Using repeatability of performance within and across contexts to validate measures of behavioral flexibility

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

2023-01-30

Open… access code peer review data

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

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

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

Preregistration: html, pdf, rmd

Post-study manuscript (submitted to PCI Ecology for post-study peer review on 3 Jan 2022): html, preprint pdf at EcoEvoRxiv, rmd

ABSTRACT

Research into animal cognitive abilities is increasing quickly and often uses methods where behavioral performance on a task is assumed to represent variation in the underlying cognitive trait. However, because these methods rely on behavioral responses as a proxy for cognitive ability, it is important to validate that the task structure does, in fact, target the cognitive trait of interest rather than non-target cognitive, personality, or motivational traits (construct validity). Although it can be difficult, or impossible, to definitively assign performance to one cognitive trait, one way to validate that task structure is more likely to elicit performance based on the target cognitive trait is to assess the temporal and contextual repeatability of performance. In other words, individual performance is likely to represent an inherent trait when it is consistent across time and across similar or different tasks that theoretically test the same trait. Here, we assessed the temporal and contextual repeatability of performance on tasks intended to test the cognitive trait behavioral flexibility in great-tailed grackles (*Quiscalus mexicanus*). For temporal repeatability, we quantified the number of trials to form a color preference after each of multiple color reversals on a serial reversal learning task. For contextual repeatability, we then compared performance on the serial color reversal task to the latency to switch among solutions on each of two different multi-access boxes. We found that the number of trials to form a preference in reversal learning was repeatable across serial color reversals and the latency to switch a preference was repeatable across color reversal learning and the multi-access box contexts. This supports the idea that the reversal learning task structure elicits performance reflective of an inherent trait, and that reversal learning and solution switching on multi-access boxes similarly reflect the inherent trait of behavioral flexibility.

KEYWORDS

Behavioral flexibility, repeatability, construct validity, animal cognition

INTRODUCTION

Research on the cognitive abilities of non-human animals is important for several reasons. By understanding animal cognitive abilities, we can clarify factors that influenced the evolution of human cognition, the mechanisms that relate cognition to ecological and evolutionary dynamics, or we can use the knowledge to facilitate more humane treatment of captive individuals (Shettleworth, 2010). In the last 50 years, comparative psychologists and behavioral ecologists have led a surge in studies innovating methods for measuring cognitive traits in animals. As a result, we have come to understand cognition as the process of acquiring information, followed by storage, retrieval, and use of that information for guiding behavior (Shettleworth, 2010). Evidence now exists that various species possess cognitive abilities in both the physical (e.g. object permanence: Salwiczek et al., 2009; causal understanding: Taylor et al., 2012) and social domains (e.g. social learning: Hoppitt et al., 2012; transitive inference: MacLean et al., 2008).

Cognitive traits are not directly observable and nearly all methods to quantify cognition use behavioral performance as a proxy for cognitive ability. Consequently, it is important to evaluate the validity of the chosen methods for quantifying a cognitive trait. To better understand whether performance on a type of task is likely to reflect a target cognitive trait (i.e., that the method has construct validity), researchers can test for repeatability in individual performance within and across tasks (Völter et al., 2018). However, while many cognitive abilities have been tested, and various methods used, it is rare for one study to repeatedly test individuals with the same method or use multiple methods to test for a given cognitive ability. This could be problematic because cognitive traits are not directly observable, so nearly all methods use behavioral performance as a proxy for cognitive ability. Using only one method to measure a cognitive trait could be problematic because it is hard to discern whether non-target cognitive, personality, or motivational factors may be the cause of variation in performance on the task (Morand-Ferron et al., 2016). For example, the success of pheasants on multiple similar and different problem-solving tasks was related to individual variation in persistence and motivation, rather than problem solving ability (Horik & Madden, 2016). Additionally, performance on cognitive tasks can be affected by different learning styles, where individuals can vary in their perception of the salience of stimuli within a task, the impact of a reward (or non-reward) on future behavior, or the propensity to sample alternative stimuli (Rowe & Healy, 2014). By assessing the temporal and contextual repeatability of performance, researchers can quantify the proportion of variation in performance that is attributable to consistent individual differences likely to reflect the level of the cognitive trait relative to other ephemeral factors that affect individual performance (Cauchoix et al., 2018).

Behavioral flexibility, the ability to change behavior when circumstances change, is a general cognitive ability that likely affects interactions with both the social and physical environment (Bond et al., 2007). Although by definition behavioral flexibility incorporates plasticity in behavior through learning, there is also evidence that the ability to change behavior could be an inherent trait that varies among individuals and species. For example, the pinyon jay - a highly social species of corvid - made fewer errors in a serial reversal learning task than the more asocial Clark's nutcracker or Woodhouse's scrub-jay, but all three species exhibited similar learning curves over successive reversals (Bond et al., 2007). This indicates that the three species differed in the level of the inherent ability, but were similar in the plasticity of performance through learning. Behavioral flexibility could be measured using a variety of methods (Mikhalevich et al., 2017), but the most popular method is reversal learning (Bond et al., 2007) where behavioral flexibility is quantified as the speed that individuals are able to switch a learned preference. However, to our knowledge, no studies have assessed the construct validity of this task by comparing performance of individuals over time and across different tasks that are predicted to require flexible behavior.

In the wild, this ability to change behavior when circumstances change is expected to result in individuals and species that adapt quickly to novelty by showing a high rate of foraging innovations. For example, cross-taxon correlational studies found that species that were “behaviorally flexible”, in that there were many documented foraging innovations, were also more likely to become invasive when introduced to novel habitats (Sol et al., 2002). The ability to innovate solutions to novel problems can also be more directly quantified using a multi-access or puzzle box task, where the subject must use new behavior patterns to solve the task to get food. While it is generally assumed that foraging innovation rate corresponds to the cognitive ability behavioral flexibility (Sol et al., 2002), few studies compare innovation performance and solution switching (a measure of flexibility) on a multi-access box task to performance on a different behavioral flexibility task like reversal learning.

We tested two hypotheses about the construct validity of the reversal learning method as a measure of behavioral flexibility in the great-tailed grackle (*Quiscalus mexicanus*; hereafter “grackle”). First, we determined whether performance on a reversal learning task represents an inherent trait by assessing the repeatability of performance across serial reversals (temporal repeatability). Secondly, we determined whether the inherent trait measured by color reversal learning is likely to represent behavioral flexibility by assessing the cross-contextual repeatability of performance on this task with another task also thought to measure flexibility. Our previous research found that behavioral flexibility does affect innovation ability on a multi-access box (C. Logan et al., 2022), so here our second hypothesis tested whether individuals show contextual repeatability of flexibility by comparing performance on the color reversal learning task to the latency of solution switching on two different multi-access boxes (Fig. 1). We chose solution switching because it requires similar attention to changing reward contingencies, thus serving as a measure of flexibility, but in a different context (e.g. the food is always visible, there is no color association learning required). In other words, in both color reversal learning and solution switching individuals learned a preferred way to obtain food, but then contingencies changed such that food can no longer be obtained with this learned preference and the grackle must be able to switch to a new method. As a human-associated species, the grackle is an ideal subject for this study because the rapid range expansion suggests that they adapted quickly in response to human-induced rapid environmental change (Summers et al., 2022; Wehtje, 2003) and the genus Quiscalus has a high rate of foraging innovations in the wild (Grabrucker & Grabrucker, 2010; Lefebvre & Sol, 2008). Therefore, as their environment may select for flexible and innovative behavior, we believe that these tasks are ecologically relevant and will elicit individual variation in performance.

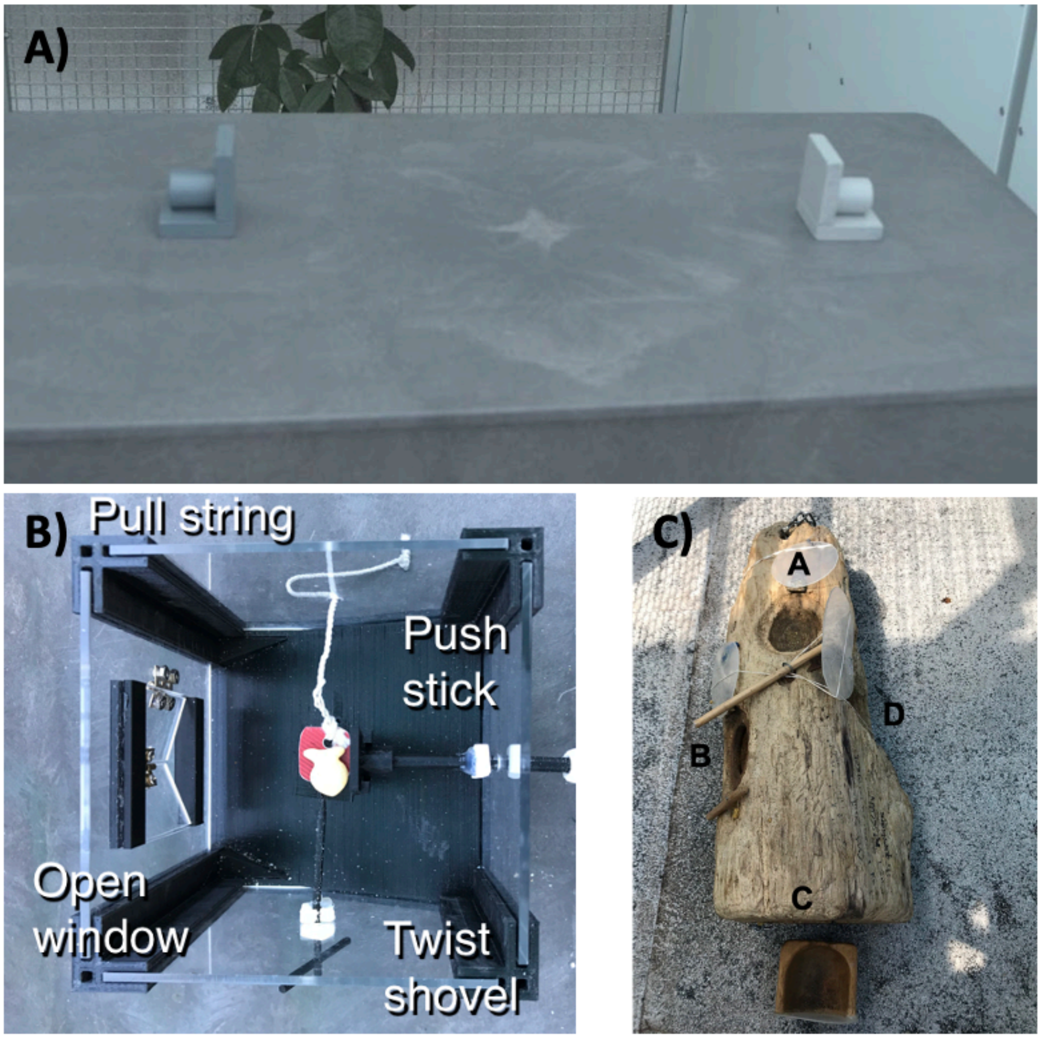

Figure 1. We assessed flexibility as the latency to switch a preference across 3 contexts illustrated here. A) We used two colored containers (tubes) in a color reversal learning task, as well as B) plastic and C) wooden multi-access boxes that each had 4 possible ways (loci) to access food. In each context, after a preference for a color/locus was formed, we made the preferred choice non-functional and then measured the latency of the grackle to switch to a new color/locus.

METHODS

The hypotheses, methods, and analysis plan for this research are described in detail in the peer-reviewed preregistration. We give a short summary of these methods here, with any changes from the preregistration summarized in the *Deviations from the preregistration* section below and further explained in the updates to the preregistration (indicated in italics).

Preregistration details

This experiment was one piece (**H3a and H3b**) of a larger project. This project is detailed in the preregistration that was written (2017), submitted to PCI Ecology for peer review (July 2018), and received the first round of peer reviews a few days before data collection began (Sep 2018). We revised and resubmitted this preregistration after data collection had started (Feb 2019) and it passed peer review (Mar 2019) before any of the planned analyses had been conducted. See the peer review history at PCI Ecology.

Summary of hypotheses

Our first hypothesis considered whether behavioral flexibility (as measured by reversal learning of a color preference) would be repeatable within individuals across serial reversals. Secondly, we hypothesized that, as an inherent trait, behavioral flexibility results in repeatable performance across other contexts (Fig. 1) that require changing behavior when circumstances change (context 1=reversal learning on colored tubes, context 2=plastic multi-access box, context 3=wooden multi-access box).

Summary of methods

Subjects Great-tailed grackles were caught in the wild in Tempe, Arizona USA using a variety of trapping methods. All individuals received color leg bands for individual identification and some individuals (n=34) were brought temporarily into aviaries. Grackles were individually housed in an aviary (each 244 cm long by 122 cm wide by 213 cm tall) for a maximum of six months where they had *ad lib* access to water at all times. During testing, we removed their maintenance diet for up to four hours per day. During this time, they had the opportunity to receive high value food items by participating in tests. Individuals were given three to four days to habituate to the aviaries before we began testing.

Serial color reversal learning We first used serial reversal learning to measure grackle behavioral flexibility. Briefly, we trained grackles to search in one of two differently colored containers for food (Fig. 1A). We used a random number generator to select the color (e.g. light gray) of the container that would consistently contain a food reward across the initial trials. Within each trial, grackles could choose only one container to look in for food. Eventually, grackles showed a significant preference for the rewarded color container (where preference was defined as a minimum of 17 out of 20 correct choices), completing the initial discrimination trials. We then switched the location of the food to the container of the other color (a reversal). The food reward was then consistently located in the container of this second color (e.g. dark gray) across trials until the grackles learned to switch their preference, after which we would again reverse the food to the original colored container (e.g. light gray) and so on back and forth until they passed the serial reversal learning experiment passing criterion [formed a preference in 2 sequential reversals in 50 or fewer trials; C. Logan et al. (2022)]. We measured behavioral flexibility on each reversal as the time it took grackles to switch their preference and search in the second colored container on a minimum of 17 out of 20 trials. See the protocol for serial reversal learning here.

Multi-access boxes We additionally used two different multi-access boxes (hereafter "MAB") to assess behavioral flexibility as the latency to switch loci when a preferred locus becomes non-functional. All grackles were given time to habituate to the MABs prior to testing. We set up the MABs in the aviary of each grackle with food in and around each apparatus in the days prior to testing. At this point all loci were absent or fixed in open, non-functional positions to prevent any early learning of how to solve each apparatus. We began testing when the grackle was eating comfortably from the MAB. For each MAB, the goal was to measure how quickly the grackle could learn to solve each locus, and then how quickly they could switch to attempting to solve a new locus. Consequently, we measured the number of trials to solve a locus and the number of trials until the grackle attempted a new locus after a previously solved locus was made non-functional (solution switching). See protocols for MAB habituation and testing here.

Plastic multi-access box This apparatus consisted of a box with transparent plastic walls (Fig. 1B). There was a pedestal within the box where the food was placed and 4 different options (loci) set within the walls for accessing the food. One locus was a window that, when opened, allowed the grackle to reach in to grab the food. The second locus was a shovel that the food was placed on such that, when turned, the food fell from the pedestal and rolled out of the box. The third locus was a string attached to a tab that the food was placed on such that, when pulled, the food fell from the pedestal and rolled out of the box. The last locus was a horizontal stick that, when pushed, would shove the food off the pedestal such that it rolled out of the box. Each trial was 10 minutes long, or until the grackle used a locus to retrieve the food item. We reset the box out of view of the grackle to begin the next trial. To pass criterion for a locus, the grackle had to get food out of the box after touching the locus only once (i.e. used a functional behavior to retrieve the food) in more than 2 trials across 2 sessions. Afterward, the locus is made non-functional to encourage the grackle to interact with the other loci.

Wooden multi-access box This apparatus consisted of a natural log that contained 4 compartments (loci) covered by transparent plastic doors (Fig. 1C). Each door opened in a different way (open up like a hatch, out to the side like a car door, pull out like a drawer, or push in). During testing, all doors were closed and food was placed in each locus. Each trial lasted 10 minutes or until the grackle opened a door. After solving a locus, the experimenter re-baited that compartment, closed the door out of view of the grackle, and the next trial began. After a grackle solved one locus 3 times, that door was fixed in the open position and the compartment left empty to encourage the grackle to attempt the other loci.

Repeatability analysis Repeatability is defined as the proportion of total variation in performance that is attributable to differences among individuals (Nakagawa & Schielzeth, 2010). In other words, performance is likely to represent an inherent trait, when variation in performance is greater among individuals than within individuals.

To measure repeatability within an individual across serial reversals of a color preference, we modeled the number of trials to pass a reversal (choosing correctly on at least 17 out of 20 sequential trials) as a function of the reversal number (i.e., first time the rewarded color is reversed, second time, third time, etc.) and a random effect for individual. The reversal number for each grackle ranged between 6 to 11 (mean = 7.6) reversals, and the range was based on when individuals were able to pass two sequential reversals in 50 or fewer trials, or (in 1 case) when we reached the maximum duration that we were permitted to keep grackles in the aviaries and they needed to be released. The variance components for the random effect and residual variance were then used to determine the proportion of variance attributable to differences among individuals.

By design in the serial reversal learning experiment, to reach the experiment ending criteria grackles became faster at switching across serial reversals. We did attempt to run a model that additionally included a random slope to test whether there were consistent individual differences in the rate that grackles switched their preferences across reversals. However, we could not get the model to converge with our sample size and the uninformative priors that were preregistered. We felt most comfortable using the preregistered methods to avoid biasing the model output. To determine the statistical significance of the repeatability value, while accounting for this non-independence of a change in reversal speed over time, we compared the actual performance on the number of trials to switch a preference in each reversal to simulated data where birds performed randomly within each reversal.

We tested for contextual repeatability by modeling the variance in latency (in seconds) to switch a preference among and within individuals across 3 behavior switching contexts. Note that the time it took to switch a colored tube preference in serial reversal learning was measured in trials, but the time it took to switch loci in the MAB experiment was measured in seconds. We used the trial start times in the serial reversal experiment to convert the latency to switch a preference from number of trials to number of seconds. Therefore, the contexts across which we measured repeatability of performance were the latency to switch a preference to a new color in the color reversal learning task and latency to switch to a new locus after a previously solved locus was made non-functional on both MABs.

We used the DHARMa package (Hartig, 2019) in R to test whether our model fit our data and was not heteroscedastic, zero-inflated or over-dispersed. We used the MCMCglmm package (Hadfield, 2010), with uninformative priors, to model the relationships of interest for our two hypotheses.

Open data

All data are available at the Knowledge Network for Biocomplexity's data repository: https://knb.ecoinformatics.org/view/doi:10.5063/F18K77JH (K. McCune et al., 2022).

Deviations from the preregistration

In the middle of data collection

1) We originally planned to use a touchscreen test of serial reversal learning as one of the contexts in this experiment. However, on 10 April 2019 we **discontinued the reversal learning experiment on the touchscreen** because it appears to measure something other than what we intended to test and it requires a huge time investment for each bird (which consequently reduces the number of other tests they are available to participate in). This is not necessarily surprising because this is the first time touchscreen tests have been conducted in this species, and also the first time (to our knowledge) this particular reversal experiment has been conducted on a touchscreen with birds. We based this decision on data from four grackles (2 in the flexibility manipulation group and 2 in the flexibility control group; 3 males and 1 female). All four of these individuals showed highly inconsistent learning curves and required hundreds more trials to form each preference when compared to the performance of these individuals on the colored tube reversal experiment. It appears that there is a confounding variable with the touchscreen such that they are extremely slow to learn a preference as indicated by passing our criterion of 17 correct trials out of the most recent 20. We will not include the data from this experiment when conducting the cross-test comparisons in the Analysis Plan section of the preregistration.

2) 16 April 2019: Because we discontinued the touchscreen reversal learning experiment, we **added an additional but distinct multi-access box** task, which allowed us to continue to measure flexibility across three different experiments. There are two main differences between the first multi-access box, which is made of plastic, and the new multi-access box, which is made of wood. First, the wooden multi-access box is a natural log in which we carved out 4 compartments. As a result, the apparatus and solving options are more comparable to what grackles experience in the wild, though each compartment is covered by a transparent plastic door that requires different behaviors to open. Furthermore, there is only one food item available in the plastic multi-access box and the bird could use any of 4 loci to reach it. In contrast, the wooden multi-access box has a piece of food in each of the 4 separate compartments.

Post data collection, pre-data analysis

3) We completed our simulation to explore the lower boundary of a minimum sample size and determined that **our sample size for the Arizona study site is above the minimum** (see details and code in Ability to detect actual effects; 17 April 2020).

4) We originally planned on testing only **adults** to have a better understanding of what the species is capable of, assuming the abilities we are testing are at their optimal levels in adulthood, and so we could increase our statistical power by eliminating the need to include age as an independent variable in the models. Because the grackles in Arizona were extremely difficult to catch, we ended up testing two juveniles in this experiment. The juveniles' performance on the three tests was similar to the adults, therefore we decided not to add age as an independent variable in the models to avoid reducing our statistical power.

Post data collection, mid-data analysis

5) The distribution of values for the "number of trials to reverse" response variable in the P3a analysis was not a good fit for the Poisson distribution because it was overdispersed and heteroscedastic. We log-transformed the data to approximate a normal distribution and it passed all of the data checks. Therefore, we used a Gaussian distribution for our model, which fits the log-transformed data well. (24 Aug 2021)

6) We realized we mis-specified the model and variables for evaluating cross-contextual repeatability P3b analysis. The dependent variable should be latency to switch to a new preference (we previously listed "number of trials to solve", which is more likely indicative of innovation rather than flexibility). Furthermore, to assess performance across contexts, this dependent variable should be the latency to switch in each of the 3 contexts. Note that the time it took to switch a colored tube preference in serial reversal learning was measured in trials, but the time it took to switch loci in the MAB experiment was measured in seconds. We used the trial start times in the serial reversal experiment to convert the latency to switch a preference from number of trials to number of seconds. In line with this change in the dependent variable, the independent variables are only Context (MAB plastic, MAB wood, reversal learning), and reversal number (the number of times individuals switched a preference when the previously preferred color/locus was made non-functional). Additionally, this dependent variable was heteroscedastic when we used a Poisson model, but passed all data checks when we log-transformed it to use a Gaussian model.

RESULTS

Our sample size was 9 for our first hypothesis testing temporal repeatability of reversal learning performance.

Performance was repeatable within individuals within the context of reversal learning (Fig. 2): we obtained a repeatability value of 0.13 (95% credible interval (CI) = 4.64x10^-16 - 0.43). We found that this repeatability value was significantly greater than expected if birds were performing randomly (p=0.003; Fig. 3; see analysis details in the R code for Analysis Plan > P3a). Consequently, and as preregistered, we did not need to conduct the analysis for the P3a alternative to determine whether a lack of repeatability was due to motivation or hunger.

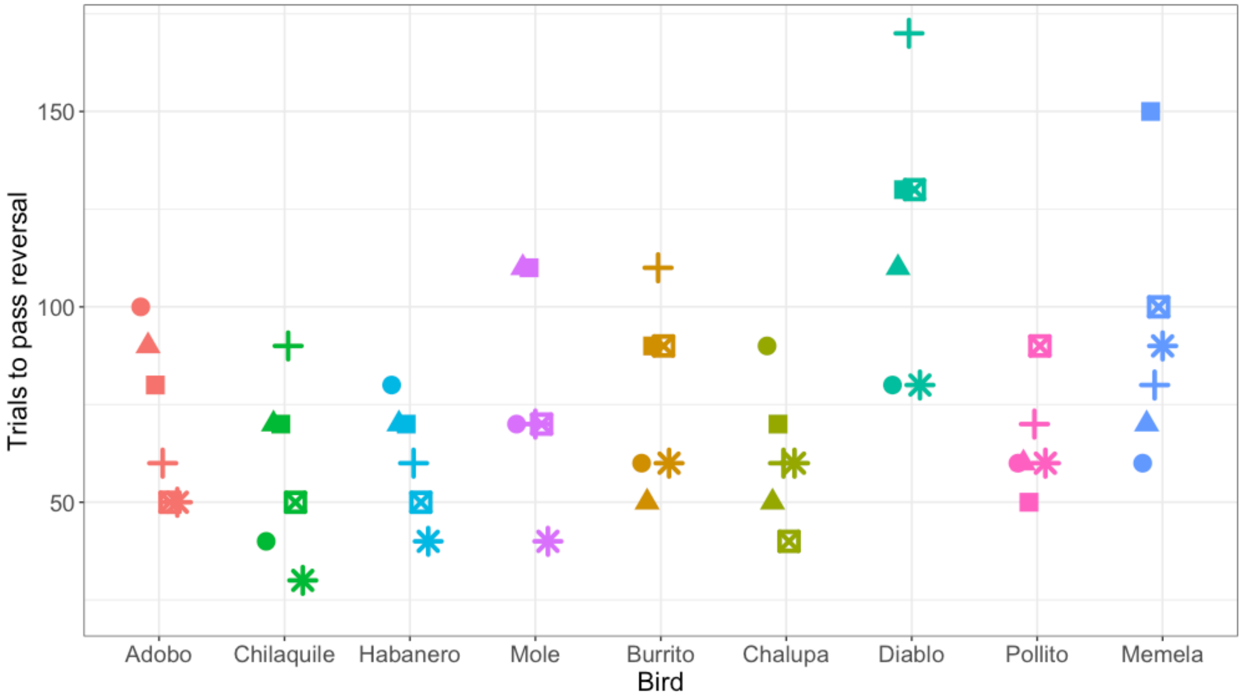

Figure 2: The number of trials each individual took to reverse a preference across serial reversals. The

clustering of data points within each individual illustrates the temporal repeatability in performance. Each reversal is indicated by a different shape. Individuals are grouped by color and arranged from fastest to slowest to complete the serial reversal experiment. Note that as per the serial reversal experimental design, data from nearly all individuals include 2 reversals at or below 50 trials. The one exception was Memela, who never increased the speed to switch her preference.

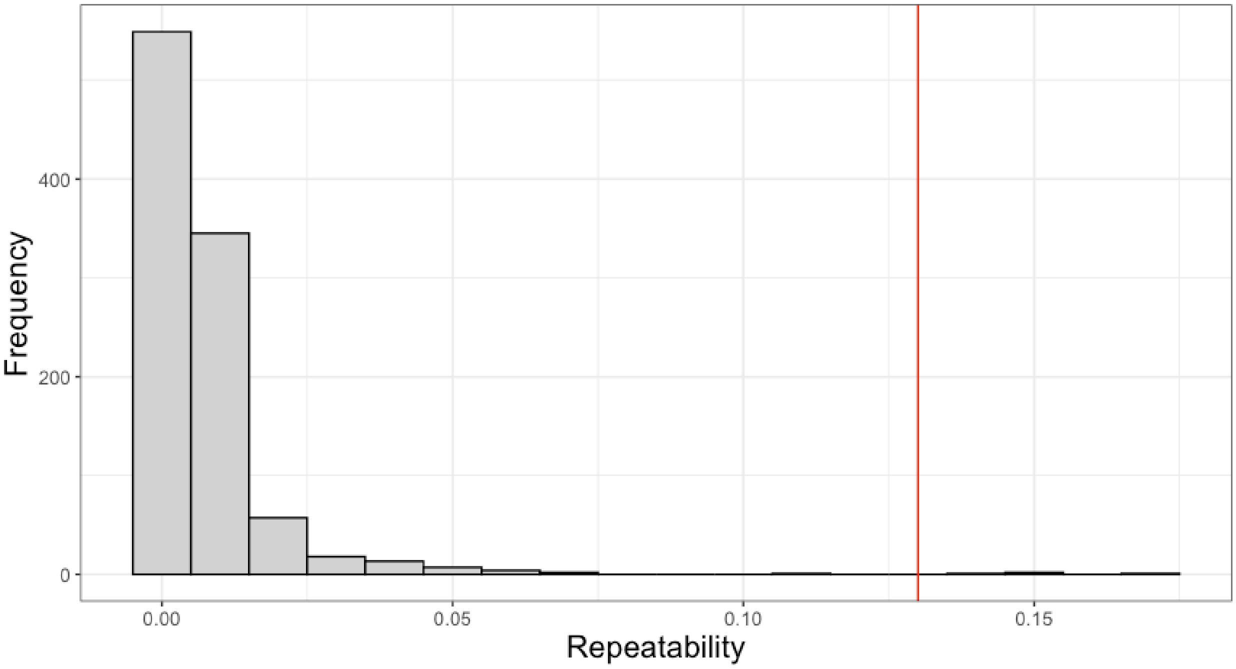

Figure 3: To determine the significance of our repeatability value while accounting for the non-independence of the serial reversal learning experimental design, we compared our repeatability value to repeatability values calculated from simulated data where birds performed randomly within each reversals. The red line indicates our observed value, and it is significantly larger than the repeatability values retrieved from the simulated data. This indicates that despite the design of the serial reversal learning experiment leading to a general increase in the speed that grackles pass each reversal, there were still consistent individual differences in performance across time.

We then assessed the repeatability of performance across contexts by quantifying whether individuals that were fast to switch a preference in the color reversal task were also fast to switch to attempting a new solution after passing criterion on a different solution on the two MAB tasks. We converted our metric of reversal speed from trials to reverse to seconds to reverse so the measures across contexts would be on the same scale. We had repeated measures across contexts for 15 grackles that participated in at least one color reversal and one solution switch on either or both MAB tasks. We found significant repeatability across contexts (R=0.36, CI = 0.10 - 0.64, p=0.01; Fig. 4), where latency to switch was consistent within individuals and different among individuals.

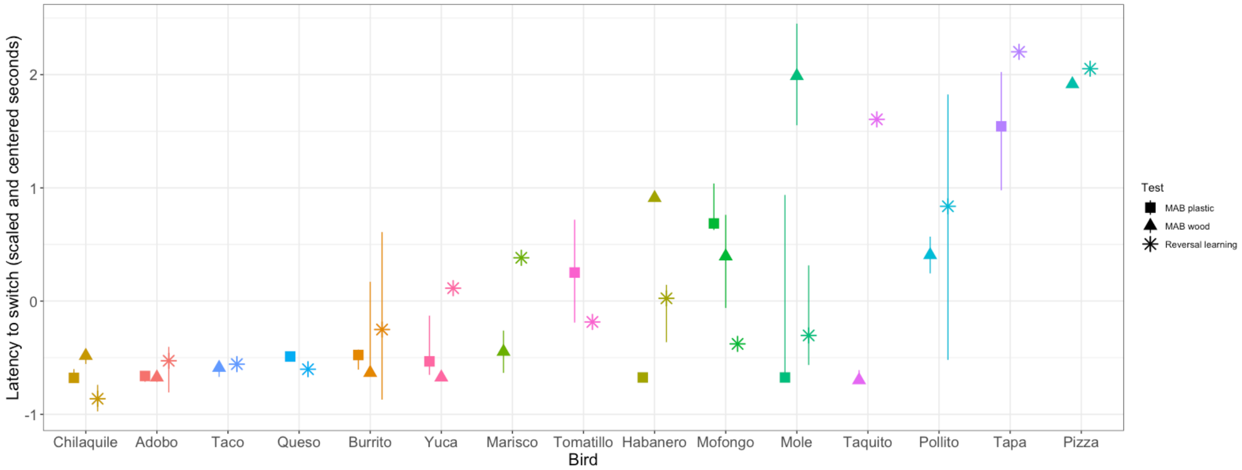

Figure 4: Grackle performance on the different contexts for measuring behavioral flexibility: multi-access box (MAB) plastic (square symbol), MAB wood (triangle symbol), and reversal learning with color tubes (star symbol). Points indicate the (scaled and centered) median performance of an individual in each context, the lines indicate the variation in performance across multiple switches within a context. Some individuals participated in a context, but did not experience multiple preference switches and so there is a point, but no line. Additionally, some individuals are missing points for a given context because they did not participate. Grackles are ordered on the x-axis from fastest (left) to slowest (right).

DISCUSSION

We found that individual grackles were consistent in their behavioral flexibility performance during multiple assessments within the same context, and across multiple assessments in different contexts. This indicates that 1) the different methods we used to measure behavioral flexibility all likely measure the same inherent trait and 2) there is consistent individual variation in behavioral flexibility, which could impact other traits such as survival and fitness in novel areas, foraging, or social behavior.

In behavioral and cognitive research on animals, it is important to determine that the chosen method measures the trait of interest (construct validity). Many experimental methods may lack construct validity because they were adapted from research on other species (e.g. from humans: Wood et al., 1980), applied to new contexts (e.g. from captive to wild animals: K. B. McCune et al., 2019), or created from an anthropomorphic perspective (e.g. mirror self recognition tasks: Kohda et al., 2022). Funding and logistical limitations result in few researchers assessing the appropriateness of their methods by testing construct validity through convergent (similar performance across similar tasks) and discriminant validity (different performance across different tasks). Although our sample size was small, which likely led to moderately large credible intervals, we still found significant temporal and contextual repeatability of switching performance. This evidence for convergent validity indicates these similar tasks are likely assessing the same latent trait of interest (Morand-Ferron et al., 2022; Völter et al., 2018). However, it is important to also test for discriminant validity by comparing performance on flexibility tasks with tasks intended to measure different cognitive abilities. For example, it is possible that performance on serial reversal learning and solution switching on the MAB tasks is reflective of a trait other than behavioral flexibility, like inhibition (MacLean et al., 2014). Indeed, we previously found that the more flexible grackles on the serial reversal learning task were also better able to inhibit responding to a non-rewarded stimulus in a go/no-go task thought to measure self-control (Logan et al., 2021). Consequently, more research is needed to interpret whether some aspect of performance on the go/no-go task reflects behavioral flexibility or whether performance on the reversal learning task is influenced by inhibition.

The functional importance of behavioral flexibility is that it is thought to facilitate invasion success by allowing individuals to quickly change their behavior when circumstances change. For example, flexible grackles may innovate new foraging techniques or generalize standard techniques to new food items in novel areas. The great-tailed grackle has rapidly expanded its range (Summers et al., 2022; Wehtje, 2003), implying that it is able to have high survival and fitness in the face of environmental change. Although grackles are a behaviorally flexible species (Logan, 2016), we show here that there are consistent individual differences among grackles in how quickly they are able to change their behavior when circumstances change in multiple different contexts. While some grackles were consistently faster at changing their behavior (e.g., Chilaquile), others were consistently slower (e.g., Yuca). This consistency in performance may seem contradictory to our previous research where we found that we are able to manipulate grackles to be more flexible using serial reversal learning (C. Logan et al., 2022). That behavioral flexibility is both repeatable within individuals across reversals, indicating it is an inherent trait, as well as being manipulatable through serial reversals, aligns with the idea of behavioral reaction norms (Sih, 2013). This idea states that individuals can show consistent individual differences in the baseline or average values of a trait of interest across time or contexts, but the plasticity in the expression of the trait can also consistently vary among individuals. Due to our small sample size, we were not able to explicitly test for behavioral reaction norms, but this is an important next step in understanding consistent individual variation in behavioral flexibility in relation to rapid environmental change. Past experience (developmental or evolutionary) with environmental change influences how plastic the individuals are able to be (Sih, 2013). To understand the implications of this individual variation in performance in this species that has experienced much environmental change during the range expansion, our future research investigates how behavioral flexibility may relate to proximity to the range edge (Logan CJ et al., 2020), and the variety of foraging techniques used in the wild (Logan CJ et al., 2019).

By first validating the experimental methods for behavioral and cognitive traits, such that we are more certain that our tests are measuring the intended trait, we are better able to understand the causes and consequences of species, population, and individual differences. Individual variation in behavioral flexibility has the potential to influence species adaptation and persistence under human-induced rapid environmental change (Sih, 2013). Consequently, we believe the results presented here are a timely addition to the field by demonstrating two potential methods for measuring behavioral flexibility that produced repeatable performance in at least one system.

ETHICS

This research is carried out in accordance with permits from the:

1) US Fish and Wildlife Service (scientific collecting permit number MB76700A-0,1,2)
2) US Geological Survey Bird Banding Laboratory (federal bird banding permit number 23872)
3) Arizona Game and Fish Department (scientific collecting license number SP594338 [2017], SP606267 [2018], and SP639866 [2019])
4) Institutional Animal Care and Use Committee at Arizona State University (protocol number 17-1594R)
5) University of Cambridge ethical review process (non-regulated use of animals in scientific procedures: zoo4/17 [2017])

AUTHOR CONTRIBUTIONS

McCune: Added MAB log experiment, hypothesis development, protocol development, data collection, data interpretation, write up, revising/editing, materials.

Blaisdell: Prediction revision, assisted with programming the reversal learning touchscreen experiment, protocol development, data interpretation, revising/editing.

Johnson-Ulrich: Prediction revision, programming, data collection, data interpretation, revising/editing.

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

MacPherson: Data collection, data interpretation, revising/editing.

Seitz: Prediction revision, programmed the reversal learning touchscreen experiment, protocol development, data interpretation, revising/editing.

Sevchik: Data collection, revising/editing.

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

FUNDING

This research is funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology (2017-current), and by a Leverhulme Early Career Research Fellowship to Logan (2017-2018).

CONFLICT OF INTEREST DISCLOSURE

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

ACKNOWLEDGEMENTS

We thank our PCI Ecology recommender, Aurelie Coulon, and reviewers, Maxime Dahirel and Andrea Griffin, for their feedback on this preregistration; Kevin Langergraber for serving as our ASU IACUC PI; Ben Trumble and Angela Bond for logistical support; Melissa Wilson for sponsoring our affiliations at Arizona State University and lending lab equipment; Kristine Johnson for technical advice on great-tailed grackles; Arizona State University School of Life Sciences Department Animal Care and Technologies for providing space for our aviaries and for their excellent support of our daily activities; Julia Cissewski for tirelessly solving problems involving financial transactions and contracts; Sophie Kaube for logistical support; Richard McElreath for project support; Aaron Blackwell and Ken Kosik for being the UCSB sponsors of the Cooperation Agreement with the Max Planck Institute for Evolutionary Anthropology; Tiana Lam, Anja Becker, and Brynna Hood for interobserver reliability video coding: Sawyer Lung for field support; Alexis Breen for coding multi-access box videos; and our research assistants: Aelin Mayer, Nancy Rodriguez, Brianna Thomas, Aldora Messinger, Elysia Mamola, Michael Guillen, Rita Barakat, Adriana Boderash, Olateju Ojekunle, August Sevchik, Justin Huynh, Jennifer Berens, Amanda Overholt, Michael Pickett, Sam Munoz, Sam Bowser, Emily Blackwell, Kaylee Delcid, Sofija Savic, Brynna Hood, Sierra Planck, and Elise Lange.

REFERENCES

SUPPLEMENTARY MATERIALS

D. PREREGISTRATION (detailed methods)

HYPOTHESES

H3a: Behavioral flexibility within a context is repeatable within individuals. Repeatability of behavioral flexibility is defined as the number of trials to reverse a color preference being strongly negatively correlated within individuals with the number of reversals.

P3a: Individuals that are faster to reverse a color preference in the first reversal will also be faster to reverse a color preference in the second, etc. reversal due to natural individual variation.

P3a alternative: There is no repeatability in behavioral flexibility within individuals, which could indicate that performance is state dependent (e.g., it depends on their fluctuating motivation, hunger levels, etc.). We will determine whether performance on colored tube reversal learning related to motivation by examining whether the latency to make a choice influenced the results. We will also determine whether performance was related to hunger levels by examining whether the number of minutes since the removal of their maintenance diet from their aviary plus the number of food rewards they received since then influenced the results.

H3b: The consistency of behavioral flexibility in individuals across contexts (context 1=reversal learning on colored tubes, context 2=multi-access boxes, context 3=reversal learning on touchscreen) indicates their ability to generalize across contexts. Individual consistency of behavioral flexibility is defined as the number of trials to reverse a color preference being strongly positively correlated within individuals with the latency to solve new loci on each of the multi-access boxes and with the number of trials to reverse a color preference on a touchscreen (total number of touchscreen reversals = 5 per bird).

If P3a is supported (repeatability of flexibility within individuals)…

P3b: …and flexibility is correlated across contexts, then the more flexible individuals are better at generalizing across contexts.

P3b alternative 1: …and flexibility is not correlated across contexts, then there is something that influences an individual's ability to discount cues in a given context. This could be the individual's reinforcement history (tested in P3a alternative), their reliance on particular learning strategies (one alternative is tested in H4), or their motivation (tested in P3a alternative) to engage with a particular task (e.g., difficulty level of the task).

DEPENDENT VARIABLES *P3a and P3a alternative 1*

Number of trials to reverse a preference. An individual is considered to have a preference if it chose the rewarded option at least 17 out of the most recent 20 trials (with a minimum of 8 or 9 correct choices out of 10 on the two most recent sets of 10 trials). We use a sliding window to look at the most recent 10 trials for a bird, regardless of when the testing sessions occurred.

P3b: additional analysis: individual consistency in flexibility across contexts + flexibility is correlated across contexts

Number of trials to solve a new locus on the multi-access boxes *NOTE: Jul 2022 we realized this variable is more likely to represent innovation, and we mean to assess flexibility here. Therefore we changed this variable to latency to attempt to switch a preference after the previously rewarded color/locus becomes non-functional.*

INDEPENDENT VARIABLES *P3a: repeatable within individuals within a context*

1) Reversal number
2) ID (random effect because repeated measures on the same individuals)

P3a alternative 1: was the potential lack of repeatability on colored tube reversal learning due to motivation or hunger?

1) Trial number
2) Latency from the beginning of the trial to when they make a choice
3) Minutes since maintenance diet was removed from the aviary
4) Cumulative number of rewards from previous trials on that day

5) ID (random effect because repeated measures on the same individuals)

6) Batch (random effect because repeated measures on the same individuals). Note: batch is a test cohort, consisting of 8 birds being tested simultaneously

P3b: repeatable across contexts

NOTE: Jul 2022 we changed the dependent variable to reflect the general latency to switch a preference (in any of the three tasks) and so IVs 3 (Latency to solve a new locus) & 4 (Number of trials to reverse a preference), below, are redundant. Furthermore, we did not include the touchscreen experiment in this manuscript (previously accounted for with IV 5; see the Deviations section). Therefore, despite being listed here in the preregistration as IVs that we proposed to include in the P3b model, in our post-study manuscript we did not include these IVs in the final model. The IVs instead consisted of: Reversal (switch) number, Context (colored tubes, plastic multi-access box, wooden multi-access box) and ID (random effect because there were repeated measures on the same individuals).

1) Reversal (switch) number

2) Context (colored tubes, plastic multi-access box, wooden multi-access box, touchscreen)

3) Latency to solve a new locus

4) Number of trials to reverse a preference (colored tubes)

5) Number of trials to reverse a preference (touchscreen)

6) ID (random effect because repeated measures on the same individuals)

ANALYSIS PLAN *P3a: repeatable within individuals within a context (reversal learning)*

Analysis: Is reversal learning (colored tubes) repeatable within individuals within a context (reversal learning)? We will obtain repeatability estimates that account for the observed and latent scales, and then compare them with the raw repeatability estimate from the null model. The repeatability estimate indicates how much of the total variance, after accounting for fixed and random effects, is explained by individual differences (ID). We will run this GLMM using the MCMCglmm function in the MCMCglmm package (Hadfield, 2010) with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors [V=1, nu=0; Hadfield (2014)]. We will ensure the GLMM shows acceptable convergence [i.e., lag time autocorrelation values <0.01; Hadfield (2010)], and adjust parameters if necessary.

NOTE (Aug 2021): our data checking process showed that the distribution of values of the data (number of trials to reverse) in this model was not a good fit for the Poisson distribution because it was overdispersed and heteroscedastic. However, when log-transformed the data approximate a normal distribution and pass all of the data checks, therefore we used a Gaussian distribution for our model, which fits the log-transformed data well.

To roughly estimate our ability to detect actual effects (because these power analyses are designed for frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings: test family=F tests, statistical test=linear multiple regression: Fixed model (R^2 deviation from zero), type of power analysis=a priori, alpha error probability=0.05. The number of predictor variables was restricted to only the fixed effects because this test was not designed for mixed models. We reduced the power to 0.70 and increased the effect size until the total sample size in the output matched our projected sample size (n=32). The protocol of the power analysis is here:

Input:

Effect size $f^2 = 0.21$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 1

Output:

Noncentrality parameter   = 6.7200000

Critical F = 4.1708768

Numerator df = 1

Denominator df = 30

Total sample size = 32

Actual power = 0.7083763

This means that, with our sample size of 32, we have a 71% chance of detecting a medium effect (approximated at f^2=0.15 by Cohen, 1988).

P3a alternative: was the potential lack of repeatability on colored tube reversal learning due to motivation or hunger?

Analysis: Because the independent variables could influence each other or measure the same variable, I will analyze them in a single model: Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCMCglmm package; Hadfield (2010)] with a binomial distribution (called categorical in MCMCglmm) and logit link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; Hadfield (2010)], and adjust parameters if necessary. The contribution of each independent variable will be evaluated using the Estimate in the full model. NOTE (Apr 2021): This analysis is restricted to data from their first reversal because this is the only reversal data that is comparable across the manipulated and control groups.

To roughly estimate our ability to detect actual effects (because these power analyses are designed for frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings: test family=F tests, statistical test=linear multiple regression: Fixed model (R^2 deviation from zero), type of power analysis=a priori, alpha error probability=0.05. We reduced the power to 0.70 and increased the effect size until the total sample size in the output matched our projected sample size (n=32). The number of predictor variables was restricted to only the fixed effects because this test was not designed for mixed models. The protocol of the power analysis is here:

Input:

Effect size f^2 = 0.31

 err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 4

Output:

Noncentrality parameter   = 11.4700000

Critical F = 2.6684369

Numerator df = 4

Denominator df = 32

Total sample size = 37

Actual power = 0.7113216

This means that, with our sample size of 32, we have a 71% chance of detecting a large effect (approximated at f^2=0.35 by Cohen, 1988).

P3b: individual consistency across contexts

Analysis: Do those individuals that are faster to reverse a color preference also have lower latencies to switch to new options on the multi-access box? A Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCMCglmm package; (Hadfield, 2010) will be used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; Hadfield (2010)], and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

To roughly estimate our ability to detect actual effects (because these power analyses are designed for frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings: test family=F tests, statistical test=linear multiple regression: Fixed model (R^2 deviation from zero), type of power analysis=a priori, alpha error probability=0.05. We reduced the power to 0.70 and increased the effect size until the total sample size in the output matched our projected sample size (n=32). The number of predictor variables was restricted to only the fixed effects because this test was not designed for mixed models. The protocol of the power analysis is here:

Input:

Effect size f^2 = 0.21

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 1

Output:

Noncentrality parameter = 6.7200000

Critical F = 4.1708768

Numerator df = 1

Denominator df = 30

Total sample size = 32

Actual power = 0.7083763

This means that, with our sample size of 32, we have a 71% chance of detecting a medium effect (approximated at f^2=0.15 by Cohen, 1988).

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

Cauchoix, M., Chow, P., Van Horik, J., Atance, C., Barbeau, E., Barragan-Jason, G., Bize, P., Boussard, A., Buechel, S. D., Cabirol, A., et al. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170281.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences 2nd edn.* Erlbaum Associates, Hillsdale.

Grabrucker, S., & Grabrucker, A. M. (2010). Rare feeding behavior of great-tailed grackles (quiscalus mexicanus) in the extreme habitat of death valley. *The Open Ornithology Journal*, *3*(1).

Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm r package. *Journal of Statistical Software*, *33*(2), 1–22. https://doi.org/10.18637/jss.v033.i02

Hadfield, J. (2014). *MCMCglmm course notes.* http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

Hartig, F. (2019). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models.* http://florianhartig.github.io/DHARMa/

Hoppitt, W., Samson, J., Laland, K. N., & Thornton, A. (2012). *Identification of learning mechanisms in a wild meerkat population*.

Horik, J. O. van, & Madden, J. R. (2016). A problem with problem solving: Motivational traits, but not cognition, predict success on novel operant foraging tasks. *Animal Behaviour*, *114*, 189–198.

Kohda, M., Sogawa, S., Jordan, A. L., Kubo, N., Awata, S., Satoh, S., Kobayashi, T., Fujita, A., & Bshary, R. (2022). Further evidence for the capacity of mirror self-recognition in cleaner fish and the significance of ecologically relevant marks. *PLoS Biology*, *20*(2), e3001529.

Lefebvre, L., & Sol, D. (2008). Brains, lifestyles and cognition: Are there general trends? *Brain, Behavior and Evolution*, *72*(2), 135–144.

Logan, C. J. (2016). Behavioral flexibility and problem solving in an invasive bird. *PeerJ*, *4*, e1975. https://doi.org/10.7717/peerj.1975

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

Logan, C., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & McCune, K. (2022). Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context. *EcoEvoRxiv*. https://doi.org/https://doi.org/10.32942/osf.io/5z8xs

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

Logan, CJ, McCune, KB, Chen, N, & Lukas, D. (2020). Implementing a rapid geographic range expansion - the role of behavior and habitat changes. *In Principle Acceptance by PCI Ecology of the Version on 6 Oct 2020*. http://corinalogan.com/Preregistrations/gxpopbehaviorhabitat.html

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., et al. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140–E2148.

MacLean, E. L., Merritt, D. J., & Brannon, E. M. (2008). Social complexity predicts transitive reasoning in prosimian primates. *Animal Behaviour*, *76*(2), 479–486.

McCune, K. B., Jablonski, P., Lee, S., & Ha, R. R. (2019). Captive jays exhibit reduced problem-solving performance compared to wild conspecifics. *Royal Society Open Science*, *6*(1), 181311.

McCune, K., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & Logan, C. (2022). Using repeatability of performance within and across contexts to validate measures of behavioral flexibility. *Knowledge Network for Biocomplexity, Data package*. https://doi.org/10.5063/F18K77JH

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

Morand-Ferron, J., Cole, E. F., & Quinn, J. L. (2016). Studying the evolutionary ecology of cognition in the wild: A review of practical and conceptual challenges. *Biological Reviews*, *91*(2), 367–389.

Morand-Ferron, J., Reichert, M. S., & Quinn, J. L. (2022). Cognitive flexibility in the wild: Individual differences in reversal learning are explained primarily by proactive interference, not by sampling strategies, in two passerine bird species. *Learning & Behavior*, *50*(1), 153–166.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for gaussian and non-gaussian data: A practical guide for biologists. *Biological Reviews*, *85*(4), 935–956.

Rowe, C., & Healy, S. D. (2014). Measuring variation in cognition. *Behavioral Ecology*, *25*(6), 1287–1292.

Salwiczek, L. H., Emery, N. J., Schlinger, B., & Clayton, N. S. (2009). The development of caching and object permanence in western scrub-jays (aphelocoma californica): Which emerges first? *Journal of Comparative Psychology*, *123*(3), 295.

Shettleworth, S. J. (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*(5), 1077–1088.

Sol, D., Timmermans, S., & Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Animal Behaviour*, *63*(3), 495–502.

Summers, J., Lukas, D., Logan, C., & Chen, N. (2022). The role of climate change and niche shifts in divergent range dynamics of a sister-species pair. *EcoEvoRxiv*. https://doi.org/10.32942/osf.io/879pe

Taylor, A. H., Knaebe, B., & Gray, R. D. (2012). An end to insight? New caledonian crows can spontaneously solve problems without planning their actions. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 4977–4981.

Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics: Establishing what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170283.

Wehtje, W. (2003). The range expansion of the great-tailed grackle (quiscalus mexicanus gmelin) in north america since 1880. *Journal of Biogeography*, *30*(10), 1593–1607. https://doi.org/10.1046/j.1365-2699.2003.00970.x

Wood, S., Moriarty, K. M., Gardner, B. T., & Gardner, R. A. (1980). Object permanence in child and chimpanzee. *Animal Learning & Behavior*, *8*(1), 3–9.