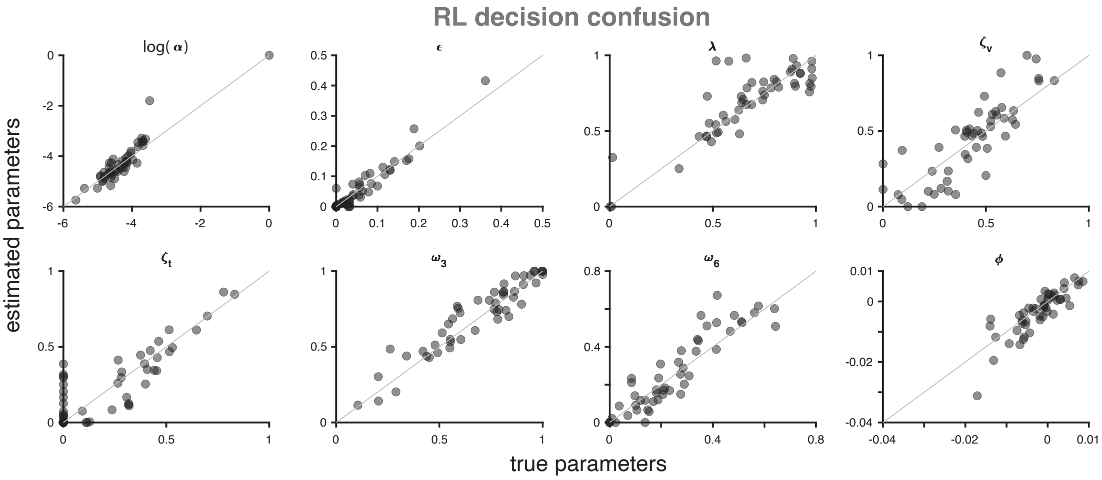

Figure 11: Parameter recovery plots for condition-specific RL decision confusion model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

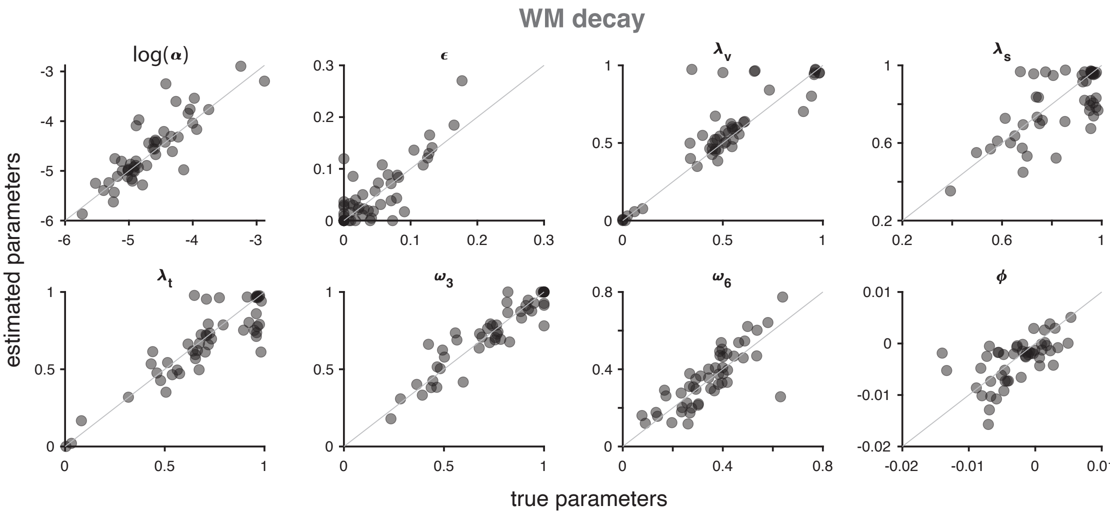

Figure 12: Parameter recovery plots for condition-specific WM decay model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

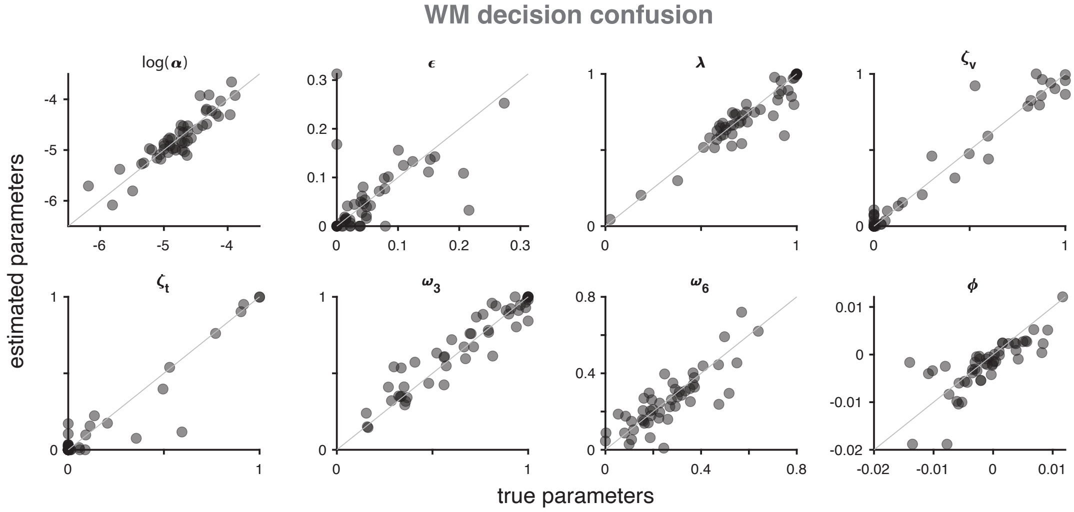

Figure 13: Parameter recovery plots for condition-specific WM decision confusion model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

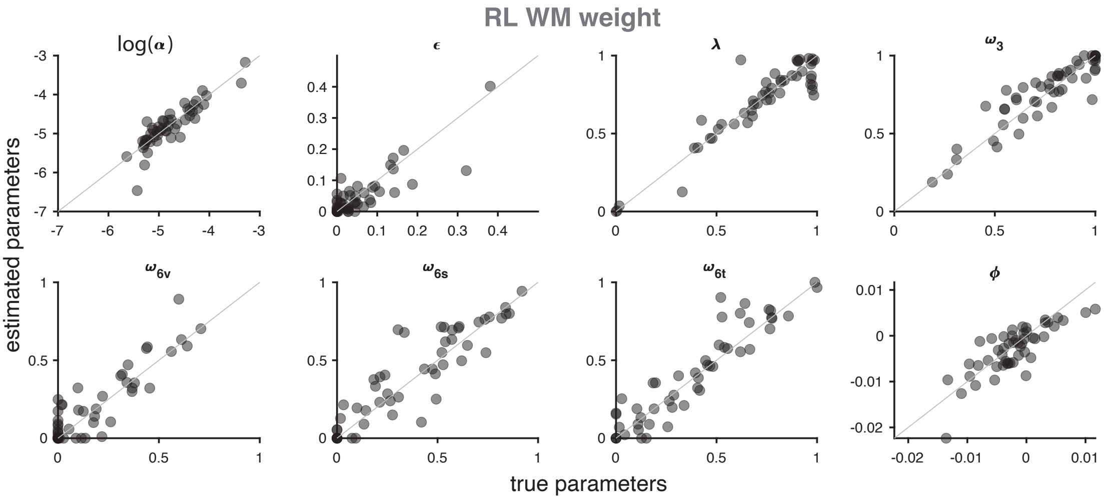

Figure 14: Parameter recovery plots for condition-specific RL WM weight model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

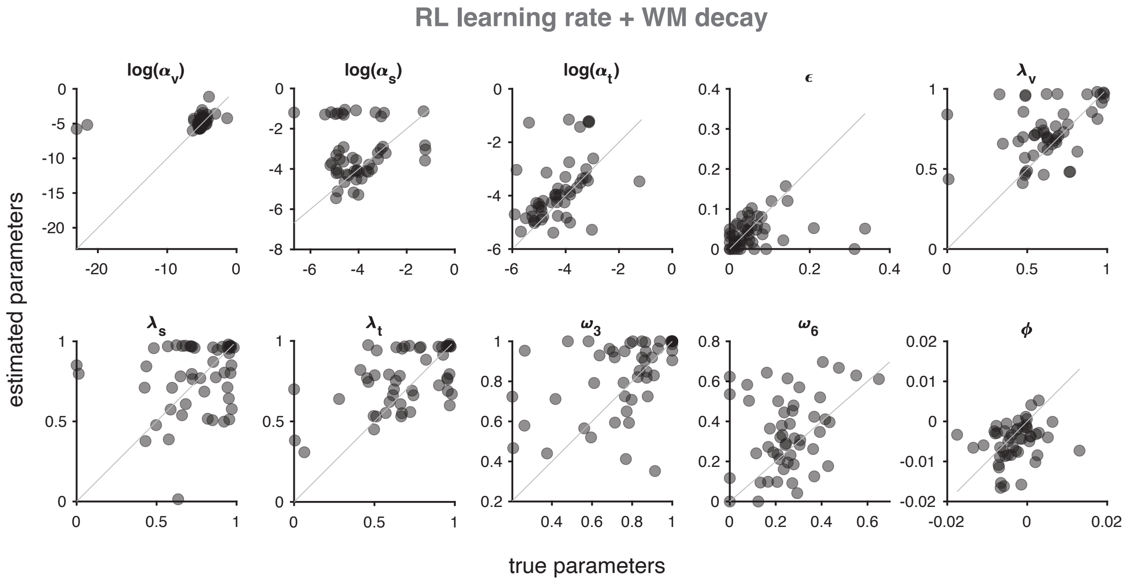

Figure 15: Parameter recovery plots for condition-specific RL learning rate + WM decay model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

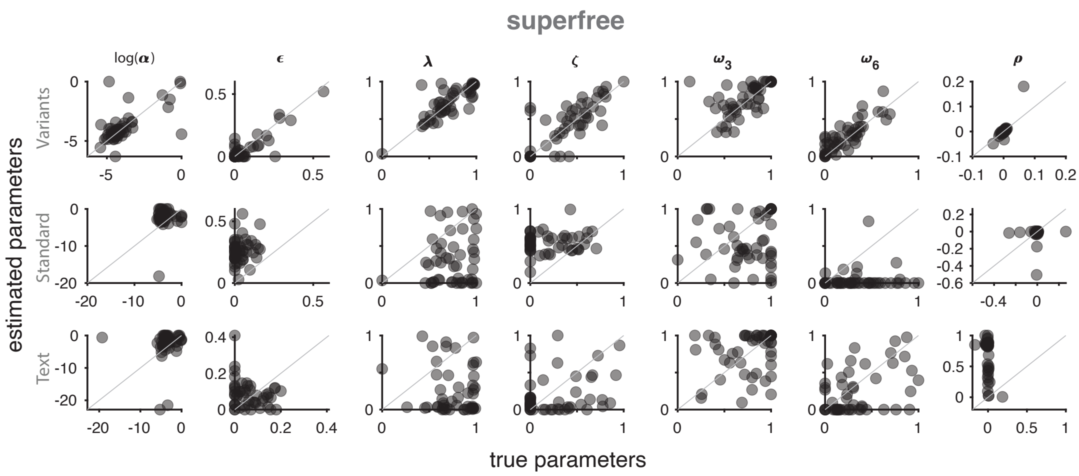

Figure 16: Parameter recovery plots for superfree model. Each subplot plots the true parameters (x-axis), which generated data, against the recovered parameter values (y-axis), estimated using MLE. Dots are individual simulated participants.

6.5 Model recovery

Model recovery is an important step before making conclusions from a quantitative model comparison (Wilson & Collins, 2019). Successful model recovery occurs when the same model that generates a data set best fits it (according to your chosen model comparison metrics), when compared to all other models in the comparison set.

For each model, we generated 50 simulated participants' data from the parameter values fitted from 50 participants, randomly sampled without replacement from both experiments. (We use the same simulated participants' data for parameter recovery). We then fit every model to each of of these (nModels x 50) simulated participants, using the same fitting methods as described in the main text.

We compared model goodness-of-fit using corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), and $\exp_r$. AICc and BIC both penalize models with more parameters, and BIC penalizes more strictly:

$$\text{AICc} = -2LL^* + 2k + \frac{2k(k+1)}{N_{\text{trials}} - k - 1}$$
$$\text{BIC} = -2LL^* + k \log N_{\text{trials}},$$

where k is the number of parameter and N_{trials} is the number of trials.

The measure exp_r is calculated using BMS spm, which explicitly assumes that the participants can be fit by different models. This value is expectation of the posterior probabilities of each model.

Successful model recovery occurs when the model that best fits a simulated data set is the same model that generated that data set. For example, if all 50 participants generated by the condition-specific RL learning rate model are best fit by the condition-specific RL learning rate model, there is successful model recovery.

For the most part, we consider our results successful model recovery (Figure 17). However, these results also indicate the RL learning rate, WM decay, and RL WM weight models are a bit more flexible than others, demonstrated by their ability to best capture data sets generated from other models. These results suggest that model comparisons favoring these three models may be do to model flexibility, rather than a genuine reflection of the underlying cognitive process. In our experimental data (see main manuscript), we do indeed find that the RL learning rate model fits the data best. However, because 1) we do not find that WM decay or RL WM weight models fit the data as well, and 2) the RL decision confusion model is able to fit the data comparably well to the RL learning rate model, we believe our interpretation of the results (i.e., that RL is specifically affected, but not committing to how) is still valid.

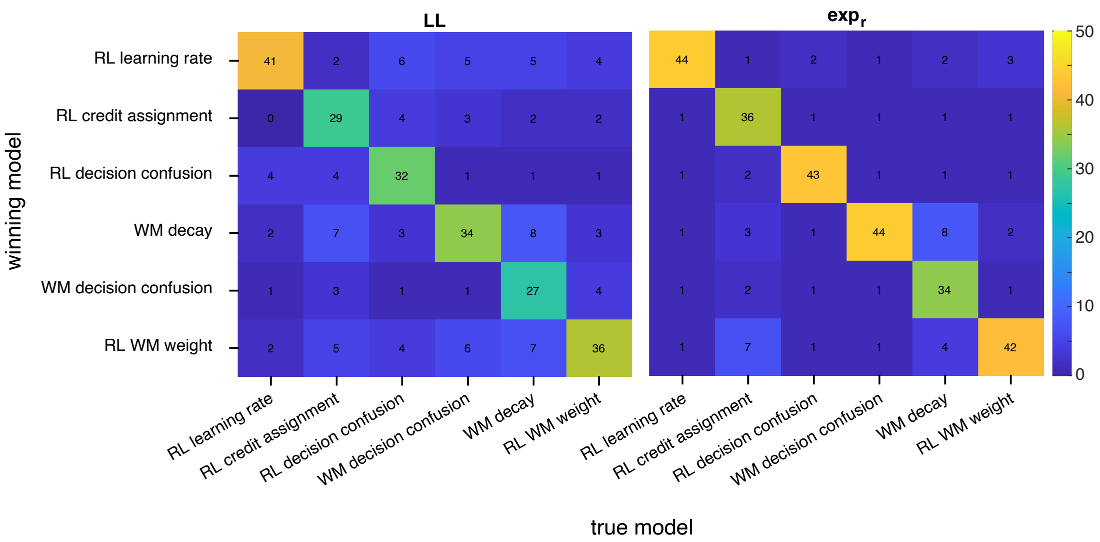

Figure 17: Model recovery when using LL^* and expected probability using BMS SPM (exp$_r$), for six main models with same number of parameters. Successful model recovery is indicated by a majority of models falling on the diagonal. Both metrics provide good model recovery, although exp$_r$ is a bit better.

Our model comparison including the additional two models (RL learning rate + WM decay, superfree) are not as simple, due to the relatively high confuseability of the RL learning rate model and the RL learning rate + WM decay model (Figure 18). We did an additional model recovery analysis between just these two models, with 500 simulated datasets, 50 parameter sets each simulated 10 times (Figure 19). Although the majority tends in the desired direction, the simpler RL learning rate model is able to account for much of the more complex RL learning rate + WM decay model. Thus, our model comparison results between these two models should be taken with a grain of salt.

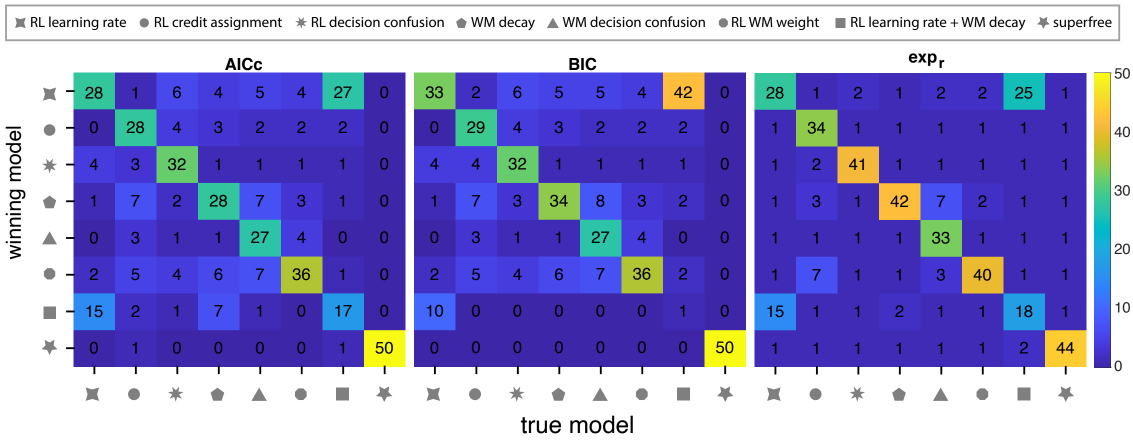

Figure 18: Model recovery when using AICc, BIC, and expected probability using BMS SPM (exp_r). Successful model recovery is indicated by a majority of models falling on the diagonal. These results generally convey reasonable model recovery, for all models except the RL learning rate + WM decay model. AICc and exp_r provide better recovery than BIC.

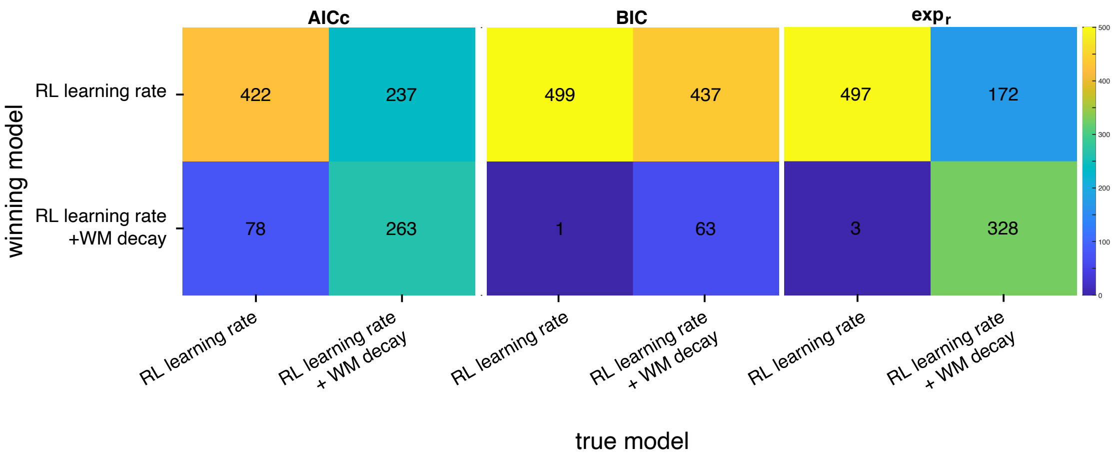

Figure 19: A follow up model recovery with more simulated data (independent from earlier datasets), with just the "RL learning rate" and "RL learning rate + WM decay" models, which had the greatest confusability in earlier model recovery plots. No metric is able to capture a desired level of model recovery, although AICc and exp_r are able to capture the correct directionality.

6.6 Parameter values

In this section, we plot the individual and group parameter values for the two winning models: the condition-specific RL learning rate model (Figure 20) and condition-specific RL decision confusion model (Figure 21).

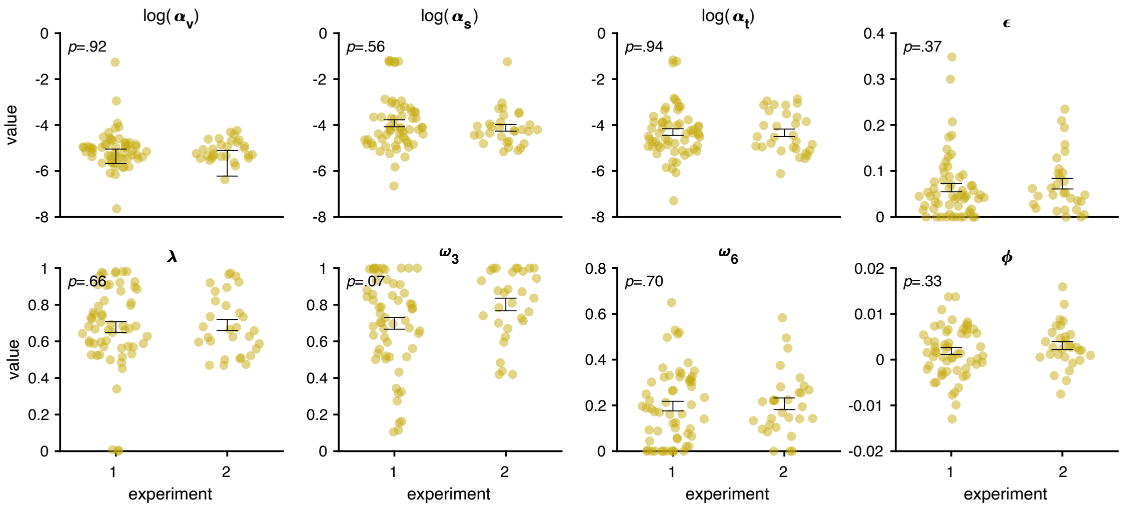

Figure 20: Parameter values (dots: individual participants. error bars: $M \pm sem$ across participants) for the condition-specific RL learning rate model for Experiment 1 and Experiment 2. Outliers for $\log(\alpha_v)$ not illustrated in plot (Exp 1: -21.66; Exp 2: 22.63). The p-values of a Wilcoxon rank sum test comparing the two participant groups, *before* any multiple comparisons corrections, displayed on the top left of each subplot.

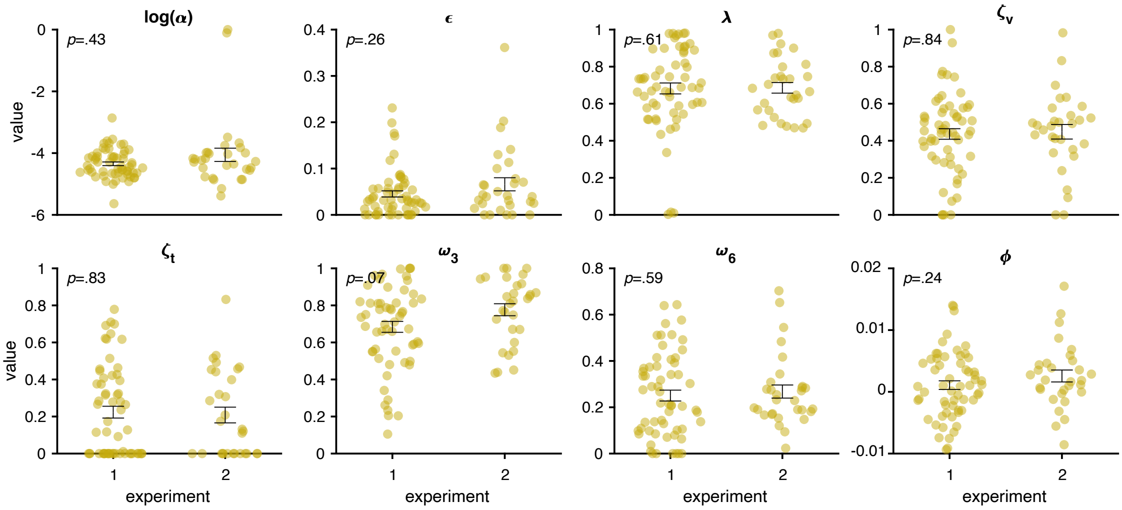

Figure 21: Parameter values (dots: individual participants. error bars: $M \pm sem$ across participants) for condition-specific RL decision confusion model for Experiment 1 and Experiment 2. The p-values of a Wilcoxon rank sum test comparing the two participant groups, *before* any multiple comparisons corrections, displayed on top left of each subplot.

6.7 Alternative Models

We tested six main models in the manuscript with the following condition-specific differences: RL learning rate, RL credit assignment, RL decision confusion, WM decay, WM decision confusion, and weight between RL and WM process contributions. There are of course an infinite amount of other models that we could have tested. This section summarizes related models that we fitted, that may be of interested to the reader. We divide this section into three parts. First, we display the results of models with only an RL component, only a WM component, and standard RLWM models without condition-dependencies. These models are common to report in similar studies, but were not reported in our main manuscript because they are obviously poorly fitting models. Second, we use factorial model comparison to test whether the goodness of fit for the eight main models we fit in the main manuscript vary with/without perseveration, and with/without a fitted negative learning rate, α_-, parameter. There are published studies suggesting the assumptions we included in the main manuscript were reasonable, but we still chose to test them directly. Third, we test if our assumption of 1-back perseveration (i.e., the time decay of perseveration) affects our modeling results, by softening this assumption. Fourth, we show model validation plots for the additional models considered in the main manuscript: the RL learning rate + WM decay model and the Superfree model. Finally, we show model validation plots for the additional models considered in Experiment 2: the RL learning rate and RL decision confusion models with condition-specific interference of WM on RL during learning.

In these sections, we compared model goodness-of-fit using AICc and BIC.

6.7.1 RL, WM, RLWM model fits

Three models that are often shown in "RLWM" papers are RL alone, WM alone, and RL+WM models. We decided not to show their fits in the main manuscript, because they explicitly do not include any condition-specific differences, and would thus obviously not fit the data well. However, for the sake of completeness and comparison, we include the model validation and model comparison plots of Experiment 1 participants, relative to the condition-specific RL learning rate model used in the main manuscript. Indeed, they are not able to capture the data (Figure **??**).

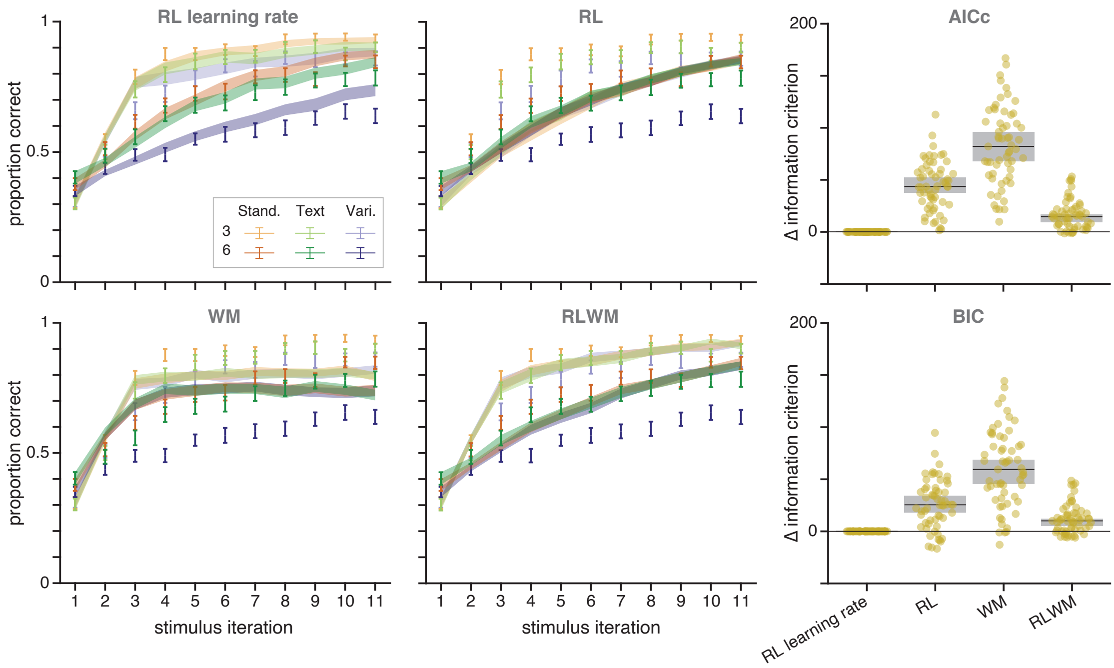

Figure 22: Model validation plots for the condition-specific RL learning rate, RL, WM, and RLWM models (left four plots) for Experiment 1 data. AICc (top) and BIC (bottom) differences between models and RL learning rate model. A smaller number indicates a better fit. The condition-specific RL learning rate clearly fit the data qualitatively and quantitatively better than these models.

6.7.2 Perseveration and negative learning rate

In our main six models, we fit a perseveration rate ϕ, and we fix negative learning rate α_- to 0. Here, we factorially compare model family (6: RL learning rate, RL credit assignment, RL decision confusion, WM decay, WM decision confusion, and RL-WM weight), perseveration (2: fixed to 0, fit as free parameter), and negative learning rate (2: fixed to 0, fit as free parameter).

Figure 23 illustrates the quantitiative comparison of all models for both AICc and BIC. We find that fitting a perseveration parameter does seem to increase the model's quantitative fit, while fitting a negative learning rate parameter does not seem to make a difference. (This is because the values are fit to 0). More importantly, we see that the ranking across model family doesn't vary no matter what perseveration / negative learning rate combination we use. In other words, our conclusion that RL learning rate and RL decision confusion models fit data best are not dependent on our specific assumptions about perseveration or negative learning rate. For simplicity, we decided in the main manuscript to include the model which keeps perseveration as a free parameter, and fixed negative learning rate $\alpha_- = 0$.

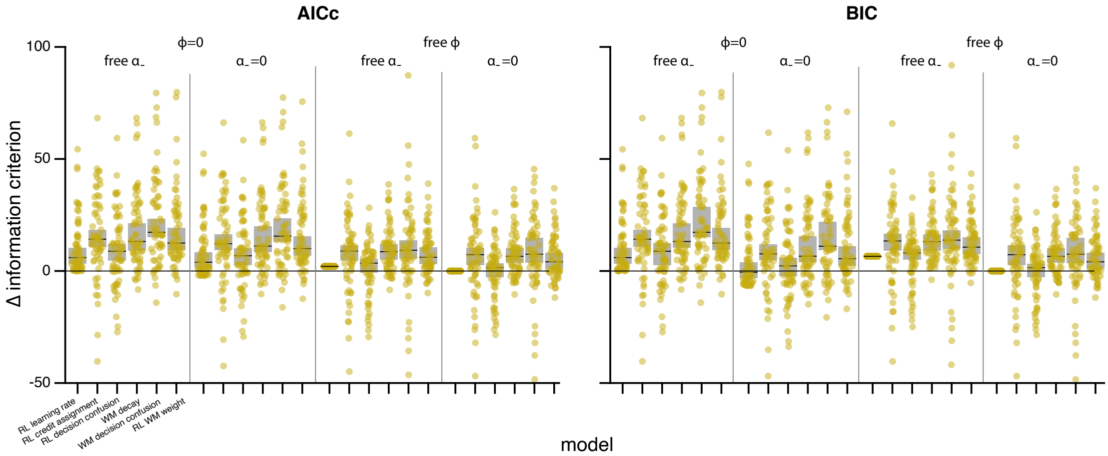

Figure 23: Quantitative results of factorial model comparison. AICc (left) and BIC (right) differences, relative to the RL learning rate model in the main manuscript. A lower number indicates a better fit. For each plot, each section of six models correspond to the respective characteristics: $\phi = 0$, fitted α_-; $\phi = 0$, $\alpha_- = 0$; fitted ϕ and α_-; fitted ϕ, $\alpha_- = 0$

6.7.3 Perseveration with free decay rate parameter

We define perseveration in Section 2.3.1 of the main manuscript, in which we fix the perseveration choice trace decay rate, τ, to 1. Thus, only the previous trial affects the current perseveration behavior. We investigate in this section whether that was a reasonable assumption, by fitting the decay rate τ as a free parameter. Freeing this parameter neither significantly increases model performance of any of our main six models nor changes model ranking.

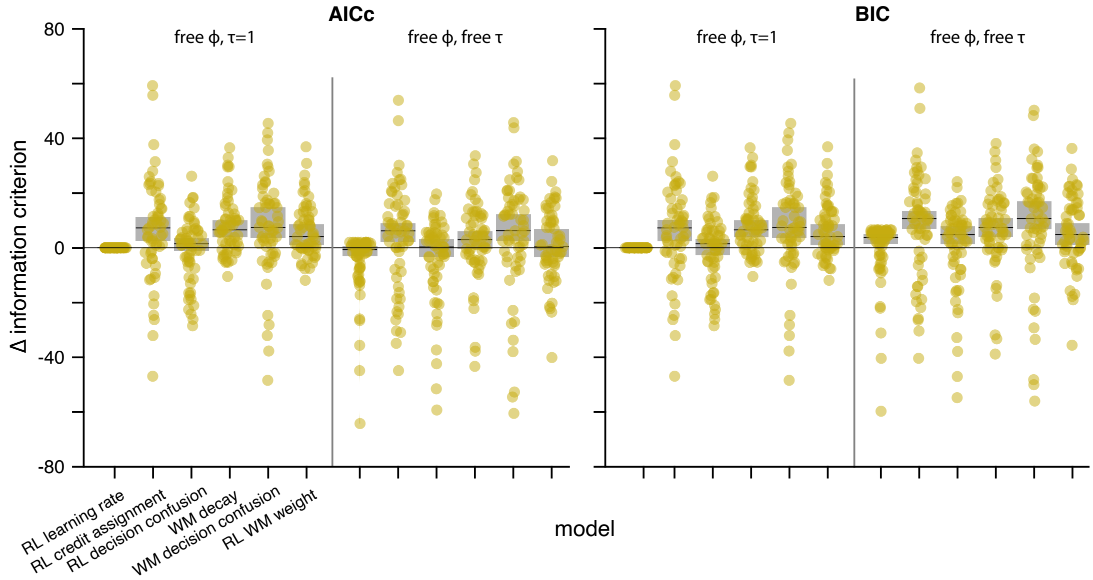

Figure 24: Factorial model comparison with perseveration parameter τ fixed to 1 (left six models on each plot) and as a free parameter (right six models on each plot). AICc (left plot) and BIC (right plot) are relative to the RL learning rate model with $\tau = 1$. A lower value indicates a better fit to data. Model differences do not change model rankings, and model fits are not noticeably improved by including a free τ parameter.

6.7.4 RL learning rate + WM decay model, Superfree model

In this section, we show the model validation and model comparison plots for the two additional models considered in the main manuscript (Section 4.2).

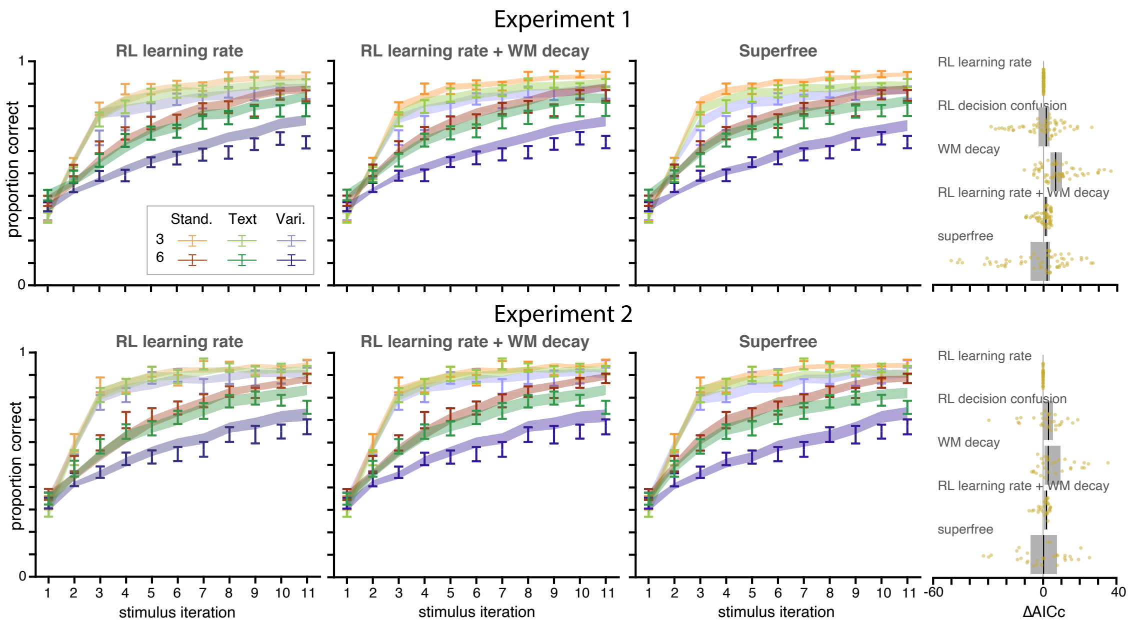

Figure 25: Model validations of RL learning rate + WM decay model and Superfree model for Experiment 1 (top row) and Experiment 2 (bottom row). We plot them next to the model validation of the RL learning rate model, which is our best fitting model. We show quantitiavie model comparison for each participant (yellow dots), with bootstrapped median 90 CI of the median (grey box). All other quantiative model comparison metrics are displayed in tables 2 and 3 in the main text.

6.7.5 Condition-specific interaction for train+test models

In this section, we describe models that were fitted with different degrees of RL/WM interference between train and test in different conditions.

The δ used in updating Q values in interference model includes the WM values, rather than just Q values (Eq: 2). For condition-specific interference, we additionally add a multipllicative term to scale the amount of interference the WM value association gives when calculating delta. We denote the condition-specific interference scalar as x_c for condition c.

$$\delta = r - (\omega_n x_c * WM(s,a) + (1 - \omega_n x_c) * Q(s,a)).$$

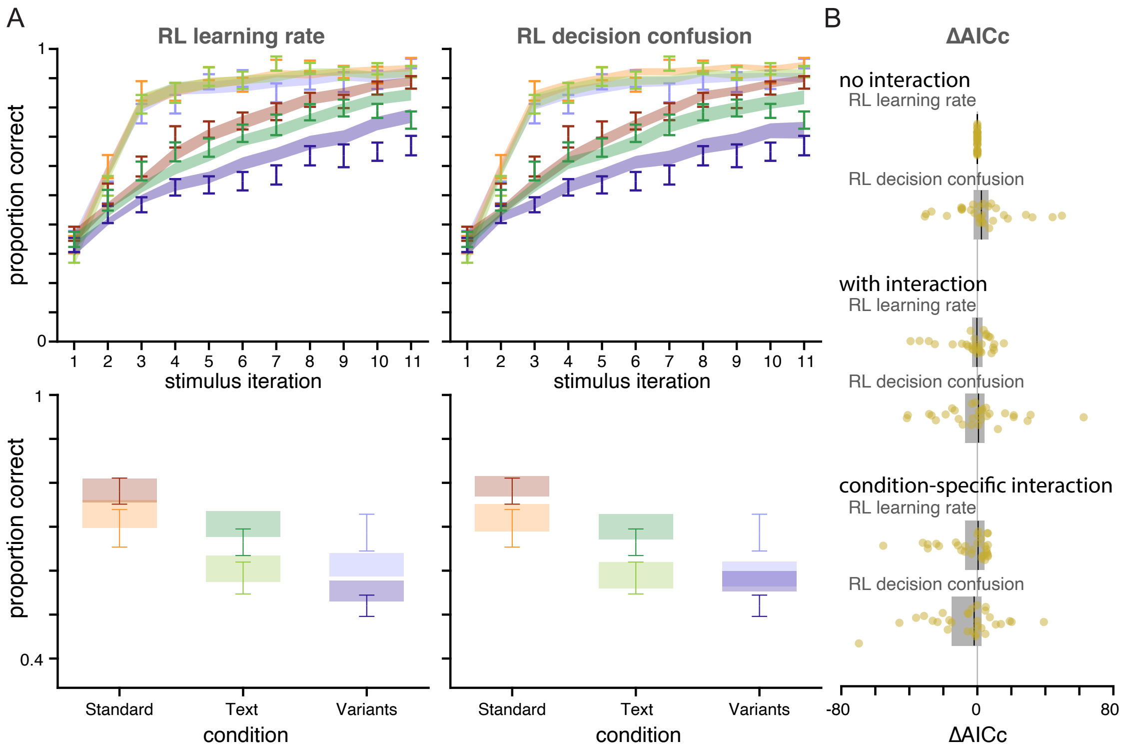

Figure 26: **Model validation and comparison for condition-specific interference models. A**. Model validation for RL learning rate (left plots) and RL decision confustion (right plots) model with condition-specific interference. Top row corresponds to learning phase, bottom row corresponds to test phase behavior (error bars) and model predictions (color fill). **B.** AICc differences of all models fit on learning and test phase data, relative to RL learning rate model with no interference. Negative values indicate better fit. Including condition-specific interference (last two) marginally improves fit, but still does not capture data perfectly.