Bayesian reinforcement learning models reveal how great-tailed grackles improve their behavioral flexibility in serial reversal learning experiments.

Lukas D[1]* McCune KB[2] Blaisdell AP[3] Johnson-Ulrich Z[2]
MacPherson M[2] Seitz B[3] Sevchik A[4] Logan CJ[1]*

2024-02-06

Open… access code peer review data

Affiliations: 1) Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2) University of California Santa Barbara, USA, 3) University of California Los Angeles, USA, 4) Arizona State University, Tempe, AZ USA. *Corresponding author: dieter_lukas@eva.mpg.de

This is one of three post-study manuscript of the preregistration that was pre-study peer reviewed and received an In Principle Recommendation on 26 Mar 2019 by:

Aurélie Coulon (2019) Can context changes improve behavioral flexibility? Towards a better understanding of species adaptability to environmental changes. *Peer Community in Ecology*, 100019. 10.24072/pci.ecology.100019. Reviewers: Maxime Dahirel and Andrea Griffin

Preregistration: html, pdf, rmd

Post-study manuscript (submitted to PCI Ecology for post-study peer review on 3 Jan 2022, revised and resubmitted Feb 2024): preprint at EcoEvoRxiv, rmd with code at github

ABSTRACT

Environments can change suddenly and unpredictably, so animals might benefit from being able to flexibly adapt their behavior through learning new associations. Reversal learning experiments, where individuals initially learn that a reward is associated with a specific cue before the reward is switched to a different cue, thus forcing individuals to reverse their learned associations, have long been used to investigate differences in behavioral flexibility among individuals and species. Here, we apply and expand newly developed Bayesian reinforcement learning models to gain additional insights into how individuals might dynamically adapt their behavioral flexibility if they experience repeated reversals in which cue is associated with a reward. Using data from simulations and great tailed grackles (Quiscalus mexicanus), we find that two parameters, the association updating rate, which reflects how much individuals weigh the most recent information relative to previously learned associations, and the sensitivity to learned associations, which reflects whether individuals no longer explore alternative options after having formed associations, are sufficient to explain the different strategies individuals display during the experiment. Individuals gain rewards more consistently if they have a higher association updating rate, because they learned that cues are reliable and they therefore can gain the reward consistently during one phase. The sensitivities to learned associations plays a role for the grackles who experienced a series of reversals, where individuals with lower sensitivities are better able to explore the alternative option after a switch. The grackles who experienced the serial reversal adapted their behavioral flexibility through two different strategies. Some individuals showed more exploration such that they can quickly change to the alternative option after a switch even if they continue to occasionally choose the unrewarded option. Others stick to the previously learned associations such that they take longer to change after a switch, but, once they have reversed their associations consistently, choose the correct option. These strategies the grackles exhibited at the end of the reversal learning experiment also relate to their performance on multi-option puzzle boxes where there are different behaviors required to access rewards. Grackles with intermediate strategies solved fewer options to access the rewards than grackles with either of the extreme strategies, and they took longer to attempt a new option. Our approach offers new insights into how individuals react to uncertainty and changes in their environment, in particular showing that they can adapt their behavioral flexibility in response to their experiences.

INTRODUCTION

Serial reversal learning experiments have long been used to understand how individuals keep track of biologically important associations in changing environments (Bitterman, 1975; Dufort et al., 1954; Mackintosh et al., 1968). Most animals live in environments that undergo changes that can affect key components of their lives, such as where to find food or which areas are safe. Accordingly, individuals are expected to be able to react to these changes. One of the ways in which animals react to changes is through behavioral flexibility, the ability to change behavior when circumstances change by updating information and making it available to other cognitive processes (Mikhalevich et al., 2017). Serial reversal learning experiments aim to measure differences in behavioral flexibility across individuals and species (Lea et al., 2020) by first presenting individuals with multiple options associated with cues, such as different colors or locations, that differ in their reward. After individuals learn the associations between rewards and cues, the rewards are reversed across cues, and individuals are observed to see how quickly they learn the changed associations. However, despite their long history, we still know little about how individuals approach these serial reversal learning tasks [Bond et al. (2007)) and what cognitive processes might lead to the observed differences in behavioral flexibility (Danwitz et al., 2022; Izquierdo et al., 2017).

A number of theoretical models have been developed to reflect the potential cognitive processes animals might rely on to make informed choices in changing environments (for a recent review see for example Frömer & Nassar (2023)). These models deconstruct the behavior of individuals making choices into two processes (Bartolo & Averbeck, 2020; Camerer & Hua Ho, 1999; P. K. Chow et al., 2015; Izquierdo et al., 2017). The first process reflects the learning about the environment, through updating associations between external cues and potential rewards (or dangers). Individuals are expected to show different rates of updating associations (which we refer to as ϕ, the greek letter phi, in the following) in different environments (Figure 1). Lower rates are expected when changes are rare and associations are not perfect such that a single absence of a reward might be an error rather than indicating a new association. Higher rates are expected when changes are frequent and associations are reliable such that individuals should update their associations when they encounter new information (Breen & Deffner, 2023; Dunlap & Stephens, 2009). The second process reflects how individuals, when presented with a set of cues, might decide between these alternative options based on their learned associations of the cues. Individuals with larger sensitivity to their learned associations (which we refer to as λ, the greek letter lambda, in the following) will quickly prefer the option that previously gave them the highest reward (or the lowest danger), while individuals with low sensitivity will continue to explore alternative options. Sensitivities are expected to show the opposite pattern to the association-updating rate (Figure 1), with larger sensitivities when cues are unreliable but environments are static such that individuals start to exploit the rare information they are learning and lower sensitivities when cues are reliable and changes are frequent such that individuals explore alternative options when conditions change (Breen & Deffner, 2023; Daw et al., 2006).

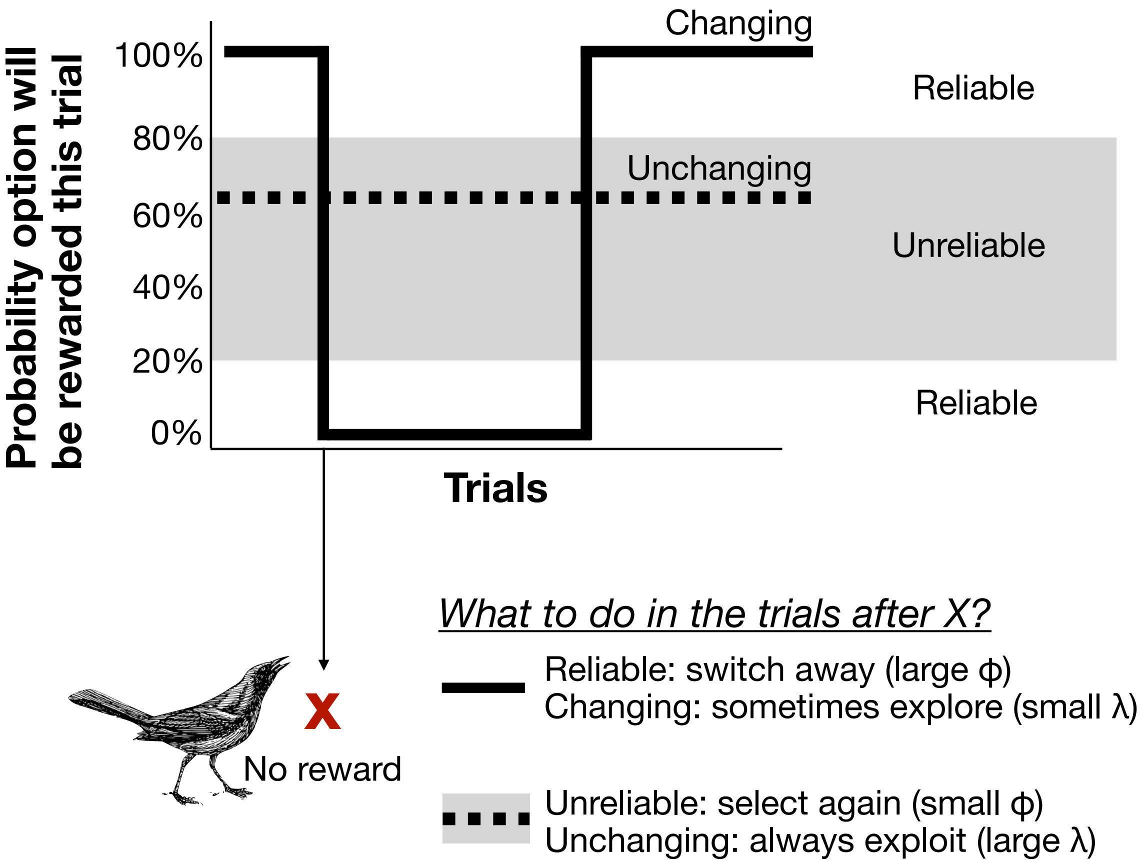
Probability option will be rewarded this trial
100%
80%
60%
40%
20%
0%
Trials
Changing
Unchanging
Reliable
Unreliable
Reliable
No reward
X
What to do in the trials after X?
Reliable: switch away (large ϕ)
Changing: sometimes explore (small λ)
Unreliable: select again (small ϕ)
Unchanging: always exploit (large λ)

Figure 1 In serial reversal learning experiments, associations are reliable, such that if an option is associated with a reward, it is rewarded during every trial (white background). However, the associations between options and the rewards change across trials (solid line). In such environments, individuals are expected to gain the most rewards if they update their associations quickly (large ϕ) to switch away from an option if it is no longer being rewarded, and if they have small sensitivities to their learned associations to continue to explore all options to check if associations have changed again (small λ). In contrast, in unchanging but unreliable environments, the probability that an option is rewarded stays constant across trials (dotted lines), but is closer to 50% (gray background). In such environments, individuals are expected to gain the most rewards if they build their associations as average across many trials (small ϕ), and have high sensitivities to learned associations to exploit the option with the highest association (large λ).

A recent development to infer the cognitive processes from the choices individuals make during reversal learning experiments are Bayesian reinforcement learning models (Bari et al., 2022; Chen et al., 2021; Danwitz et al., 2022; Deffner et al., 2020). These Bayesian models estimate the association-updating rate and the sensitivity to learned associations by modeling the likelihood of the subsequent choices individuals were observed to make based on how the underlying reward associations would predict each choice. The learning of information is reflected by the Rescorla-Wagner rule (Rescorla & Wagner, 1972), which includes the association-updating rate (the rate's label differs across authors) which weights the most recent information proportionally to the previously accumulated information for that cue (as a proportion, the rate can range between 0 and 1, see below for equation). The decision between different options is reflected by relative probabilities (Agrawal & Goyal, 2012; Danwitz et al., 2022; Daw et al., 2006), where the sensitivity to learned associations (again, the label can differ by author) modifies the relative difference in learned rewards to generate the probabilities to choose each option. A value of zero means individuals do not pay attention to their learned associations, but choose randomly, whereas increasingly larger values mean that individuals show strong biases in choice as soon as there are small differences in their learned associations. These static models have, for example, recently been used to indicate sex differences in exploration, with individuals of one sex on average showing lower sensitivities to learned associations (Breen & Deffner, 2023; Chen et al., 2021). More generally, they support the prediction that individuals with higher association-updating rates are more successful in reversal learning experiments (Bari et al., 2022; Danwitz et al., 2022). However, the application of these models has thus far however been static, rather than inferring whether and how individuals might adapt their strategies over time (Tello-Ramos et al., 2019). We need an understanding of the dynamic changes individuals might undergo in their processes to describe the improvement in performance that occurs through the serial reversal learning experiments to gain a full better of behavioral flexibility.

In serial reversal learning experiments, there are potentially three types of information individuals might pay attention to when adjusting their cognitive processes. First, in most reversal learning designs, there are two options differentiated by a cue, of which only one has the reward. Accordingly, exploring one option still provides information about the presence or absence of a reward in the other option. Second, linked to this, the association between a cue and a reward can be perfect such that one option is always rewarded during a reversal, but it could also be probabilistic, where both options contain a reward that differs in amount or frequency. In most animal experiments, the former is used where only one option contains a reward, so the association is perfect. In contrast, experiments in humans often introduce uncertainty in the associations by providing rewards only in a certain percentage of trials or by assigning rewards as draws from distributions (multi-armed bandit experiments). Third, reversals in the association between a cue and the reward can occur more or less frequently depending on the experimental design. At the extreme, when an individuals' previous experience suggests that rewards are only at one of the options during any given trial, associations are highly reliable, and changes are frequent, they might switch to an abstract rule, where the choice in the next trial is completely determined by the most recent experience (win-stay/lose-shift one-shot strategy, Mackintosh et al. (1968); Jang et al. (2015)). In experiments, such switches in strategy seem to appear in individuals living in the highly stable conditions of captivity (Metha et al., 2020; Rayburn-Reeves et al., 2013), especially if these individuals have been over-trained (Bartolo & Averbeck, 2020), and for highly reliable cues such as the location of a tree (Liu et al., 2016). However, most associations that animals have to learn however often have a probabilistic association between the cue and the outcome, the relationship between options is not necessarily straightforward, and the initial learning phase introduces a period of stability. Accordingly, most animals tested on serial reversal learning experiments do not show switches to abstract strategies, but rather improvements in their flexibility (Bitterman, 1975; Bond et al., 2007). In the classic two-choice serial reversal learning experiments given to animals, these improvements likely reflect how individuals adjust their association-updating rate and their sensitivity to learned associations depending on their experience of the frequency of the change and of the reliability of the association between the cue and the reward (Leimar et al., 2024; Neftci & Averbeck, 2019). Based on the static theoretical models, we would predict that individuals increase their association-updating rate because cues are highly reliable, and reduce their sensitivity to the learned associations because the option that is rewarded switches frequently.

Here, we applied and modified the Bayesian reinforcement learning models to data from our great-tailed grackle (*Quiscalus mexicanus*, hereafter grackle) research on behavioral flexibility to assess how the two parameters of the model interact and dynamically change to shape the behavior of individuals. We previously found that the model can predict the performance of grackles in a static reversal learning task with a single switch of a color preference using two differently colored tubes (one light gray and one dark gray Blaisdell et al., 2021). Here, we build on this work with additional data from another population (Logan et al., 2023a), where we conducted a flexibility manipulation using serial reversal learning. The serial reversal manipulation consisted of switching the rewarded color whenever individuals chose the rewarded option more than expected by chance (passing criterion of choosing correctly in 17 out of the last 20 trials), until their reversal speeds were consistently fast (reaching criterion in 50 trials or less in two consecutive reversals). We randomly assigned individuals to a manipulated group who received serial reversals, or to a control group who received one reversal and then a similar amount of experience in making choices between two yellow tubes that both contained the same reward (Logan et al., 2023a). After the reversal learning experiment, both the manipulated and the control grackles were given a different flexibility test using multi-option puzzle boxes. Grackles who experienced the serial reversal learning experiment subsequently also appeared to show improved behavioral flexibility in this different context because they required less time to switch to a new option to access a food reward when the previously learned option was blocked. They also solved a larger number of the four options presented in the multi-option puzzle boxes (Logan et al., 2023a).

RESEARCH QUESTIONS

1) Are the Bayesian reinforcement learning models sufficiently sensitive to detect changes that occur across the limited number of serial reversals that individuals participated in?
The models infer two parameters, the association updating rate ϕ and the sensitivity to learned associations λ, from the behavior of individuals, from across the traditional single outcome of the number of trials needed to reach the criterion. In theory, multiple combinations of the two parameters could lead to the same number of trials during a given reversal. Whether information from a single or few reversals is sufficient to infer these values for individuals at different time points throughout a serial reversal experiment has not been systematically addressed before. Therefore we used simulations to assess whether these models work on the samples that people usually work with. Determining the minimum number of choices per individual necessary to correctly infer their underlying parameters is necessary to reveal the dynamic changes in these parameters as individuals adjust their expectation of change throughout the serial reversal learning experiments and react accordingly.
Prediction 1: We predicted that the Bayesian reinforcement learning model can reliably infer the two parameters based on the choices individuals make during reversal learning, and that it can detect changes in these parameters that might occur during the series of reversals that individuals usually experience (4-6 reversals).

2) Is a strategy of high association-updating (ϕ) and low sensitivity to learned associations (λ) best to reduce errors in the serial reversal learning experiment?
Previous modeling work predicts that in situations in which changes are abrupt, but information is reliable, individuals learning in accordance with a Bayesian reinforcement model should show a high association-updating rate and a low sensitivity to learned associations (Breen & Deffner, 2023; Dunlap & Stephens, 2009). However, the modeled situations were abstract and the inferred optimal association updating rates and sensitivities were higher than what is usually observed in reversal learning experiments. Therefore, we perform simulations of the specific behavior exhibited in serial reversal learning experiments to assess how changes in the choices individuals make link to their ϕ and λ values. In addition, previous studies were only focused on the optimal values for the two parameters in different situations rather than looking at how ϕ and λ interact to explain variation among individuals. Therefore, we also use the simulations to determine whether one of the two parameters, ϕ or λ, might explain more of the variation in the number of trials individuals need to reach the criterion of choosing the correct option 17 out of 20 times during a reversal.
Prediction 2: We predicted that both ϕ and λ influence the performance of individuals in a reversal learning task, with higher ϕ values (faster learning with a higher association-updating rate) and lower λ values (more exploration with less sensitivity to learned associations) leading to individuals more quickly reaching the passing criterion after a reversal in the color of the rewarded option.

3) Which of the two parameters ϕ or λ explains more of the variation in the reversal learning experiment performance of the tested grackles?
Across both the manipulated and control grackles, we assessed whether variation in the number of trials an individual needs to reach the criterion in a given reversal is better explained by their inferred association updating rate or by their sensitivity to learned associations.
Prediction 3: We predicted that both ϕ and λ explain variation in the reversal performance of the grackles.

4) Which of the two parameters ϕ or λ changes more for the grackles that improved their performance through the serial reversal experiment?
If individuals learn the contingencies of the serial reversal experiment, they should be reducing their sensitivity to learned associations λ to explore the alternative option when rewards change, and increase their association-updating rate ϕ to quickly exploit the new reliably rewarded option.
Prediction 4: We predicted that individuals have higher ϕ and lower λ values during their last reversal of the serial reversal experiment than during their first reversal.

5) Are some individuals better than others at adapting to the serial reversals?
In previous work, we found that there are individual differences that persist throughout the experiment, with individuals who required fewer trials to solve the initial reversal also requiring fewer trials in the final reversal after their manipulation [mccune2023flexmanippeerj]. We could expect that these individual differences are guided by consistency in how individuals solve the reversal learning paradigm, meaning they are reflected in

individual consistency in ϕ and λ that persist through the serial reversal manipulation. In addition, it is not clear whether some grackles change their behavior more than others: for example, it could be that individuals who have a higher association-updating rate ϕ at the beginning of the experiment might also be better able to quickly change their behavior to match the particular conditions of the serial reversal learning experiment. Therefore, we also analyze whether the ϕ and λ values of individuals at the beginning predict how much they changed throughout the serial reversal learning experiment. Alternatively, given that the prediction for which sensitivity to learned association is best during a reversal (high sensitivity to stick to the learned associations) is different from the prediction for what is best right after a reversal (low sensitivity to explore the alternative option), the individuals who improved the most might end up with different strategies.
Prediction 5: We predicted that differences in ϕ and λ among individuals persist through the serial reversal learning experiment, or that they might even increase as some individuals change their learning more than others.

6) Can the ϕ or λ from the performance of the grackles during their final reversal predict variation in the performance on the multi-option puzzle boxes?

We previously found that grackles who needed fewer trials to reach the criterion in their last reversal on the color tube test were also better at performing on the two (plastic and wooden) multi-access boxes. This association could potentially be explained by either of the parameters underlying flexibility, or by an interaction between the parameters. With the multi-option puzzle boxes, grackles would be expected to gain more rewards if they quickly update their previously learned associations with the options (high ϕ) and/or if they are less sensitive to previously learned associations and instead continue to explore alternative options (low λ).
Prediction 6: We predicted that grackles that are more flexible, those who have a high ϕ and/or a low λ, have shorter latencies to attempt a new option and solve more options on the two multi-option puzzle boxes. Given that grackles are expected to change both their ϕ and their λ through the serial reversal (see prediction 2), we also explore whether the relationship between ϕ or λ and the performance on the multi-access boxes is non-linear.

METHODS

The Bayesian reinforcement learning model

We used the version of the Bayesian model that was developed in Blaisdell et al. (2021) and modified in Logan CJ et al. (2023) (see their Analysis Plan > "Flexibility analysis" for model specifications and validation). This model uses data from every trial of reversal learning (rather than only using the total number of trials to pass criterion) and represents behavioral flexibility using two parameters: the association-updating rate (ϕ) and the sensitivity to learned associations (λ). The model transforms the series of choices each grackle made based on two equations to estimate the most likely ϕ and λ that generated the observed behavior.

Equation 1 (attraction and ϕ): $A_{j,i,t+1}{=}(1{-}\phi_j)A_{j,i,t}{+}\phi_j\ \pi_{j,i,t}$

Equation 1 estimates how the associations A that individual j forms between the two different options (i {1, 2} and their expected rewards change from one trial to the next (time t+1) as a function of their previously formed associations $A_{j,i,t}$ (how preferable option i is to grackle j at time t) and recently experienced payoff π (in our case, $\pi = 1$ when they chose the correct option and received a reward in a given trial, and 0 when they chose the unrewarded option). The parameter ϕ_j modifies how much individual j updates its associations based on its most recent experience. The higher the value of ϕ_j, the faster the individual updates its associations, paying more attention to recent experiences, whereas when ϕ_j is lower, a grackle's associations reflect averages across many trials. Association scores thus reflect the accumulated learning history up to this point. The association with the option that is not explored in a given trial remains unchanged. At the beginning of the experiment, we assume that individuals have the same low association between both options and rewards ($A_{j,1} = A_{j,2} = 0.1$).

Equation 2 (choice and λ): $P(j,i)_{t+1}{=}\dfrac{exp(\lambda_j A_{j,i,t})}{\sum_{i=1}^{2} exp(\lambda_j A_{j,i,t})}$

Equation 2 expresses the probability P that an individual j chooses option i in the next trial, t+1, based on their learned associations of the two options with rewards. The parameter λ_j represents the sensitivity of a given grackle j to how different its associations to the two options are. As λ_j gets larger, choices become more deterministic and individuals consistently choose the option with the higher association even if associations are very similar. As λ_j gets smaller, choices become more exploratory, with individuals choosing randomly between the two options independently of their learned associations if λ_j is 0.

Equation 2 expresses the probability *P that an individual j chooses option i in the next trial, t+1, based on the attractions. The parameter λ_j represents the rate of deviating from learned attractions of an individual. It controls how sensitive choices are to differences in attraction scores. As λ_j gets larger, choices become more deterministic and individuals consistently choose the option with the higher attraction even if attractions are very similar, as λ_j gets smaller, choices become more exploratory (random choice independent of the attractions if λ_j=0).

We implemented the Bayesian reinforcement learning model in the statistical language Stan (Stan Development Team, 2023), calling the model and analyzing its output in *R* (version 4.2.2) (R Core Team, 2023). The model takes the full series of choices individuals make (which of the two options did they choose, which option was rewarded, did they make the correct choice) across all their trials to find the ϕ and λ values that best fit these choices given the two equations: whether or not individuals chose the rewarded option was reflected as a categorical likelihood (yes or no) with probability P as estimated from equation 2, before updating the associations using equation 1. The model was fit across all choices, with individual ϕ and λ values estimated as varying effects. In the model, ϕ is estimated on the logit-scale to force the values to be positive before being converted back for equation 1 to update the associations, and λ is estimated on the log-scale to account for the exponentiation that occurs in equation 2. We set the priors for ϕ and λ to come from a normal distribution with a mean of zero and a standard deviation of one. We set the initial associations with both options for all individuals at the beginning of the experiment to 0.1 to indicate that they do not have an initial preference for either option but are likely to be somewhat curious about exploring the tubes because they underwent habituation with a differently colored tube (see below). For estimations at the end of the serial reversal learning experiment, we set the association with the option that was rewarded before the switch to 0.7 and to the option that was previously not rewarded to 0.1. Note that when applying equation 1 in the context of the reversal learning experiment as most commonly used, where there are only rewards (positive association) or no rewards (zero association) but no punishment (negative association), associations can never reach zero because they change proportionally.

We used functions in the package "posterior" (Vehtari et al., 2021) to draw 4000 samples from the posterior (the default in the functions). We report the estimates for ϕ and λ for each individual (simulated or grackle) as the mean from these samples from the posterior. For the subsequent analyses where the estimated ϕ and λ values were response or predictor variables, we ran the analyses both with the single mean per individual as well as looping over the full 4000 samples from the posterior to reflect the uncertainty in the estimates. The analyses with the samples from the posterior provided the same estimates as the analyses with the single mean values, though with larger confidence estimates because of the increased uncertainty. In the results, we report the estimates from the analyses with the mean values. The estimates with the samples from the posterior can be found in the code in the rmd file at the repository. In analyses where ϕ and λ are predictor variables, we standardized the values that went into each analysis (either the means, or the respective samples from the posterior) by subtracting the average from each value and then dividing by the standard deviation. We did this to define the priors for the relationship on a more standard scale and to be able to more directly compare their respective influence on the outcome variable.

We also used the two equations analytically to more directly make predictions about how a specific ϕ and λ would influence the choices individuals make during the reversal learning. To derive the learning curves for individuals with different ϕ and λ, we incorporated the dynamic aspect of change over time by inserting the probabilities of choosing either the rewarded or the non-rewarded option from time t-1 as the likelihood for the changes in associations at time t:

Equation 3a (dynamic association): $AssociationRewarded_{t+1}$ = ((1-ϕ) * $AssociationRewarded_t$ + ϕ * Reward) * $ProbabilityRewarded_t$ + (1-$ProbabilityRewarded_t$) * $AssociationRewarded_t$

Equation 3b (dynamic association): $AssociationNonrewarded_{t+1}$ = (1-$ProbabilityRewarded_t$) * (1-ϕ) * $AssociationNonrewarded_t$ + $ProbabilityRewarded_t$ + (1-$ProbabilityRewarded_t$) * $AssociationNonrewarded_t$

1) Using simulations to determine whether the Bayesian serial reinforcement learning models have sufficient power to detect changes through the serial reversal learning experiment

We re-analyzed data we previously simulated for power analyses to estimate sample sizes for population comparisons (Logan CJ et al., 2023). In brief, we simulated 20 individuals each from 32 different populations (640 individuals). The ϕ and λ values for each individual were drawn from a distribution representing that population, with different mean ϕ (8 different means) and mean λ (4 different values) for each population (32 populations as the combination of each ϕ and λ). The range for ϕ and λ values assigned to the artificial individuals in the simulations were based on the previous analysis of the single reversal data from grackles in a different population (Santa Barbara, California, USA, Blaisdell et al. (2021)) to reflect the likely expected behavior. Based on their assigned ϕ and λ values, each individual was simulated to pass first through the initial association learning phase and, after they reached criterion (chose the correct option 17 out of the last 20 times), the rewarded option switched and simulated individuals went through the reversal learning phase until they again reached criterion. Each choice that each individual made was simulated consecutively, updating their internal associations with the two options based on their ϕ values and setting the probability of their next choice based on how their λ value weighted their associations to the two options. We excluded simulated individuals from the further analyses if they did not reach criterion either during the initial association or the reversal within 300 trials, the maximum that was also set for the experiments with the grackles.

We ran the Bayesian reinforcement learning model on these simulated data to understand the minimum number of choices per individual that would be necessary to recover the association-updating rate ϕ and the sensitivity to learned association λ values assigned to each individual.

To determine whether the Bayesian reinforcement learning model can accurately recover the simulated ϕ and λ values from limited data, we applied the model first to only the choices from the initial association learning phase, next to only the choices from the first reversal learning phase, and finally from both phases combined. To estimate whether the Bayesian reinforcement learning model can recover the simulated ϕ and λ values without bias from either of the single or from the combined datasets, we correlated the estimated values with the values individuals were initially assigned:

Assigned value of ϕ or λ ~ Normal(μ, σ)
μ = a + b * Estimated value of ϕ or λ
a ~ Normal(0,0.1)
b ~ Normal(1,1)
σ ~ Exponential(1)

A slope *b* between the assigned and estimated values close to 1 would indicate that the estimated values matched the assigned values.

This, and all following statistical models, were implemented using functions of the package 'rethinking' (McElreath, 2020) in R to estimate the association with stan. Following the social convention set in (McElreath, 2020), we report the mean estimate and the 89% confidence interval from the posterior estimate from these models. For each model, we ran four chains with 10,000 iterations each (half of which were burn-in, and half samples for the posterior). We checked that the number of effective samples was sufficiently high and evenly distributed across parameters such that auto-correlation did not influence the estimates. We also confirmed that in all cases the Gelman-Rubin convergence diagnostic, R, was 1.01 or smaller indicating that the chains had converged on the final estimates (Gelman & Rubin, 1995). In all cases, we also linked the model inferences back to the distribution of the raw data to confirm that the estimated predictions matched the observed patterns.

2) Using simulations to determine whether variation in ϕ or in λ has a stronger influence on the number of trials individuals might need to reach criterion in reversal learning experiments

We determined how the ϕ and λ values that were assigned to the simulated individuals influenced their performance in the reversal learning trials, building a regression model to determine which of the two parameters had a more direct influence on the number of trials individuals needed to reach criterion. We assumed that the number of trials followed a Poisson distribution because the number of trials to reach criterion is a count that is bounded at smaller numbers (individuals need at least 20 trials to reach the criterion), with a log-linear link, because we expect there are diminishing influences of further increases in ϕ or λ.

Number of trials to reverse ~ Poisson(μ)
log μ = a + b * ϕ + c * λ
a ~ Normal(4.5,1)
b ~ Normal(0,1)
c ~ Normal(0,1)

The prior for the intercept *a* was based on the average number of trials (90) grackles in Santa Barbara were observed to need to reach the criterion during the reversal (mean of 4.5 is equal to logarithm of 90, standard deviation set to 1 to constrain the estimate to the range observed across individuals). The priors for the relationships *b* and *c* with ϕ and λ were centered on zero, indicating that, a-priori, we do not bias it toward a relationship.

3) Estimating ϕ and λ from the observed reversal learning performances of great-tailed grackles to determine which has more influence on variation in how many trials individuals needed to reach the passing criterion

The collection of the great-tailed grackle data is described in detail in (Logan et al., 2023a). The data collection was based on our preregistration that received in principle acceptance at PCI Ecology (Coulon, 2023). All of the analyses reported here were not part of the original preregistration.

The research on the great-tailed grackles followed established ethical guidelines for the involvement and treatment of animals in experiments and received institutional approval prior to conducting the study (US Fish and Wildlife Service scientific collecting permit number MB76700A-0,1,2; US Geological Survey Bird Banding Laboratory federal bird banding permit number 23872; Arizona Game and Fish Department scientific collecting license number SP594338 [2017], SP606267 [2018], and SP639866 [2019]; California Department of Fish and Wildlife scientific collecting permit number S-192100001-19210-001; Institutional Animal Care and Use Committee at Arizona State University protocol number 17-1594R; Institutional Animal Care and Use Committee at the University of California Santa Barbara protocol number 958; University of Cambridge ethical review process non-regulated use of animals in scientific procedures: zoo4/17 [2017]).

The data we use here were published as part of an earlier article (Logan et al., 2023b) and are available at the Knowledge Network for Biocomplexity's data repository: https://knb.ecoinformatics.org/view/corina_logan.84.42.

Great-tailed grackles were caught in the wild in Tempe, Arizona, USA for individual identification (colored leg bands in unique combinations), and brought temporarily into aviaries for testing, before being released back to the wild. After training individuals to gain food from a yellow-colored tube, individuals then participated in the reversal learning tasks. A subset of individuals was part of the control group, where they learned the association of the reward with one color before experiencing one reversal to learn that the other color is rewarded (initial reward option was randomly assigned to either a dark-gray or a light-gray tube). The rewarded option was switched when grackles passed the criterion of choosing the rewarded option during 17 of the most recent 20 trials. This criterion was set based on earlier serial reversal learning studies, and is based on the chi-square test which indicates that 17 out of 20 represents a significant association. With this criterion, individuals can be assumed to have learned the association between the cue and the reward (Logan et al., 2022). After their single reversal, the 11 control grackles participated in a number of trials with two identically colored tubes (yellow) which both contained a reward. This matched their general experiment participation to that of the manipulated group. The other subset of 8 individuals in the manipulated group went through a series of reversals until they reached the criterion of having formed an association (17 out of 20 choices correct) in less than 50 trials in two consecutive reversals. The individuals in the manipulated group needed between 6-8 reversals to consistently reach this threshold, with the number of reversals not being linked to their performance at the beginning or at the end of the experiment.

We fit the Bayesian reinforcement learning model to the data of both the control and the manipulated grackles. Based on the simulation results indicating that the minimum sample required for accurate estimation are two learning phases across one switch (see below), we fit the model first to only the choices from the initial association learning phase and the first reversal learning phase for both control and manipulated individuals. For the control grackles, these estimated ϕ and λ values also reflect their behavioral flexibility at the end of the reversal learning experiment. For the manipulated grackles, we additionally calculated ϕ and λ separately for their final two reversals at the end of the manipulation to infer the potential changes in the parameters . We fit the same regression model as with the simulated data to determine how ϕ and λ link to the number of trials grackles needed during their reversals.

4) Comparing ϕ and λ from the beginning and the end of the observed serial reversal learning performances to assess which changes more as grackles improve their performance

For the subset of grackles that were part of the manipulated group, we calculated how much their ϕ and λ changed from their first to their last reversal.

ϕ or λ ~ Normal(μ , σ)
$\mu = a_{bird} + b_{bird}$ * reversal $[a_{bird}, b_{bird}]$ ~ MVNormal([a,b],S)
S = (δ_{bird} ,0) Rho (δ_{bird} ,0)
Rho ~ LKJcorr(2)
a ~ Normal(5,2)
b ~ Normal(-1,0.5)
δ_{bird} ~ Exponential(1)
σ ~ Exponential(1)

where each grackle has two ϕ or λ values, one from the beginning ('reversal' equals 1) and one from the end of the serial reversal experiment ('reversal' equals 2). We assume that there are individual differences that persist through the experiment (intercept a_{bird}) and that how much individuals change might also depend on their values at the beginning (multi-normal matrix correlation between the bird specific intercepts a and the bird specific changes between the reversals b).

We also fit a model to assess whether how much individuals improved in the number of trials from their first to their last reversal was linked more to their change in ϕ or to their change in λ.

Improvement in number of trials ~ Normal(μ, σ)
μ = a + b * change in ϕ + c * change in λ a ~ Normal(40, 10)
b ~ Normal(0, 10)
c ~ Normal(0, 10)
σ ~ Exponential(1)

where *Improvement in the number of trials* is the difference in the number of trials between the first and the last reversal and *change in* ϕ and *change in* λ are the respective differences in these parameters between the beginning and the end of the serial reversal experiment.

5) Calculating whether individual differences in ϕ and λ persist throughout the serial reversal learning experiment and whether individuals differ in how much they change throughout the experiment

We checked whether the ϕ or λ values of individuals at the beginning (*first*) was associated with how much they changed (*change*, difference in values between beginning or end) or with the values they had at the end (*last*).

ϕ *change* or λ *change* ~ Normal(μ , σ)
μ = a + b * ϕ *first* or λ *first*
a ~ Normal(0,1)
b ~ Normal(0,1)
σ ~ Exponential(1)

ϕ *last* or λ *last* ~ Normal(μ , σ) μ = a + b * ϕ *first* or λ *first*
a ~ Normal(0,1)
b ~ Normal(0,1)
σ ~ Exponential(1)

In addition, we assessed whether grackles at the end show the potential trade-off between ϕ and λ that could be expected in the serial reversal experiment.

ϕ *last* ~ Normal(μ , σ)
μ = a + b * λ *last*
a ~ Normal(0,1)
b ~ Normal(0,1)
σ ~ Exponential(1)

6) Linking ϕ and λ from the observed serial reversal learning performances to the performance on the multi-access boxes

After the individuals had completed the reversal learning experiment, they were provided access to two multi-access puzzle boxes, one made of wood and one made of plastic. The two boxes were designed with slight differences to explore how general the performance of the grackles was. The wooden box was made from a natural log, so was more representative of something the grackles might encounter in the wild. In addition, while both boxes had 4 possible ways (options) to access food, the four options on the wooden box were distinct compartments, each containing rewards, while the four options on the plastic box all led to the same reward. Grackles were tested sequentially on both boxes, where individuals could initially explore all options. After proficiency at an option was achieved (gaining food from this locus three times in a row), this option became non-functional by closing access to the option, and then the latency of the grackle to switch to attempting a different option was measured. If they again successfully solved another option, this second options was also made non-functional, and so on. The outcome measures for each individual with each box were the average latency it took to switch to a new option and the total number of options they successfully solved. For details see (Logan et al., 2023a).

We modified the models in the original article (Logan et al., 2023a) that linked performance on the serial reversal learning tasks to performance on the multi-access boxes, replacing the previously used independent variable of number of trials needed to reach criterion in the last reversal with the estimated ϕ and λ values from the last two reversals (manipulated grackles) or the initial discrimination and the first reversal (control grackles) (see below for explanation of these choices). With our expectation that ϕ and λ could be negatively correlated , we realized that grackles might be using different strategies when facing a situation in which cues change: some grackles might quickly discard previous information and rely on what they recently experienced (high ϕ and low λ), or they might rely on earlier information and continue to explore other options (low ϕ and high λ). Accordingly, we assumed that there also might be non-linear, U-shaped relationships between ϕ and/or λ and the performance on the multi-access box. For the number of options solved, we fit a binomial model with a logit link:

options solved ~ Binomial(4, p)
logit(p) ~ a + b * ϕ + c * ϕ^2 + d * λ + e * λ^2
a ~ dnorm(1, 1)
b ~ dnorm(0, 1)
c ~ dnorm(0, 1)
d ~ dnorm(0, 1)
e ~ dnorm(0, 1)

where *options solved* is the number of options solved on the multi-access puzzle box, 4 is the total number of options, *p* is the probability of solving any one option across the whole experiment, *a* is the intercept, *b* is the expected linear amount of change in *options solved* for every one unit change in ϕ in the reversal learning experiments, *c* is the expected non-linear amount of change in *options solved* for every one unit change in ϕ squared, *d* the expected linear amount of change for changes in λ, and *e* the expected non-linear amount of change for changes in λ squared.

For the average latency to attempt a new option on the multi-access puzzle box as it relates to trials to reverse (both are measures of flexibility), we fit a Gamma-Poisson model with a log-link:

latency ~ Gamma-Poisson(μ_i, σ)
log(μ_i) ~ a + b * ϕ + c * ϕ^2 + d * λ + e * λ^2
a ~ dnorm(1, 1)
b ~ dnorm(0, 1)
c ~ dnorm(0, 1)
d ~ dnorm(0, 1)
e ~ dnorm(0, 1)
σ ~ Exponential(1)

latency is the average latency to attempt a new option on the multi-access box, μ_i is the rate (probability of attempting an option in each second) per grackles (and we take the log of it to make sure it is always positive; grackles with a higher rate have a smaller latency), σ is the dispersion of the rates across grackles, *a* is the intercept, *b* is the expected linear amount of change in latency for every one unit change in ϕ , *c* is the expected non-linear amount of change in latency for every one unit change in ϕ squared, *d* the expected linear amount of change for changes in λ, and *e* the expected non-linear amount of change for changes in λ squared.

RESULTS

1) Power of the Bayesian reinforcement learning model to detect short-term changes in the association-updating rate ϕ and the sensitivity to learned associations λ

Applying the Bayesian reinforcement learning model to simulated data from only a single phase (initial association or first reversal) revealed that, while the model recovered the differences among individuals, the estimated ϕ and λ values did not match those the individuals had been assigned (Figure 2 shows the relationship between the assigned and estimated ϕ values when estimated from only the first reversal as an illustration). We realized that ϕ and λ values were consistently shifted,with the Bayesian estimation adjusting both parameters towards the mean and away from extreme values. Simulated individuals who were assigned large λ values were estimated to have a smaller λ values but in turn estimated to have ϕ values such that they would reach criterion in a similar number of trials because while the model assumed that they were more exploratory the model also assumed that they updated their associations more quickly. Similarly, individuals with large assigned ϕ values were estimated to have smaller ϕ values, but in turn were estimated to have larger λ values than those λ they were assigned. Because the estimation from a single reversal did not accurately recover large values for either parameter, both the estimated ϕ values (slope of the correlation between the estimated and the assigned ϕ +0.15, confidence interval +0.06 to +0.23, n=626 simulated individuals) and the estimated λ values (slope of the correlation between the estimated and the assigned λ +0.58, confidence interval +0.48 to +0.68, n=626 simulated individuals) were underestimates of the assigned values. In addition, this shift means that, even though simulated individuals were assigned ϕ and λ values randomly from across all possible combinations, the estimated values showed a strong positive correlation as the model had to make up the shifts in estimates of one parameter through shifting the estimate of the other parameter (slope of the correlation between the estimated λ and estimated ϕ values +505, confidence interval +435 to +570, n=626 simulated individuals).

In contrast, when we combined data from across the initial discrimination learning and the first reversal, the model accurately recovered the ϕ and λ values that the simulated individuals had been assigned (slope of the correlation between the estimated and the assigned ϕ +0.96, confidence interval +0.70 to +1.21, n=626 simulated individuals; slope of the correlation between the estimated and the assigned λ +0.98, confidence interval +0.92 to +1.05, n=626 simulated individuals) (Figure 2). While different combinations of ϕ and λ could potentially explain the series of choices during a single phase (initial discrimination and single reversal), these different combinations lead to different assumptions about how an individual would behave right after a reversal when the reward is switched to the alternative option, making it possible to infer the assigned value when combining behavioral choices from two phases (initial learning plus first reversal, or two subsequent reversals).

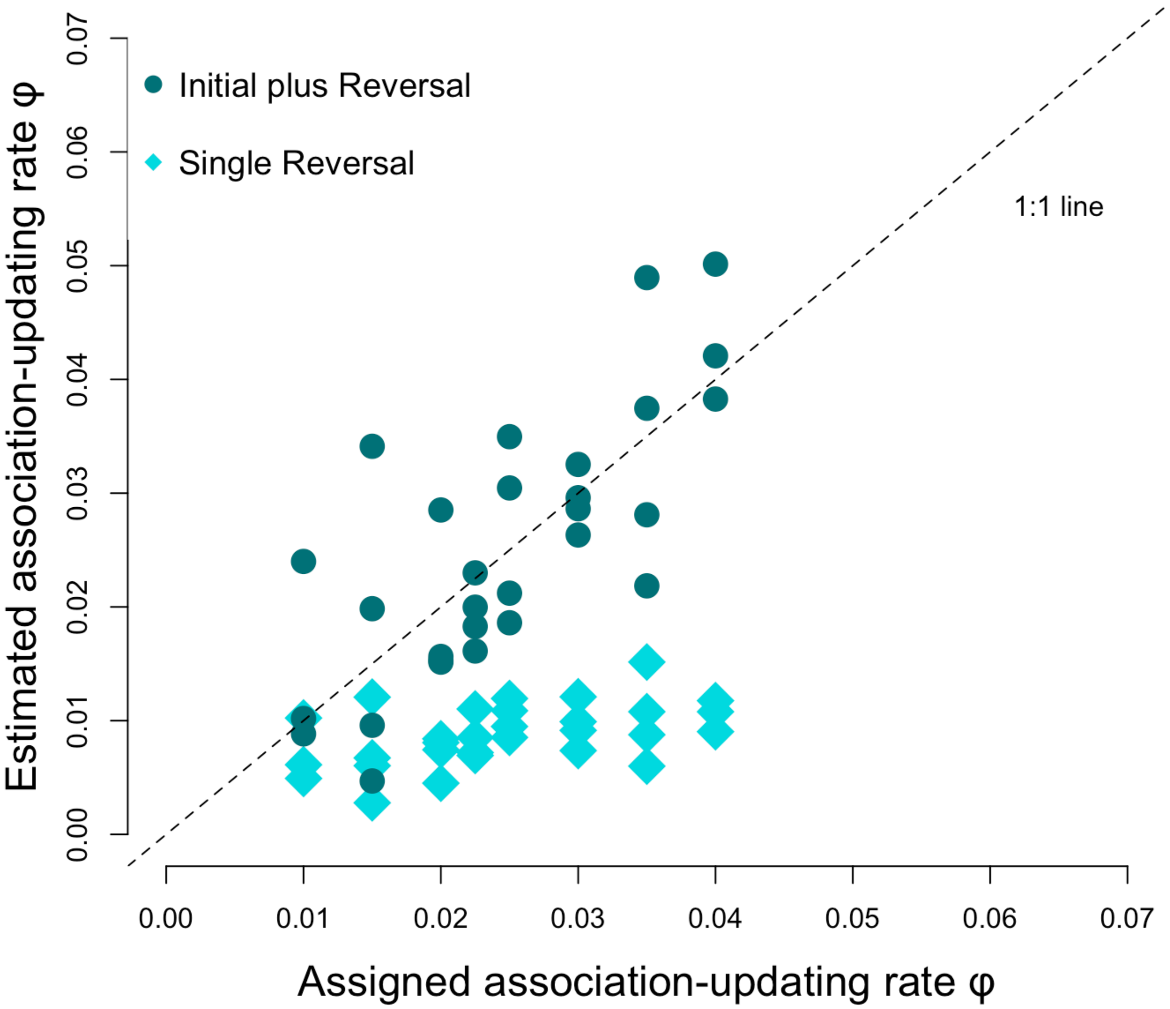

Figure 2: The ϕ values estimated by the model based on the choices made by 30 of the simulated individuals (y-axis) versus the ϕ values assigned to them (x-axis). Individuals were assigned a range of ϕ values, their choices were simulated and these values were used to back-estimate the ϕ. When ϕ was estimated based on the choices made only during the first reversal, the estimates were consistently lower than the assigned values, particularly for large ϕ values (lightblue squares). However, when ϕ was estimated based on the choices made during the initial association and the first reversal, the estimates were close to the assigned values (darkgreen circles). Patterns are similar for the relationship between the estimated and assigned λ values, and when ϕ and λ are estimated only from the trials during the initial association learning. Lines around the points indicate the confidence intervals of the estimated values.

2) Predicted role of ϕ and λ on performance in the serial reversal learning task based on simulations

In terms of the influence of the two parameters ϕ and λ on the number of trials grackles needed to reverse a color preference, the ϕ values assigned to simulated individuals had a stronger influence than the λ values (estimated association of number of trials with standardized values of ϕ: -0.23, confidence interval: -0.24 to -0.23; with standardized values of λ: -0.17, confidence interval: -0.18 to -0.16, n = 626 simulated individuals).

In line with the prediction, there was a linear negative relationship between ϕ and the number of trials to reverse, with simulated individuals needing fewer trials the more they updated their association based on their most recent experience. There also was, as predicted, an overall negative relationship between λ and the number of trials to reverse. Individuals generally needed few trials to reach the criterion if they were assigned a high λ value because they acted even on small differences in their learned associations. However, while individuals with small λ values can show large numbers of 150 or more trials to reach criterion because they are not sensitive to the differences in their learned associations, individuals with small λ values can also reach the criterion in small numbers of trials if they simultaneously quickly update their association because of their high ϕ values (Figure 3).

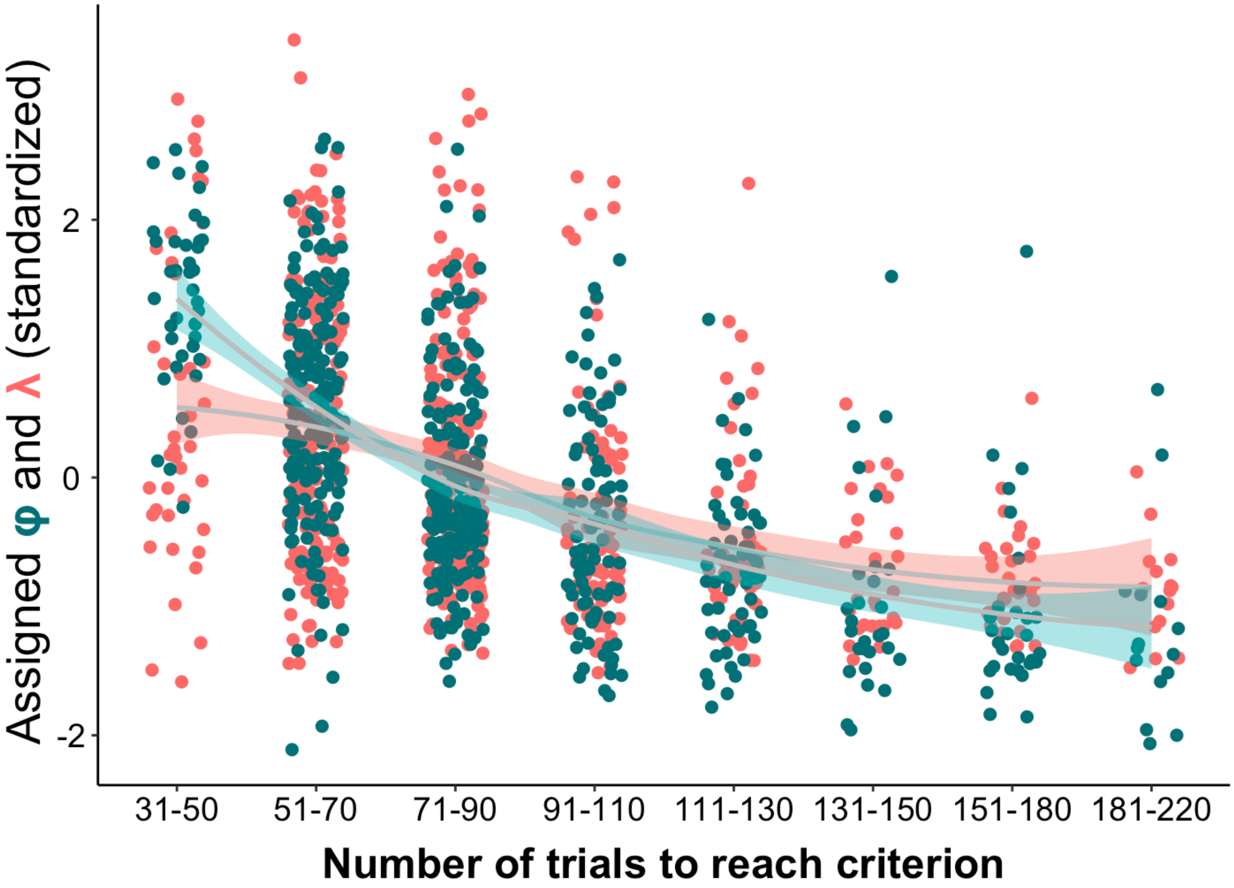

Figure 3. In the simulations, the ϕ values assigned to individuals (green) had a larger influence on the number of trials these individuals needed to reverse than their λ values (red). In general, individuals needed fewer trials to reverse if they had larger ϕ and λ values. However, relatively small λ values could be found across the range of reversal performances, whereas there was a more clear distinction with ϕ values (shaded lines represent confidence intervals of the estimated relationship for these data). ϕ and The to reach criterion are grouped into discrete blocks for easier illustration, but the analyses were performed on the raw values for each individual.

We performed an analytical assessment of this likely trade-off between the association updating rate ϕ and the sensitivity to the learned associations λ to identify the range of values we could expect in the serial reversal learning experiment. We assigned an hypothetical individual one of nine potential ϕ values in the range of 0.02 to 0.10 (steps differ by 0.01), assumed that this individual initially had the same association of the reward with both of the options (associations of 0.10 for light gray and 0.10 for dark gray), and assumed that this individual would choose each options 10 times during its first 20 trials. We calculated the associations to both options after the first 20 trials given the respective ϕ (e.g. with a ϕ of 0.10, the association with the rewarded option increases to 0.69 while the association with the unrewarded option declines to 0.03). Based on the differences in the two associations, we estimated the λ value necessary for individuals to choose the rewarded option 85% in the next 20 trials (to reach the criterion of choosing the rewarded option in 17 out of 20 trials). We detected a clear negative, and exponential, trade-off between the necessary ϕ and λ values to reach the criterion (Figure 4): individuals with the highest ϕ value of 0.10 only need a λ of 2.7 to reach the criterion, whereas individuals with a ϕ value of 0.02 need a λ of 9.5. This trade-off, where individuals can reach criterion during a reversal in few trials by either quickly updating their associations or by being highly sensitive to even small differences in their learned associations, means that in the serial reversal learning experiment individuals are expected to choose a strategy from across this range, and that doing so means they can also react to the sudden reversals in the reward location. In the serial reversal learning experiments, individuals will be able to reach the criterion more quickly during subsequent trials if they have, as predicted, a high ϕ and a low λ value. First, even if individuals were to choose randomly during the first trials after a reversal, individuals with a low ϕ need exponentially more trials to reverse their bias in associations between the two options. If an individual after one reversal has an association to the no longer rewarded option of 0.70 and to the now rewarded option of 0.10, with a ϕ of 0.02 it will take 48 random trials until their association to the now rewarded options is higher than their association to the no longer rewarded option. In contrast, with a ϕ of 0.08 it will only take them 10 trials. Second, individuals with a high λ value will keep on choosing the previously rewarded option in almost all of their trials until this switch in associations occurs, further delaying the learning of the new associations. Individuals that have an association of 0.70 with the no longer rewarded option and 0.10 with the now rewarded option will choose the now rewarded option in 14% of cases if their λ is only 3, but only in 0.8% of cases if their λ is 8.

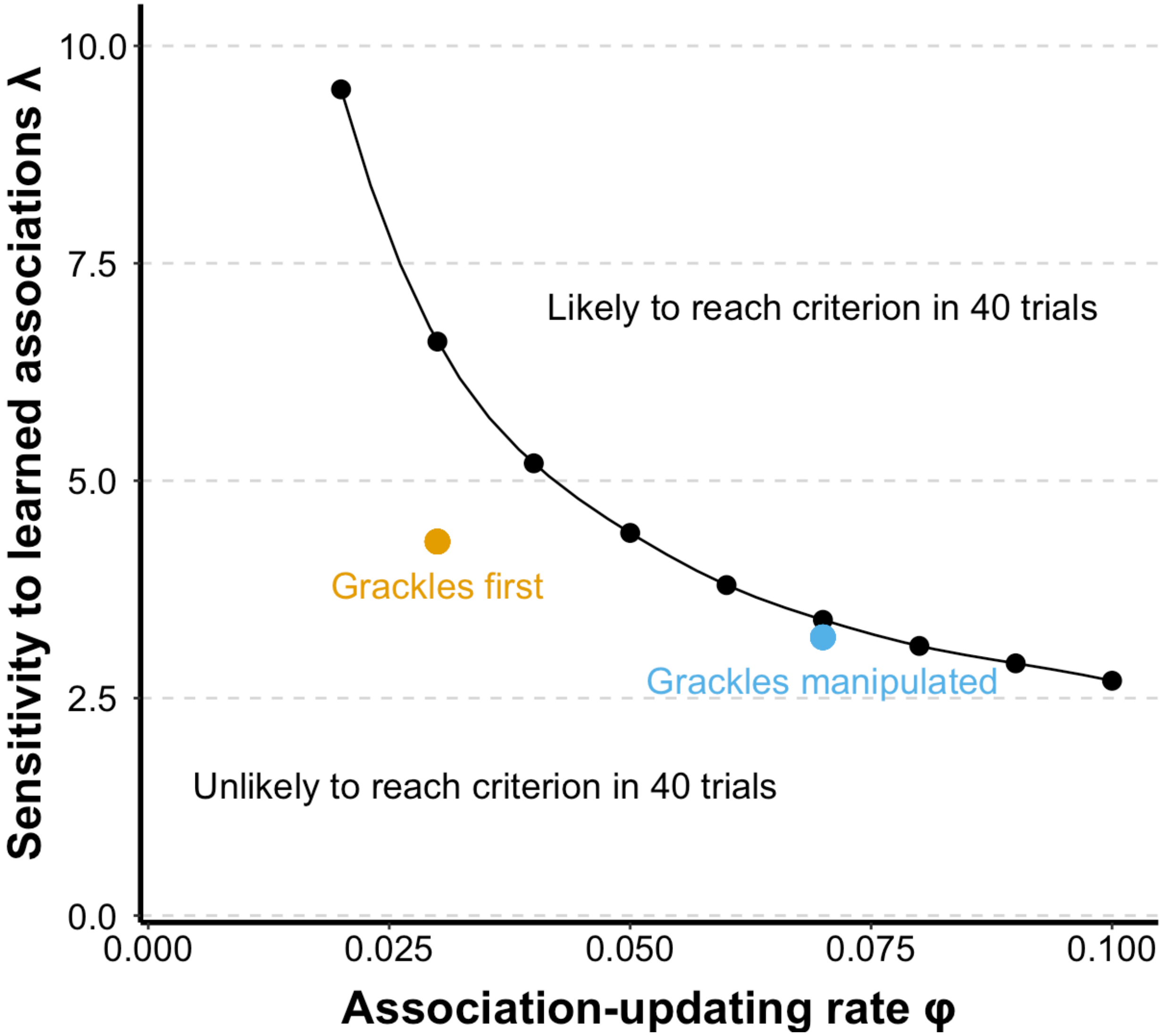

Figure 4. Individuals are more likely to reach the criterion of choosing the correct option 17 out of 20 times during the reversal trials if they update their associations quickly (high ϕ) and/or are sensitive to even small differences in their learned associations (high λ), because, during a reversal, recent information accurately predicts where the reward can be found. The figure shows this trade-off of individuals needing either high ϕ or high λ values to reach the criterion in a hypothetical situation where all individuals reach the criterion in 40 trials. This also means that if an individual has, for example, a high ϕ, their λ value becomes less important for reaching the criterion quickly. In this example, individuals with a ϕ of 0.10 will reach the criterion in 40 trials if their λ is at least 3.3. The figure also shows the median ϕ and λ values estimated for the grackles during their first reversal (yellow) when they needed about 70 trials to reach criterion and for the manipulated individuals during their last reversal (blue) when they did needed about 40 trials to reach criterion. During the manipulation, grackles increased their ϕ to become efficient at gaining the reward and reaching the criterion, despite the concordant decline in λ.

3) Observed role of ϕ and λ on performance of grackles in the reversal learning task

For the grackles, we estimated ϕ and λ after the first reversal for all individuals, and additionally after the final reversal for the individuals who experienced the serial reversal learning experiment. The findings from the simulated data indicated that λ and ϕ can only be estimated accurately when calculated across at least one switch. In the simulation, we could combine the performance of individuals during the initial learning with the first reversal to estimate the parameters because the behavior during those two phases in the simulations was determined in the same way by the ϕ and λ values that individuals were assigned. We determined that we can also combine the first two phases for the grackles, because we found that the performance of the great-tailed grackles during the initial learning and the first reversal learning is correlated, with grackles needing about 28 trials more to reach criterion during the first reversal than they needed during the initial association learning (estimate of the association between number of trials in initial learning and first reversal +1.61, confidence interval +1.53 to +1.69, n=19 grackles). Therefore, we estimated ϕ and λ for the great-tailed grackles based on their performance in the initial discrimination plus first reversal, and for the manipulated grackles additionally based on their performance in the final two reversals. The inferred ϕ values for the grackles in Arizona range between 0.01 and 0.10, and the λ values between 2.1 and 6.5 (Figure 5).

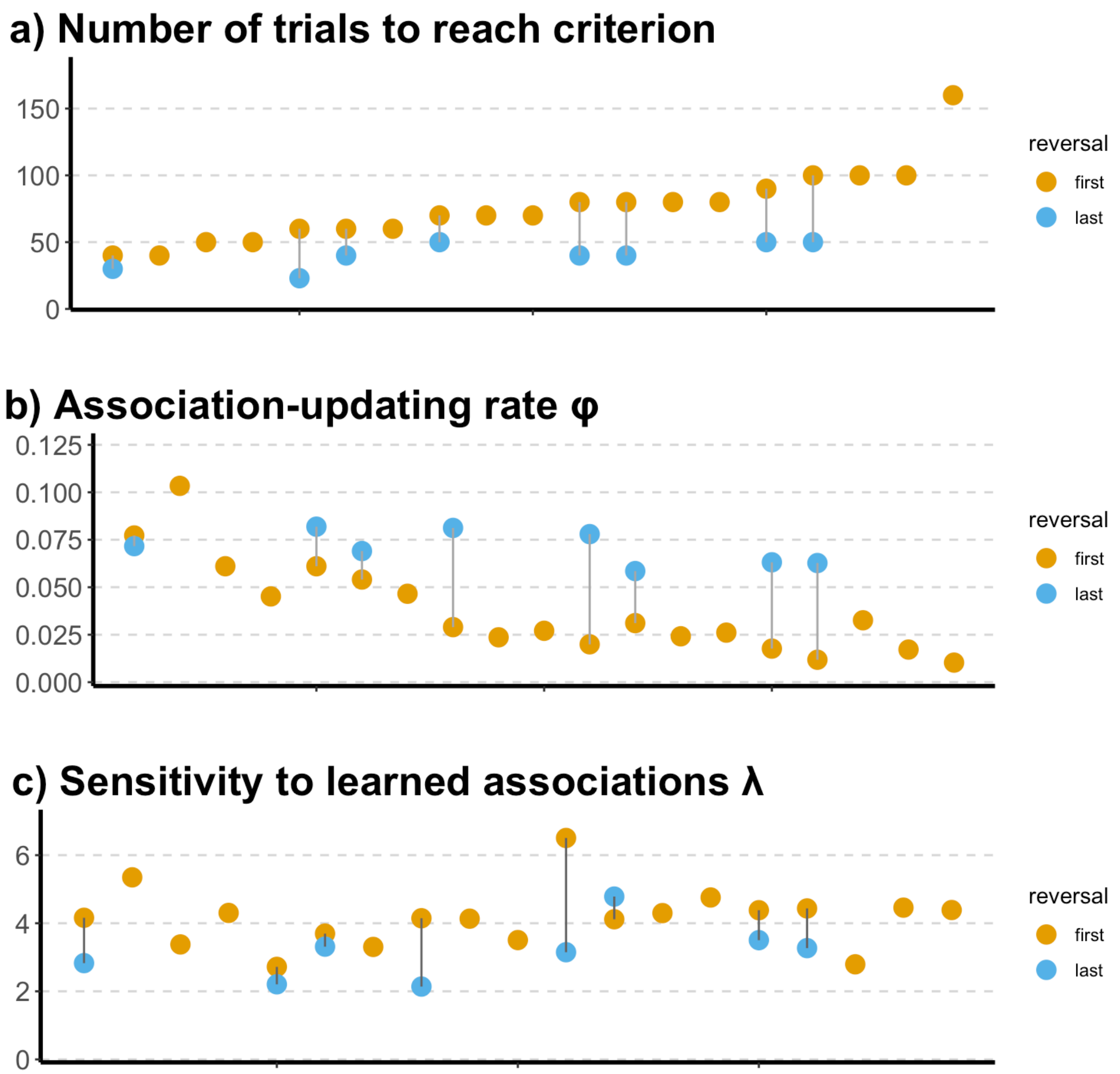

Figure 5. Comparisons of the different measures of ability in the reversal task for each of the 19 great-tailed grackles. The figure shows a) the number of trials to pass criterion for the first reversal (orange; all grackles) and the last reversal (blue; only manipulated grackles); b) the ϕ values reflecting the rate of updating associations with the two options inferred from the initial discrimination and first reversal (orange; all grackles) and from the last two reversals (blue; manipulated grackles); and c) the λ values reflecting the sensitivity to the learned associations inferred from the initial discrimination and first reversal (orange; all grackles) and from the last two reversals (blue; manipulated grackles). Individual grackles have the same position along the x-axis in all three panels. Grackles that needed fewer trials to reverse their preference generally had higher ϕ values, whereas λ appeared unrelated to the number of trials grackles needed during the first reversal. For the manipulated grackles, their ϕ values changed more consistently than their λ values, and the ϕ values of the manipulated individuals were generally higher than those observed in the control individuals, while their λ values remained within the range observed in the control group.

For the 19 grackles that finished the initial learning and the first reversal, only their ϕ, but not their λ, predicted the number of trials they needed to reach criterion during their first reversal (mean estimate of correlation between number of trials and: standardized ϕ: -20.69, confidence interval -26.17 to -15.13; standardized λ: -0.22, confidence interval -5.66 to 5.26, n=19 grackles)(Figure 6). A grackle with a 0.01 higher ϕ than another individual needed about 10 fewer trials to reach the criterion. The slope between ϕ and the number of trials for the grackles was essentially identical to that observed in the simulations (-21.21

vs -20.48, Figure 6). The number of trials grackles needed to reach the criterion given their ϕ values fell right into the range observed in the relationship between the ϕ and the number of trials observed among the simulated individuals (Figure 6) Even though the 8 manipulated grackles also appeared to need slightly fewer trials to reach criterion in their final two reversals if they had a higher ϕ, the limited variation in the number of trials and in ϕ and λ values among individuals means that there is no clear association (mean estimate of correlation between number of trials and: standardized ϕ: -7.38, confidence interval -15.97 to 1.28; standardized λ: -4.00, 89% confidence interval 12.53 to 4.61, n=8 grackles).

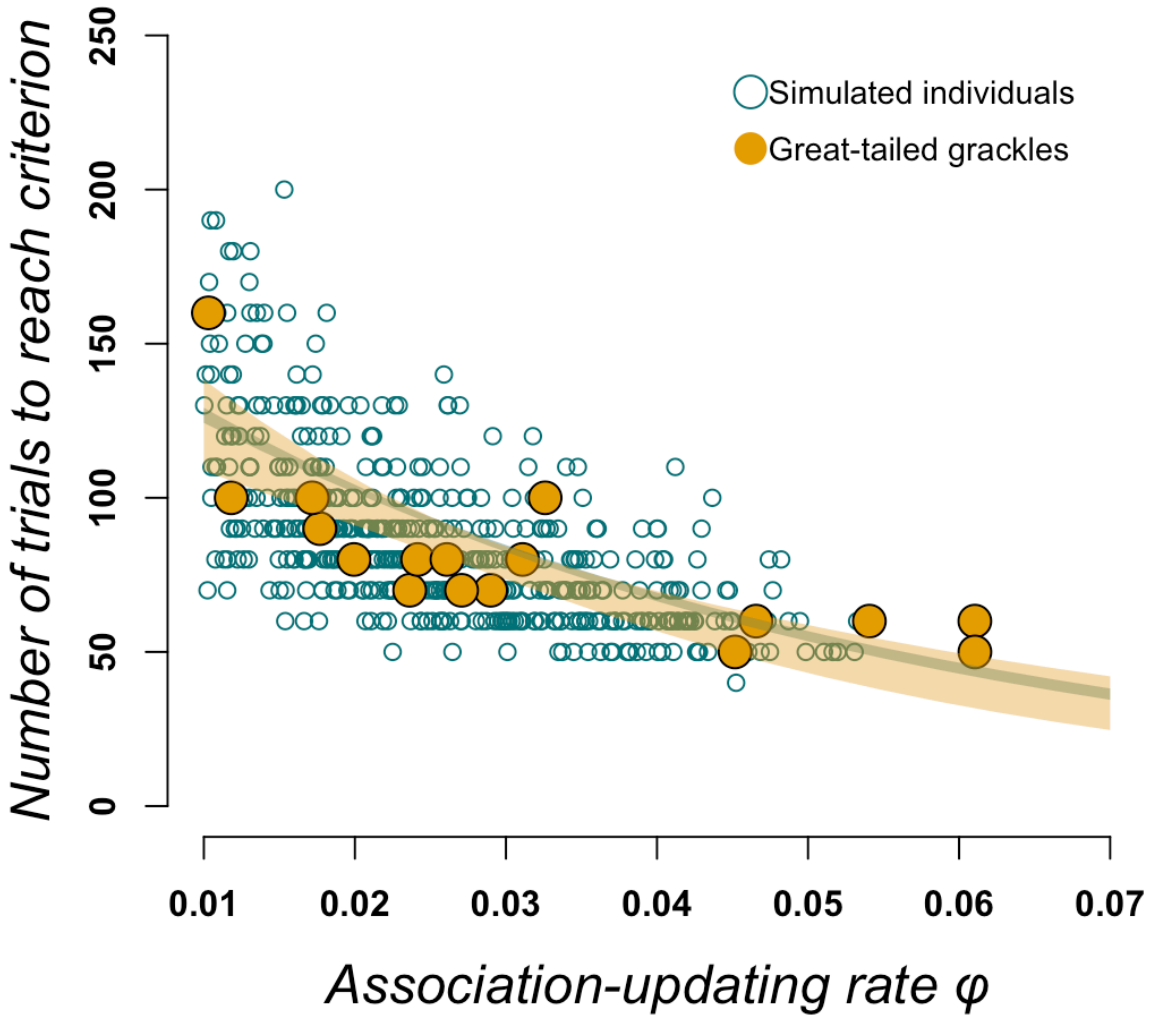

Figure 6. Relationship between ϕ and the number of trials grackles (yellow points) and simulated individuals (green circles) needed to reach criterion in their first trial. The observed grackle data falls within the range of the number of trials individuals with a given ϕ value are expected to need, and shows the same negative correlation between their ϕ and the number of trials as the simulated individuals (lines display the confidence interval of the estimated relationships).

4) Changes in ϕ and λ through the serial reversal learning task

Great-tailed grackles who experienced the serial reversal learning manipulation reduced the number of trials they needed to reach the criterion from an average of 75 to an average of 40 (estimate of change in number of trials -30.02, confidence interval -36.05 to -24.16, n=8 grackles). For the manipulated grackles, the estimated ϕ values more than doubled from 0.03 in their initial discrimination and first reversal (which is identical to the average observed among the control grackles who did not experience the manipulation) to 0.07 in their last two reversals (estimate of expected average change: +0.03, confidence interval +0.02 to +0.05, n=8). The λ values of the manipulated grackles went slightly down from 4.2 (again, identical to control grackles) to 3.2 (estimate of average change: -1.07, confidence interval -1.63 to -0.56, n=8 grackles) (Figure 5). The values we observed after the manipulation in the last reversal for the number of trials to reverse, as well as the ϕ and λ values estimated from the last reversal, all fall within the range of variation we observed among the control grackles in their first and only reversal (Figure 5). This means that the manipulation did not push grackles to new levels, but changed them within the boundaries of their natural abilities observed in the population.

As predicted, the increase in ϕ during the manipulation fits with the outcome from the simulations: larger ϕ values were associated with fewer trials to reverse. The improvement the grackles showed in the number of trials they needed to reach the criterion from the first to the last reversal matched the changes in their ϕ values (confidence interval +1.54 to +14.22, n=8 grackles). The improvement did not match the change in their λ values (confidence interval -4.66 to 9.46, n=8 grackles), because, as predicted, the grackles in the manipulation showed a decreased λ in their last reversal. This decrease in λ meant that grackles quickly found the rewarded option after a switch in which option was rewarded. In their first reversal grackles chose the newly rewarded option in 25% of the first 20 trials, in their final reversal the manipulated grackles chose correctly in 35% of the first 20 trials. Despite their low λ values, manipulated grackles still chose the rewarded option consistently because the increase in ϕ compensated for this reduced sensitivity (Figure 4; also see below).

5) Individual consistency in the serial reversal learning task

While we had previously found that differences among grackles in whether they needed many or few trials persisted through the manipulation, we did not find similar consistency in either ϕ or λ. We found a negative correlation between the ϕ estimated from an individual's performance in the first reversal and how much their ϕ changed toward the value for their performance in the last reversal (-0.84, confidence interval -1.14 to -0.52, n=8 grackles) such that individuals ended up with similar ϕ values to each other at the end of the manipulation and their beginning and end ϕ values were not correlated (-0.21, confidence interval -1.55 to 1.35, n=8 grackles). Similarly, individuals who started with a high λ changed more than individuals who already had lower a λ during the first reversal (-0.44, confidence interval -0.76 to -0.10, n=8 grackles) and these changes were not consistent such that individual differences in λ did not remain through the serial reversal learning task (+0.17confidence interval -0.67 to +0.97, n=8 grackles). Individuals appeared to use different adjustments to their strategies to improve their performance through the manipulation. There was a negative correlation between an individual's ϕ and λ after their last reversal (-0.39, 89% confidence interval: -0.72 to -0.06, n=8 grackles), indicating that they ended up with different strategies from along the range of potential solutions. Some individuals quickly learn the new reward structure after a switch, but continue to explore the alternative option even after they have learned the new associations (high association-updating rate and low sensitivity to learned associations). Other individuals take longer to learn that the reward has switched but once they have reversed their associations they rarely choose the unrewarded option (Figure 7). Together, this suggests that all individuals improved by the same extent through the manipulation such that the differences in their performances persisted, but they ended up with different strategies for how to quickly reach the criterion after a reversal by either having a high association updating rate or a low sensitivity to their learned associations.

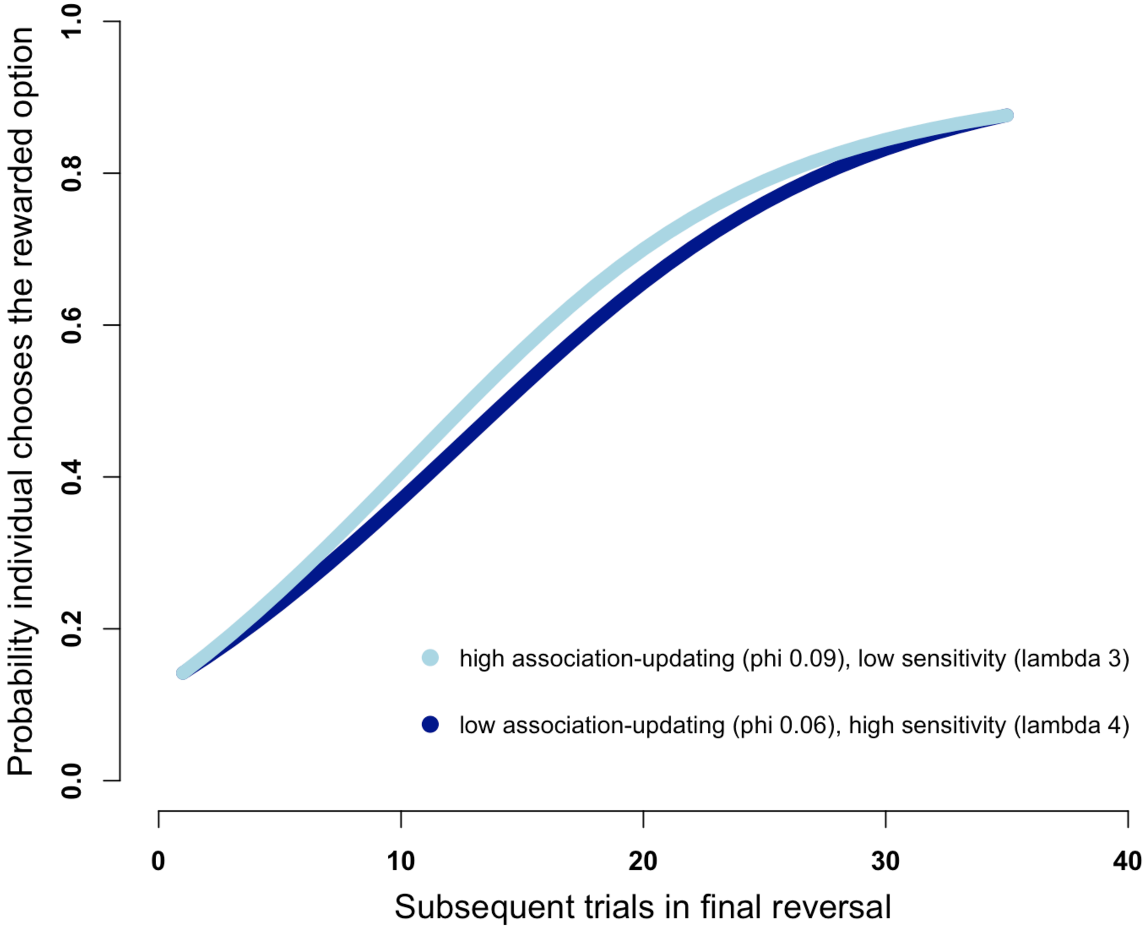

Figure 7. Predicted performance curves of individuals with different ϕ and λ values at the end of the serial reversal learning experiment based on the analytical formulas. We observed that, among the grackles who completed the serial reversal learning experiment, there was a negative correlation between their ϕ and λ, indicating that individuals used slightly different strategies to reach the criterion (choosing the rewarded option in 85% or more of trials) at equally few number of trials after the reward switched (when they had chosen the now rewarded option in 15% or less of trials). Individuals with a higher ϕ and lower λ (light blue line) quickly learn the new associations, but continue to explore the unrewarded option even after they have learned the association, leading to a curve with a more gradual increase throughout the trials. Individuals with a lower ϕ and higher λ (dark blue line) take longer to switch their associations, but once they do, they only rarely choose the non-rewarded option, leading to a more S-shaped curve where the initial increase in probability is lower and a more rapid rise later.

6) Association between ϕ and λ with performance on the multi-access boxes

We previously found that three measures of performance in the two multi-access puzzle boxes (number of options solved for both the wooden and the plastic multi-access puzzle box, latency to solve a new option on the plastic multi-access puzzle box) were correlated with the number of trials grackles needed to reach the criterion in the color tube reversal. We find that these measures also correlate with the underlying flexibility parameters ϕ and λ. In particular, the number of options solved on both the plastic and the wooden multi-access puzzle boxes had a U-shaped association with the λ values individuals had at the end in their last reversal (estimate of association between number of options solved on plastic box and: ϕ = +0.03, confidence interval -0.38 to +0.43; squared ϕ = -0.16, confidence interval -0.59 to +0.28; $lambda = +0.17, confidence interval -0.27 to +0.61; squared λ = +0.59, confidence interval +0.18 to +1.02; n=15 grackles; estimate of association between number of options solved on wooden box and: ϕ = -0.08, confidence interval -0.62 to +0.47; ϕ squared = +0.43, confidence interval -0.08 to +0.97; λ = +0.03, confidence interval -0.50 to +0.59; squared λ = +0.63, confidence interval +0.12 to +1.19; n=12 grackles). Grackles who had either particularly low or particularly high sensitivities to their previously learned associations were more likely to solve all four options than grackles with intermediate values of λ (Figure 8). For the latency to attempt a new option on the plastic box there was also a U-shaped association, but with ϕ (estimate of association between latency to attempt new option on plastic box and: ϕ = -0.66, confidence interval -1.30 to +0.0.06; squared ϕ = +0.58, confidence interval -0.06 to +1.30; $lambda = +0.14, confidence interval -0.45 to +0.70; squared λ = +1.09, confidence interval +0.28 to +1.87; n=11 grackles; estimate of association between latency to attempt new option on wooden box and: ϕ = -0.62, confidence interval -1.46 to +0.14; ϕ squared = +0.39, confidence interval -0.47 to +1.26; λ = +0.13, confidence interval -0.66 to +0.86; squared λ = +0.32, confidence interval -0.62 to +1.35; n=11 grackles). Grackles with either particularly high or particularly low rates of updating their associations took longer to attempt a new option than grackles with intermediate values of ϕ (Figure 8).

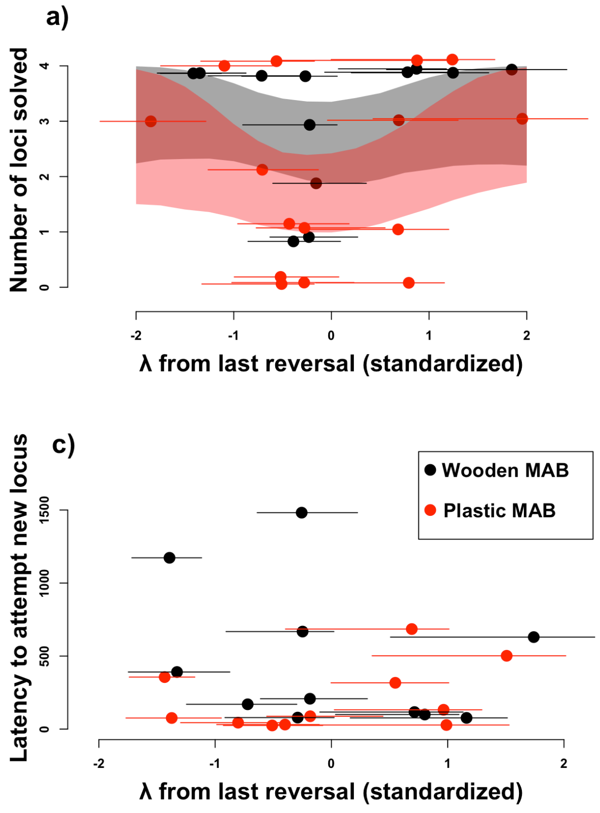

Figure 8. Relationships between ϕ and λ from the last reversal and performance on the wooden (black dots) and plastic (red dots) multi-access puzzle boxes. Grackles with intermediate λ values in their last reversal (a) were less likely to solve all four options on both boxes than grackles with either high or low λ values. Grackles with intermediate ϕ values have a shorter latency to attempt a new option on the plastic box (d). There are no clear relationships between ϕ and the number of options solved on either box (b), λ and the latency to attempt an option on either box (c), or (d) ϕ and the latency to attempt a new option on the wooden box. The ϕ and λ values change slightly between the top and bottom rows because the sample differs between boxes, and values were standardized for each plot.

DISCUSSION

Our analyses indicate that applying a more mechanistic model to understand the behavior of great-tailed grackles in a serial reversal learning experiment can provide additional insights into the potential components of behavioral flexibility and their dynamic changes. First, the simulations showed that the Bayesian reinforcement learning model accurately captures variation in the behavior of individuals in the serial reversal learning experiment and that the two key parameters ϕ, the association-updating rate, and λ, the sensitivity to learned associations, can be reliably inferred if we combine at least two association learning periods across a switch in the rewarded options. This provides the opportunity to also infer whether and how individuals who experience the serial reversal learning experiment dynamically change their behavioral flexibility. Second, in line with our prediction, the simulations indicate that higher ϕ and lower λ mean that individuals should reach the reversal learning criterion in fewer trials. However, we observe that for a single reversal ϕ is more important and that λ simply sets a threshold on the number of trials individuals need to consistently choose the rewarded option. Third, post-hoc analyses of grackle serial reversal learning data revealed that, contrary to our prediction but in line with the simulation results, ϕ but not λ explained more of the interindividual variation in how many trials individuals needed to reach criterion during a reversal. Fourth, matching these observations, we found that the primary component of flexibility that was manipulated during the serial reversal experiments was ϕ, which more than doubled between the first and last reversals, whereas λ slightly declined, as expected based on the simulations. Fifth, while individual differences in performance persist across the manipulation, the underlying changes in ϕ and λ are not predictable based on their initial values. Grackles appear to use different strategies to improve their performance during the serial reversal experiment, with some individuals showing more changes in their association-updating rate but less in their sensitivity to learned associations, while others show the opposite, leading to a negative correlation between the inferred ϕ and λ values among the individuals at the end of the serial reversal learning experiment.. Finally, these different strategies to improve their behavioral flexibility that individuals revealed in the serial reversal learning experiment subsequently also influenced their behavior in a different experimental test of behavioral flexibility. Grackles with intermediate values of λ (and phi) solved fewer options on both multi-access puzzle boxes than grackles with either high or low λ (and low or high ϕ), and grackles with intermediate values of ϕ have shorter latencies to attempt a new option. Accordingly, the grackles appeared to react to the predictability of the associations and the frequent switches of the reward location that they experienced during the serial reversal learning experiment to adjust their behavioral flexibility.

Previous analyses of reversal learning performance in wild-caught animals have often focused on summaries of the choices individuals make (e.g. Bond et al., 2007), setting criteria to define success and how much individuals sample or explore the different options versus acquire or exploit the reward (e.g. Federspiel et al., 2017). These approaches are more descriptive, making it difficult to link the differences to specific processes and to predict how variation in behavior might transfer to other tasks. While there have been attempts to identify potential rules that individuals might learn during serial reversal learning (Minh Le et al., 2023; Spence, 1936; J. Warren, 1965; J. M. Warren, 1965), these rules were often about abstract switches to extreme strategies (e.g. win-stay / lose-shift) and therefore could not account for the full variation in the behavior. In contrast, the Bayesian reinforcement learning model with its two parameters of the association-updating rate and the sensitivity to learned associations has a clear theoretical foundation and appears to be sufficient to accurately represent the behavior of grackles in the serial reversal experiment. The previously described rules, including dramatic shifts in strategies, can be recovered with the dynamic Bayesian reinforcement learning model, including the different 'learning curves' that we observe among individuals (e.g. Gallistel et al. (2004)). Applying the Bayesian reinforcement model to (serial) reversal learning experiments can provide several benefits to our understanding of behavioral flexibility. First, it highlights the key pieces of information that individuals likely pay attention to when adjusting their behavior. This provides ways to also link their performances and inferred cognitive abilities to how they experience and react to their natural environments. In particular, literature on foraging behavior that focuses on the likely trade-offs between the exploration versus exploitation of different options has a similar focus on gaining information (exploration) versus decision making (exploitation) (Addicott et al., 2017; Berger-Tal et al., 2014; Kramer & Weary, 1991). Having a mechanistic model for the behavioral choices can also help to design better and alternative experiments. Simulating the likely behavioral choices of individuals can help to decide how to track the progress of individuals and when to switch rewards (Logan et al., 2023a). Deciding on which external conditions might matter most to a given group of individuals can help to determine which parameters to vary and can help to adapt the model further. For example, it has been extended to allow for unpredictability in the association between the cue and the reward (Danwitz et al., 2022; Gershman, 2018) or to assume that experiencing a reward will update the association more than not experiencing a reward (Metha et al., 2020). Our advance here was to make the model dynamic to determine how individuals adjust their behavior during the serial reversal learning experiment.

The dynamic model shows that behavioral flexibility in the grackles is not a fixed trait, but individuals can change their flexibility in response to their experiences. Grackles coming into the experiment already had different strategies, suggesting that they had different experiences of how predictable cues are and how frequently their environment changes. In general, the association-updating rate ϕ appears to explain more of the variation in how many trials individuals need to reach the criterion of consistently choosing the rewarded option during a single phase. The importance of the association-updating rate for the performance of the grackles in the reversal learning experiment matches what has been reported for squirrel monkeys (Bari et al., 2022). In contrast, the sensitivity to learned associations λ appears to set a threshold on the performance during a single phase, but appears more important as the rewards switch more frequently. In the serial reversal learning experiments, we observed an initial decline in performance, with most grackles needing more trials in the second and third reversal compared to the first, before improving and reaching the criterion in 50 trials or less (Logan et al., 2023a). This initial increase likely reflects that grackles need to distinguish between the absence of a reward at the previously rewarded location reflects stochastic variation in the association between the cue and the reward or an actual switch in reward structure. In a stochastic environment, individuals can gain more reward if they do not update their associations quickly, but stick with an option that previously gave them high rewards (Woo et al., 2023). In their natural environment, most cues are presumably not perfect such that their initial expectation might be that the particular tube just did not have a reward that time, but should still provide rewards frequently, thus explaining their initial decline in performance. Only after several switches is there sufficient information for the grackles to infer that the cues are highly reliable and the switches are relatively frequent. This is when they show the increase in their association-updating rate ϕ, which on average doubled across individuals, changing more for individuals who started off with lower ϕ values. IGrackles also changed their sensitivity to the learned associations during the manipulation, in line with the prediction that they benefit from being open to exploring the alternative option when the reward structure frequently switches.

Most animals that have been tested in serial reversal learning experiments thus far show improvements throughout the consecutive reversals, suggesting that most species can adapt their behavioral flexibility in response to the predictability and stability of their environments (e.g. J. Warren & Warren (1962); Komischke et al. (2002); Bond et al. (2007); Strang & Sherry (2014); P. K. Chow et al. (2015); Cauchoix et al. (2017); Erdsack et al. (2022); Degrande et al. (2022)]. For the grackles, the manipulation pushed individuals to levels that were already observed in some individuals at the beginning of the experiment, meaning that the change within the experiment is within the natural range of abilities also observed in the wild. While there were individual differences in how individuals performed (McCune et al., 2023), all individuals changed depending on their experiences. Among the manipulated grackles, who all quickly switched to consistently gain the reward, we observed different strategies. On the one side, there are grackles who change gradually throughout an association phase, already choosing the newly rewarded option at the beginning but continuing to explore the alternative non-rewarded option throughout. These are the individuals with a high association-updating rate and low sensitivity to learned associations. On the other side are grackles who take longer to choose the newly rewarded option after a switch, but once they discover which option is rewarded, quickly reverse their preference. These are the individuals with low association-updating rates and high sensitivities to learned associations. With the variables we measured here, we could not predict which strategies ended up with after the manipulation. We observed additional strategies with different combinations of ϕ and λ across the grackles during their first reversal, but these are not efficient in the serial reversal learning experiment and instead are more suited to unpredictable and less frequently changing environments. How frequently and how quickly individuals change their behavioral flexibility in their natural environments is unclear. Individual differences might persist if their different behavioral flexibility leads them to continue to experience their environment differently. For the grackles, we have some indication that after releasing them back to their original environments, differences in behavioral flexibility between the manipulated and control individuals persisted for at least several months, with individuals who had changed their ϕ and λ appearing to switch more frequently between food types and foraging techniques (Logan CJ et al., 2019, results are in prep.).

The analyses linking ϕ and λ to the performance on the multi-access boxes show that the different strategies grackles ended up with to improve their performance during the serial reversal learning experiment subsequently appeared to influence how they solved the multi-access box. The negative correlation between ϕ and λ prompted us to explore whether the relationship between these two variables and the performance on the multi-access boxes could be non-linear. We detected U-shaped relationships between ϕ and λ and how individuals performed on the multi-access puzzle boxes. First, grackles with intermediate ϕ values showed shorter latencies to attempt a new option. This could reflect that grackles with high ϕ values take longer because they formed very strong associations with the previously rewarded option, while grackles with small ϕ values take longer because they do not update their associations even though the first option is no longer rewarded or because they do not explore as much because of their small λ. Second, we found that grackles with intermediate values of λ solved fewer options. This could indicate that grackles with a small λ are more likely to explore new options while grackles with a large λ, and low ϕ are less likely to return to an option that is no longer rewarded. Given that there was also a positive correlation between the number of options solved and the latency to attempt a new options, there might be a trade-off, where grackles with extreme ϕ and λ values solve more options, but need more time, whereas grackles with intermediate values have shorter latencies, but solve fewer options. We are limited though in our interpretation by the small sample sizes. More detailed studies would be needed in order to fully understand how the association-updating rate and the sensitivity to learned associations might shape performance on the multi-access puzzle boxes. In addition, it is also possible that performance on the multi-access boxes relies on other cognitive abilities in which individuals may differ. For example, we previously found that grackles who are faster to complete an inhibition task, where they had to learn to not react to a cue in order to wait for a trial in which a different cue could result in gaining a a reward, were slower to switch options on the boxes (Logan et al., 2021). As such, variation in self control may affect performance on flexibility and innovation tasks by decreasing exploratory behaviors. However, all these analyses are exploratory and based on a small sample, so these interpretations are speculative and further investigation is needed to understand how potential cognitive abilities shape performance on such tasks.

Overall, these findings indicate the potential benefits of applying more mechanistic models to psychological experiments. Inferring the cognitive processes potentially underlying behavior can allow us to make clearer predictions about how the performance in one experiment might translate to other paradigms and to behavior in the wild. For the serial reversal learning paradigm, we could expect that the previously observed differences in whether performance links with performance in other experiments like innovation or inhibition Logan (2016) could be linked to differences in whether the association-updating rate or the sensitivity to learned associations plays a larger role in the reversal performance in a given species and in particular for the other trait. The advanced capabilities of reflecting behavioral choices directly in a Bayesian framework offers an opportunity for the field of comparative cognition to implement more informed assessments of cognitive abilities and the factors shaping them.

AUTHOR CONTRIBUTIONS

Lukas: Hypothesis development, simulation development, data analyses, data interpretation, write up, revising/editing.

McCune: Added MAB log experiment, protocol development, data collection, revising/editing.

Blaisdell: Prediction revision, revising/editing.

Johnson-Ulrich: Data collection, revising/editing.

MacPherson: Data collection, revising/editing.

Seitz: Prediction revision, revising/editing.

Sevchik: Data collection, revising/editing.

Logan: Hypothesis development, protocol development, data collection, data analysis, data interpretation, revising/editing.

FUNDING

This research is funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology.

CONFLICT OF INTEREST DISCLOSURE

We, the authors, declare that we have no financial conflicts of interest with the content of this article. CJ Logan is a Recommender and, until 2022, was on the Managing Board at PCI Ecology. D Lukas is a Recommender at PCI Ecology.

ACKNOWLEDGEMENTS

We thank our PCI Ecology recommender, Aurelie Coulon, and reviewers, Maxime Dahirel and Andrea Griffin, for their feedback on this preregistration; and the reviewers of this manuscript for their constructive feedback that helped with the framing of the study; Julia Cissewski for tirelessly solving problems involving financial transactions and contracts; Sophie Kaube for logistical support; and Richard McElreath for project support.

REFERENCES

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, *42*(10), 1931–1939. https://doi.org/https://doi.org/10.1038/npp.2017.108

Agrawal, S., & Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory*, 39–31. https://doi.org/https://proceedings.mlr.press/v23/agrawal12.html

Bari, B. A., Moerke, M. J., Jedema, H. P., Effinger, D. P., Cohen, J. Y., & Bradberry, C. W. (2022). Reinforcement learning modeling reveals a reward-history-dependent strategy underlying reversal learning in squirrel monkeys. *Behavioral Neuroscience*, *136*(1), 46. https://doi.org/https://doi.org/10.1037/bne0000492

Bartolo, R., & Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron*, *106*(6), 1044–1054. https://doi.org/https://doi.org/10.1016/j.neuron.2020.03.024

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PloS One*, *9*(4), e95693. https://doi.org/https://doi.org/10.1371/journal.pone.0095693

Bitterman, M. E. (1975). The comparative analysis of learning: Are the laws of learning the same in all animals? *Science*, *188*(4189), 699–709. https://doi.org/https://doi.org/10.1126/science.188.4189.699

Blaisdell, A., Seitz, B., Rowney, C., Folsom, M., MacPherson, M., Deffner, D., & Logan, C. J. (2021). Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles. *Peer Community Journal*, *1*. https://doi.org/https://doi.org/10.24072/pcjournal.44

Bond, A. B., Kamil, A. C., & Balda, R. P. (2007). Serial reversal learning and the evolution of behavioral flexibility in three species of north american corvids (gymnorhinus cyanocephalus, nucifraga columbiana, aphelocoma californica). *Journal of Comparative Psychology*, *121*(4), 372. https://doi.org/https://doi.org/10.1037/0735-7036.121.4.372

Breen, A. J., & Deffner, D. (2023). Leading an urban invasion: Risk-sensitive learning is a winning strategy. *eLife*, *12*, RP89315. https://doi.org/10.1101/2023.03.19.533319

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874. https://doi.org/https://doi.org/10.1111/1468-0262.00054

Cauchoix, M., Hermer, E., Chaine, A., & Morand-Ferron, J. (2017). Cognition in the field: Comparison of reversal learning performance in captive and wild passerines. *Scientific Reports*, *7*(1), 12945. https://doi.org/https://doi.org/10.1038/s41598-017-13179-5

Chen, C. S., Knep, E., Han, A., Ebitz, R. B., & Grissom, N. M. (2021). Sex differences in learning from exploration. *Elife*, *10*, e69748. https://doi.org/https://doi.org/10.7554/elife.69748

Chow, P. K. Y., Lea, S. E., & Leaver, L. A. (2016). How practice makes perfect: The role of persistence, flexibility and learning in problem-solving efficiency. *Animal Behaviour*, *112*, 273–283. https://doi.org/10.1016/j.anbehav.2015.11.014

Chow, P. K., Leaver, L. A., Wang, M., & Lea, S. E. (2015). Serial reversal learning in gray squirrels: Learning efficiency as a function of learning and change of tactics. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(4), 343. https://doi.org/https://doi.org/10.1037/xan0000072

Coulon, A. (2023). An experiment to improve our understanding of the link between behavioral flexibility and innovativeness. *Peer Community in Ecology*, *1*, 100407. https://doi.org/https://doi.org/10.24072/pci.ecology.100407

Danwitz, L., Mathar, D., Smith, E., Tuzsus, D., & Peters, J. (2022). Parameter and model recovery of reinforcement learning models for restless bandit problems. *Computational Brain & Behavior*, *5*(4), 547–563. https://doi.org/https://doi.org/10.1007/s42113-022-00139-0

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/https://doi.org/10.1038/nature04766

Deffner, D., Kleinow, V., & McElreath, R. (2020). Dynamic social learning in temporally and spatially variable environments. *Royal Society Open Science*, *7*(12), 200734. https://doi.org/https://doi.org/10.1098/rsos.200734

Degrande, R., Cornilleau, F., Lansade, L., Jardat, P., Colson, V., & Calandreau, L. (2022). Domestic hens succeed at serial reversal learning and perceptual concept generalisation using a new automated touchscreen device. *Animal*, *16*(8), 100607. https://doi.org/https://doi.org/10.1016/j.animal.2022.100607

Dufort, R. H., Guttman, N., & Kimble, G. A. (1954). One-trial discrimination reversal in the white rat. *Journal of Comparative and Physiological Psychology*, *47*(3), 248. https://doi.org/https://doi.org/10.1037/h0057856

Dunlap, A. S., & Stephens, D. W. (2009). Components of change in the evolution of learning and unlearned preference. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1670), 3201–3208. https://doi.org/https://doi.org/10.1098/rspb.2009.0602

Erdsack, N., Dehnhardt, G., & Hanke, F. D. (2022). Serial visual reversal learning in harbor seals (phoca vitulina). *Animal Cognition*, *25*(5), 1183–1193. https://doi.org/https://doi.org/10.1007/s10071-022-01653-1

Federspiel, I. G., Garland, A., Guez, D., Bugnyar, T., Healy, S. D., Güntürkün, O., & Griffin, A. S. (2017). Adjusting foraging strategies: A comparison of rural and urban common mynas (acridotheres tristis). *Animal Cognition*, *20*(1), 65–74. https://doi.org/10.1007/s10071-016-1045-7

Frömer, R., & Nassar, M. (2023). *Belief updates, learning and adaptive decision making*. https://doi.org/https://doi.org/10.31234/osf.io/qndba

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. https://doi.org/https://doi.org/10.1073/pnas.0404965101

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in bayesian social research. *Sociological Methodology*, *25*, 165–173. https://doi.org/https://doi.org/10.2307/271064

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/https://doi.org/10.1016/j.cognition.2017.12.014

Griffin, A. S., Guez, D., Lermite, F., & Patience, M. (2013). Tracking changing environments: Innovators are fast, but not flexible learners. *PloS One*, *8*(12), e84907. https://doi.org/10.1371/journal.pone.0084907

Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., & Holmes, A. (2017). The neural basis of reversal learning: An updated perspective. *Neuroscience*, *345*, 12–26. https://doi.org/https://doi.org/10.1016/j.neuroscience.2016.03.021

Jang, A. I., Costa, V. D., Rudebeck, P. H., Chudasama, Y., Murray, E. A., & Averbeck, B. B. (2015). The role of frontal cortical and medial-temporal lobe brain areas in learning a bayesian prior belief on reversals. *Journal of Neuroscience*, *35*(33), 11751–11760.

Komischke, B., Giurfa, M., Lachnit, H., & Malun, D. (2002). Successive olfactory reversal learning in honeybees. *Learning & Memory*, *9*(3), 122–129. https://doi.org/https://doi.org/10.1101/lm.44602

Kramer, D. L., & Weary, D. M. (1991). Exploration versus exploitation: A field study of time allocation to environmental tracking by foraging chipmunks. *Animal Behaviour*, *41*(3), 443–449. https://doi.org/https://doi.org/10.1016/s0003-3472(05)80846-2

Lea, S. E., Chow, P. K., Leaver, L. A., & McLaren, I. P. (2020). Behavioral flexibility: A review, a model, and some exploratory tests. *Learning & Behavior*, *48*(1), 173–187. https://doi.org/10.3758/s13420-020-00421-w

Leimar, O., Quiñones, A. E., & Bshary, R. (2024). Flexible learning in complex worlds. *Behavioral Ecology*, *35*(1), arad109. https://doi.org/https://doi.org/10.1093/beheco/arad109

Liu, Y., Day, L. B., Summers, K., & Burmeister, S. S. (2016). Learning to learn: Advanced behavioural flexibility in a poison frog. *Animal Behaviour*, *111*, 167–172. https://doi.org/https://doi.org/10.1016/j.anbehav.2015.10.018

Logan, C. J. (2016). Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ*, *4*, e2215. https://doi.org/10.7717/peerj.2215

Logan, C. J., McCune, K., MacPherson, M., Johnson-Ulrich, Z., Rowney, C., Seitz, B., Blaisdell, A., Deffner, D., & Wascher, C. (2021). Are the more flexible great-tailed grackles also better at behavioral inhibition? *PsyArXiv*. https://doi.org/10.31234/osf.io/vpc39

Logan, C. J., Shaw, R., Lukas, D., & McCune, K. B. (2022). How to succeed in human modified environments. *In Principle Acceptance by PCI Ecology of the Version on 8 Sep 2022*. https://doi.org/https://doi.org/10.17605/OSF.IO/346AF

Logan, CJ, Lukas D, Bergeron L, Folsom M, & McCune, K. (2019). Is behavioral flexibility related to foraging and social behavior in a rapidly expanding species? *In Principle Acceptance by PCI Ecology of the Version on 6 Aug 2019*. http://corinalogan.com/Preregistrations/g_flexforaging.html

Logan, CJ, McCune, KB, LeGrande-Rolls C, Marfori Z, Hubbard J, & Lukas, D. (2023). Implementing a rapid geographic range expansion - the role of behavior changes. *Peer Community Journal*. https://doi.org/10.24072/pcjournal.320

Logan, C., Lukas, D., Blaisdell, A., Johnson-Ulrich, Z., MacPherson, M., Seitz, B., Sevchik, A., & McCune, K. (2023a). Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context. *Peer Community Journal*, *3*. https://doi.org/10.24072/pcjournal.284

Logan, C., Lukas, D., Blaisdell, A., Johnson-Ulrich, Z., MacPherson, M., Seitz, B., Sevchik, A., & McCune, K. (2023b). Data: Behavioral flexibility is manipulable and it improves flexibility and problem solving in a new context. *Knowledge Network for Biocomplexity, Data package*. https://doi.org/10.5063/F1BR8QNC

Mackintosh, N., McGonigle, B., & Holgate, V. (1968). Factors underlying improvement in serial reversal learning. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *22*(2), 85. https://doi.org/https://doi.org/10.1037/h0082753

McCune, K., Blaisdell, A., Johnson-Ulrich, Z., Sevchik, A., Lukas, D., MacPherson, M., Seitz, B., & Logan, C. (2023). Repeatability of performance within and across contexts measuring behavioral flexibility. *PeerJ*. https://doi.org/10.7717/peerj.15773

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC, Boca Raton, FL. https://doi.org/10.1201/9780429029608

Metha, J. A., Brian, M. L., Oberrauch, S., Barnes, S. A., Featherby, T. J., Bossaerts, P., Murawski, C., Hoyer, D., & Jacobson, L. H. (2020). Separating probability and reversal learning in a novel probabilistic reversal learning task for mice. *Frontiers in Behavioral Neuroscience*, *13*, 270.

Mikhalevich, I., Powell, R., & Logan, C. (2017). Is behavioural flexibility evidence of cognitive complexity? How evolution can inform comparative cognition. *Interface Focus*, *7*(3), 20160121. https://doi.org/10.1098/rsfs.2016.0121

Minh Le, N., Yildirim, M., Wang, Y., Sugihara, H., Jazayeri, M., & Sur, M. (2023). Mixtures of strategies underlie rodent behavior during reversal learning. *PLOS Computational Biology*, *19*(9), e1011430. https://doi.org/https://doi.org/10.1371/journal.pcbi.1011430

Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, *1*(3), 133–143.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rayburn-Reeves, R. M., Stagner, J. P., Kirk, C. R., & Zentall, T. R. (2013). Reversal learning in rats (rattus norvegicus) and pigeons (columba livia): Qualitative differences in behavioral flexibility. *Journal of Comparative Psychology*, *127*(2), 202. https://doi.org/https://doi.org/10.1037/a0026311

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prosy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). Appleton-Century-Crofts, New York.

Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review*, *43*(5), 427. https://doi.org/10.1037/h0056975

Stan Development Team. (2023). *Stan modeling language users guide and reference manual, version 2.32.0, https://mc-stan.org/* (Version 2.32.0). https://mc-stan.org/

Strang, C. G., & Sherry, D. F. (2014). Serial reversal learning in bumblebees (bombus impatiens). *Animal Cognition*, *17*, 723–734. https://doi.org/https://doi.org/10.1007/s10071-013-0704-1

Tello-Ramos, M. C., Branch, C. L., Kozlovsky, D. Y., Pitera, A. M., & Pravosudov, V. V. (2019). Spatial memory and cognitive flexibility trade-offs: To be or not to be flexible, that is the question. *Animal Behaviour*, *147*, 129–136. https://doi.org/https://doi.org/10.1016/j.anbehav.2018.02.019

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved rhat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*. https://doi.org/https://doi.org/10.1214/20-BA1221

Warren, J. (1965). Primate learning in comparative perspective. *Behavior of Nonhuman Primates*, *1*, 249–281. https://doi.org/10.1016/B978-1-4832-2820-4.50014-7

Warren, J. M. (1965). The comparative psychology of learning. *Annual Review of Psychology*, *16*(1), 95–118. https://doi.org/10.1146/annurev.ps.16.020165.000523

Warren, J., & Warren, H. B. (1962). Reversal learning by horse and raccoon. *The Journal of Genetic Psychology*, *100*(2), 215–220. https://doi.org/https://doi.org/10.1080/00221325.1962.10533590

Woo, J. H., Aguirre, C. G., Bari, B. A., Tsutsui, K.-I., Grabenhorst, F., Cohen, J. Y., Schultz, W., Izquierdo, A., & Soltani, A. (2023). Mechanisms of adjustments to different types of uncertainty in the reward environment across mice and monkeys. *Cognitive, Affective, & Behavioral Neuroscience*, 1–20. https://doi.org/https://doi.org/10.1101/2022.10.01.510477