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[1,2]

2024-06-20

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 the third of of three post-study manuscripts from 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 and July 2024): preprint at EcoEvoRxiv, rmd with code at github

Abstract

Environments can change suddenly and unpredictably and animals might benefit from being able to flexibly adapt their behavior through learning new associations. Serial (repeated) reversal learning experiments have long been used to investigate differences in behavioral flexibility among individuals and species. In these experiments, individuals initially learn that a reward is associated with a specific cue before the reward is reversed back and forth between cues, forcing individuals to reverse their learned associations. Cues are reliably associated with a reward, but the association between the reward and the cue frequently changes. Here, we apply and expand newly developed Bayesian reinforcement learning models to gain additional insights into how individuals might dynamically modulate their behavioral flexibility if they experience serial reversals. We derive mathematical predictions that, in this experiment, individuals will gain the most rewards if they 1) increase their *rate of updating associations* between cues and the reward to quickly change to a new option after a reversal, and 2) decrease their *sensitivity* to their learned association to explore the alternative option after a reversal. We reanalyzed reversal learning data from 19 wild-caught great-tailed grackles (*Quiscalus mexicanus*), eight of whom participated in serial reversal learning experiment, and found that these predictions were supported. Their estimated association-updating rate was more than twice as high at the end of the serial reversal learning experiment than at the beginning, and their estimated sensitivities to their learned associations declined by about a third. The changes in behavioral flexibility that grackles showed in their experience of the serial reversals also influenced their behavior in a subsequent experiment, where individuals with more extreme rates or sensitivities solved more options on a multi-option

puzzle box. Our findings offer new insights into how individuals react to uncertainty and changes in their environment, in particular, showing how they can modulate their behavioral flexibility in response to their experiences.

Introduction

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 that cannot react to these changes should have reduced survival and/or reproductive success (Boyce et al., 2006; Starrfelt & Kokko, 2012). One of the ways animals react to changes is through behavioral flexibility, the ability to change behavior when circumstances change (Shettleworth, 2010). The level of behavioral flexibility present in a given species is often assumed to have been shaped by selection, with past levels of change in the environment determining how well species might be able to cope with more rapidly changing (Sih, 2013) or novel environments (Sol et al., 2002). However, in another conception, behavioral flexibility is itself plastic (Wright et al., 2010). Behavioral flexibility arises because individuals update their information about the environment through personal experience and make that information available to other cognitive processes (Mikhalevich et al., 2017). Such modulation of behavioral flexibility is presumably relevant if the rate and extent of environmental change is variable and unpredictable (Donaldson-Matasci et al., 2013; Tello-Ramos et al., 2019). We are still limited in our understanding of when and how individuals might react to their experiences of environmental change.

Evidence that animals can change their behavioral flexibility based on their recent experience comes from serial reversal learning experiments. Serial reversal learning experiments have long been used to understand how individuals keep track of biologically important associations in changing environments (Dufort et al., 1954; Mackintosh et al., 1968; Bitterman, 1975). In these experiments, individuals are presented with multiple options associated with cues, such as different colors or locations, that differ in their reward. Individuals can repeatedly choose among the options to learn the associations between rewards and cues. After they show a clear preference for the most rewarded option, the rewards are reversed across cues, and individuals are observed to see how quickly they learn the changed associations. When they have reversed their preference, the reward is changed back to the other option, until the individual reverses their preference again, and these reversals continue in a process called serial reversals. Their performance during the reversal task is taken as a measure of their behavioral flexibility, with the more flexible individuals being those that need fewer trials to consistently choose the rewarded option after a reversal (Bond et al., 2007). While the primary focus of these serial reversal learning experiments has been to measure differences in behavioral flexibility across individuals and species (Lea et al., 2020), several of these experiments show that behavioral flexibility is not a fixed trait, but that individuals can improve their performance if they experience repeated reversals (Bond et al., 2007; Liu et al., 2016; Cauchoix et al., 2017). Here, we investigate how individuals might change their behavioral flexibility during serial reversal learning experiments to better understand what cognitive processes could lead to the observed differences and adjustments in behavioral flexibility (Izquierdo et al., 2017; Danwitz et al., 2022).

We recently found that great-tailed grackles (*Quiscalus mexicanus*; hereafter grackles) can be trained to improve how quickly they learn to change associations in a serial reversal learning experiment (Logan et al., 2023a). After training birds to search for food in a yellow tube, the reversal learning experiment consisted of presenting birds with a light gray and a dark gray tube, only one of which contained a reward. After individuals chose one of the tubes, thus experiencing whether this color was rewarded or not, the experiment was reset, with the reward being in the same colored tube as before. Once an individual chose the rewarded color more than expected by chance (passing criterion of choosing correctly in at least 17 out of the last 20 trials, which represents a significant association according to the chi-square test), the reward was switched to the other color. Again, individuals made choices until they chose the now rewarded tube above the passing criterion. For one set of individuals, the trained group, we repeated the reversal of rewards from one color to the other until the birds reached the serial reversal passing criterion of forming a preference in 50 trials or less in two consecutive reversals. The median number of trials birds in this trained group needed to reach the passing criterion during their first reversal was 75, which improved to 40 trials in their final reversal.

Importantly, we found that, in comparison to a control group who only experienced a single reversal, trained grackles who experienced serial reversals also showed increased behavioral flexibility and innovativeness in other contexts. In particular, trained grackles performed better on multi-option puzzle boxes than control grackles, being faster to switch to a new access option on a box if the previous option was closed, and they solved more of the available access options (Logan et al., 2023a). This indicates that individuals did not just learn an abstract rule about the serial reversal learning experiment, but rather changed their overall behavioral flexibility in response to their experience.

Previous analyses of serial reversal learning experiments were limited in understanding the potential changes in behavioral flexibility because they focused on summaries of the choices that individuals make (e.g. Bond et al., 2007). These approaches are more descriptive, making it difficult to link flexibility 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 (Spence, 1936; Warren, 1965a; Warren, 1965b; Minh Le et al., 2023), these rules were often about abstract switches to extreme behaviors (e.g. win-stay / lose-shift) and therefore could not account for the full variation of behavior. A number of theoretical models have recently been developed that appear to reflect the potential cognitive processes individuals seem to rely on when making choices in reversal learning experiments (for a recent review see, for example, Frömer & Nassar, 2023). These theoretical models deconstruct the behavior of individuals in a reversal learning task into two primary parameters (Camerer & Hua Ho, 1999; Chow et al., 2015; Izquierdo et al., 2017; Bartolo & Averbeck, 2020). Importantly, in the Bayesian reinforcement learning models there are now also statistical approaches to infer these underlying parameters from the behavior of individuals (Camerer & Hua Ho, 1999; Lloyd & Leslie, 2013). The first process reflects the *rate of updating associations* (which we refer to hereafter as ϕ, the Greek letter phi), or how quickly individuals learn about the associations between the cues and potential rewards (or dangers). In the reinforcement learning models, this rate is reflected by the Rescorla-Wagner rule (Rescorla & Wagner, 1972). The rate 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 equations 1-3). Individuals are expected to show different rates in different environments, particularly in response to the reliability of the cues (Figure 1). Lower updating rates are expected when associations are not perfect such that a single absence of a reward might be an error rather than indicating a new association. Higher updating rates are expected when associations are reliable such that individuals should update their associations quickly when they encounter new information (Dunlap & Stephens, 2009; Breen & Deffner, 2023). The second process, the *sensitivity to their learned associations* (which we hereafter refer to as λ, the Greek letter lambda) reflects how individuals, when presented with a set of cues, might decide between these alternative options based on their learned associations of the cues. In the reinforcement learning model, the sensitivity to learned associations modifies the relative difference in learned rewards to generate the probabilities of choosing either option (Daw et al., 2006; Agrawal & Goyal, 2012; Danwitz et al., 2022). 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 biases in choice as soon as there are small differences in their learned associations (see equations 1-3). Individuals with larger sensitivities will quickly prefer the option that previously gave them the highest reward (or the lowest danger), while individuals with lower sensitivities will continue to explore alternative options. Sensitivities are expected to reflect the rate of change in the environment (Figure 1), with larger sensitivities occurring when environments are static such that individuals start to exploit any differences they recognise as soon as possible. Lower sensitivities are expected when changes are frequent, such that individuals continue to explore alternative options when conditions change (Daw et al., 2006; Breen & Deffner, 2023).

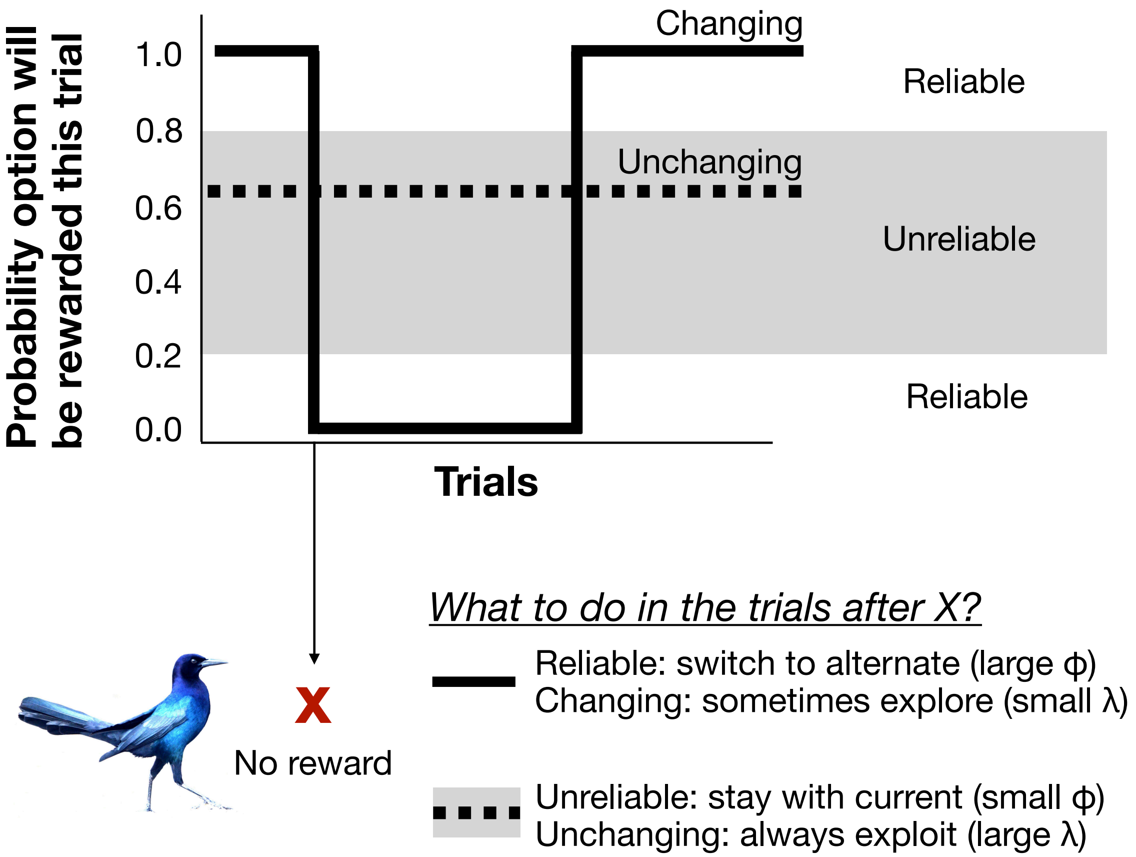
Probability option will be rewarded this trial
1.0
0.8
0.6
0.4
0.2
0.0
Changing
Unchanging
Reliable
Unreliable
Reliable
Trials
X
No reward
What to do in the trials after X?
Reliable: switch to alternate (large ϕ)
Changing: sometimes explore (small λ)
Unreliable: stay with current (small ϕ)
Unchanging: always exploit (large λ)

Figure 1. Individuals are expected to update their associations and make decisions differently depending on the environment they experience. 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 these reliable, but changing conditions, 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, but to 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 and unreliable conditions, the probability that an option is rewarded stays constant across trials (dotted lines), but is closer to 50% (gray background). In these conditions, individuals are expected to gain the most rewards if they build their associations by averaging information across many trials (small ϕ), and have high sensitivities to learned associations to exploit the option with the highest association (large λ). Grackle picture credit (CC BY 4.0): Dieter Lukas.

Here, we applied and modified the Bayesian reinforcement learning models to data from our grackle research on behavioral flexibility to assess if and how the cognitive processes might have changed as individuals experienced the serial reversal learning experiment. We previously found that the model can predict the performance of grackles in a reversal learning task with a single reversal of a color preference (Blaisdell et al., 2021). Grackles experiencing the serial reversal learning experiment are expected to infer that associations can frequently change but that, before and after a change, cues reliably indicate whether a reward is present or not. Based on the theoretical models, we predict that individuals increase their association-updating rate because cues are highly reliable, such that they can change their associations as soon as there is a change in the reward (Dunlap & Stephens, 2009; Breen & Deffner, 2023). In addition, we predict that individuals reduce their sensitivity to the learned associations, because the option that is rewarded reverses frequently, requiring individuals to explore alternative options (Neftci & Averbeck, 2019; Leimar et al., 2024). Given that reversals in the associations are not very frequent, we also expect some variation in individuals in whether they switch to the newly rewarded option because they find the reward quickly through continued exploration (somewhat lower λ and higher ϕ) or because they quickly move away from the option that is no longer rewarded (somewhat higher λ and lower ϕ). To assess these predictions, we addressed the following six research questions. With the first research question, we determined the feasibility and validity of our approach using simulations. As far as we were aware, Bayesian reinforcement learning models had not been used to investigate temporal changes in behavior. We therefore used simulations as a proof-of-concept assessment to show their sensitivity and ability to answer our questions. With the second research question, we derive mathematically specific predictions about the role of ϕ and λ in the serial reversal learning experiment. With the other four questions, we analyzed the grackle data to determine how the association-updating rate and the sensitivity to learned associations reflect the variation and changes in behavioral flexibility in grackles.

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?
We used agent-based simulations to answer this question, where simulated individuals made choices based on assigned ϕ and λ values. We determined how to apply the Bayesian reinforcement learning models to recover the assigned values from the choices in each trial. Previous applications of the Bayesian reinforcement learning models always combined the full sample of observations, so it is not clear whether these models are sufficiently sensitive to detect the changes over time that we are interested in. Two problems arise when trying to infer the underlying processes from a limited number of trials. The stochasticity in which option an individual chooses based on a given set of associations introduces differences in the set of choices across trials even among individuals with the same ϕ and λ values. On the flip-side, because of the probabilistic decisions, a given series of specific choices during a short number of trials can occur even if individuals have different ϕ and λ values. We varied the number of trials we analyzed to determine how many trials per individual are necessary to recover the assigned ϕ and λ values in light of this noise.

2) Is a high rate of association-updating (ϕ) and a low sensitivity to learned associations (λ) best to reduce errors in the serial reversal learning experiment?
We used analytical approaches to systematically vary ϕ and λ to determine how the interaction of the two processes shapes the behavior of individuals throughout the serial reversal learning experiment. Previous studies made general predictions about the role of ϕ and λ in different environments (Dunlap & Stephens, 2009; Breen & Deffner, 2023). We assessed here whether, under the specific conditions in the serial reversal experiments, where information is reliable and changes occur frequently, the best approach for individuals is to show high ϕ and low λ.

3) Which of the two parameters ϕ or λ explains more of the variation in the serial reversal learning experiment performance of the tested grackles?
Across both the trained (experienced serial reversals) and control (experienced a single reversal) grackles, we assessed whether variation in the number of trials an individual needed to reach the criterion in a given reversal is better explained by their inferred association-updating rate or by their inferred sensitivity to learned associations.

4) Do the grackles who improved their performance through the serial reversal learning experiment show the predicted changes in ϕ and λ?
If individuals learn the contingencies of the serial reversal experiment, they should reduce 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.

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 training (McCune et al., 2023). 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 reversals. 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 analyzed whether the ϕ and λ values of individuals at the beginning predict how much they changed throughout the serial reversal learning experiment.

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?
Grackles would be expected to solve more options on the multi-option puzzle boxes if they quickly update their previously learned associations when a previous option becomes unavailable (high ϕ). Given that, in the puzzle box experiment, individuals only receive a reward at any given option a few times, instead of repeatedly as in the reversal learning task, we predict that those individuals who are less sensitive to previously learned associations and instead continue to explore alternative options (low λ) can also gain more rewards.

Materials and Methods

Data

For question 1, we re-analyzed data we previously simulated for power analyses to estimate sample sizes for population comparisons (Logan et al., 2023c). In brief, we simulated choices in an initial association learning and single reversal experiment for a total of 640 individuals. The ϕ and λ values for each individual were drawn from a distribution representing one of 32 populations, with different mean ϕ (8 different means) and mean λ (4 different values) values for each population (32 populations is the combination of each ϕ and λ). We simulated 20 individuals in each of the 32 populations. The range for the ϕ and λ values assigned to the artificial individuals in the simulations were based on the previous analysis of 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. Choices during trials were based on equation 2 (see below). The first choice a simulated individual made in the initial association learning was random because we assumed individuals had no information about the rewards and therefore set the initial attractions to both options to be equally low. Based on their choices, individuals updated their internal associations with the two options based on equation 1. We excluded simulated individuals from 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. For each simulated individual, we recorded their assigned ϕ and λ values, as well as the series of choices they made during the initial association and the first reversal. For a given ϕ and λ, the stochasticity in which option a simulated individual chooses based on their attractions, plus the experience of either receiving a reward or not during previous choices, can lead to differences in the actual choices individuals make. The aim was to see what sample is needed to correctly infer the assigned ϕ and λ given the noise in the choice data. We also used the simulated data for question 3, to compare the influence of ϕ and λ on the behavior of the simulated individuals with that of the grackles.

To address question 2, we used an analytical approach and did not analyze any data.

For the empirical questions 3-6, we re-analyzed data on the performance of grackles in serial reversal learning and multi-option puzzle box experiments (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 data we use here were published as part of the earlier article and are available at the Knowledge Network for Biocomplexity's data repository (Logan et al., 2023b).

In brief, 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. The first experiment individuals participated in in the aviaries was the reversal learning experiment, as described in the introduction. A total of 19 grackles participated in the serial reversal learning experiment, where they learned to associate a reward with one color before experiencing one reversal to learn that the other color was rewarded (initial rewarded option was counterbalanced and randomly assigned as either a dark gray or a light gray tube). The rewarded option was switched when grackles passed the criterion of choosing the rewarded option in 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 rather than having randomly chosen one option more than the other (Logan et al., 2022). A subset of 8 individuals were randomly assigned to the trained group and went through a series of reversals until they reached the criterion of having formed an association (17 out of 20 choices correct) in 50 trials or less in two consecutive reversals. The individuals in the trained 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. A subset of 11 grackles were part of the control group, who experienced only a single reversal, before participating in trials with two identically colored tubes (yellow) where both contained a reward. The number of yellow tube trials was set to the average number of trials it took a bird in the trained group to pass their serial reversals.

For question 6, we additionally used data from an experiment the grackles participated in after they had completed the reversal learning experiment. Both the control and trained individuals were provided access to two multi-option puzzle boxes, one made of wood and one made of plastic. The two boxes were designed with slight differences to explore how general their performance was. The wooden box was made from a natural log, thus was more representative of something the grackles might encounter in the wild. In addition, while both boxes had four 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, in a counterbalanced order, 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 option was also made non-functional, and so on. The outcome measures for each individual on each box were the average latency it took to switch to a new option and the total number of options they successfully solved.

The Bayesian reinforcement learning model

For both the simulated and the observed grackle data, we used the Bayesian reinforcement learning model to estimate for each individual their ϕ and λ values based on the choices they made during the reversal learning experiments. The estimated ϕ and λ values were then used as outcome and/or predictor variables in the statistical models built to assess questions 3-6. We used the version of the Bayesian model that was developed in Blaisdell et al. (2021) and modified in Logan et al. (2023c) (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 (learning and ϕ): $A_{b,o,t+1}$=$(1-\phi_b)A_{b,o,t}$+$\phi_b\ \pi_{b,o,t}$.

Equation 1 estimates how the associations A, that individual b forms between the two different options (o, option 1 or 2) and their expected rewards, change from one trial to the next (trial t+1) as a function of their previously formed associations $A_{b,o,t}$ (how preferable option o is to grackle b at trial 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_b modifies how much individual b updates its associations based on its most recent experience. The higher the value of ϕ_b, the faster the individual updates its associations, paying more attention to recent experiences, whereas when ϕ_b is lower, a grackle's associations reflect averages across many trials. Association scores thus reflect the accumulated learning history up to trial t. The association with the option that is not explored in a given trial remains unchanged. At the beginning of the experiment (trial t equals 0), we assumed that individuals had the same low association between both options and rewards ($A_{b,1,0} = A_{b,2,0} = 0.1$).

Equation 2 (choice and λ): $P_{b,o,t} = \frac{exp(\lambda_b A_{b,o,t})}{\sum_{o=1}^{2} exp(\lambda_b A_{b,o,t})}$.

Equation 2 is a normalized exponential (softmax) function to convert the learned associations of the two options with rewards into the probability, P, that an individual, b, chooses one of the two options, o, in the current trial, t. The parameter λ_b represents the sensitivity of a given grackle, b, to how different its associations to the two options are. As λ_b gets larger, choices become more deterministic and individuals consistently choose the option with the higher association even if associations are very similar. As λ_b gets smaller, choices become more exploratory, with individuals choosing randomly between the two options independently of their learned associations if λ_b is 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 4.3.2) (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. Which option individuals chose was estimated with a categorical distribution with the probability, P, as estimated from equation 2 for each of the two options (categories), 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 reflect that it is a proportion (can only take values between 0 and 1), and λ is estimated on the log-scale to reflect that values have to be positive (there is no upper bound). 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 to 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 and training with a differently colored tube (see below). For estimations at the end of each reversal, we set the association with the option that was rewarded before the reversal 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 is 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.

For each estimation (simulated and observed grackle data), we ran four chains with 2000 samples each (half of which were warm up). We used functions in the package "posterior" (Vehtari et al., 2021) to draw 4000 samples from the posterior (the default). We report the estimates for ϕ and λ for each individual (simulated or observed 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 compatibility intervals 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 https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip2post.Rmd. 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 dividing by the standard deviation. We did this to define the priors for the relationships on a more standard scale and to be able to more directly compare the respective influence of ϕ and λ on the outcome variable.

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 ran the Bayesian reinforcement learning model on the 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 the learned associations λ 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 the single or the combined phases, we correlated the estimated values with the values individuals were initially assigned:

$\phi_{b,1}$ or $\lambda_{b,1}$ ~ Normal(μ_b, σ),
$\mu_b = \alpha + \beta$ x $\phi_{b,0}$ or $\lambda_{b,0}$,
α ~ Normal(0,0.1),
β ~ Normal(1,1),
σ ~ Exponential(1),

where $\phi_{b,1}$ or $\lambda_{b,1}$, the values estimated for each bird, indexed by b, from the simulated behavior are assumed to come from a normal distribution with a mean that can vary for each bird, mu_b, and overall variance, σ. The mean for each bird is constructed from an overall intercept, α, and the change in expectation, the slope, β, depending on the values assigned to each bird at the beginning of the simulation ($\phi_{b,0}$ or $\lambda_{b,0}$). The combination of α close to 0 and of β 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 call Stan and estimate the relationships. Following the social convention set in (McElreath, 2020), we report the mean estimates and the 89% compatibility intervals from the posterior estimates from these models. For each model, we ran four chains with 10,000 iterations each (half of which were warm up). We checked that the number of effective samples was sufficiently high and evenly distributed across all estimated variables such that autocorrelation did not influence the estimates. We also confirmed that in all cases the Gelman-Rubin convergence diagnostic, $\hat{R}$, was 1.01 or smaller, indicating that the chains had converged on the final estimates (Gelman & Rubin, 1995). In all cases, we also plotted the model inferences onto the distribution of the raw data to confirm that the estimated predictions matched the observed patterns.

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

We mathematically derived predictions about the choice behavior of individuals using equations 1-3. We determined the values for ϕ and λ that individuals would need to reach the passing criterion in 50 trials or fewer in the serial reversal learning experiment. 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 trial t as the likelihood for the changes in associations at trial t+1.

Equation 3a (dynamic association for the rewarded option):

$A_{r,t+1}$ = ((1-ϕ) x $A_{r,t}$ + ϕ x π) x P_t + (1-P_t) x $A_{r,t}$.

Equation 3b (dynamic association for the non-rewarded option):

$A_{n,t+1}$ = (1-P_t) x (1-ϕ) x $A_{n,t}$ + P_t + (1-P_t) x $A_{n,t}$.

In equations 3a and 3b, the association with both the rewarded, A_r, and the non-rewarded, A_n, options change from trial t to trial t+1 depending on the association updating rate ϕ and the probability, P, that the association was chosen during trial t. The probability, P, is calculated using equation 2. The reward π is set to 1. We used these equations to explore which combinations of ϕ and λ would lead to an individual choosing the rewarded option above the passing criterion in 50 trials or less after a reversal in the rewarded option. We assumed serial reversals, and therefore set the initial associations after the reversal to 0.1 for the now rewarded option (previously unrewarded, so low association) and to 0.7 for the now unrewarded option (previously rewarded, so high association). We obtained these associations from the end of the reversal learning simulation in question 1. For a given combination of ϕ and λ, we first used equation 2 to calculate the probability that an individual would choose the rewarded option during this first trial after the reversal (where the remaining probability reflects the individual choosing the non-rewarded option). We then used equations 3a and 3b to update the associations. We repeated the calculations of the probabilities and the updates of the associations 50 times to determine whether individuals with a given combination of ϕ and λ would reach the passing criterion within either 50 (the serial reversal passing criterion) or 40 trials (the average observed among the trained grackles). For ϕ ranging between 0.02 and 0.10, we manually explored which λ would be needed such that an individual would choose the rewarded option with more than 50% probability at trial 31 (or 21) and with more than 85% probability at trial 50 (or 40), to match the passing criterion of 17 correct out of the last 20 trials (17/20=0.85).

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

We fit the Bayesian reinforcement learning model to the data of both the control and the trained grackles. Based on the simulation results indicating that the minimum sample per individual required for accurate estimation are two learning phases across one reversal (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 trained individuals. For the control grackles, these estimated ϕ and λ values also reflected their behavioral flexibility at the end of the reversal learning experiment. For the trained grackles, we additionally calculated ϕ and λ separately for their final two reversals at the end of the serial reversal to infer the potential changes in the parameters.

We determined how the ϕ and λ values influenced the number of trials individuals needed during a reversal by 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 the passing criterion. We fit this model to the data from the simulated individuals, as well as to the data from the grackles. 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. The model is as follows:

v_b ~ Poisson(μ),
log $\mu = \alpha + \beta_1$ x ϕ_b + β_2 x λ_b,
α ~ Normal(4.5,1),
β_1 ~ Normal(0,1),
β_2 ~ Normal(0,1),

where the number of trials each individual needed during their reversal, v_b, was linked with separate slopes, β_1 and β_2, to both the ϕ and λ of each individual. The mean of the prior distribution for the intercept, α, was based on the average number of trials (90) grackles in Santa Barbara were observed to need to reach the criterion during their one 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 β_1 and β_2 with ϕ and λ were centered on zero, indicating that, *a priori*, we did not bias these toward a relationship.

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

For the subset of grackles that were part of the serial reversal group, we calculated how much their ϕ and λ changed from their first to their last reversal. The model is as follows:

$\phi_{b,r}$ or $\lambda_{b,r}$ ~ Normal(μ_b , σ),
$\mu_b = \alpha_b + \beta_b$ x r,

$$\begin{bmatrix}\alpha_b\\ \beta_b\end{bmatrix} \sim MVNormal\left(\begin{bmatrix}\alpha\\ \beta\end{bmatrix}, S\right),$$

$$S = \begin{pmatrix}\sigma_\alpha & 0\\ 0 & \sigma_\beta\end{pmatrix} Z \begin{pmatrix}\sigma_\alpha & 0\\ 0 & \sigma_\beta\end{pmatrix},$$

Z ~ LKJcorr(2),
α ~ Normal(5,2),
β ~ Normal(-1,0.5),
δ_b ~ Exponential(1),
σ ~ Exponential(1),

where each grackle, b, has two ϕ and λ values, one from the beginning ($r = 0$) and one from the end of the serial reversal experiment ($r = 1$). We assume that there are individual differences that persist through the experiment (intercept α_b), and that how much individuals change from the first to the last reversal, r, estimated by β_b, might also depend on their values at the beginning. Each bird has an intercept and slope with a prior distribution defined by the two dimensional Gaussian distribution ($MVNormal$) with means, σ_α and σ_β, and covariance matrix, S. The covariance matrix, S, is factored into separate standard deviations, δ_b, and a correlation matrix, Z. The prior for the correlation matrix is set to come from the Lewandowski-Kurowicka-Joe ($LKJcorr$) distribution, and is set to be weakly informative and skeptical of extreme correlations near -1 or 1.

We also fit a model to assess whether individual improvement 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. The model is as follows:

Δv_b ~ Normal(μ_b, σ),
$\mu_b = \alpha + \beta_1$ x $\Delta\phi_b + \beta_2$ x $\Delta\lambda_b$,
α_b ~ Normal(40, 10),
β_1 ~ Normal(0, 10),
β_2 ~ Normal(0, 10),
σ ~ Exponential(1),

where Δv_b, the improvement in the number of trials, is the difference in the number of trials between the first and the last reversal, and $\Delta\phi_b$ and $\Delta\lambda_b$ are the respective differences in these parameters between the beginning and the end of the serial reversal experiment. The remaining parameters in the model are as defined above.

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

We checked whether the ϕ and λ values of grackles at the beginning were associated with how much they changed (difference in values between beginning and end):

$\Delta\phi_b$ or $\Delta\lambda_b$ ~ Normal(μ_b , σ),
$\mu_b = \alpha + \beta$ x $\phi_{b,0}$ or $\lambda_{b,0}$,
α ~ Normal(0,1),
β ~ Normal(0,1),
σ ~ Exponential(1),

where $\Delta\phi_b$ and $\Delta\lambda_b$ are the changes in these values, and $\phi_{b,0}$ and $\lambda_{b,0}$ are the bird's values from their first reversal. The remaining parameters are as defined above.

We also checked whether the ϕ or λ values of grackles at the beginning were associated with the values they had at the end:

$\phi_{b,1}$ or $\lambda_{b,1}$ ~ Normal(μ_b , σ), $\mu_b = \alpha + \beta$ x $\phi_{b,0}$ or $\lambda_{b,0}$,
α ~ Normal(0,1),
β ~ Normal(0,1),
σ ~ Exponential(1),

where $\phi_{b,1}$ and $\lambda_{b,1}$ are from the last reversal. The remaining parameters are as defined above.

In addition, we assessed whether grackles at the end of the serial reversal experiment focused more on one of the processes, ϕ or λ, than the other. The model is as follows:

$\phi_{b,1}$ ~ Normal(μ_b , σ),
$\mu_b = \alpha + \beta$ x $\lambda_{b,1}$,
α ~ Normal(0,1),
β ~ Normal(0,1),
σ ~ Exponential(1),

where the values estimated for birds from their last reversal are assessed for an association. All parameters as defined above.

We used the ϕ and λ values estimated from individuals after they completed the serial reversal learning experiment to better understand how individuals behave after a reversal in which option is rewarded. We chose two combinations of ϕ and λ from the end of the range of values observed among the individuals who completed the serial reversal learning experiment. The first combines a slightly higher ϕ (0.09) with a slightly lower λ (3), and the second combines a slightly lower ϕ (0.06) with a slightly higher λ (4). We entered these values in equations 2, 3a, and 3b. We plotted the change in the probability that an individual will choose the rewarded option across the first 40 trials after a switch. As above, we set the initial associations to the now rewarded option to 0.1 and to the now non-rewarded option to 0.7.

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

We modified the statistical models in the original article (Logan et al., 2023a) that linked performance on the serial reversal learning tasks to performance on the multi-option puzzle boxes, replacing the previously used independent variable of the number of trials needed to reach criterion in the last reversal with the estimated ϕ and λ values from the last two reversals (trained grackles) or the initial discrimination and the first reversal (control grackles). We assumed that there also might be non-linear, U-shaped relationships between ϕ and/or λ and the performance on the multi-option puzzle box. For the number of options solved, we fit a binomial model with a logit link:

o_b ~ Binomial(4, p),
logit(p) ~ $\alpha + \beta_1$ x $\phi + \beta_2$ x $\phi^2 + \beta_3$ x $\lambda + \beta_4$ x λ^2,
α ~ Normal(1, 1),
β_1 ~ Normal(0, 1),
β_2 ~ Normal(0, 1),
β_3 ~ Normal(0, 1),
β_4 ~ Normal(0, 1),

where o_b is the number of options solved on the multi-option puzzle box, 4 is the total number of options on the multi-option puzzle box, p is the probability of solving any one option across the whole experiment, α is the intercept, β_1 is the expected linear amount of change in p for every one unit change in ϕ in the reversal learning experiments, β_2 is the expected non-linear amount of change in p for every one unit change in ϕ^2, β_3 the expected linear amount of change for changes in λ, and β_4 is the expected non-linear amount of change for changes in λ^2.

For the average latency to attempt a new option on the multi-option puzzle box as it relates to ϕ and λ, we fit a Gamma-Poisson model with a log-link:

n_b ~ Gamma-Poisson(m_b, s),
log(m_b) ~ α + β_1 x ϕ + β_2 x ϕ^2 + β_3 x λ + β_4 x λ^2,
α ~ Normal(1, 1),
β_1 ~ Normal(0, 1),
β_2 ~ Normal(0, 1),
β_3 ~ Normal(0, 1),
β_4 ~ Normal(0, 1),
s ~ Exponential(1),

where n_b is the average latency, counted as the number of seconds, to attempt a new option on the multi-option puzzle box, m_b reflects the tendency of each grackle to wait (if they have a higher tendency to wait, they have a longer latency), s controls the variance (larger values mean the overall distribution is more like a pure Poisson process in which all grackles have the same tendency to wait), α is the intercept, β_1 is the expected linear amount of change in latency for every one unit change in ϕ, β_2 is the expected non-linear amount of change in latency for every one unit change in ϕ^2, β_3 the expected linear amount of change for changes in λ, and β_4 is the expected non-linear amount of change for changes in λ^2.

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). The estimated ϕ and λ values were consistently shifted away from the values assigned to the simulated individuals. The estimated ϕ values were consistently smaller than those assigned to the simulated individuals (here and hereafter, we report the posterior mean slope of the association, the β factor in the statistical models, with the 89% compatibility interval; +0.15, +0.06 to +0.23, n=626 simulated individuals), while the estimated λ values were consistently estimated to be larger than the assigned λ values (+6.04, +5.86 to +6.22, n=626 simulated individuals)(Figure 2). The model assumed that, during the initial association learning, individuals only needed to experience each option once to learn which of the two options to choose. This would lead to a difference in the associations between the two options. The model assumed that the simulated individuals would not require a large ϕ because a small difference in the associations would already be informative. Individuals would then be expected to consistently choose the option that was just rewarded, and they would because of their large λ. In addition, these shifts mean that ϕ and λ are no longer estimated independently. The model estimated that, if an individual had a particularly low ϕ value, it would require a particularly high λ value. This dependency (which was due to inaccurate estimation) between ϕ and λ led to a strong positive correlation in the estimated values of ϕ and λ (+505, +435 to +570, n=626 simulated individuals). This correlation is erroneous because individuals were assigned their λ values independent of their ϕ values, with the different combinations across the populations meaning that high and low values of λ were assigned to individuals with both high and with low ϕ values.

In contrast, when we combined data from across the initial discrimination learning and the first reversal, the model recovered the ϕ and λ values that the simulated individuals had been assigned (ϕ: intercept 0.00, -0.01 to +0.01, slope +0.96, +0.70 to +1.21, n=626 simulated individuals; λ: intercept +0.01, -0.15 to +0.16, slope +0.98, +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. In combination, the choices before and after a reversal make it possible to infer the assigned values (initial learning plus first reversal, or two subsequent reversals). Given that the choices individuals make during any given trial are probabilistic, the estimation can show slight deviations from the assigned values. However, this was also reflected in the uncertainties of the estimated values, and the compatibility intervals of the estimated values included the value assigned to the simulated individuals (Figure 2).

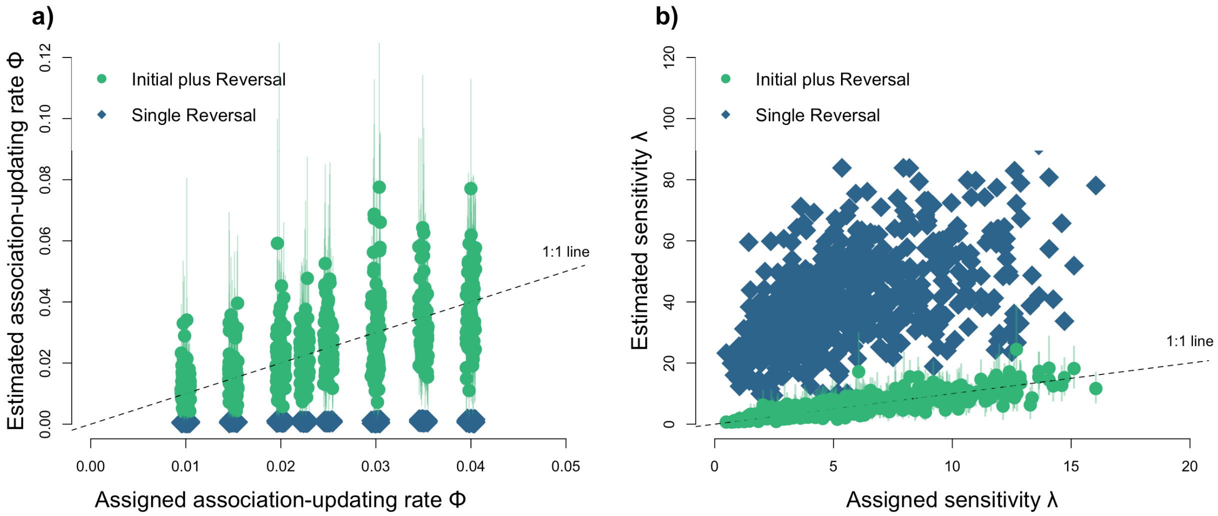

Figure 2. Both the ϕ (a) and the λ (b) values are only estimated correctly by the Bayesian reinforcement model when the choices from the simulated reversal learning are combined with the previous initial association learning (green circles). When ϕ was estimated based on the choices made only during the first reversal, the estimates were consistently lower than the assigned values, particularly for large ϕ values (a, blue diamonds). The model assumed that the simulated individuals chose the rewarded option consistently not because they updated their associations, but because they consistently chose the rewarded option as soon as they had learned which option was rewarded. Accordingly, the model wrongly assigned individuals very high λ values (b, blue diamonds). Lines around the points indicate the 89% compatibility intervals of the estimated values and are only shown for the estimation from the combined choices from the initial and reversal learning - the approach we ended up using for the remaining analyses.

2) Role of ϕ and λ on performance in the serial reversal learning task based on analytical predictions

To determine how ϕ and λ influence behavior during the serial reversals, we performed a mathematical derivation using equations 2, 3a, and 3b. We identified the range of values for ϕ and for λ that we would expect in individuals who quickly change their behavior after a revrsal in the serial reversal learning experiment. We found that ϕ needs to be 0.04 or larger for individuals to be able to reach the passing criterion in 40 or 50 trials after a reversal (Figure 3). With smaller ϕ values, individuals are expected to take longer before switching to the newly rewarded option because they would not update their associations fast enough. We also found that, as ϕ values increased beyond 0.04, individuals could have a larger range of λ values and still reach the passing criterion in 40 or 50 trials. However, the λ values are expected to be small, less than 10 and as low as 2.4.

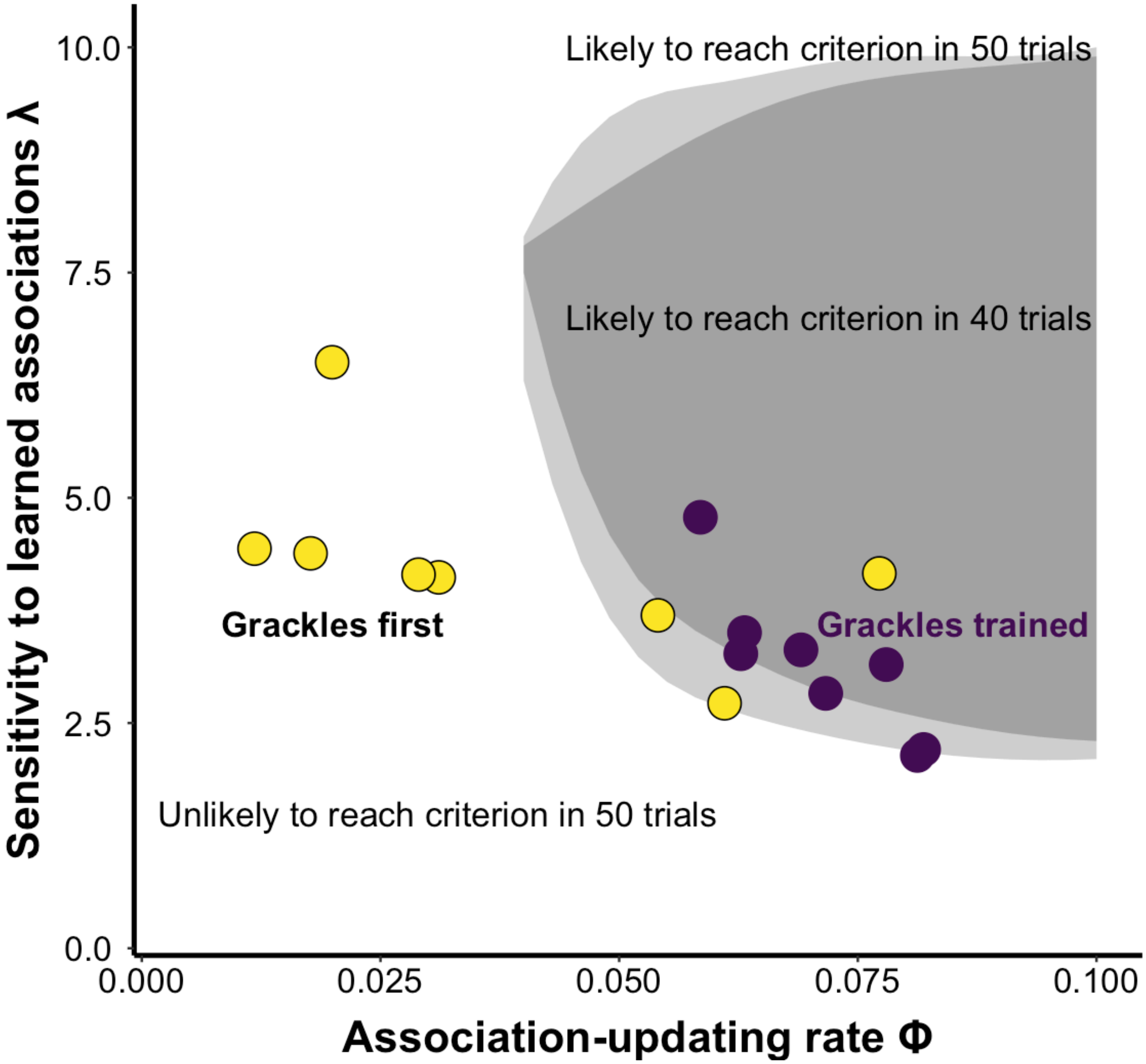

Figure 3. Individuals are more likely to reach the criterion of choosing the correct option 17 out of 20 times during the serial reversal trials if they update their associations quickly (high ϕ). Using the equations, we found the space of values individuals are predicted to need to reach the passing criterion in 40 trials or less (dark gray shading) or 50 trials or less (light gray shading). Individuals are predicted to need a large ϕ to completely reverse their associations with the two options presented in the serial reversal learning experiment. The predicted λ values are expected to be relatively small. The figure also shows the median ϕ and λ values estimated for the trained grackles during their first reversal (yellow), when they needed on average 70 trials to reach criterion, and during their last reversal (purple) when they needed on average 40 trials to reach criterion. During the training, grackles increased their ϕ to become efficient at gaining the reward and reaching the criterion. They also showed a slight decline in their λ, allowing them to explore the alternative option after a reversal.

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

We estimated ϕ and λ after the first reversal for all grackles, 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 reversal. 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 observed grackle data because we found that the number of trials grackles needed to reach criterion during the initial learning and the first reversal learning were correlated (+1.61, +1.53 to +1.69, n=19 grackles), where grackles needed about 28 trials more to reach criterion during the first reversal than they needed during the initial association learning. Therefore, we estimated ϕ and λ for the grackles based on their performance in the initial discrimination plus first reversal, and for the trained grackles additionally based on their performance in their final two reversals. The inferred ϕ values for the grackles in Arizona ranged between 0.01 and 0.10, and the λ values between 2.1 and 6.5 (Figure 4).

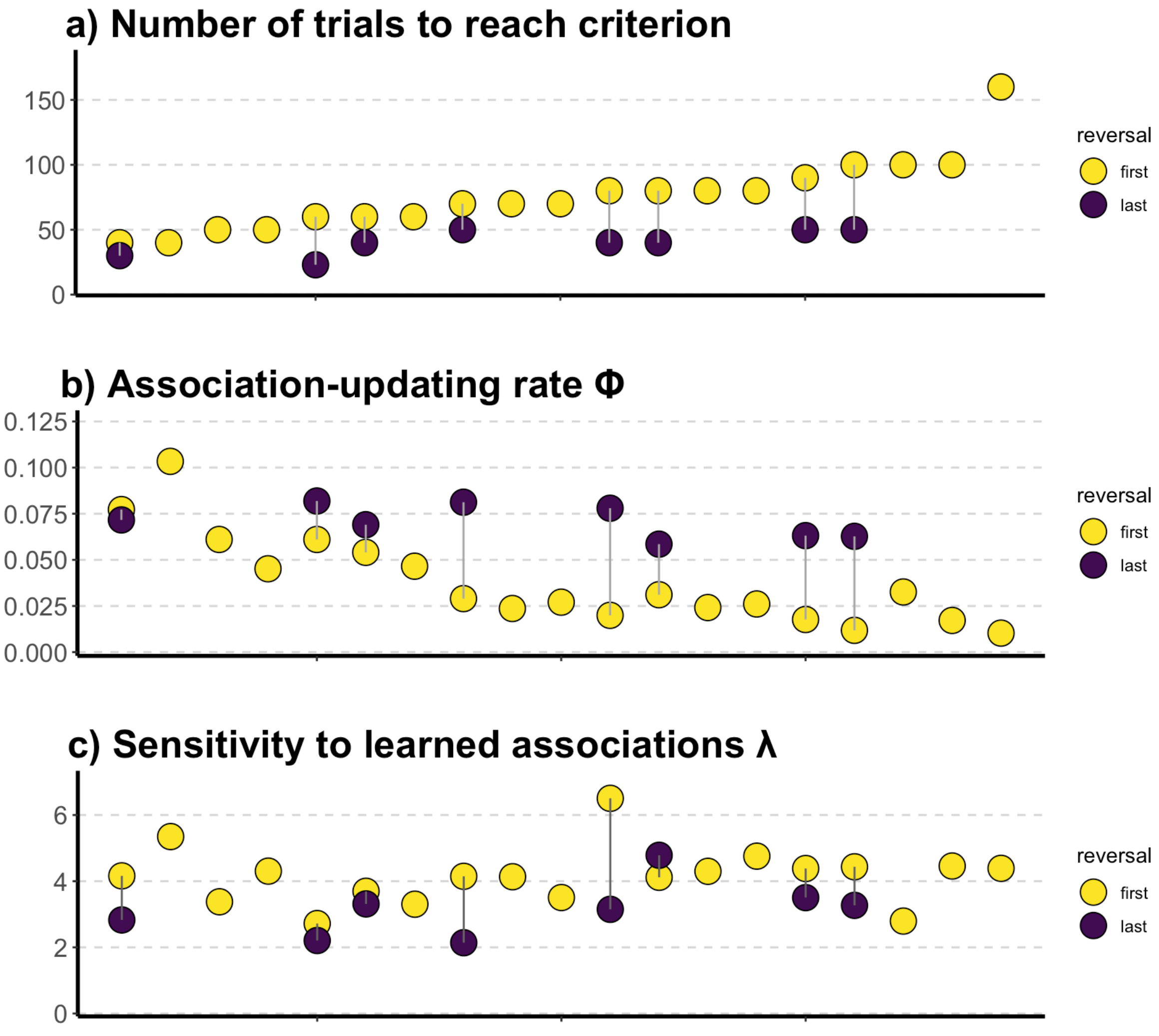

Figure 4. Comparisons of the parameters estimated from the behavior of 19 grackles in the serial reversal task. The figure shows a) the number of trials to pass criterion for the first reversal (yellow; all grackles) and

the last reversal (purple; only trained grackles); b) the ϕ values reflecting the rate of updating associations with the two options inferred from the initial discrimination and first reversal (yellow; all grackles) and from the last two reversals (purple; trained grackles); and c) the λ values reflecting the sensitivity to the learned associations inferred from the initial discrimination and first reversal (yellow; all grackles) and from the last two reversals (purple; trained 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 trained grackles, their ϕ values changed more consistently than their λ values: their ϕ values were generally higher than those observed in the control individuals, while their λ values remained within the range observed for the control group.

For the 19 grackles that finished the initial learning and the first reversal, only their ϕ (-20.69, -26.17 to -15.13; n=19 grackles), but not their λ (-0.22, -5.66 to +5.26, n=19 grackles), predicted the number of trials they needed to reach criterion during their first reversal (Figure 4). A grackle with a ϕ of 0.01 higher 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 the same as the slope from the simulations (-20.69 vs -20.48, Figure 5). The number of trials grackles needed to reach the criterion given their ϕ values fell right into the range for the relationship between ϕ and the number of trials for simulated individuals (Figure 5). Even though the 8 trained 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 between the number of trials and either parameter in the last reversals (ϕ: -7.38, -15.97 to +1.28; λ: -4.00, -12.53 to +4.61, n=8 grackles).

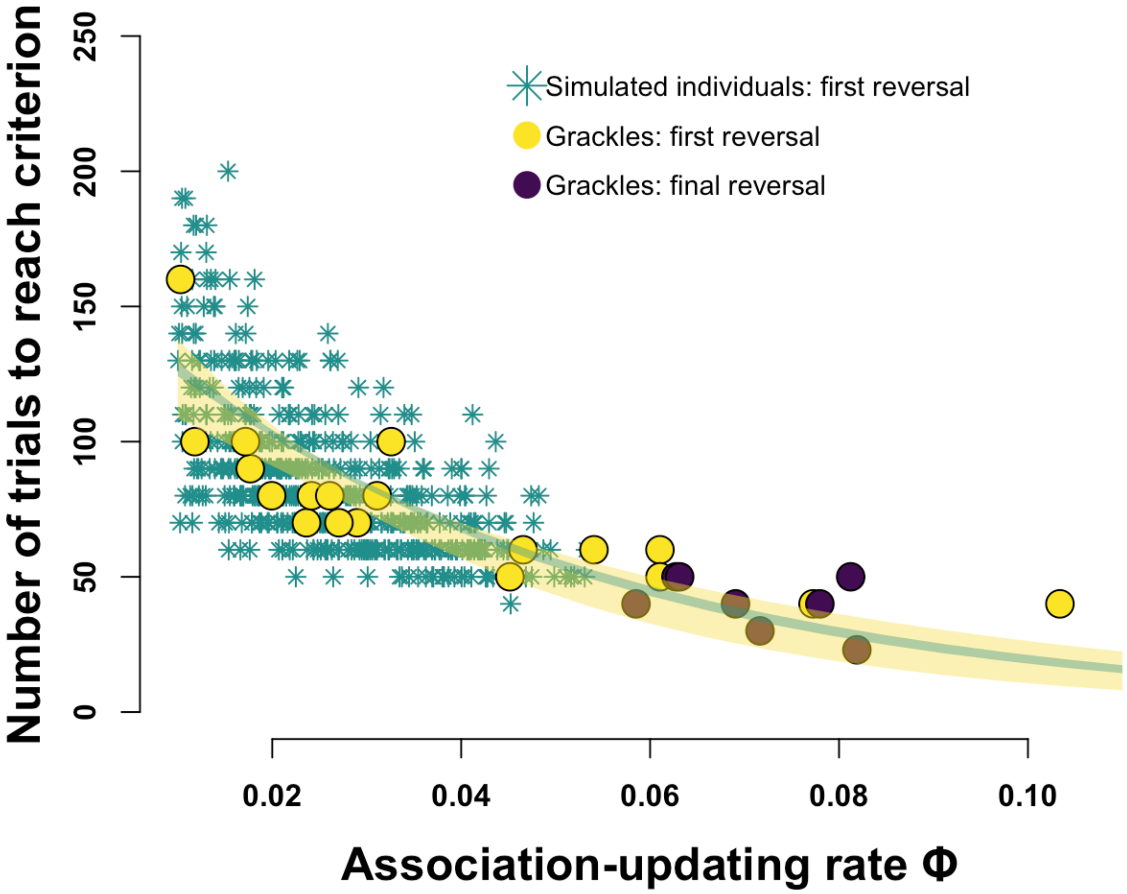

Figure 5. Relationship between ϕ and the number of trials needed to reach criterion observed among grackles during their first reversal (yellow points; all grackles) and last reversal (purple points; trained grackles), as well as for the first reversal for the simulated individuals (green stars). The observed grackle data falls within the range of the number of trials individuals with a given ϕ value are expected to need. Grackles show the same negative correlation between their ϕ and the number of trials needed to reach criterion as the simulated individuals (the shaded lines display the 89% compatibility interval of the estimated relationships between ϕ and the number of trials for both the simulated individuals, green line, and for the grackles during their first reversal, yellow line). We did not simulate individuals with ϕ values larger than 0.05 because we did not observe larger values among grackles in the Santa Barbara population, which we used to parameterize the simulations.

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

Grackles who experienced the serial reversal learning reduced the number of trials they needed to reach the criterion from an average of 75 to an average of 40 by the end of their experiment (-30.02, -36.05 to -24.16, n=8 grackles). For the trained 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 serial reversals) to 0.07 in their last two reversals (+0.03, +0.02 to +0.05, n=8). The λ values of the trained grackles went slightly down from 4.2 (again, similar to control grackles) to 3.2 (-1.07,

-1.63 to -0.56, n=8 grackles) (Figure 4). The number of trials to reverse that we observed in the last reversal, 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 training 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 training fits with the outcome from the mathematical predictions: 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 increase in their ϕ values (+7.59,l +1.54 to +14.22, n=8 grackles). The improvement did not match the change in their λ values (+2.17, -4.66 to +9.46, n=8 grackles) because, as predicted, the trained grackles showed a decreased λ in their last reversal. This decrease in λ meant that grackles quickly found the rewarded option after a reversal in which option was rewarded. Across all grackles, in their first reversal, grackles chose the newly rewarded option in 25% of the first 20 trials, while the trained grackles in their final reversal chose correctly in 35% of the first 20 trials. Despite their low λ values, trained grackles still chose the rewarded option consistently because the increase in ϕ compensated for this reduced sensitivity (Figure 3; also see below).

5) Individual consistency in the serial reversal learning task

We found a negative correlation between the ϕ estimated from an individual's performance in the first reversal and how much their ϕ changed through the serial reversals (-0.84, -1.14 to -0.52, n=8 grackles). The larger increases in ϕ for individuals who had smaller ϕ values at the beginning made it so that individuals ended up with similar ϕ values at the end of the serial reversals. We did not find consistent individual variation among grackles in ϕ: their beginning and end ϕ values were not correlated (-0.21, -1.55 to +1.35, n=8 grackles). Similarly, individuals who started with a high λ changed more than individuals who already had a lower λ during the first reversal (-0.44, -0.76 to -0.10, n=8 grackles). Individuals changed to different degrees, such that those with higher λ values in the beginning did not necessarily have higher λ values than other individuals at the end of the serial reversal learning: their values at the beginning and end were not associated (+0.17, -0.67 to +0.97, n=8 grackles).

Individuals appeared to adjust their behavior differently to improve their performance through the serial reversals. There was a negative correlation between an individual's ϕ and λ after their last reversal (-0.39, -0.72 to -0.06, n=8 grackles). While, as predicted, essentially all grackles who experienced the serial reversal learning experiments increased their ϕ and decreased their λ (Figure 5), individuals ended up with different combinations of the two parameters and all combinations allowed them to switch to the newly rewarded option in 50 trials or less. Individuals ended up along the lower (on the y-axis) side of the space of values that are needed to reach criterion in the serial reversal learning experiment (the lower edge of the light gray shading in Figure 3).

We used the values from the two individuals at the ends of the spectrum: the one with the highest ϕ and lowest λ, and the one with the lowest ϕ and highest λ. Based on equations 1-3, individuals with a slightly higher ϕ and slightly lower λ are expected to learn the new reward associations after a reversal more quickly. However, they continue to explore the alternative option even after they learned the new association and therefore do not exclusively choose the rewarded option (red line in Figure 6). Individuals with a slightly lower ϕ and a slightly higher λ are expected to take slightly longer to learn that the reward has switched, but once they reversed their association, they rarely choose the unrewarded option (purple line in Figure 6). Together, this suggests that all individuals improved by the same extent through the training such that the differences in their performances persisted, but they utilized slightly different behaviors to quickly reach criterion after a reversal.

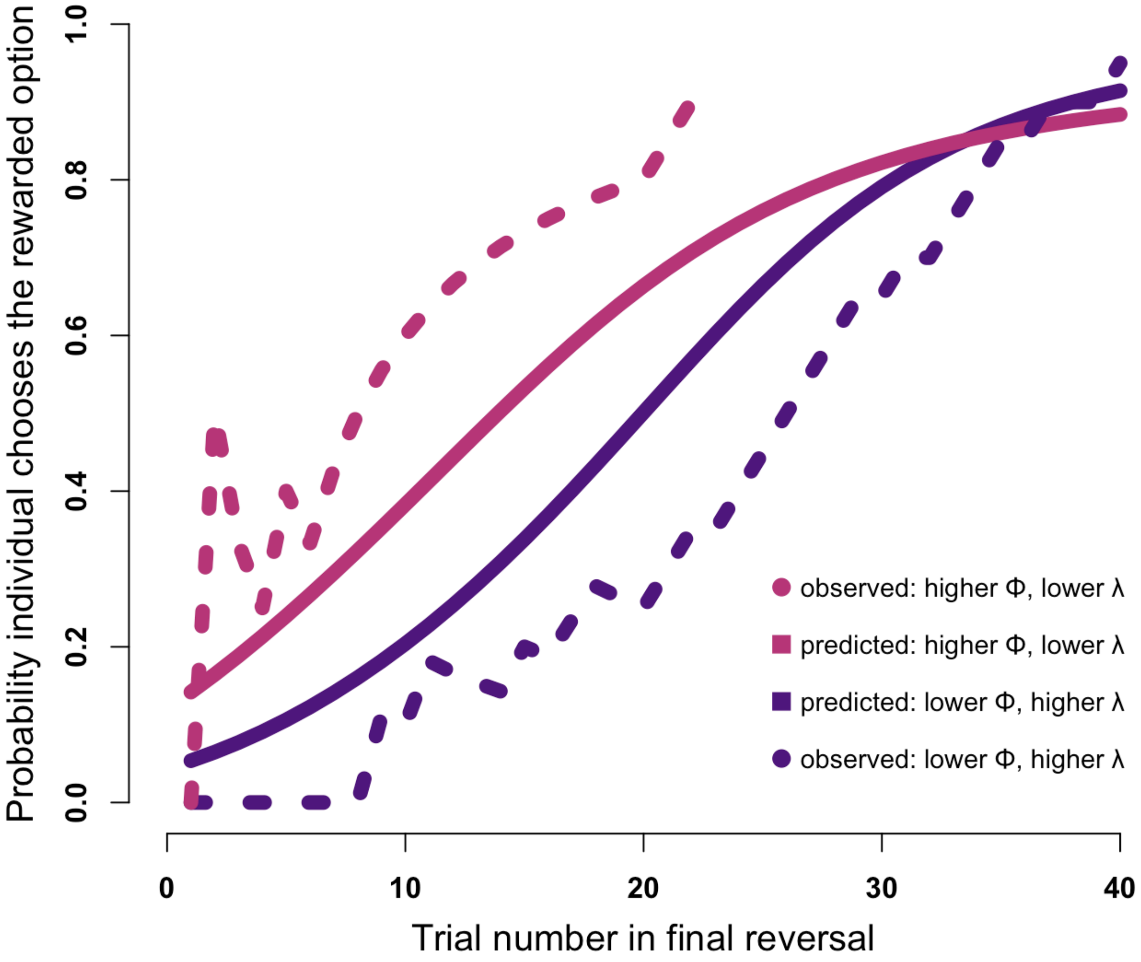

Figure 6. Predicted and observed performance curves of individuals with different ϕ and λ values in their last reversal in the serial reversal learning experiment. The dotted lines present the behavior of the grackles Burrito (red on the top, $\phi = 0.08$, $\lambda = 2.1$) and Habanero (purple on the bottom, $\phi = 0.06$, $\lambda = 4.8$) during their last reversal. The dotted lines show the probability with which they chose the rewarded option during their last 20 trials. We used their ϕ and λ values in the analytical equations 2, 3a, and 3b to derive the predicted curves (solid lines) of the probability that an individual will choose the option that is currently rewarded for each trial number. Individuals with a higher ϕ and lower λ (red lines on the top) are expected and observed to quickly learn the new association, but continue to explore the unrewarded option even after they learned the association, leading to a curve with a more gradual increase through the trials. Individuals with a lower ϕ and higher λ (purple lines on the bottom) are expected and observed to take longer to switch their association, but, once they do, they rarely choose the non-rewarded option, leading to a more S-shaped curve where the initial increase in probability is lower and more rapid later.

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

We found that the number of options solved for both the wooden and the plastic multi-option puzzle boxes as well as the latency to solve a new option on both boxes correlated with the underlying flexibility parameters ϕ and λ. In particular, the λ values individuals had after their last reversal had a U-shaped relationship with the number of options solved on both the plastic (λ +0.17, -0.27 to +0.61; λ^2 +0.59, +0.18 to +1.02; n=15 grackles) and the wooden multi-option puzzle boxes (λ +0.03, -0.50 to +0.59; λ^2 +0.63, +0.12 to +1.19; n=12 grackles). There was no association between the number of options solved on either box and ϕ (plastic box: ϕ +0.03, -0.38 to +0.43; ϕ^2 -0.16, -0.59 to +0.28, n=15 grackles; wooden box: ϕ -0.08, -0.62 to +0.47, ϕ^2 +0.43, -0.08 to +0.97, 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 7).

For the latency to attempt a new option on the plastic box, there was also a U-shaped association, but only with ϕ (ϕ -0.66, -1.30 to +0.0.06; ϕ^2 +0.58, -0.06 to +1.30; λ +0.14, -0.45 to +0.70; λ^2 +1.09, +0.28 to +1.87; 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). There was no association between the latency to attempt a new option on the wooden box with either ϕ (-0.62, -1.46 to +0.14; ϕ^2 +0.39, -0.47 to +1.26; 11 grackles) or λ (+0.13, -0.66 to +0.86; λ^2 +0.32, -0.62 to +1.35; n=11 grackles).

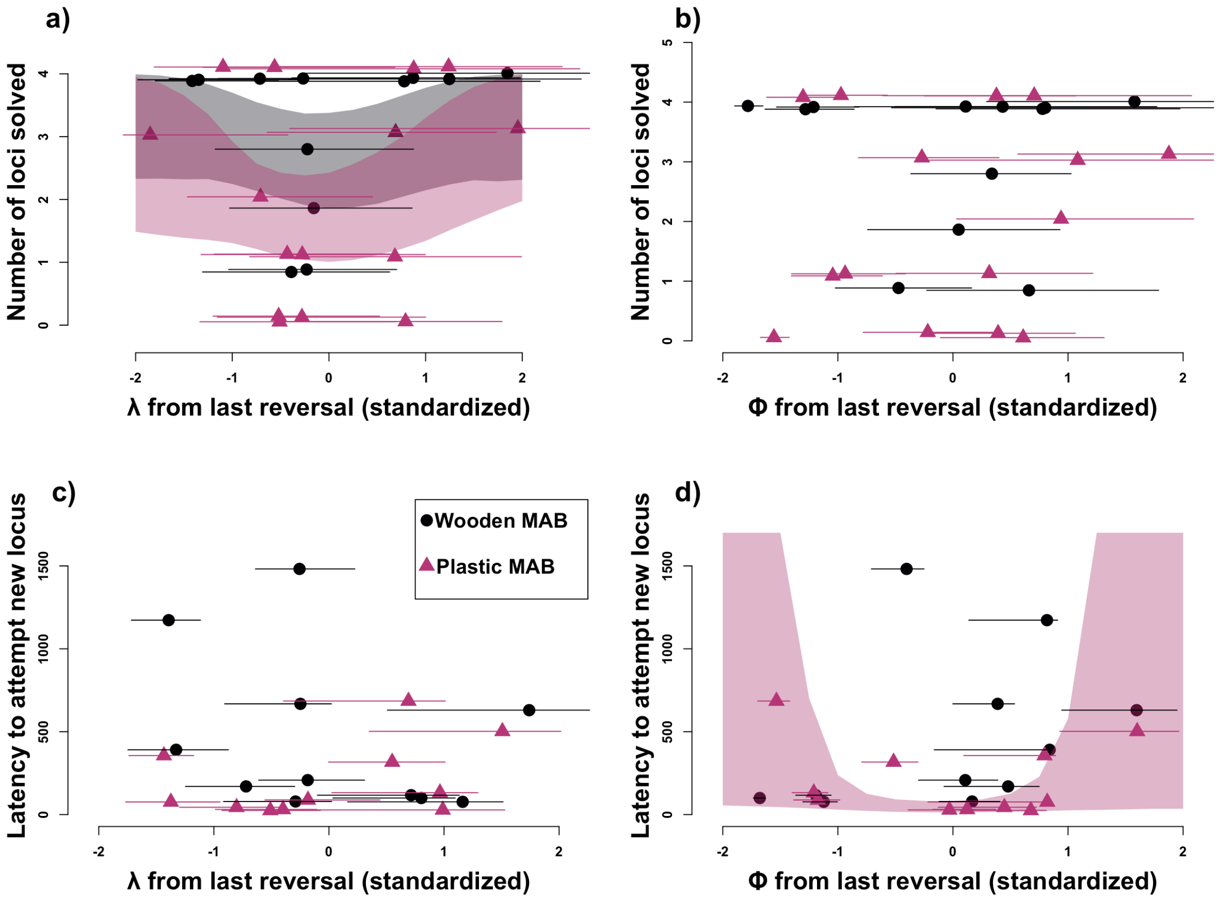

Figure 7. Relationships between ϕ and λ from the last reversal and performance on the wooden (black dots) and plastic (red triangles) multi-option puzzle boxes. Grackles with intermediate λ values in their

last reversal (a) were less likely to solve all four options on both multi-option puzzle boxes than grackles with either high or low λ values. Grackles with intermediate ϕ values had a shorter latency to attempt a new option on the plastic box (d). There were 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 ϕ and the latency to attempt a new option on the wooden box (d). An individual's ϕ and λ values changed slightly between the top and bottom rows because values were standardized for each plot and not all individuals were tested on both boxes, therefore values changed relative to the mean of the points included in each plot. Lines around each point indicate the 89% compatibility intervals for the estimated ϕ and λ values.

Discussion

Our analyses show that grackles change their behavioral flexibility to match the reliability and stability of the environment they experience. The application of the Bayesian reinforcement learning model to the grackle serial reversal learning data revealed that the association-updating rate, ϕ, explained more of the interindividual variation in how many trials individuals needed to reach criterion during a reversal than the sensitivity to learned associations, λ. We found that, as predicted given the reliability of cues and frequent switches in the serial reversal learning experiment, ϕ more than doubled between the first and last reversals, whereas λ slightly declined. Even though all grackles changed their behavior in the expected direction by the end of the serial reversal learning experiment, we found that these trained individuals used slightly different approaches from across the range of possible behaviors. Finally, these changes in how the trained individuals explored alternative options and switched preferences in light of recent information subsequently also influenced their behavior in a different experimental test of behavioral flexibility and innovativeness. Grackles with intermediate sensitivities to learned associations solved fewer options on both multi-option puzzle boxes than grackles with either low or high sensitivities. Accordingly, the trained grackles not only changed their behavior within the specific serial reversal learning task, they also more generally changed their behavior across contexts in response to their training. Our findings show that grackles modulate their behavioral flexibility in response to the high reliability of cues and frequent changes in associations they experienced in the serial reversal learning experiment.

Applying the Bayesian reinforcement learning model to serial reversal data shows that participating in the serial reversal learning experiment made grackles change how much they value new information over old to update their associations, and how much they continue to explore alternative options or whether they are sensitive to the reward they are receiving at their current choice. Grackles coming into the experiment already had different rates of updating their associations and different sensitivities to learned associations, suggesting they had different experiences of how predictable cues are and how frequently their environment changes. In the urban environment they live in, changes are presumably frequent, so they would be expected to change their associations frequently (Lee & Thornton, 2021; Breen & Deffner, 2023). In line with this, the association-updating rate, ϕ, appeared to explain more of the variation in how many trials individuals needed to reach the criterion of consistently choosing the rewarded option during a single phase as early as in their first reversal. Other recent applications of the Bayesian reinforcement learning model to serial reversal learning experiments also found that the association-updating rate explains more of the variation in the number of trials to pass criterion (squirrel monkeys Bari et al., 2022; mice Metha et al., 2020; Woo et al., 2023). In response to learning that the cues are highly reliable and the reversals are relatively frequent, the grackles increased their association-updating rate, ϕ, which on average doubled across individuals, changing more for individuals who started off with lower ϕ values. Grackles also changed their sensitivity to the learned associations, λ, during the serial reversals in line with the prediction that they benefit from being open to exploring the alternative option when the associations between cues and rewards switch frequently. Individuals changed their ϕ and λ more if their initial values were further from those necessary to reach the passing criterion quickly. Individuals who passed their first reversal in 50 trials or less, changed ϕ and λ only slightly by the end of the serial reversal learning experiment. Among the trained grackles, who all required very few trials to consistently reach the criterion by the end of the experiment, we observed different approaches (see also Chen et al., 2021). Some individuals seemed more focused on the frequent changes, such that they kept exploring the alternative options and changed their associations as soon as they encountered new information. These individuals reached the passing criterion quickly because they switched to the newly rewarded option soon after a reversal. However, their continued exploration of the alternative option meant that they still needed several trials to reach the criterion. Other individuals seemed to place more emphasis on the reliability of the cues, focusing on the rewarded option after they learned that the cues had reversed. These individuals reached the passing criterion quickly because they consistently chose the rewarded option. However, these grackles needed a few more trials after a reversal began to switch to the new option. At the beginning of the experiment, the grackles showed a diversity of ϕ and λ values and, because they had no prior experience, they did not show specific approaches to quickly reach the criterion. With the variables we measured at the beginning of the serial reversal learning experiment, we could not predict which approach grackles ended up with after the serial reversals.

The changes in behavioral flexibility that the grackles showed during the serial reversal learning experiment influenced their subsequent behavior in other tasks. The analyses linking ϕ and λ to the performance on the multi-option puzzle boxes show that the different approaches grackles utilized to improve their performance during the serial reversal learning experiment subsequently appeared to influence how they solved the multi-option puzzle box. 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 either do not update their associations even though the first option is no longer rewarded or they do not explore as much due to their small λ. We also found that grackles with intermediate values of λ solved fewer puzzle box 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. We are limited in our interpretation by the small sample sizes for the multi-option puzzle boxes. We have some indication that experiencing the serial reversal learning experiment continued to shape the behavior of the grackles after releasing them back to the wild. Individuals who changed their ϕ and λ more during the serial reversal learning experiment appeared to switch more frequently between food types and foraging techniques (Logan et al., 2024). It took a grackle on average one month to pass the serial reversal learning experiment (Logan et al., 2023a), and the observations of the foraging behavior in the wild continued for up to 8 months after individuals were released (Logan et al., 2024). This indicates that the effects of enhancing flexibility are durable and generalize to other contexts. In grackles, behavioral flexibility does not change within days or only during certain critical periods. Our results suggest that individuals change their behavioral flexibility to match their environment if they experience the same conditions repeatedly across weeks.

Most individuals that have been tested in serial reversal learning experiments thus far show improvements throughout the reversals, suggesting that most species can modulate their behavioral flexibility in response to the predictability and stability of their environments (e.g. Warren & Warren, 1962; Komischke et al., 2002; Bond et al., 2007; Strang & Sherry, 2014; Chow et al., 2015; Cauchoix et al., 2017; Degrande et al., 2022; Erdsack et al., 2022). Previous studies used summary statistics to describe how the behavior of individuals changes during the serial reversal learning experiment (e.g. Federspiel et al., 2017) or show changes in learning curves (e.g. Gallistel et al., 2004). As shown in Figure 6, we can recreate these learning curves from the inferred association-updating rates and sensitivities to learned associations. The advantage of the Bayesian reinforcement learning model with its two parameters of the association-updating rate and the sensitivity to learned associations is that it has a clear theoretical foundation of what aspects of the experimental setting should lead to changes in the behavior (Gershman, 2018; Metha et al., 2020; Danwitz et al., 2022; Woo et al., 2023). Based on our application here, the model appears to be sufficient to accurately represent the behavior of grackles in the serial reversal experiment. This suggests that the stability and reliability of the environment has a large influence on how individuals learn about rewards. The importance of experiencing stable and predictable environments potentially explains the difference between lab-raised and wild-caught animals in how they change their behavior during the serial reversal learning experiment. Many lab-raised animals were observed to switch to a “win-stay versus lose-shift” strategy, where only their most recent experience guided their behavior and they no longer explored alternative options (Mackintosh et al., 1968; Rayburn-Reeves et al., 2013). These animals generally experience very stable conditions during their lives, and often participate in large numbers of trials in an experiment. Accordingly, cues are reliable and changes are rare, so individuals would be expected to show the high association-updating rates and high sensitivities to learned associations that would lead to the “win stay versus lose shift” strategy. In contrast, wild-caught animals, including grackles, only slowly move away when an option is no longer rewarded and they continue to explore alternative options (Chow et al., 2015; Cauchoix et al., 2017). These individuals probably experience environments in which associations are not perfectly reliable and changes occur more gradually. These individuals are expected to show smaller sensitivities to their associations and therefore continue to explore their environment. This focus on the key pieces of information that individuals likely pay attention to when adjusting their behavior also provides ways to link their performances and inferred cognitive abilities to their natural behavior. We found that, for the grackles, the behavioral flexibility they exhibited at the end of the serial reversal learning experiment linked to their foraging behavior in the wild (Logan et al., 2024). The existing literature on foraging behavior, investigating trade-offs between the exploration versus exploitation of different options, has a similar focus on gaining information (exploration) versus decision making (exploitation) (Kramer & Weary, 1991; Berger-Tal et al., 2014; Addicott et al., 2017). Linking this framework to the concepts of reinforcement learning and decision making could provide further insights into the cognitive processes that are involved and the information that individuals might pay attention to. The approach we established here to study behavioral flexibility, linking the theoretical framework of the Bayesian reinforcement learning model to the specific experimental task of the serial reversal learning experiment and the natural behavior of individuals, offers opportunities to better understand cognition in the wild (Rosati et al., 2022).

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.

Ethics

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]).

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 the preregistration; Coulon for serving as the recommender on the post-study manuscript, and reviewers, Maxime Dahirel and an anonymous person, for their constructive feedback this manuscript 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 MA, Pearson JM, Sweitzer MM, Barack DL, Platt ML (2017) A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, **42**, 1931–1939. https://doi.org/10.1038/npp.2017.108

Agrawal S, Goyal N (2012) Analysis of thompson sampling for the multi-armed bandit problem. In: *Conference on learning theory*, pp. 39–1. JMLR Workshop; Conference Proceedings.

Bari BA, Moerke MJ, Jedema HP, Effinger DP, Cohen JY, Bradberry CW (2022) Reinforcement learning modeling reveals a reward-history-dependent strategy underlying reversal learning in squirrel monkeys. *Behavioral neuroscience*, **136**, 46. https://doi.org/10.1037/bne0000492

Bartolo R, Averbeck BB (2020) Prefrontal cortex predicts state switches during reversal learning. *Neuron*, **106**, 1044–1054. 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**, e95693. https://doi.org/10.1371/journal.pone.0095693

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

Blaisdell A, Seitz B, Rowney C, Folsom M, MacPherson M, Deffner D, Logan CJ (2021) Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles (version 5 of this preprint has been peer reviewed and recommended by peer community in ecology [https://doi.org/10.24072/pci.ecology.100076]). https://doi.org/10.31234/osf.io/z4p6s

Bond AB, Kamil AC, Balda RP (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**, 372. https://doi.org/10.1037/0735-7036.121.4.372

Boyce MS, Haridas CV, Lee CT, Group NSDW, *et al.* (2006) Demography in an increasingly variable world. *Trends in Ecology & Evolution*, **21**, 141–148.

Breen AJ, 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**, 827–874. 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**, 12945. https://doi.org/10.1038/s41598-017-13179-5

Chen CS, Knep E, Han A, Ebitz RB, Grissom NM (2021) Sex differences in learning from exploration. *Elife*, **10**, e69748. https://doi.org/10.7554/elife.69748

Chow PK, Leaver LA, Wang M, Lea SE (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**, 343. 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/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**, 547–563. https://doi.org/10.1007/s42113-022-00139-0

Daw ND, O'doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876–879. https://doi.org/10.1038/nature04766

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**, 100607. https://doi.org/10.1016/j.animal.2022.100607

Donaldson-Matasci MC, Bergstrom CT, Lachmann M (2013) When unreliable cues are good enough. *The American Naturalist*, **182**, 313–327.

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

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

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

Federspiel IG, Garland A, Guez D, Bugnyar T, Healy SD, Güntürkün O, Griffin AS (2017) Adjusting foraging strategies: A comparison of rural and urban common mynas (acridotheres tristis). *Animal cognition*, **20**, 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/10.31234/osf.io/qndba

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

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

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

Izquierdo A, Brigman JL, Radke AK, Rudebeck PH, Holmes A (2017) The neural basis of reversal learning: An updated perspective. *Neuroscience*, **345**, 12–26. https://doi.org/10.1016/j.neuroscience.2016.03.021

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

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

Lea SE, Chow PK, Leaver LA, McLaren IP (2020) Behavioral flexibility: A review, a model, and some exploratory tests. *Learning & Behavior*, **48**, 173–187. https://doi.org/10.3758/s13420-020-00421-w

Lee VE, Thornton A (2021) Animal cognition in an urbanised world. *Frontiers in Ecology and Evolution*, **9**, 120.

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

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

Lloyd K, Leslie DS (2013) Context-dependent decision-making: A simple bayesian model. *Journal of The Royal Society Interface*, **10**, 20130069.

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

Logan C, Lukas D, Geng X, LeGrande-Rolls C, Marfori Z, MacPherson M, Rowney C, Smith C, McCune

K (2024) Behavioral flexibility is related to foraging, but not social or habitat use behaviors in a species that is rapidly expanding its range. *EcoEvoRxiv*. https://doi.org/10.32942/X2T036

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

Logan CJ, Shaw R, Lukas D, McCune KB (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

Mackintosh N, McGonigle B, Holgate V (1968) Factors underlying improvement in serial reversal learning. *Canadian Journal of Psychology/Revue canadienne de psychologie*, **22**, 85. 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) *Rethinking: Statistical rethinking book package.*

Metha JA, Brian ML, Oberrauch S, Barnes SA, Featherby TJ, Bossaerts P, Murawski C, Hoyer D, Jacobson LH (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**, 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**, e1011430. https://doi.org/10.1371/journal.pcbi.1011430

Neftci EO, Averbeck BB (2019) Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, **1**, 133–143.

R Core Team (2023) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rayburn-Reeves RM, Stagner JP, Kirk CR, Zentall TR (2013) Reversal learning in rats (rattus norvegicus) and pigeons (columba livia): Qualitative differences in behavioral flexibility. *Journal of Comparative Psychology*, **127**, 202. https://doi.org/10.1037/a0026311

Rescorla RA, Wagner AR (1972) A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning II: Current theory and research* (eds Black AH, Prosy WF), pp. 64–99. Appleton-Century-Crofts, New York.

Rosati AG, Machanda ZP, Slocombe KE (2022) Cognition in the wild: Understanding animal thought in its natural context. *Current Opinion in Behavioral Sciences*, **47**.

Shettleworth SJ (2010) *Cognition, evolution, and behavior.* Oxford university press.

Sih A (2013) Understanding variation in behavioural responses to human-induced rapid environmental change: A conceptual overview. *Animal Behaviour*, **85**, 1077–1088.

Sol D, Timmermans S, Lefebvre L (2002) Behavioural flexibility and invasion success in birds. *Animal behaviour*, **63**, 495–502. https://doi.org/10.1006/anbe.2001.1953

Spence KW (1936) The nature of discrimination learning in animals. *Psychological review*, **43**, 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/.

Starrfelt J, Kokko H (2012) Bet-hedging—a triple trade-off between means, variances and correlations. *Biological Reviews*, **87**, 742–755.

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

Tello-Ramos MC, Branch CL, Kozlovsky DY, Pitera AM, Pravosudov VV (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/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/10.1214/20-BA1221

Warren J (1965a) 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 JM (1965b) The comparative psychology of learning. *Annual review of psychology*, **16**, 95–118. https://doi.org/10.1146/annurev.ps.16.020165.000523

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

Woo JH, Aguirre CG, Bari BA, Tsutsui K-I, Grabenhorst F, Cohen JY, 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/10.1101/2022.10.01.510477

Wright TF, Eberhard JR, Hobson EA, Avery ML, Russello MA (2010) Behavioral flexibility and species invasions: The adaptive flexibility hypothesis. *Ethology Ecology & Evolution*, **22**, 393–404. https://doi.org/10.1080/03949370.2010.505580