The importance of linguistic information in human reinforcement learning

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

University of California, Berkeley. Department of Psychology[1]

University of California, Berkeley. Helen Wills Neuroscience Institute [2]

Brown University, Department of Cognitive, Linguistic Psychological Sciences[3]

Abstract

How does the nature of a stimulus affect our ability to learn appropriate response associations? In typical laboratory experiments learning is investigated under somewhat ideal circumstances, where stimuli are easily discriminable visually and linguistically. This is not representative of most real-life learning, where visually or linguistically overlapping "stimuli" can result in different "rewards" (e.g., you may learn over time that you can pet one specific dog that is friendly, but that you should avoid a very similar looking one that isn't). With two experiments, we test how humans learn in three stimulus conditions: stimuli with distinct visual representations but overlapping linguistic representations, stimuli with distinct linguistic representations but overlapping visual representations, and stimuli with distinct visual and linguistic representations. We find that decreasing linguistic and visual distinctness both decrease performance, substantially more for the lowered linguistic distinctness condition. We develop computational models to test different hypotheses about how reinforcement learning (RL) and working memory (WM) processes are affected by different stimulus conditions. Interestingly, we find that only RL, and not WM, is affected by stimulus condition: people learn slower and have higher across-stimulus value confusion at decision when linguistic information overlaps relative to when it is distinct. These results demonstrate strong effects of stimulus type on learning, and highlight the importance of considering the parallel contributions of different cognitive processes when studying behavior.

1 Introduction

Reinforcement learning (RL) broadly refers to the process that characterizes how people learn incrementally through valenced feedback. RL also denotes a set of algorithms describing this process (Sutton & Barto, 1998), that can be remarkably simple and simultaneously flexible in a variety of contexts, including instrumental learning (Eckstein, Wilbrecht, & Collins, 2021). Laboratory RL tasks sometimes employ stimuli that are visually and linguistically easy to discriminate (such as photos of common objects, shapes, and colors; Collins & Frank, 2012; Collins, Brown, Gold, Waltz, & Frank, 2014; Frank et al., 2015; Collins, Albrecht, Waltz, Gold, & Frank, 2017; Collins, 2018). This is in contrast to studies that use stimuli which are difficult to name 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). Some studies explicitly manipulate the amount of information in their stimuli, for example limiting the number of possible 'features' (dimensions) which could be attached to value or that the agent should pay attention to, to investigate properties of the learning algorithms which the agent is employing (for example, feature based vs object based learning). However, most of this prior research is more concerned with the amount of stimulus information available during learning or on how attention is focused on relevant dimensions, rather than the type of information (i.e., linguistic, visual).

Many RL studies actively select non-verbalizable 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 linguistic domain may impact learning, and in particular the balance of RL processes and higher level processes such as inference or memory. However, the hypothesis that linguistic information is important for instrumental learning and may affect different processes supporting it distinctly is untested.

Additionally, even with the goal of limiting contributions of other memory processes, there is ample evidence that working memory (WM), the process involved in actively maintaining no-longer-present information over a short period of time, plays a nontrivial role in stimulus-response association learning (Collins & Frank, 2012; Collins, 2018; Collins et al., 2017, 2014; Jafarpour, Buffalo, Knight, & Collins, 2022; McDougle & Collins, 2020; Viejo, Khamassi, Brovelli, & Girard, 2015). These studies show that computational models that include WM in associative learning tasks fit human data better than those with RL alone. In contrast to RL, WM is often conceptualized as being explicit memoranda being consciously accessible and manipulable by the observer. There is, however, evidence that information maintained in WM and used for subsequent decisions need not be explicit (i.e., can be implicit; Yoo, Klyszejko, Curtis, & Ma, 2018; Honig, Ma, & Fougnie, 2020; Yoo, Acerbi, & Ma, 2021). Additionally, because WM was traditionally theorized to contain separate linguistic and visual storage units (Baddeley & Hitch, 1974), the WM field often

uses stimuli designed to probe only one of the "two" WM storage units (for example, using color, orientation, or spatial frequency to investigate visual WM, e.g., Wilken & Ma, 2004; Galeano Weber, Keglovits, Fisher, & Bunge, 2020; spoken letters, digits, or words to investigate verbal WM, e.g., Conrad, 1964) or uses control manipulations to ensure the other storage unit cannot provide assistance (e.g., concurrent verbal tasks while doing a visual WM task; Brady, Störmer, & Alvarez, 2016; Starr, Srinivasan, & Bunge, 2020).

In this study, we wanted to examine directly how different types of stimuli affect the contributions of RL and WM processes during an associative learning task. We designed and collected data on two stimulus-response association learning experiments, manipulating the information available to the learner. Subjects learned in conditions where stimuli had only distinct visual information but were linguistically poorly discriminable (the "Visual" condition), distinct linguistic information but were visually poorly discriminable (the "Linguistic" condition), or both rich visual and rich linguistic information (the "Conjunctive" condition). We investigated these questions through behavioral comparisons of learning behavior across the three conditions, as well as computational modeling to try to understand changes in the underlying RL and WM processes across conditions.

Generally, we predicted that there would be a difference in the processes involved in learning in the three stimuli conditions. However, due to the potentially competing effects between RL and WM, it is difficult to predict exactly how changes in RL, WM, and their interplay, will affect the ultimate behavioral performance. Take, for example, the Visual condition. The implicit assumption in the RL literature that non-linguistic stimuli limit access to higher level cognitive processes would suggest that WM would be the primary process impacted in the Visual condition. However, the WM literature may suggest the opposite result; because WM representations need not be verbalizable (visual WM literature) and fidelity for naturalistic visual stimuli is fairly high (Brady et al., 2016), WM might not be impaired in the Visual condition. Similarly, if RL is a mostly implicit process, as often hinted in the literature, then stimulus condition should not impact it much. However, if RL instead relies heavily on explicit linguistic information, performance should suffer when only salient visual information is available. 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, and 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 Visual only and Linguistic only conditions relative to the Conjunctive condition, demonstrating that overall discriminability is important to RL. This deficit was particularly pronounced in the Visual only condition, demonstrating the particular importance of linguistic information in stimulus-response association learning. Through computational modeling, we found that stimulus conditions specifically affected RL, and largely spared the WM processes supporting learning.

2 Experiment 1

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. Participants received $0.50 base payment for participating, and earned bonus payments for the time the spent on the task and their accuracy. 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 et al., 2014). At the beginning of a block, participants were shown a new set of stimuli, a group of images. 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. Each stimulus was presented approximately 13 times within a block (stimuli were presented as few as 11 and as many as 14 times). Stimulus sets corresponded to a different category and differed for each block (i.e., if you saw photos of vegetables in one block, you would not see broccoli in another block). Subjects 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):

- Conjunctive: stimuli are images of different subcategory members belonging to the same

category (e.g., farm animals: horse, cow, chicken), and easily discriminable both linguistically and visually.

- Linguistic: stimuli are words printed in black letters on a white background, corresponding to subcategory name (e.g., the words "horse", "cow", "chicken"), and provide discriminable linguistic information but limited distinct visual information.

- Visual: stimuli are different images of the same subcategory (e.g., different images of horses), and provide distinct visual information, but limited distinct linguistic information – each image set was designed to call to mind the same word and limit the ability to have different verbal labels as much as possible.

Each block had a unique category, so a participant would not see, for example, stimuli corresponding to "farm animals" in both the Conjunctive and Visual conditions. Which category was assigned to each stimulus condition was counterbalanced across participants, so participants saw different subsets of the entire stimulus set. 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). 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 ($M = 0.79$, $SEM = 0.02$) than set size 6 blocks ($M = 0.66$, $SEM = 0.02, F(1, 58) = 106.2, p = 10^{-14}$, Figure 1C), supporting the involvement of WM in learning. There was a significant main effect of condition ($F(2, 116) = 43.95, p = 6 \times 10^{-15}$), such that performance in the Visual condition ($M = .66, SEM.02$) was significantly lower than both Conjunctive ($M = .78, SEM = .02, p = 10^{-7}$) and Linguistic conditions ($M = .74, SEM = .02, p = .0009$). Conjunctive and Linguistic 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 ($F(2, 116) = 38.8, 10^{-13}$) than set size 3 blocks ($F(2, 116) = 8.707, p = 3 \times 10^{-4}$).

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 correct choices 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, and negatively related to set size and delay in all conditions (Figure 1D). This analysis suggests that participants in all conditions were more likely to get the current trial correct if they had gotten the stimulus correct more time previously (as expected from incremental RL-like learning), if there was a shorter delay since the last correct feedback for the corresponding stimulus, and if there were fewer other stimuli they had to simultaneously learn (both as expected from contributions of WM to choice).

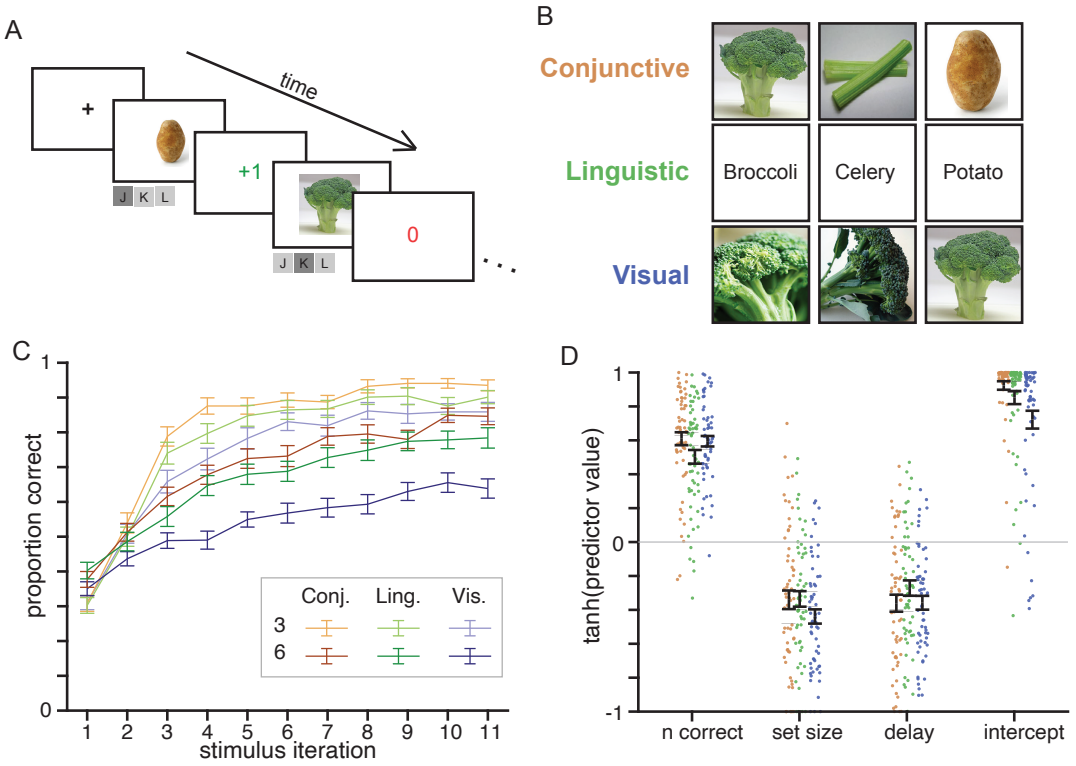

Figure 1: **Experiment 1 task and learning curves.** A. Behavioral task. Participants learn through trial and error, with truthful, deterministic feedback, the correct response to each stimulus. B. Example "vegetable" stimuli, for the three different stimulus conditions: Conjunctive, Linguistic, Visual. Stimulus categories were different for each block, so participants would never see (for example) a broccoli in multiple learning blocks. C. Learning curves (error bars: M ± SEM over participants) for stimulus condition (color) and set size (value/saturation). Each learning curve shows the proportion of correct choices as a function of the number of times a stimulus has been encountered within a block (stimulus iteration). Increasing learning curves indicates participants learn the correct response with increasing stimulus iteration. D. Logistic regression weights (hyperbolic tangent transformed) for each condition (colors) and participant (dots; error bars indicate M ± 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 RL and WM processes produce these behavioral differences across conditions. For this, we turn to behavioral modeling. In this section, we describe the computational models we developed and compared to better understand human behavior. Like previous publications using similar tasks and models (e.g., Collins & Frank, 2012; Viejo et al., 2015; Jafarpour et al., 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. Individual RL and WM models were tested, but they, as in previous studies using this behavioral paradigm, had poor fits to the data (see Supplementary 5.5 for exclusive RL and WM models).

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, $\text{WM}(s,a)$. When we refer 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. We use two different learning rates for positive (correct) and negative (incorrect) rewards. We fit the positive-reward α, and fix the negative learning rate $\alpha_- = 0$. (Relaxing this assumption did not improve model fit and did not change our main results or conclusions; Supplementary 5.5.2).

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:

$$\forall s, a \, \text{WM}_0(s,a) = \frac{1}{N_a}$$
$$\text{WM}_{t+1}(s,a) \leftarrow r_{t+1},$$

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 return to their starting value:

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

where λ is the decay rate. With this formulation, WM's stored values regress to uninformative values, $\text{WM}_0(s, a)$, 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 actionHere we fix $\beta = 100$; fixing β to high values is common in "RLWM" models (e.g., Jafarpour et al., 2022; McDougle & Collins, 2020) as this improves the models' interpretability (enforcing close to perfect one-back WM policy under low load), as well as parameter recovery for remaining parameters. $V_t(s, a_i)$ 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.

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

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 5.5.3).

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.

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

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

We considered six models, each making a specific prediction about how the RL process, WM process, or interaction between the two is affected by the block's stimulus condition. 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. All models are explained in more detail below, and alternative assumptions, such as RL only, WM only, different specifications for perseveration, or nonzero negative learning rate α_- are presented in Supplementary Materials 5.5.

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 RL updating. We implement this assumption by allowing the learning parameter in the RL process, α, to vary as a function of stimulus condition. We denote the learning parameter for Conjunctive, Linguistic, and Visual stimuli as α_c, α_l, 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 η:

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

We fit credit assignment confusion parameters to Linguistic and Visual conditions only, denoted η_l and η_v, respectively. We did attempt to fit a model with credit assignment confusion in the Conjunctive condition, η_c, 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 Conjunctive condition would be generally captured by noise parameters, and that the **additional** confusion in the Linguistic and Visual 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 Linguistic and Visual conditions, denoted ζ_l and ζ_v, respectively. Like in the RL credit assignment model, we implicitly assume there is no RL decision confusion in the Conjunctive condition, $\zeta_c = 0$, 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 Linguistic and Visual 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 Conjunctive Linguistic, and Visual stimuli as λ_c, λ_l, 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. In this model, the weights ωs differ across condition and set size, and are denoted with subscript. For example, ω_{6c} corresponds to the RLWM weight of a set size 6 Conjunctive stimulus condition. We consider 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, $\omega_3, \omega_{6c}, \omega_{6l}$, and ω_{6v}.

2.3.3 Parameters and estimation

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

Model	**Base parameters**	**Model-specific parameters**
RL learning rate	$\alpha_c, \lambda, \phi, \omega_3, \omega_6, \epsilon$	α_l, α_v
RL credit assignment	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	η_l, η_v
RL decision confusion	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	ζ_l, ζ_v
WM decay	$\alpha, \lambda_c, \phi, \omega_3, \omega_6, \epsilon$	λ_l, λ_v
WM decision confusion	$\alpha, \lambda, \phi, \omega_3, \omega_6, \epsilon$	ζ_l, ζ_v
RL WM weight	$\alpha, \lambda, \phi, \omega_3, \omega_{6c}, \epsilon$	ω_{6l}, ω_{6v}

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

2.3.4 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.

2.3.5 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 (distinguishable). 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 (Nilsson, Rieskamp, Wagenmakers, & Nilsson, 2011; Wilson & Collins, 2019). Successful parameter recovery exists when one is able to "recover" the same (or similar) parameter values that generated the data.

Model recovery is important in making conclusions from quantitative model comparisons. 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. We obtained reasonable parameter and model recovery; details and figures for both analyses are in Supplementary sections 5.2 and 5.3, respectively.

2.4 Modeling Results

We compared models quantitatively using two metrics, and qualitatively using posterior predictive checks. First, we compared models' ability to fit the existing data quantitatively using the obtained maximum LL, LL^*. We compared fits across participants in two ways: through summed ΔLL^* and through group Bayesian Model Selection (BMS; Stephan, Penny, Daunizeau, Moran, & Friston, 2009; Rigoux, Stephan, Friston, & Daunizeau, 2014). While summed ΔLL^* implicitly assumes all participants are generated by the same model, BMS explicitly assumes that participants can be best fit by different models. BMS uses the log marginal likelihoods for each model and participant to infer the probability of each model over participants. From this, we can compute the protected exceedance probability (pxp), how likely a given model is to be more frequent than the other models in the comparison set, above and beyond chance. A lower ΔLL^* and higher pxp indicate better model fit to data. 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, differences of 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. Both model comparisons are important to assess model fits, particularly for data with sequential trial dependencies. For example, 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 (Palminteri, Wyart, & Koechlin, 2017). We find that the qualitative fits to the data (Figure 2A) reflect the quantitative model comparison; the model that features condition-specific learning provides a better fit to the true data). These results suggests that different stimulus conditions affect exclusively the RL process, by how quickly it learns from RPE.

2.5 Interim conclusions

In Experiment 1, we asked how limiting discriminability in stimuli linguistic or visual information 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 performance. Second, and to our main question, we found that limiting either discriminable visual or linguistic information across stimuli detrimented performance, but taking away linguistic information severely lowered learning performance. Additionally, 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. These behavioral

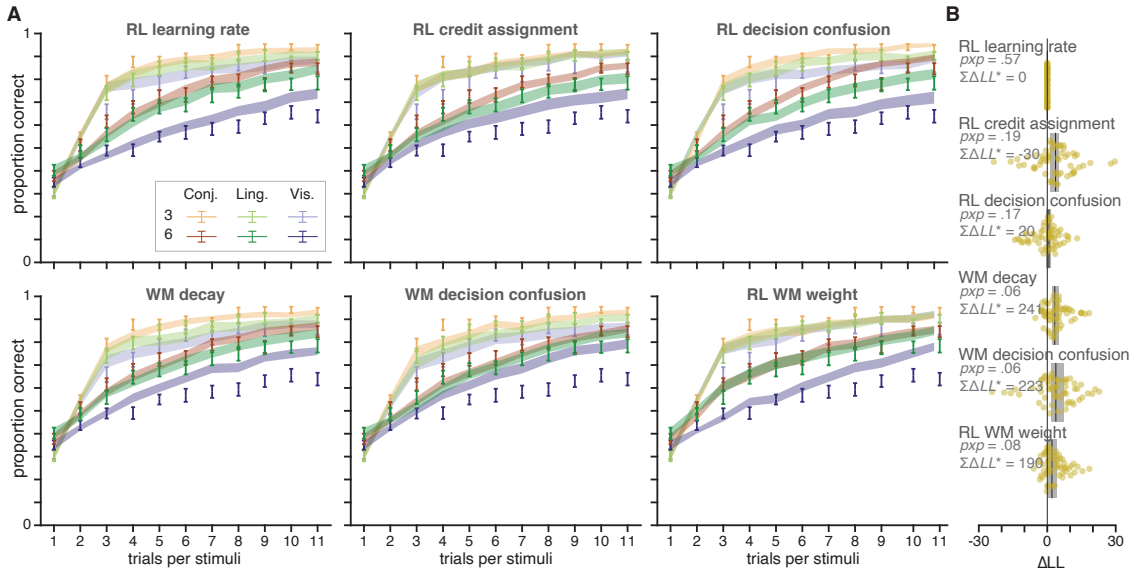

Figure 2: **Experiment 1 Modeling Results.** A. Learning curves for each condition (color) and set size (value/saturation) across participants for data (errorbars, M ± sem) and model predictions (fills, M ± sem). B. Difference in LL scores for each model, relative to the RL learning rate model. Dots indicate individual participants, horizontal 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. For all values, lower scores indicate better model fit.

results suggest, perhaps surprisingly, that linguistic information is essential to the RL component of the learning process, a finding that goes against the oft held intuition that RL processes are more implicit and thus less reliant on explicit mechanisms.

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 over-perform slightly in the Visual condition set size 6 (Figure 2). Could we design an experiment that more directly tests the contribution of RL and WM processes, and thus strengthens the conclusion that solely RL is affected by lowering either visual or linguistic information in stimuli?

3 Experiment 2

In this section, we describe our second experiment, which was designed to replicate and extend the behavioral and modeling results of the first experiment. First, we replicated the same experiments with an in-person sample, and extended the experiment to test specifically the RL process's contribution to learning, to see if people's RL process is selectively impaired in the Visual condition relative to Conjunctive and Linguistic conditions.

Collins and others (2018) demonstrated an interaction between WM and RL processes for long-term retention of the correct stimulus-action pair. Items in lower set size blocks had better performance during learning phase compared to higher set size blocks, but interestingly, worse performance in long-term storage. 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. This data provides an additional way to potentially tease apart the contribution of RL and WM processes across stimulus conditions. If RL is selectively impaired for certain stimulus conditions, then this may be further revealed in the test phase: there would be a larger deficit in long-term retention (above and beyond any effects of lower learning performance). In other words, if two stimuli in Visual and Conjunctive had the same asymptotic performance in the Learning phase, the performance of the Visual stimulus could still be worse in the Test phase.

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. 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. Seven participants were excluded for 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 paradigm as MTurk participants did. 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 5.1. It took approximately 10 minutes to complete.

Lastly, participants completed a Test Phase, in which all stimuli from all Learning phase blocks were presented again in random order. This phase probed how well stimulus-response pairs were learned by a RL process, without the aid of WM. 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 (Fig. 3A, middle). Like before, we conducted a 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 ($F(1, 29) = 185.1$, $p = 4 \times 10^{-14}$), condition ($F(2, 58) = 24.66$, $p = 2 \times 10^{-8}$), and interaction between set size and condition ($F(2, 58) = 11.90$, $p = 4 \times 10^{-5}$). For condition, performance in the Visual condition ($M = .69, SEM = .03$) was significantly lower than that of the Conjunctive ($M = .79, SEM = .02, p = .0001$) and Linguistic ($M = .76, SEM = .02, p = 0.02$) conditions. The p-values for posthoc tests are Bonferroni corrected. The interaction was driven by a not-quite-significant condition effect in set size 3 blocks ($F(2, 58) = 2.44$, $p = .10$) but a strong condition effect in set size 6 blocks ($F(2, 58) = 27.07$, $p = 5 \times 5 \times 10^{-9}$).

We then conducted the same logistic regression as in Experiment 1: we investigated 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 consistent

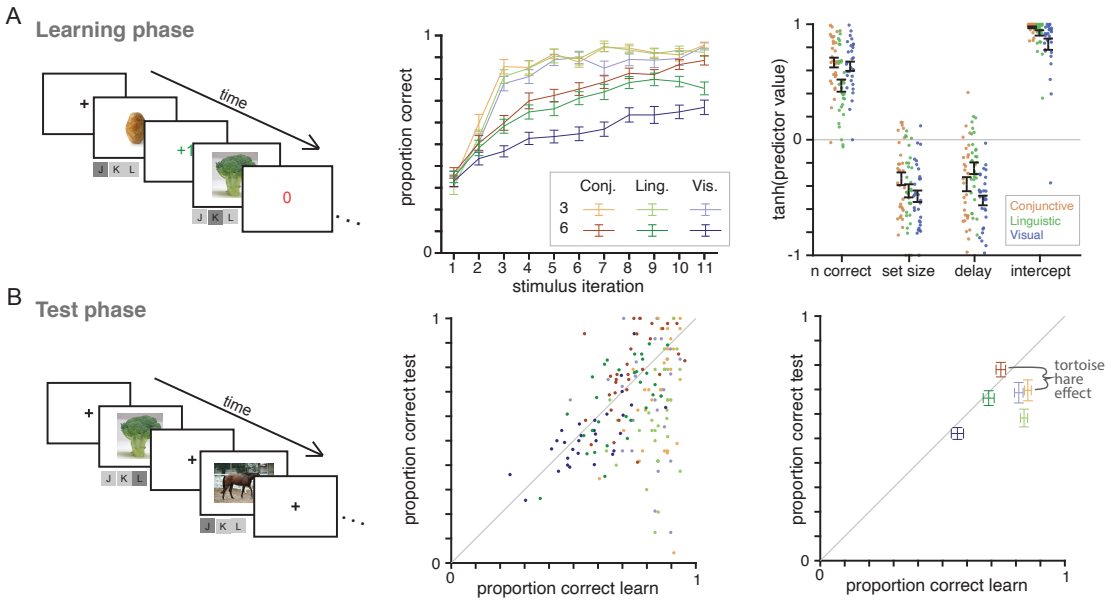

Figure 3: **Experiment 2 task and results.** A. Learning phase. *Left*: task design. *Middle*: Proportion of correct choices increases as a function of stimulus iteration for all stimulus and set size conditions but slower for set size 6, especially in the Visual condition. *Right*: logistic regression. 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, Right*: Proportion correct in training (x-axis) and testing (y-axis) phase for condition (color), showing individual participants (middle) or $M \pm sem$ across participants (right). There is a larger deficit in long-term retention (indicated by the deviation from the identity line) with stimuli learned in set size 3 blocks (lighter) than set size 6 blocks (darker), demonstrating the tortoise-hare effect.

results 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).

Second, we analyzed the participants' performance on the Test phase. For all conditions and set sizes, performance was above chance ($t(29) > 5.97, p < 10^{-6}$), 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 ($t(29) = 6.41, p = 5 \times 10^{-7}$), replicating (Collins, 2018) and 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 ($F(2, 58) = 2.207, p = .12$).

3.3 Modeling methods

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.

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 2.3.2-2.3.4.

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 cooperate 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:

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

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

3.4.1 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).

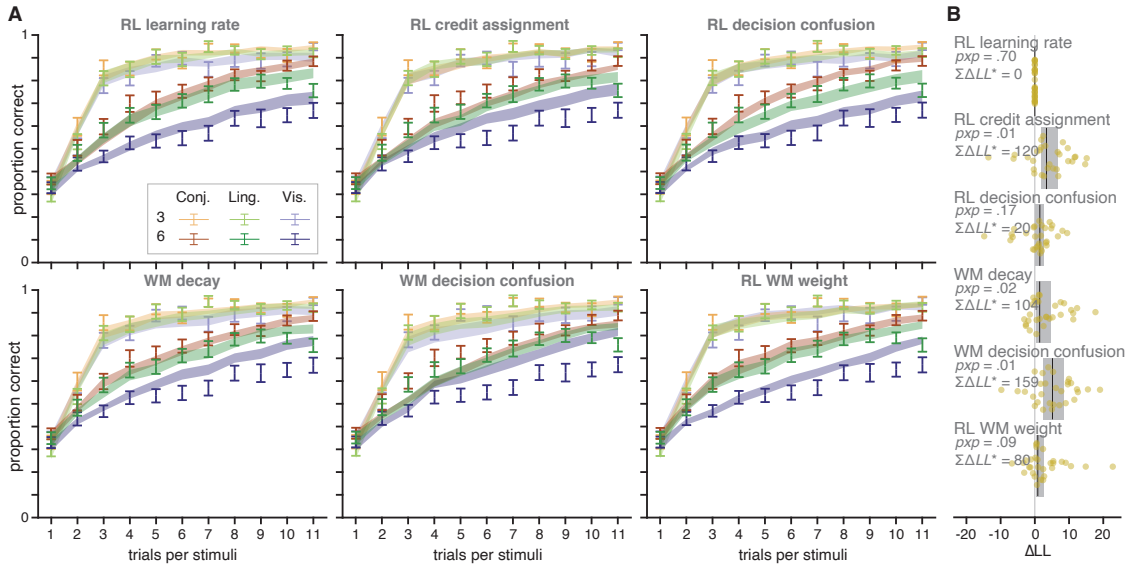

Figure 4: **Experiment 2 modeling results** A. Learning curves for each condition (legend at top) across participants for data (errorbars, M ± sem) and model predictions (fills, M ± sem). B. Difference in LL for each model relative to the RL learning rate model. Dots indicate individual participants, horizontal 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. For all values, lower scores indicate better model fit.

3.4.2 Investigating Test Phase

Model validation plots are illustrated in Figure 5. Models that assume an interaction between RL and WM during learning were able to capture Test phase data better for the Conjunctive and Linguistic condition (orange and green), but models that assume no interaction were able to capture Test phase data better in the Visual condition (blue).

Model fits for all four models were quantitatively very similar (lower ΔLL^* and higher pxp indicates better model fits to data. RL learning rate $\Delta LL^* = 0$, $pxp = .25$; RL decision confusion $\Delta LL^* = 49$, $pxp = .23$; RL learning rate + interaction $\Delta LL^* = -44$, $pxp = .27$; RL decision confusion + interaction $\Delta LL^* = -8$, $pxp = .25$).

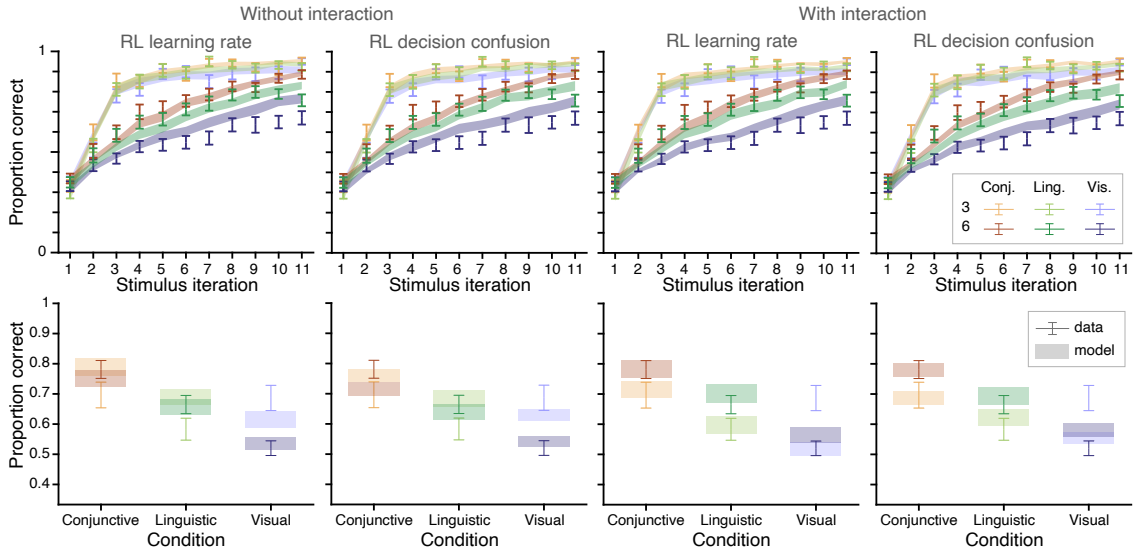

Figure 5: **Exp 2 train and test phase model validation**. Model validation for RL learning rate and RL decision confusion models without (left two plots) and with (right two plots) an interaction between RL and WM processes. Model predictions (fill) and data (error bars) for models jointly fitted on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

3.5 Interpreting model parameters

Finally, 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 winning models displayed in Supplementary Figure 14 15).

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 ($p > .41$). The contribution of WM in set size 3, however, was marginally significant before multiple comparisons in both models. The numerical differences in ω_3, although not significant after correcting for

multiple comparisons, could explain subtle differences we see across experiments. Specifically, we observed a condition effect in set size 3 in Experiment 1, but not Experiment 2. This behavioral difference may be due to higher RL contribution; since RL seems to be the main factor affecting difference in behavior across conditions, a larger contribution of RL may result in more salient condition differences. Indeed, our estimated parameters are consistent, indicating lower WM contribution (and thus higher RL contribution) in Experiment 1 than Experiment 2 for both winning models. For all following analyses, we combine participant parameters across experiments.

To investigate the condition 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 across conditions differ in the condition-specific RL learning rate model. The learning rate for Visual condition (α_v: $M = .01$, $SEM = .003$) was significantly lower than that of Linguistic condition (α_l: $M = .03$, $SEM = .006$, $z = -7.40$, $p = 4 \times 10^{-14}$) and Conjunctive condition (α_c: $M = .04$, $SEM = .008$, $z = -6.37$, $p = 5 \times 10^{-10}$). The difference in learning rates for Conjunctive and Linguistic condition were approaching significance ($z = 2.25$, $p = .07$). 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 Visual (no interaction model: $M = .01, SEM = .001$, interaction model: $M = .008, SEM = .0008$) condition is lower than that of Conjunctive (no interaction: $M = .04, SEM = .03, z = -4.37, p = 4 \times 10^{-5}$; interaction: $M = .04, SEM = .02, z = 4.41, p = 3 \times 10^{-5}$) and Linguistic (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 significant difference between Linguistic and Conjunctive 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 Visual condition (ζ_v: $M = .44$, $SEM = .02$) was significantly higher than that of the Linguistic condition (ζ_l: $M = .22$, $SEM = .03$, $z = 6.02$, $p = 2 \times 10^{-9}$). This effect is also true for the models fitted on Learning and Test phase of Experiment 2; decision confusion is greater in the Visual condition than the Linguistic condition in both the models that assume no interaction between RL and WM (Visual: $M = .36, SEM = .04$, Linguistic: $M = .18, SEM = .04$, $z = 2.95, p = .003$) and those that do (Visual: $M = .40, SEM = .04$, Linguistic: $M = .20, SEM = .04$, $z = 3.38, p = .001$).

4 Discussion

In this study, we investigated how the type of information in a stimulus affected learning. Participants learned the correct response to stimuli that had linguistic and visual disciminability (Conjunctive), only Linguistic discriminability, or only Visual discriminability, relative to other stimuli in the same block. We found, behaviorally from two experiments, that linguistic information was particularly important for learning, demonstrated by a large deficit in performance, particularly in set size 6 blocks, for stimuli that only were visually discriminable, but linguistically similar.

Anecdotally, this fits with participants' comments on the task: participants in this experiment and other similar ones sometimes report using a single word label for the images they see, so preventing this strategy may have particular detriment to performance.

We used computational modeling to investigate what exactly in the process of learning was being affected by lowering the amount of information in the Linguistic and Visual conditions. We used RLWM models, which assumes RL and WM both contribute to support learning. We found in both experiments that the data was best described by a model in which the RL process was detrimented for Linguistic and Visual conditions, particularly in the Visual condition. Additionally, Experiment 2 supported the idea that, while WM assists RL during learning, this in turn leads to an interference in long-term retention, leading to the "tortoise and hare" effect (Collins, 2018).

What 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). In the Visual condition, the visual disciminability was decreased in addition to the loss of linguistic information, due to the nature of constraining stimuli to the same "linguistic cue." Thus, participants may have had to rely on other systems like executive function (Rmus, McDougle, & Collins, 2021) and attention Radulescu et al., 2019; Niv et al., 2015 to distinguish stimuli in the Visual conditions. What features are important to pay attention to itself is something that must learned (e.g., Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017), and can affect later 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 exceptionally poor performance in the Visual condition could have been because the relevant disciminating features in the Visual 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 it is placed on a table. 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 difficultly successfully using these features to discriminate between stimuli for RL. Regardless of exact cognitive mechanism at play, these results demonstrate the importance of considering how a learning state is defined. Other studies corroborate this conclusion, finding stimulus type (naturalistic stimuli learned better than abstract stimuli; Farashahi, Xu, Wu, & Soltani, 2020) and response "state" (motor responses learned better than stimulus responses; Rmus & Collins, 2020) affect learning.

In remarkable contrast to RL, our 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. 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 et al., 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 et al., 2020) (even when doing a simultaneous verbal task, to ensure verbal WM is not assisting). Additionally, some studies of WM suggest a more unified store of information across modality (e.g., Morey & Cowan, 2004; Uittenhove, Chaabi, Camos, & Barrouillet, 2019), so the traditional hard line separation of verbal and visual information may not be so relevant to behavior. This literature together, along with our results, suggest that, unlike RL, Visual or Linguistic information alone is sufficient for WM to discriminate stimuli and maintain stimulus-response associations.

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, behavior in the Visual condition in set size 6 was lower than that of any model predictions. Perhaps learning detriments in the Visual condition is a combination of multiple processes being affected, an assumption we did not test. Another possibility is that participants focused on a subset of stimuli, which the model cannot account for easily, but has previously been shown to lead to similar validation mismatch (Wilson & Collins, 2019). Second and relatedly, as in all modeling papers, there are an infinite number of models that can be defined and tested, and we focused on a small subset of this model space. In our paper, for the sake of simplicity, we do not test models which assume that removing a linguistic or visual information affects RL and WM processes differently or in multiple ways. In other words, we assume that one specific detriment occurs in learning by removing information, and that detriment is modality-unspecific. This is likely not be a reasonable assumption, and this is something for future studies to investigate. However, it is unlikely that more complex models would be identifiable within this experimental framework, justifying our limited exploration of the model space.

Finally, 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 5.3), 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, we can confidently conclude to an impact on the RL learning process, if not on the exact nature of this impact.

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 ex-

periments (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 consistent 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, particularly the amount of linguistic information. 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 choice of stimuli affect 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/LinguisticInfoInRL.

References

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

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

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

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. doi: 10.1162/jocn_a_01238

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. doi: 10.1016/j.biopsych.2017.05.017

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., & 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

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

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., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137.

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., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *Journal of Neuroscience*, *35*(2), 485–494. doi: 10.1523/JNEUROSCI.2036-14.2015

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.

Galeano Weber, E. M., Keglovits, H., Fisher, A., & Bunge, S. A. (2020). Insights into visual working memory precision at the feature- and object-level from a hemispheric encoding manipulation. *Quarterly Journal of Experimental Psychology*, *73*(11), 1949–1968. doi: 10.1177/1747021820934990

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

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences*, *117*(15), 8391–8397. doi: 10.1073/pnas.1918143117

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

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

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review*, *11*(2), 296–301. doi: 10.3758/BF03196573

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.*

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

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.

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

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

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

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

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

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

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

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is working memory storage intrinsically domain-specific? *Journal of Experimental Psychology: General*, *148*(11), 2027–2057. doi: http://dx.doi.org/10.1037/xge0000566

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

Yoo, A. H., Acerbi, L., & Ma, W. J. (2021). Uncertainty is maintained and used in working memory. *Journal of Vision*, *21*(8), 13. doi: 10.1167/jov.21.8.13

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

5 Appendix

In the Supplementary Materials, we include additional analyses that broadly support the main text. We include details on the N-back distractor task, 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.

5.1 N-back distractor task

The first block was a practice block with N=2, then the following four blocks were N=2-5 sequentially. 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). We additionally received and independent measure of WM strength from this task. Code for the N-back task can be found here.

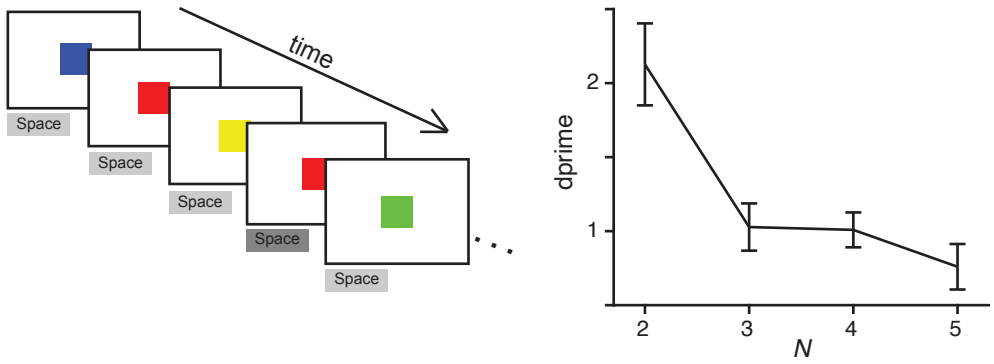

Figure 6: N-back task. *Left*: task design. Participants viewed a series of colors and made a key press every time the color N trials ago was the same as the color of the current block. This illustration demonstrates all correct responses on a $N = 2$ back task. *Right*: d' decreases a function of N, indicating worse performance with increasing set size.

5.2 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. Without successful parameter recovery, one cannot be certain of the meaningfulness of the actual values of the parameters.

For each model, we generated parameters by sampling the fitted parameter vectors from participants of 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 distribution informed by the fitted parameter 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.

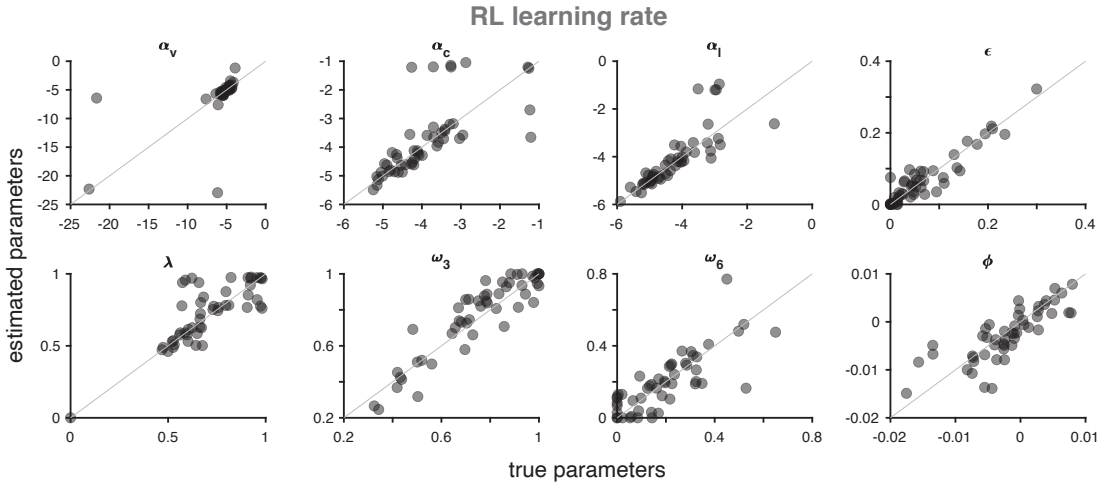

Figure 7: Parameter recovery plots for condition-specific RL learning rate model. Each subplot plots the true parameters, which generated data, against the recovered parameter values, estimated using MLE. Dots are 50 simulated subjects.

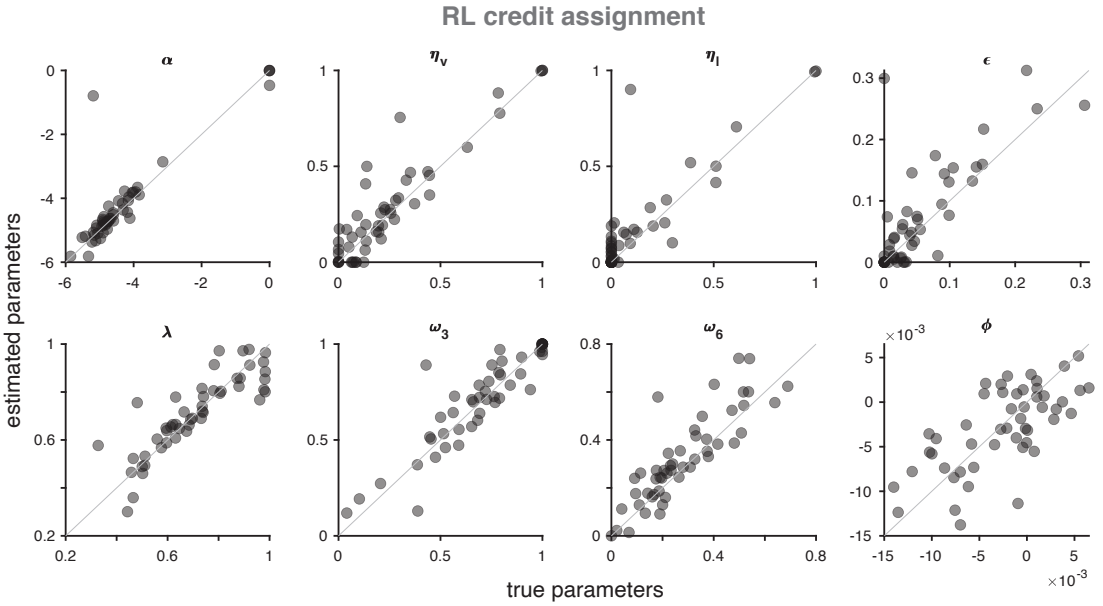

Figure 8:

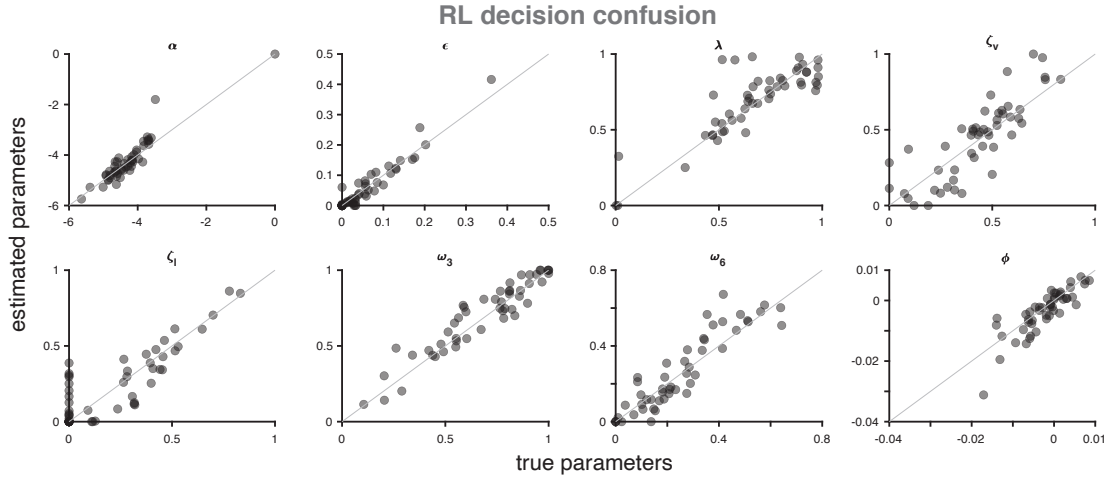

Figure 9:

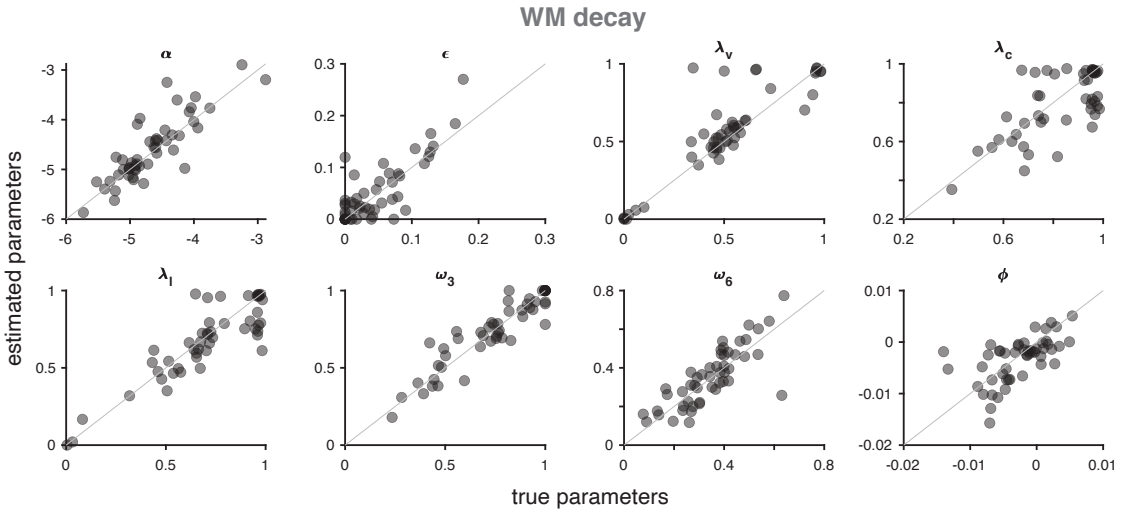

Figure 10:

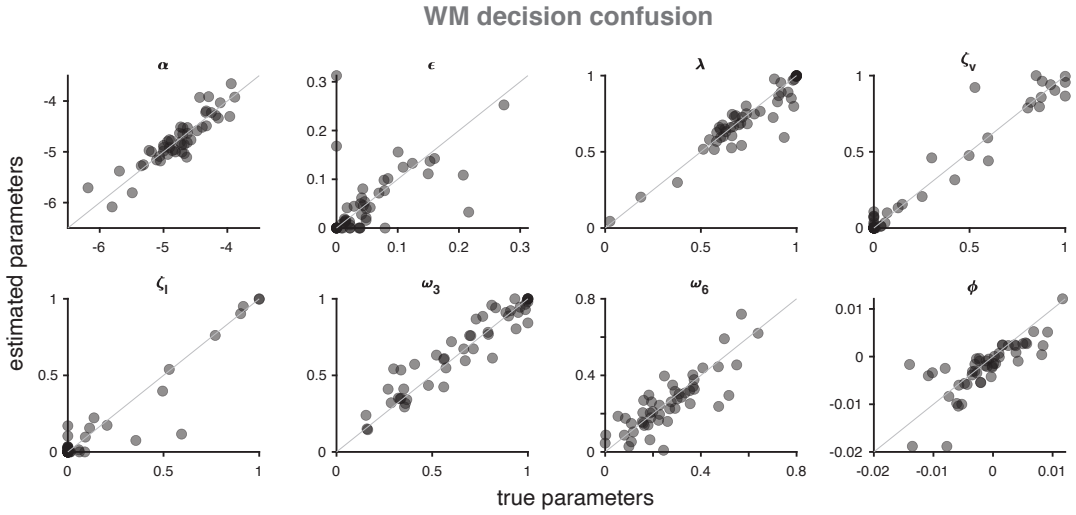

Figure 11:

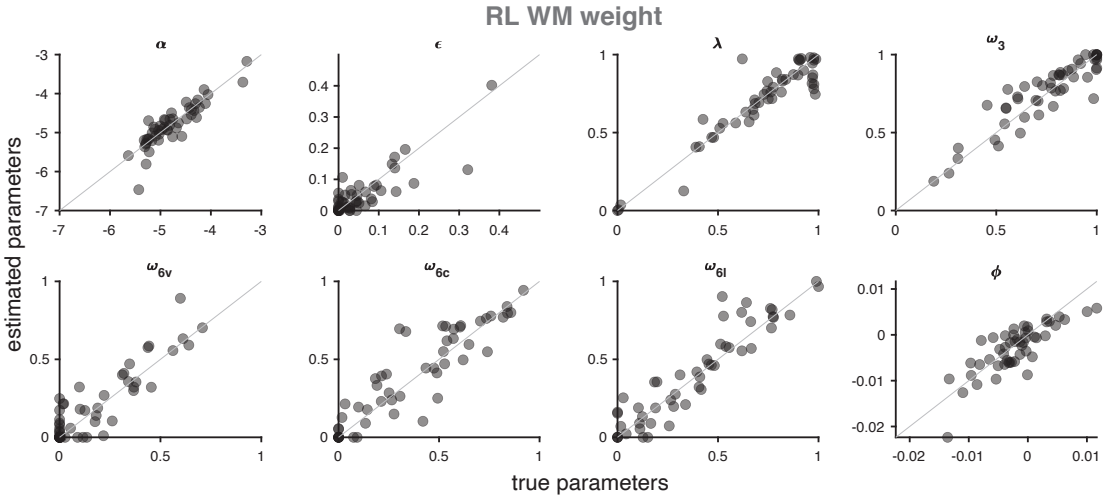

Figure 12:

5.3 Model recovery

Model recovery is important in making conclusions from 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 then calculate the AICc and BIC for every fitted model participant. Successful model recovery occurs when the model that best fits the simulated data is the same model that generated that data. 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.

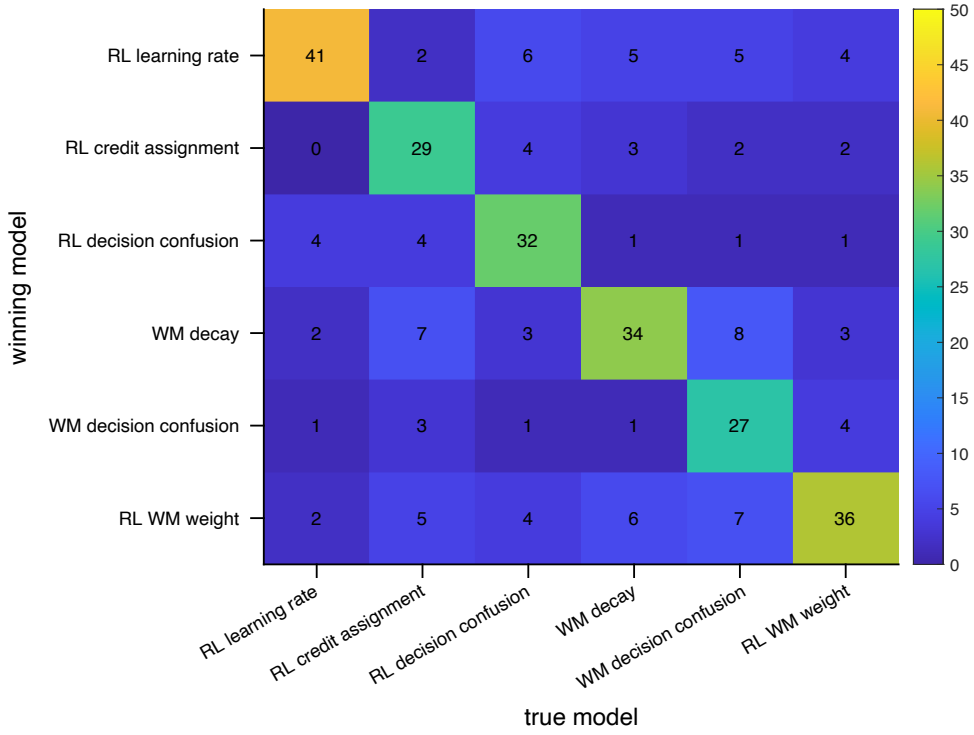

Figure 13: Model recovery when using LL to compare models. Successful model recovery is indicated by a majority of models falling on the identity. In these results, it appears that condition-specific RL learning rate and WM decay models fit some participants better than the model that actually generated the model, suggesting that if either of these models are best fitting to experimental data, we should take those results with a grain of salt.

5.4 Parameter values

RL learning rate model

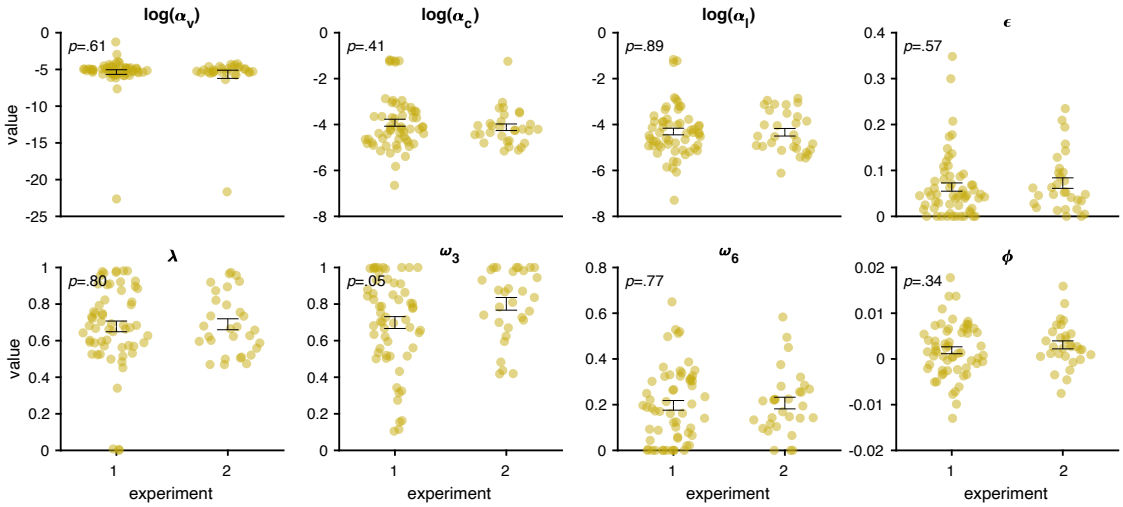

Figure 14: Parameter values (dots: individual participants. error bars: $M \pm sem$ across participants) for condition-specific learning rate model for condition-specific decision confusion model for Experiment 1 and Experiment 2. The p-values comparing the two participant groups, before any multiple comparisons corrections, displayed on the top left of each subplot. There were no significant parameter differences across participant populations after correction for multiple comparisons.

RL decision confusion model

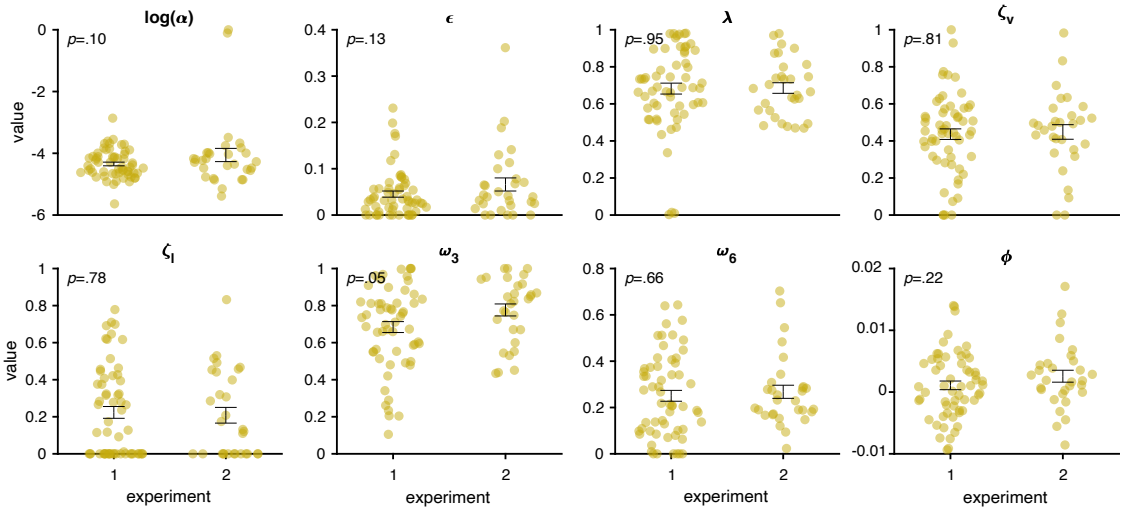

Figure 15: Parameter values (dots: individual participants. error bars: $M \pm sem$ across participants) for condition-specific decision confusion model for Experiment 1 and Experiment 2. The p-values comparing the two participant groups, before any multiple comparisons corrections, displayed on top left of each subplot. There were no significant parameter differences across participant populations after correction for multiple comparison.

5.5 Alternative Models

We tested six main RLWM models in main manuscript with the following condition-specific differences: RL learning rate, RL credit assignment confusion, 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. Second, we use factorial model comparison to test whether the goodness of fit for the six main models we fit in the main manuscript vary with/without perseveration, and with/without a fitted negative learning rate, α_-, parameter. Finally, we test if our assumption of 1-back perseveration affects our modeling results, by softening this assumption.

In these sections, we compared model goodness-of-fit using corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). Both measures 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.

5.5.1 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 here, relative to the condition-specific RL learning rate model used in the main manuscript.

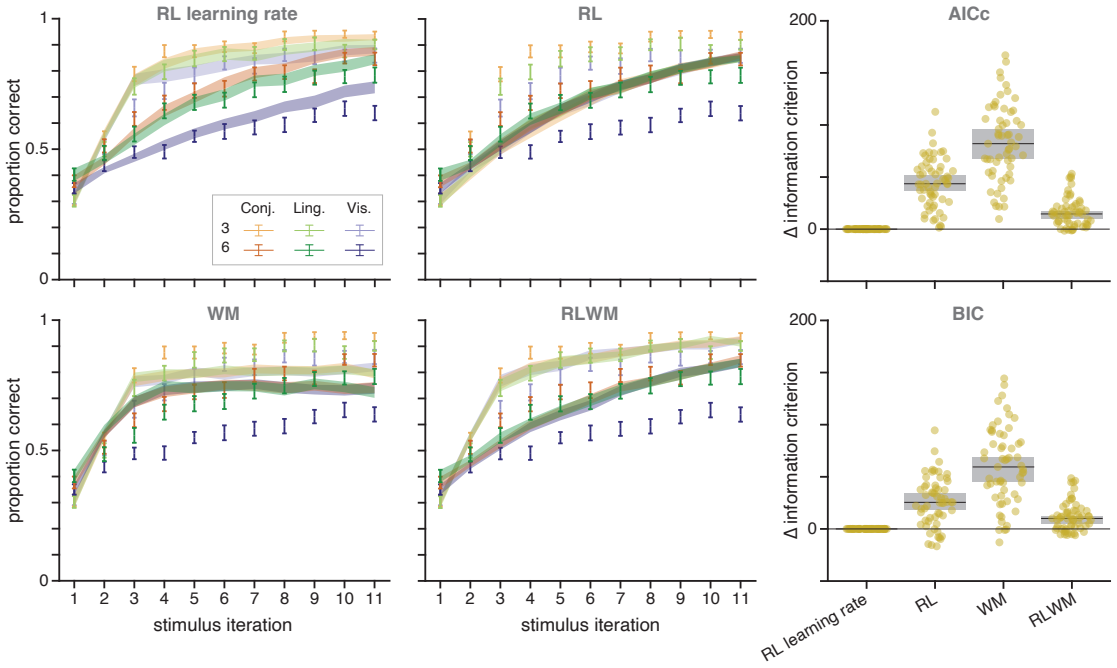

Figure 16: Model validation plots for the condition-specific RL learning rate, RL, WM, and RLWM models (left four plots). AICc (top) and BIC (bottom) differences between models and RL learning rate model. A smaller number indicates a better fit. The condition-specific RL learning rate clearly fit the data qualitatively and quantitatively better than these models.

5.5.2 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), and negative learning rate (2: fixed to 0, fit).

Figure 17 illustrates the quantitiative comparison of all models for both AICc and BIC. We see that the model ranking doesn't vary no matter what perseveration / negative learning rate combination we use. However, we find that fitting a perseveration paramter 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). Thus, we decided in the main manuscript to

include the model which keeps perseveration as a free parameter, and fixed negative learning rate $\alpha_- = 0$.

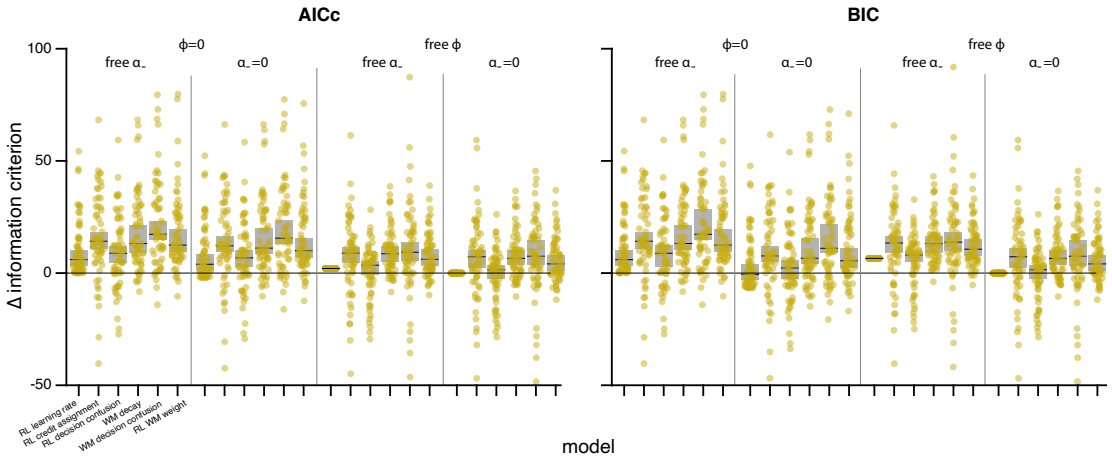

Figure 17: Quantitative results of factorial model comparison. AICc (left) and BIC (right) differences, relative to the RL learning rate model in the main manuscript. A lower number indicates a better fit. For each plot, each section of six models correspond to the respective characteristics: $\phi = 0$, fitted α_-; ϕ=0,$\alpha_- = 0$; fitted ϕ and α_-; fitted ϕ, $\alpha_- = 0$

5.5.3 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 of 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.

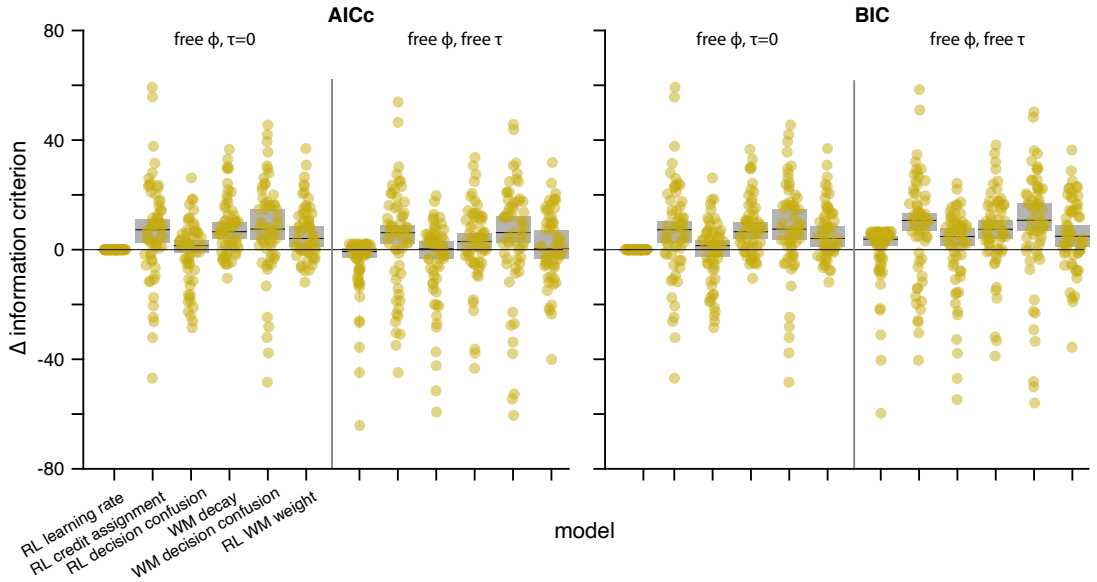

Figure 18: Factorial model comparison with perseveration parameter τ fixed to 1 (left six models on each plot) and as a free parameter (right six models on each plot). AICc (left plot) and BIC (right plot) are relative to the RL learning rate model with a free τ parameter. A lower value indicates a better fit to data. Model differences do not change model rankings, and model fits are not noticeably improved by including a free τ parameter.