Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context.

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

2021-12-30

Open… access code peer review data

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

This is 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, pdf (at EcoEvoRxiv), rmd

ABSTRACT

Behavioral flexibility, the ability to adapt behavior to new circumstances, is thought to play an important role in a species' ability to successfully adapt to new environments and expand its geographic range. However, flexibility is rarely directly tested in species in a way that would allow us to determine how flexibility works and predictions a species' ability to adapt their behavior to new environments. We use great-tailed grackles (a bird species) as a model to investigate this question because they have rapidly expanded their range into North America over the past 140 years. We attempted to manipulate grackle flexibility using colored tube reversal learning to determine whether flexibility is generalizable across contexts (touchscreen reversal learning and multi-access box), whether it is repeatable within individuals and across contexts, and what learning strategies grackles employ. We found that we were able to manipulate flexibility: birds in the manipulated group took fewer trials to pass criterion with increasing reversal number, and they reversed a color preference in fewer trials by the end of their serial reversals compared to control birds who had only one reversal. Flexibility was repeatable within individuals (reversal), but not across contexts (from reversal to multi-access box). The touchscreen reversal experiment did not appear to measure what was measured in the reversal learning experiment with the tubes, and we speculate as to why. One third of the grackles in the manipulated reversal learning group switched from one learning strategy (epsilon-decreasing

where they have a long exploration period) to a different strategy (epsilon-first where they quickly shift their preference). A separate analysis showed that the grackles did not use a particular strategy earlier or later in their serial reversals. Posthoc analyses using a model that breaks down performance on the reversal learning task into different components showed that learning to be attracted to an option (phi) more consistently correlated with reversal performance than the rate of deviating from learned attractions that were rewarded (lambda). This result held in simulations and in the data from the grackles: learning rates in the manipulated grackles doubled by the end of the manipulation compared to control grackles, while the rate of deviation slightly decreased. Grackles with intermediate rates of deviation in their last reversal, independently of whether they had gone through the serial reversal manipulation, solved fewer loci on the plastic and wooden multi-access boxes, and those with intermediate learning rates in their last reversal were faster to attempt a new locus on both multi-access boxes. This investigation allowed us to make causal conclusions rather than relying only on correlations: we manipulated reversal learning, which caused changes in a different flexibility measure (multi-access box switch times) and in an innovativeness measure (multi-access box loci solved), as well as validating that the manipulation had an effect on the cognitive ability we think of as flexibility. Understanding how behavioral flexibility causally relates to other traits will allow researchers to develop robust theory about what behavioral flexibility is and when to invoke it as a primary driver in a given context, such as a rapid geographic range expansion. Given our results, flexibility manipulations could be useful in training threatened and endangered species in how to be more flexible. If such a flexibility manipulation was successful, it could then change their behavior in this and other domains, giving them a better chance of succeeding in human modified environments.

Video summary

INTRODUCTION

Behavioral flexibility, the ability to adapt behavior to new circumstances (see Mikhalevich et al., 2017 for the theoretical background on this definition), is thought to play an important role in a species' ability to successfully adapt to new environments and expand its geographic range (e.g., Lefebvre et al., 1997; Sol et al., 2002, 2005, 2007; Sol & Lefebvre, 2000). This research predicts that behavioral flexibility (hereafter referred to as flexibility) should positively relate with innovativeness. However, these predictions are based on species-level data and proxies for flexibility and for innovation when examining such relationships (see Logan et al., 2018). Flexibility is rarely directly tested in species that are rapidly expanding their geographic ranges in a way that would allow us to determine how flexibility works and predict a species' ability to adapt their behavior to new areas. Those investigations that examine the relationship between flexibility and innovation (or problem solving) in species that are expanding their range show mixed results, with these variables correlating positively (e.g., grey squirrels: Chow et al., 2016), negatively (e.g., Indian mynas: Griffin et al., 2013), or not at all (e.g., stick tool use and string pulling in great-tailed grackles: Logan, 2016). One way to improve our understanding of whether and how flexibility relates to innovativeness is to perform a manipulative experiment on one of the variables to determine whether there is an associated change in the other.

We focused our study on great-tailed grackles (*Quiscalus mexicanus*, hereafter grackles), a bird species that is flexible (Logan, 2016) and rapidly expanding its geographic range (Wehtje, 2003). We attempted to manipulate grackle flexibility using serial reversals of a color preference to determine whether their flexibility is generalizable across additional experimental contexts (touchscreen reversal learning and multi-access box solution switching), whether improving flexibility also improves innovativeness (number of loci solved on a multi-access box), whether it is repeatable within individuals and across contexts, and what learning strategies grackles employ (Figure 1).

If grackle flexibility is manipulatable using serial reversals, this could provide conservation managers with an important tool for managing at-risk populations. If the manipulation works in grackles, it has the potential to be effective in other species as well. This could be particularly useful for endangered species conservation efforts, such as when selecting individuals for captive breeding programs, because individuals that are more flexible might be able to adapt better to new environments. If the flexibility manipulation is not successful,

this could indicate either that we did not manipulate the right aspect of flexibility (e.g., perhaps training them to solve a variety of different types of tasks quickly would be more effective) or that grackle flexibility is not a trait that is trainable.

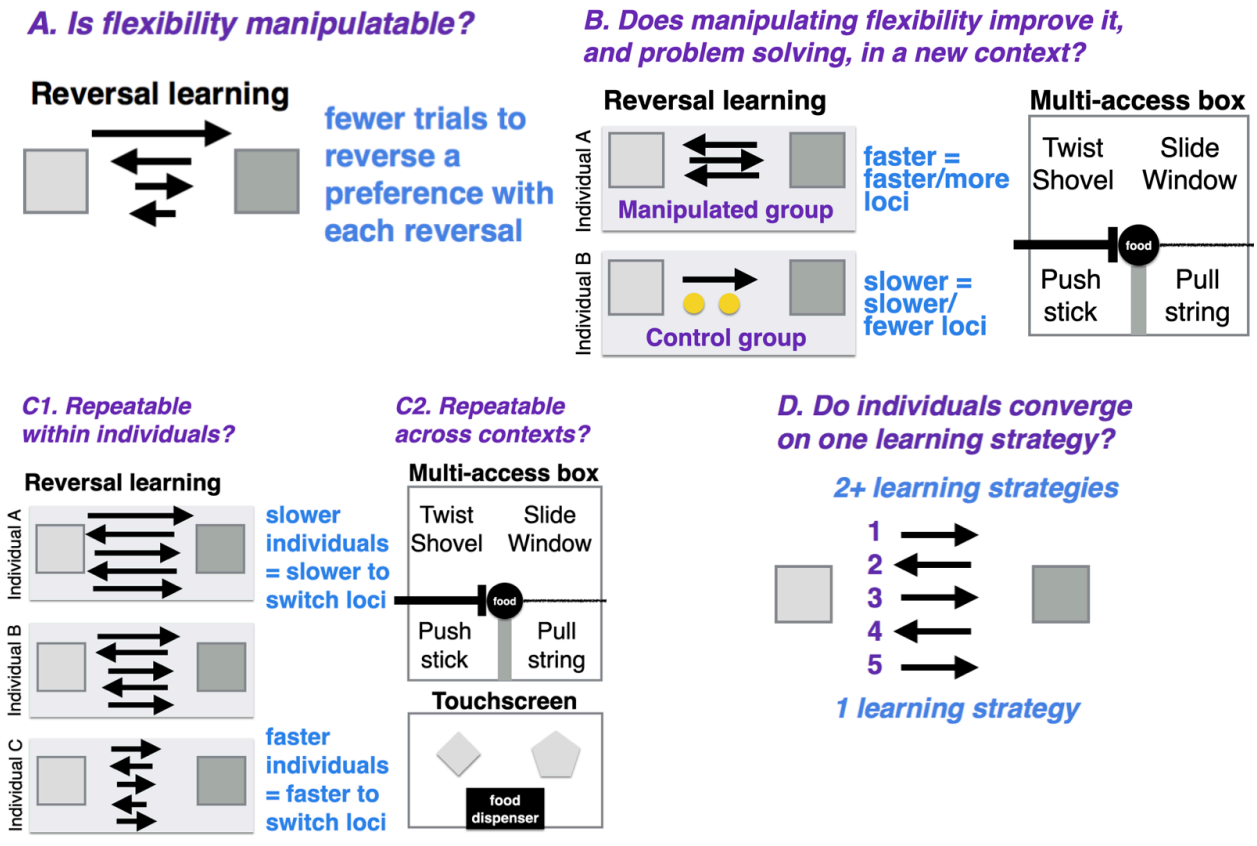


Figure 1. A visual illustration of Hypothesis 1 (A), Hypothesis 2 (B), Hypothesis 3 (C1 and C2), and Hypothesis 4 (D). Longer black arrows indicate slower reversal times, the two yellow circles represent experience with the two yellow tubes that both contained food for the control group.

HYPOTHESES

H1: Behavioral flexibility, as measured by reversal learning using colored tubes, is manipulatable. Prediction 1: Individuals improve their flexibility on a serial reversal learning task using colored tubes by generally requiring fewer trials to reverse a preference as the number of reversals increases (manipulation condition). Their flexibility on this test will have been manipulated relative to control birds who do not undergo serial reversals. Instead, individuals in the control condition will be matched to manipulated birds for experience (they will experience a similar number of trials), but there will be no possibility of a functional tube preference because both tubes will be the same color and both will contain food, therefore either choice will be correct.

P1 alternative 1: If the number of trials to reverse a preference does not correlate with or positively correlates with reversal number, which would account for all potential correlation outcomes, this suggests that some individuals may prefer to rely on information acquired previously (i.e., they are slow to reverse) rather than relying on current cues (e.g., the food is in a new location) (Griffin & Guez, 2014; Liu et al., 2016; e.g., Manrique et al., 2013; but see Homberg et al., 2007).

H2: Manipulating behavioral flexibility (improving reversal learning speed through serial reversals using colored tubes) improves flexibility (rule learning and/or switching) and problem

solving in a new context (two distinct multi-access boxes and serial reversals on a touchscreen). **P2:** Individuals that have improved their flexibility on a serial reversal learning task using colored tubes (requiring fewer trials to reverse a preference as the number of reversals increases) are faster to switch between new methods of solving (latency to solve or attempt to solve a new way of accessing the food [locus]), and learn more new loci (higher total number of solved loci) on multi-access box flexibility tasks, and are faster to reverse preferences in a serial reversal task using a touchscreen than individuals in the control group where flexibility has not been manipulated. The positive correlation between reversal learning performance using colored tubes and a touchscreen (faster birds have fewer trials) and the multi-access boxes (faster birds have lower latencies) indicates that all three tests measure the same ability even though the multi-access boxes require inventing new rules to solve new loci (while potentially learning a rule about switching: "when an option becomes non-functional, try a different option") while reversal learning requires switching between two rules ("choose light gray" or "choose dark gray") or learning the rule to "switch when the previously rewarded option no longer contains a reward." Serial reversals eliminate the confounds of exploration, inhibition, and persistence in explaining reversal learning speed because, after multiple reversals, what is being measured is the ability to learn one or more rules. If the manipulation works, this indicates that flexibility can be influenced by previous experience and might indicate that any individual has the potential to move into new environments (see relevant hypotheses in preregistrations on genetics (R1) and expansion (H1)).

P2 alternative 1: If the manipulation does not work in that those individuals in the experimental condition do not decrease their reversal speeds more than control individuals, then this experiment will elucidate whether general individual variation in flexibility relates to flexibility in new contexts (two distinct multi-access boxes and serial reversals on a touchscreen) as well as problem solving ability (multi-access boxes). The prediction is the same in P2, but in this case variation in flexibility is constrained by traits inherent to the individual (some of which will be tested in McCune et al., 2019), which suggests that certain individuals will be more likely to move into new environments.

P2 alternative 2: If there is no correlation between reversal learning speed (colored tubes) and the latency to solve/attempt a new locus on the multi-access boxes, this could be because the latency to solve not only measures flexibility but also innovativeness. In this case, an additional analysis will be run with the latency to solve as the response variable, to determine whether the fit of the model (as determined by the lower AIC value) with reversal learning as an explanatory variable is improved if motor diversity (the number of different motor actions used when attempting to solve the multi-access box) is included as an explanatory variable (see Diquelou et al., 2015; Griffin et al., 2016). If the inclusion of motor diversity improves the model fit, then this indicates that the latency to solve a new locus on the multi-access box is influenced by flexibility (reversal learning speed) and innovation (motor diversity).

P2 alternative 3: If there is a negative correlation or no correlation between reversal learning speed on colored tubes and reversal learning speed on the touchscreen, then this indicates that it may be difficult for individuals to perceive and/or understand images on the touchscreen in contrast with physical objects (colored tubes) (e.g., O'Hara et al., 2015).

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).

H4: Individuals should converge on an epsilon-first learning strategy (learn the correct choice after one trial) as they progress through serial reversals. P4: Individuals will prefer a mixture of learning strategies in the first serial reversals (an *epsilon-decreasing* strategy where individuals explore both options extensively before learning to prefer the rewarded option, and an *epsilon-first* strategy where the correct choice is consistently made after the first trial), and then move toward the epsilon-first learning strategy. The epsilon-first strategy works better later in the serial reversals where the reward is all or nothing because individuals will have learned the environment is changing in predictable ways (Bergstrom & Lachmann, 2004): only one option is consistently rewarded, and if the reward isn't in the previously rewarded option, it must be in the other option.

P4 alternative 1: Individuals will continue to prefer a mixture of learning strategies, and/or they do not converge on the more functional epsilon-first learning strategy, regardless of how many reversals they participate in. This pattern could suggest that the grackles do not attend to functional meta-strategies, that is, they do not learn the overarching rule (once food is found in the non-preferred tube, one must switch to preferring that tube color), but rather they learn each preference change as if it was new.

ASSOCIATED PREREGISTRATION

Our methods and analysis plans are described in the peer-reviewed preregistration of this article that received in principle recommendation from PCI Ecology, which is included below as the Methods. We moved the hypotheses from the preregistration to the section above to improve flow for the reader.

DEVIATIONS FROM THE PREREGISTRATION

In the middle of data collection

1) 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. Instead, in the Results section, we provide summary results for this experiment and, in the Discussion, qualitatively compare it with performance on the colored tube reversal test to explain what might have confounded the touchscreen experiment.

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) Please see our Alternative Analyses section where we describe how we **changed the analysis for P2** and that we are replacing this analysis with the new models in the Ability to detect actual effects section (14 May 2020). We also describe here that we realized that Condition (manipulated or control) does not need to be a variable in our models because the manipulated birds have, by definition, faster reversal speeds.

5) 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: Taco and Chilaquile. We did not conduct the full test battery with Taco or put him in the flexibility manipulation or control groups (he received 1 reversal and then moved on to the next test) because he was the first juvenile and we wanted to see whether his performance was different from adult performances. His performances were similar to the adults, therefore we decided to put Chilaquile in the full test battery. Chilaquile's performances were also 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

6) We **log transformed** the response variable and changed the GLMM distribution from Poisson to Gaussian in the P3a analysis (24 Aug 2021).

7) The original model for P2 (Table 4: Model 1) included the covariate aviary batch, however this ended up confounding the analysis because control and manipulated individuals, while randomly assigned to these conditions, ended up in particular batches as a result of their willingness to participate in tests offered during their time in the aviary (Table 4: Model 3). Several grackles never passed habituation or training such that their first experiment could begin, therefore we replaced these grackles in the aviaries with others who were willing to participate. This means that batch did not indicate a particular temporal period. Therefore, we **removed batch from the model**.

8) **Bayesian data analysis**: we conducted post-hoc exploratory analyses on the serial reversal learning data to better understand the effect the flexibility manipulation had on performance. We used the version of the Bayesian model that was developed by A. Blaisdell et al. (2021) and modified by Logan CJ et al. (2020) [see Analysis Plan > mance. We used the version of the Bayesian model that was developed by A. Blaisdell et al. (2021) and modified by Logan CJ et al. (2020, see Analysis Plan > Flexibility analysis in 2020 for model specifications and validation). This model estimates two components to describe the behavior of individuals in the serial reversal learning experiments (the rate of updating previously learned attractions and the rate of deviating from the learned attractions), and we also relate these components to the data from the other experiments. See model details in Methods > Analysis Plan > Unregistered analyses: Bayesian flexibility models. We report our results at the end of the Results section.

RESULTS

Data are publicly available at the Knowledge Network for Biocomplexity (Logan, Blaisdell, et al., 2021). Please see the data sheet titled g_flexmanip_data_AllGrackleExpOrder at KNB for an overview of all color marked grackles at the Arizona field site (2018-2021), which of the aviary experiments they participated in, and whether data for the variables that were collected in the wild are present.

Although 22 grackles completed their initial colored tube discrimination, only 20 grackles participated in one or more reversals (Table 1). The rest of the tests began only after a bird's reversal experiment was complete (see the order of tests for each bird at the data sheet titled g_flexmanip_data_AllGrackleExpOrder at Logan, Blaisdell, et al. (2021)). Interobserver reliability analyses (unregistered) showed that the reversal learning and multi-access box (plastic and wooden) experiments were highly repeatable across live coders and video coders (see details in Analysis Plan > Interobserver reliability).

Table 1. Summarized results per bird in the reversal learning (tube and touchscreen) and multi-access box (plastic and wooden) experiments. Reversals to pass indicates how many serial reversals it took a bird to pass criterion if they were in the flexibility manipulation condition. Note: Tapa did not finish the MAB log experiment; Marisco's MAB log experiment ended too early due to experimenter error (timed out on 2 consecutive sessions, not 3); Mole and Habanero: do not count MAB plastic number of options solved because they were given the box fully put together for habituation due to experimenter error; Taco was the first juvenile we tested and we did not put him in the flexibility experiment: he received 1 reversal and moved on to his next test, therefore he was essentially a control bird without the matched yellow tube experience.

Bird	Batch	Sex	Trials to learn (tube)	Trials to first reversal (tube)	Trials to last reversal (tube)	Reversals to pass	Total loci solved (MAB plastic)	Total loci solved (MAB wooden)	Average latency to attempt new locus (MAB plastic)	Average latency to attempt new locus (MAB wooden)	Trials to learn (touch-screen)	Trials to first reversal (touch-screen)	Motor actions (MAB plastic)	Motor actions (MAB wooden)
Tomatillo	1	M	40	50	50	Control	3	NA	317	NA	NA	NA	13	NA
Queso	1	M	50	70	70	Control	1	NA	88	NA	330	460	8	NA
Tapa	1	F	30	100	100	Control	4	NA	685	NA	450	(629+)	13	NA
Yuca	3	F	40	80	80	Control	4	4	132	77	NA	NA	13	11
Marisco	3	M	40	50	50	Control	1	2	NA	208	NA	NA	4	7
Pizza	3	M	50	60	60	Control	0	1	NA	1482	NA	NA	0	8
Mofongo	4	M	20	40	40	Control	3	4	502	630	NA	NA	13	14
Taquito	4	M	90	160	160	Control	0	4	NA	100	NA	NA	11	10
Chalupa	1	F	50	90	50	8	0	NA	NA	NA	NA	NA	6	NA
Mole	1	M	30	70	50	7	4	4	356	1173	431	307	14	15
Habanero	1	M	50	80	40	6	4	NA	28	NA	350	290	15	NA
Diablo	3	M	20	80	40	8	2	1	25	NA	NA	NA	10	2
Burrito	3	M	40	60	23	8	3	4	76	391	NA	NA	17	18
Adobo	3	M	50	100	50	6	4	4	31	79	NA	NA	16	18
Chilaquile	3	JM	30	40	30	6	4	4	44	170	NA	NA	19	11
Pollito	4	M	40	60	40	8	0	3	NA	668	NA	NA	0	11
Taco	3a	JM	50	80	80	(Control)	1	4	NA	117	NA	NA	3	19
Memela	1	F	50	60	80	X (11+)	NA	NA	NA	NA	NA	NA	NA	NA
Fideo	2	M	60	70	70	Control	NA	NA	NA	NA	NA	NA	NA	NA
Avocada	1	F	50	100	100	Control	NA	NA	NA	NA	NA	NA	NA	NA
Huachinago	3	M	70	NA	NA	Control	NA	NA	NA	NA	NA	NA	NA	NA
Guacamole	4	M	30	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Because the wooden multi-access box was added after in principle recommendation, we conducted an unregistered analysis to determine whether the plastic and wooden multi-access box results correlated with each other, which would indicate that these tests are interchangeable. We found that they did not correlate with each other on either variable measured: the average latency to attempt a new locus (switching; Pearson's r=0.74, 95% CI=-0.19-0.97, t=2.18, df=4, p=0.09) or the total number of loci solved (problem solving; Pearson's r=0.51, 95% CI=-0.09-0.84, t=1.86, df=10, p=0.09). Therefore, these two tests are not interchangeable and we analyzed them separately.

P1: reversal speed gets faster with serial reversals

The birds in the manipulated group required a similar number of trials during their first reversal (R1 median=75 trials) as the birds in the control group needed during their first and only reversal (R1 median=70 trials). The manipulated birds improved during the reversal manipulation to a median of 40 trials in their last reversal. There was a significant negative correlation between the number of trials to reverse (average=71 trials, standard deviation (sd)=28) and the reversal number for those grackles in the flexibility manipulation condition (n=9, which included Memela who did not pass the manipulation condition; Figure 2).

Unregistered analysis: There was additionally a difference between manipulated and control reversal speeds when comparing their last reversals (Figure 3; for the control birds, their last reversal was their first reversal): the Akaike weight of the full model was 0.94, which means that including condition in the model explains the bulk of the variation in the number of trials to reverse in the last reversal (Table 3). This analysis includes 19 grackles (8 manipulated condition - only those who actually passed the manipulation, 11 control condition) who had an overall average of 62 trials in their last reversal (sd=32).

Table 2. The number of trials to reverse decreases with increasing reversal number.

	Posterior mean	Lower 95% confidence interval	Upper 95% confidence interval	Effective sample size	pMCMC	Significance code: **=0.01
(Intercept)	4.43921	4.24323	4.63401	420	<0.002	**
Reverse Number	-0.05558	-0.09386	-0.01920	420	<0.002	**

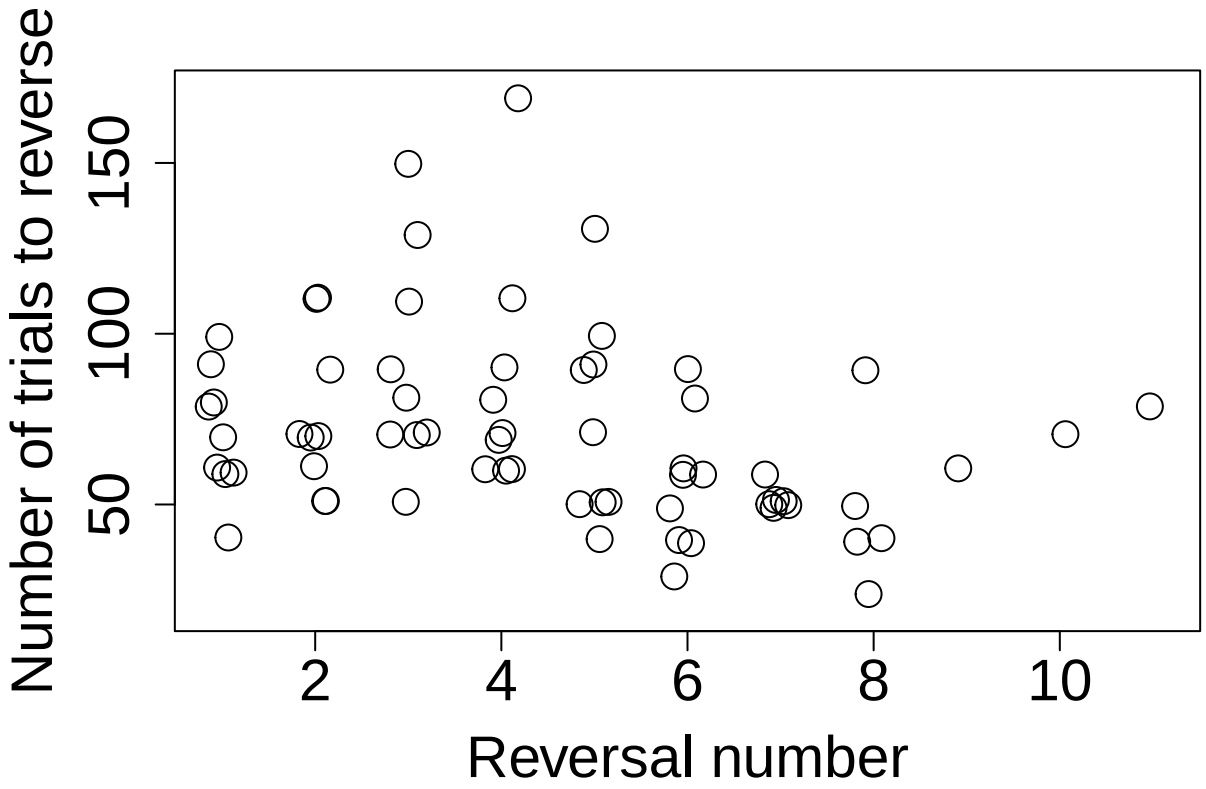


Figure 2. Individuals in the manipulated condition (who received serial reversals) did not linearly decrease their reversal passing speeds with increasing reversal number (n=9 grackles).

Table 3. Individuals in the manipulated condition pass their last reversal in fewer trials than control individuals. The Akaike weight of the full model was >0.89, indicating that it is more reliable than the null model.

	(Intercept)	d$ReversalsToPass	df	logLik	AICc	delta	weight
2	78.18182	+	3	-88.09966	183.7993	0.000000	0.94218449
1	62.26316	NA	2	-92.31561	189.3812	5.581888	0.05781551

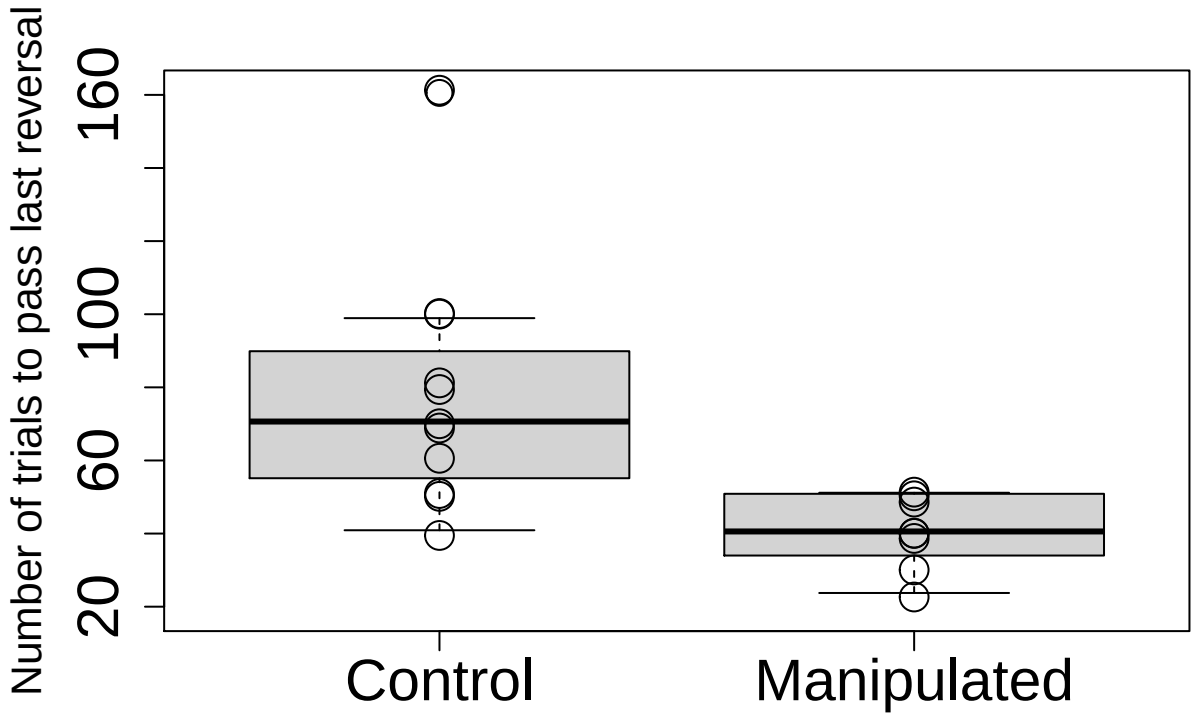


Figure 3. Individuals in the manipulated condition (who received serial reversals) passed their last reversal in fewer trials than individuals in the control condition (who only received 1 reversal). n=19 grackles: 11=control, 8=manipulated.

P2: serial reversals improve rule switching and problem solving on the MAB

To determine whether the serial reversal manipulation affected flexibility generally, we compared performance (the number of trials to reverse a preference in the first and last color reversal, performance of the manipulated group relative to the control group) to speed of solution switching on two multi-access boxes. Furthermore, we assessed whether flexibility measured through these serial reversals related to innovativeness by comparing performance to the number of loci solved on the multi-access boxes. The results for each of these comparisons are described in detail below and an overview is provided in Figure 4.

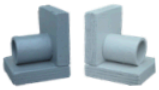
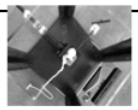
P2: How does flexibility, measured via performance on serial reversals, relate to flexibility in another context and innovativeness?	**Flexibility** (number of trials to pass in serial reversals) 			**Phi** (rate of updating attractions)	**Lambda** (deviation from attractions)
	First Reversal	Last Reversal	Manipulated relative to Control	Last Reversal	Last Reversal
Flexibility in a new context (latency to switch loci) 	+	+	+	U	U
	-	0	0	U	0*
Innovativeness (number of loci solved) 	0	+	0	0*	U
	0	0*	+	0*	U

Figure 4. Overview of the results from the P2 analyses with the multi-access boxes (plastic and wooden). An effect of natural variation in flexibility on performance on the multi-access box tasks would result in correlations in the first reversal. An effect of the flexibility manipulation would result in a change in correlations from the first to last reversals. A plus sign (+) indicates a positive correlation, a minus sign (-) indicates a negative correlation, the letter U indicates that birds with intermediate values perform worse, and a 0 indicates no correlation between the two variables. The asterisks (*) indicate that a small sample size decreases the reliability of this result.

Rule switching: latency to attempt a new locus on the multi-access box (plastic) ~ trials to reverse Grackles that were faster to reverse a preference in their **last reversal** (average 52 trials, sd=23), where grackles in the control condition received only one reversal which served as their first and last reversal, were also faster to attempt to solve a new locus on the plastic multi-access box (after just having passed criterion on a different locus; average=208 seconds, sd=226; Figure 5a; Table 4: Model 9; n=11 grackles: 6 in manipulated condition, 5 in control condition; 6 subjects completed this experiment but solved 0 loci or 1 locus and so did not have switching times). We also found that individuals in the flexibility manipulation had faster switch latencies than those in the control condition (Table 4: Model 10). There was a positive correlation between the number of trials to reverse in the **first reversal** (average=70 trials, sd=21) and the average switch latency on the plastic multi-access box (Table 4: Model 11). A correlation was determined to be present if the prediction interval for the slope (b) in the model output did not cross zero (Table 4). This criterion was used throughout the analyses for P2.

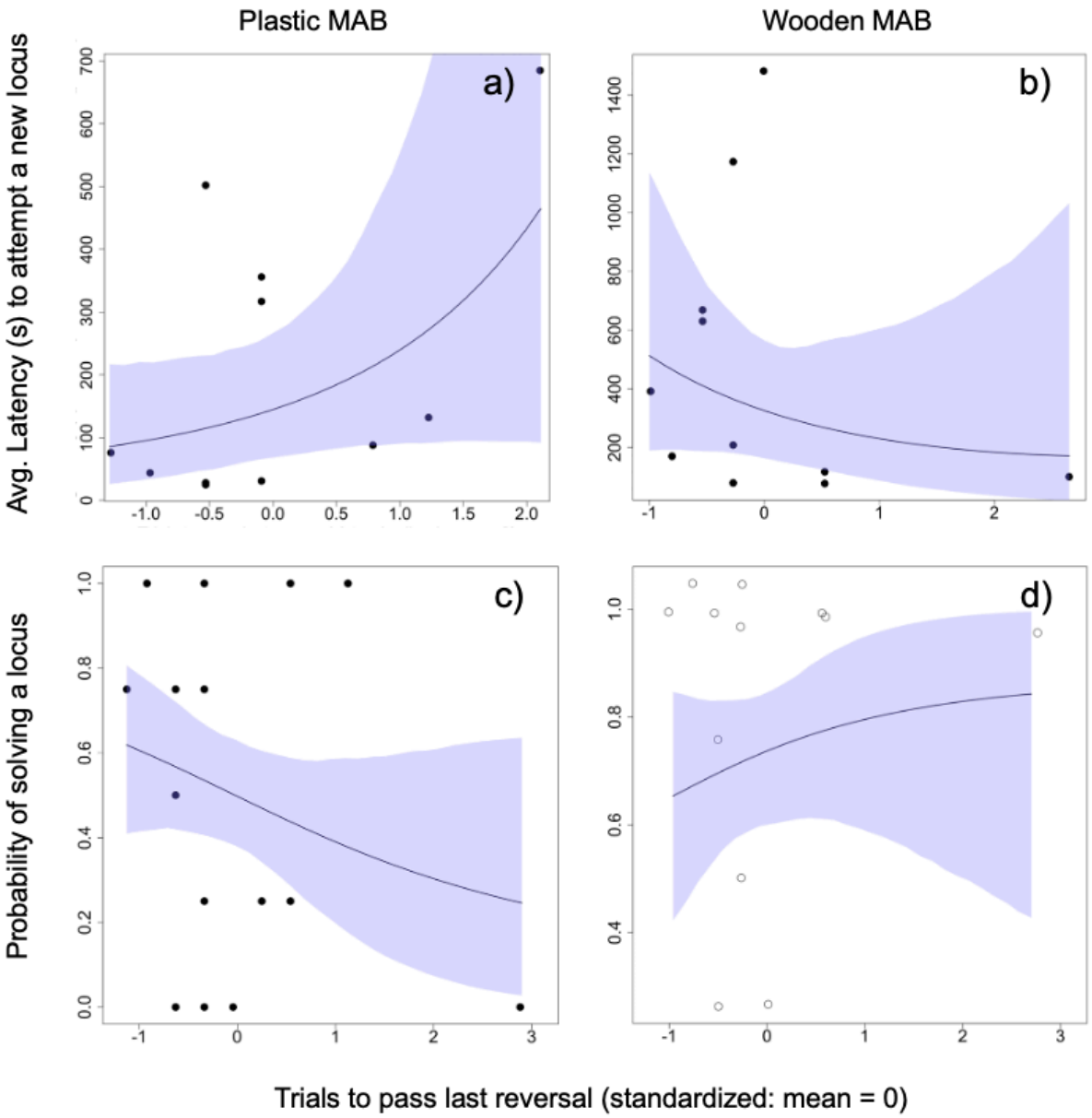


Figure 5. The average latency (seconds) to attempt to solve a different locus after having previously successfully solved a locus on a) the plastic multi-access box (MAB) is positively correlated with the number of trials to pass their last reversal (n = 11 grackles), but on b) the wooden MAB it is not correlated with the number of trials to pass their last reversal (n = 11 grackles). Additionally, the probability of solving a locus on c) the plastic MAB is negatively correlated with the number of trials to pass their last reversal (n = 15 grackles), but on d) the wooden MAB it is not correlated with the number of trials to pass their last reversal (n = 12 grackles, estimate of slope includes zero). Shading represents the 97% prediction intervals.

Rule switching: latency to attempt a new locus on the multi-access box (wooden) ~ trials to reverse (unregistered analysis) There was no correlation between the number of trials to reverse a preference in their **last reversal** (average 60 trials, sd=38) and the latency to attempt to solve a new locus on the wooden multi-access box (after just having passed criterion on a different locus; average=463 seconds, sd=481; Figure 5b; Table 4: Model 12; n=11 grackles: 5 in manipulated condition, 6 in control condition; Diablo also completed this experiment and solved 1 locus, but did not attempt another locus after that,

thus he does not have any switching times to analyze). We additionally found that there was no difference in the average latency to switch between individuals in the flexibility manipulation and those in the control condition (Table 4: Model 13). There was a negative correlation between the number of trials to reverse in the **first reversal** (average=73 trials, sd=34) and the average switch latency on the multi-access box (Table 4: Model 14).

Innovativeness: number of loci solved on the multi-access box (plastic) ~ trials to reverse Grackles that were faster to reverse a preference in their **last reversal** (average 62 trials, sd=34) solved more loci on the plastic multi-access box (average=2 loci, sd=1.6; Figure 5c; Table 4: Model 2; n=15 grackles: 6 in manipulated condition, 9 in control condition; this number excludes Mole and Habanero who were, due to experimenter error, given the fully put together box during habituation and could have learned how to solve the loci at that time). There was no correlation between the number of loci solved and which reversal condition a grackle was randomly assigned to (Table 4: Model 4). There was also no correlation between the number of trials to reverse in the **first reversal** (average=75 trials, sd=31) and the number of loci solved on the multi-access box (Table 4: Model 5).

Table 4. Model outputs for the number of loci solved and the latency to switch loci after passing criterion on a different locus on the plastic (models 1-5 and 9-11) and wooden (models 6-8 and 12-14) multi-access boxes. SD=standard deviation, the 89% prediction intervals are shown, n_eff=effective sample size, Rhat4=an indicator of model convergence (1 is ideal), b=the slope of the relationship between loci solved or average switch latency and the number of trials to pass the reversal.

	Mean	SD	5.5%	94.5%	n_eff	Rhat4
MODEL 1 (last reversal): loci solved plastic ~ a[batch] + b*trials						
a[1]	0.04	0.46	-0.70	0.78	2304	1.00
a[2]	0.29	0.36	-0.30	0.87	2456	1.00
a[3]	-0.78	0.55	-1.65	0.08	2510	1.00
b	-0.22	0.25	-0.63	0.18	2364	1.00
MODEL 2 (last reversal): loci solved plastic ~ a + b*trials						
a	-0.02	0.24	-0.40	0.35	1466	1.00
b	-0.46	0.31	-0.97	-0.01	1383	1.00
MODEL 3 (last reversal): trials ~ a[batch]						
a[1]	0.09	0.37	-0.48	0.69	2095	1.00
a[2]	-0.21	0.29	-0.68	0.25	1715	1.00
a[3]	0.25	0.39	-0.38	0.86	2161	1.00
sigma	1.03	0.21	0.75	1.39	2049	1.00
MODEL 4: loci solved ~ a[condition]						
a[1] control	-0.11	0.32	-0.62	0.40	1311	1.00
a[2] manipulated	0.15	0.39	-0.46	0.80	1222	1.00
MODEL 5 (first reversal): loci solved plastic ~ a + b*trials						
a	0.00	0.24	-0.37	0.39	1208	1.00
b	-0.44	0.30	-0.94	0.02	1273	1.00
MODEL 6 (last reversal): loci solved wooden ~ a + b*trials						
a	1.06	0.27	0.63	1.50	1255	1.00
b	0.41	0.43	-0.21	1.13	1107	1.00
MODEL 7: loci solved ~ a[condition]						
a[1] control	-0.45	0.40	-1.10	0.18	1161	1.00
a[2] manipulated	0.77	0.41	0.13	1.44	1302	1.00
MODEL 8 (first reversal): loci solved wooden ~ a + b*trials						
a	0.11	0.26	-0.30	0.52	1221	1.00
b	-0.50	0.35	-1.09	0.04	1234	1.00
MODEL 9 (last reversal): avg switch latency plastic ~ a + b*trials						
a	4.93	0.30	4.45	5.41	1235	1.01
b	0.46	0.29	0.00	0.92	1363	1.00
phi	0.93	0.35	0.44	1.55	1476	1.00
MODEL 10: avg switch latency plastic ~ a[condition]						
a[1] manipulated	4.07	0.39	3.46	4.68	1027	1.00
a[2] control	5.18	0.39	4.50	5.76	1006	1.00
phi	0.91	0.41	0.37	1.63	925	1.01
MODEL 11 (first reversal): avg switch latency plastic ~ a + b*trials						
a	4.93	0.29	4.46	5.39	1488	1.00
b	0.46	0.28	0.02	0.93	1211	1.00
phi	0.94	0.36	0.44	1.60	1447	1.00
MODEL 12 (last reversal): avg switch latency wooden ~ a + b*trials						
a	5.75	0.28	5.28	6.18	1049	1.00
b	-0.41	0.32	-0.86	0.15	1281	1.01
phi	1.04	0.42	0.48	1.77	1456	1.00
MODEL 13: avg switch latency wooden ~ a[condition]						
a[1] control	5.31	0.42	4.61	5.95	701	1.00
a[2] manipulated	5.34	0.44	4.61	6.00	620	1.01
phi	0.66	0.32	0.25	1.25	806	1.00
MODEL 14 (first reversal): avg switch latency wooden ~ a + b*trials						
a	5.71	0.26	5.28	6.12	1109	1.00
b	-0.50	0.28	-0.89	-0.01	1308	1.00
phi	1.08	0.41	0.53	1.80	1347	1.00

Innovativeness: number of loci solved on the multi-access box (wooden) ~ trials to reverse (unregistered analysis) The prediction interval for the estimate for the association between the number of loci solved on the wooden multi-access box (average=3.2, sd=1.3) and the number of trials to reverse a preference in their **last reversal** (average=59 trials, sd=38) crossed zero (Figure 5d; Model 6, Table 4; n=12 grackles: 6 in manipulated condition, 6 in control condition). This could mean that there is no association, however our simulations showed that we would not be able to reliably distinguish whether a small effect is different from zero with our sample size (correlation test suggests effect size of 0.2; Table M2). We did find a correlation between the number of loci solved and which reversal condition a grackle was randomly assigned to, indicating the reversal manipulation appears to have affected performance on the wooden multi-access box. The model estimates that manipulated birds solved on average 1.2 more loci than birds in the control condition (Table 4: Model 7, wooden; 89% prediction intervals=0.34-2.14; n=12 grackles: 6 in manipulated condition, 6 in control condition). However, there is no association between the number of trials to reverse in the first reversal (average=74 trials, sd=34) and the number of loci solved on the multi-access box (Table 4: Model 8, wooden).

Reversal learning experiments: discriminating shapes on the touchscreen compared with color using tubes In the tube experiment, it took four grackles an average of 40 trials (sd=12) in the initial discrimination phase to learn to prefer a color, while it took the same individuals an average of 390 trials (sd=59) to learn to prefer a shape using the touchscreen (Queso, Mole, Habanero, and Tapa). The two individuals who were faster to learn in the tube experiment were slower to learn in the touchscreen experiment. For the reversal, it took three of these individuals (Queso, Mole, and Habanero) an average of 80 trials (sd=14) to reverse their colored tube preference, and an average of 362 trials (sd=111) to reverse their shape preference on the touchscreen (Tapa had to be released back to the wild before finishing the experiment, but was on trial 629 in reversal one of the touchscreen experiment at the time of release. In the tube experiment, she was also the slowest of the four to reverse at 100 trials). All three individuals were about equally fast at the reversal in the tube experiment, while their reversal learning speeds differed on the touchscreen.

P2 alternative 2 (additional analysis): latency and motor diversity

Because there was no correlation between the number of trials to reverse in the last reversal and the latency to attempt a different locus on the wooden multi-access box, we conducted this additional analysis to determine whether the model fit was improved when adding the number of motor actions as an explanatory variable. Adding the number of motor actions (wooden: average=13, sd=4) did not improve the model fit when examining the relationship between the latency to switch loci on the wooden multi-access box (wooden: average=463, sd=481) and the number of trials to reverse in the last reversal (wooden: average=60, sd=38) because the Akaike weights were similar for both models (wooden: n=11 grackles: 5 in the manipulated group, 6 in the control group; Table 5).

Table 5. Adding the number of motor actions used to the analysis of the average latency to attempt a new option on the wooden multi-access box and the number of trials to reverse in the last reversal does not improve the model fit.

	(Intercept)	dw$MotorActions	dw$TrialsLastReversal	df	logLik	AICc	delta	weight
1	463.1818			2	-83.02521	171.5504	0.000000	0.70712147
3	665.8320		-3.362220	3	-82.63113	174.6908	3.140406	0.14708333
2	783.9748	-24.85016		3	-82.76565	174.9599	3.409451	0.12857047
4	1136.8430	-32.86188	-4.138591	4	-82.15674	178.9801	7.429713	0.01722472

P3a: reversal is repeatable within individuals within a context

Performance was repeatable within individuals within the context of reversal learning. We obtained a repeatability value of 0.13, which is significantly greater than that expected if birds are performing randomly in each reversal (p=0.001; 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.

P3b: not repeatable across contexts

There was no consistency of flexibility in individuals across contexts: the latency to attempt a different locus on both multi-access boxes did not correlate within individuals with the number of trials to reverse a preference in each reversal (Table 6; n=8 grackles: only those in the manipulated condition because only they experienced more than one reversal; Memela was not included because she did not complete the reversal experiment and therefore was not offered the multi-access box experiments).

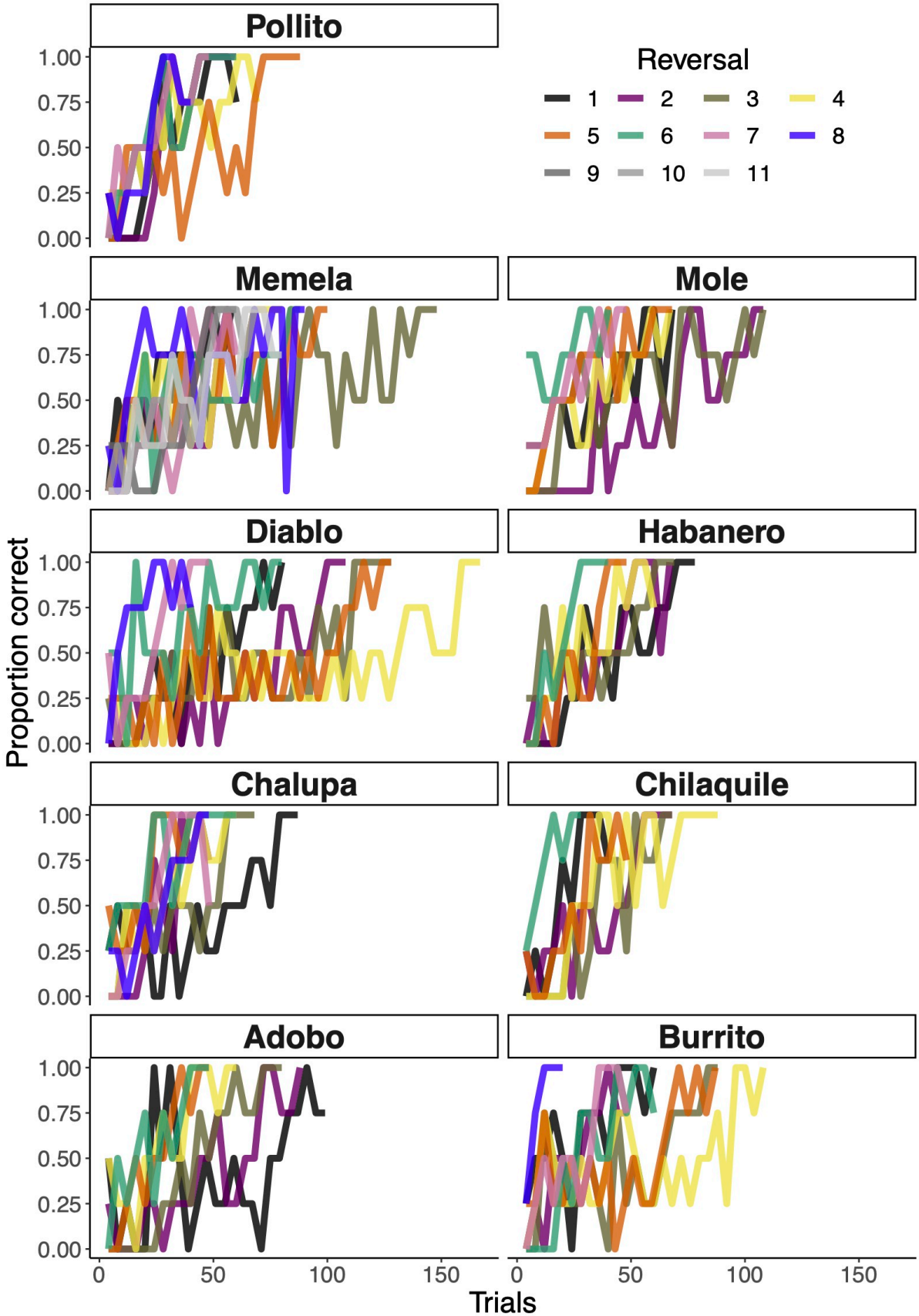
Table 6. No repeatability across contexts. MCMCglmm output for the multi-access box plastic and wooden models.

Table 1:

	variable	post.mean	l.95..CI	u.95..CI	eff.samp	pMCMC	effect	modelNa
1	(Intercept)	2.3	-5.5	11.3	100	0.6	fixed	Plasti
2	ReverseNumber	1.0	-2.3	6.1	100	0.6	fixed	Plasti
3	TrialsToReverse	0.01	-0.1	0.1	100	0.8	fixed	Plasti
4	ReverseNumber:TrialsToReverse	-0.01	-0.1	0.04	100	0.7	fixed	Plasti
5	ID	0.1	0	0.3	100		random	Plasti
6	units	1.9	0.7	3.7	100		residual	Plasti
7	(Intercept)	4.8	0.5	9.9	28.4	0.02	fixed	Woode
8	ReverseNumber	-0.4	-2.8	2.4	49.4	0.9	fixed	Woode
9	TrialsToReverse	0.02	-0.04	0.1	31.8	0.5	fixed	Woode
10	ReverseNumber:TrialsToReverse	0.002	-0.03	0.03	51.7	0.8	fixed	Woode
11	ID	1.3	0	5.1	100		random	Woode
12	units	0.5	0.1	1.8	69.7		residual	Woode

P4: serial reversal learning strategy

Three out of nine grackles switched from an epsilon-decreasing to an epsilon-first strategy in their last reversal (Diablo reversal 8, Burrito reversal 8, and Chilaquile reversal 6; Figure 6). The rest continued to rely on an epsilon-decreasing strategy throughout their reversals.


Pollito
Memela
Mole
Diablo
Habanero
Chalupa
Chilaquile
Adobo
Burrito
Reversal
1
2
3
4
5
6
7
8
9
10
11
Proportion correct
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Trials

429

Figure 6. The proportion of trials correct by trial number and reversal for each bird.

We additionally quantitatively determined to what degree each bird used the exploration versus exploitation strategy using methods in Federspiel et al. (2017) by calculating the number of 10-trial blocks where birds were choosing "randomly" (2-9 correct choices; called sampling blocks; akin to the exploration strategy) divided by the total number of blocks to reach criterion per bird. This ratio was also calculated for "acquisition" blocks where birds made primarily correct choices (9-10 correct choices; akin to the exploitation strategy). There was no correlation between exploration (sampling ratio) or exploitation (acquisition ratio) and reversal number (sampling: reversal estimate=-0.09, SE=0.11, z=-0.86, p=0.39; acquisition: reversal estimate=0.00, SE=0.00, z=-0, p=1.00), indicating that the grackles did not use a particular strategy earlier or later in their serial reversals.

Post-hoc, unregistered exploratory analyses to investigate the effect the flexibility manipulation had on performance

In addition to the planned analyses, we conducted post-hoc exploratory analyses on the serial reversal learning data to better understand the effect the flexibility manipulation had on performance. We used the version of the Bayesian model that was developed by A. Blaisdell et al. (2021) and modified by Logan CJ et al. (2020see 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 learning rate of attraction to either option (ϕ) and the rate of deviating from learned attractions (λ). We wanted to address the following questions: 1) What did the manipulation change? 2) Do the manipulations shift birds beyond what is naturally observed and does it make them more similar? 3) Are ϕ or λ, the two components of flexibility in reversal learning, associated with performance on the multi-access boxes across control and manipulated birds?

1) Observed effects of the manipulation on reversal performance, ϕ, and λ A pooled model of performance across all reversals estimates that birds can expect to improve by about 30 trials (89% prediction interval (PI): 25-36; Table 7: Model 15) after completing the serial reversals. While all manipulated birds improved, those birds that were already fast to reverse in their first reversal improved less than the birds that required many trials to reverse in their first reversal (posterior peak indicates a correlation of +0.64, with highest posterior density intervals (HPDI) all positive, between the first reversal value and the improvement achieved by the last reversal; Table 7: Model 16). However, the birds who were the fastest in the first reversal, were also the fastest in the last reversal, but the difference between the slower and faster reversers is reduced (Figure 7).

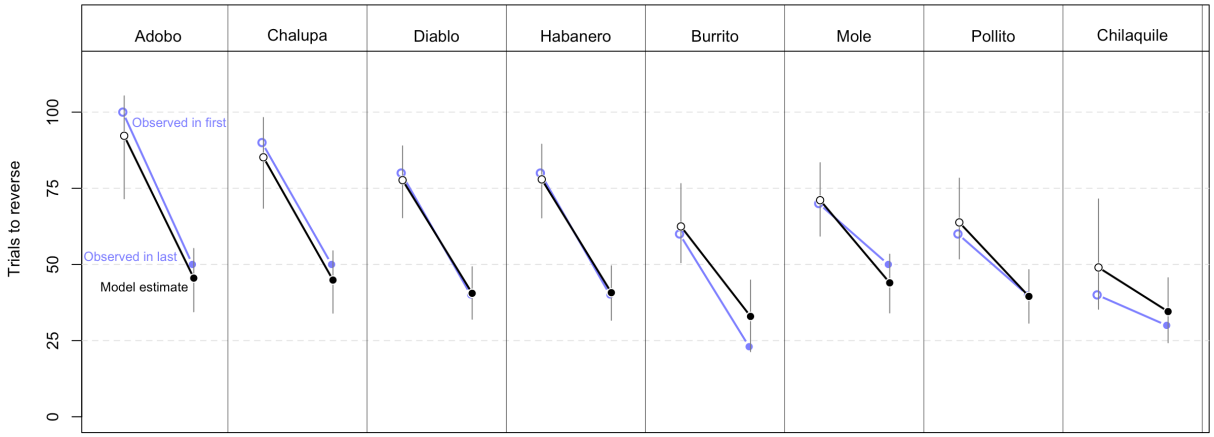


Figure 7. All eight manipulated birds needed fewer trials to reverse in their last reversal than in their first. Their improvement depended on their starting value, with steeper slopes for those birds that needed more trials to reverse in the first reversal (blue = observed values and changes, black = model estimates). However, birds who needed more trials in the first reversal did not completely catch up, such that the birds that needed more trials in their first reversal also needed more trials in their last reversal relative to other grackles.

The findings from the simulated data indicated that λ and ϕ can only be estimated accurately when calculated across at least one switch (initial discrimination plus first reversal or final two reversals). For the manipulated birds, the estimated ϕ more than doubled from 0.03 (for reference, control grackles=0.03) in the beginning to 0.07 in their last two reversals (model estimate of expected average change: +0.02 to +0.05; Table 7: Model 17), while their λ went slightly down from 4.2 (for reference, control grackles=4.3) to 3.2 (model estimate of average change -1.63 to -0.56; Table 7: Model 18). For ϕ, this pattern fits with the observations in the simulations: larger ϕ values are associated with fewer trials to reverse. However, while in the simulations individuals needed fewer trials to reverse when we increased λ (less deviation from the learned association), the birds in the manipulation showed an increased λ in their last reversal when they needed fewer trials to reverse. This suggests that λ is a constraint rather than having a direct linear influence on the number of trials to reverse: birds with low λ still can reach the criterion in a small number of trials as long as they have a sufficiently high value of ϕ (see Figure M1 in the Methods).

For the ϕ values, we also observed a correlation between the ϕ estimated from an individual's performance in the first reversal and how much their ϕ changed toward the value for their performance in the last reversal (-0.4; 50% highest posterior density intervals (HPDI) all negative;Table 7: Model 17), while there is no such obvious relationship for λ (-0.15; 50% HPDI crosses zero; Table 7: Model 18). For both ϕ and λ, unlike for the number of trials to reverse, we did not see that the individuals who had the largest values during the first reversal also always had the largest values during the last reversal. The manipulation changed both ϕ and λ, such that, across all birds, there was a negative correlation between ϕ and λ.

2) Variation in reversal performance, ϕ, and λ The values we observed after the manipulation in the last reversal for the number of trials to reverse, as well as the ϕ and λ values estimated from the last reversal, all fall within the range of variation we observed among the control birds in their first and only reversal (Figure 8). This means that the manipulation did not push birds to new levels, but changed them within the boundaries of their natural environment. Some birds in the control group already had similar flexibility measures to the manipulated birds after going through serial reversal learning, presumably because some birds have had experiences in their natural environments that made them more flexible. Accordingly, birds

in the manipulated group were not automatically all better performers than all of the birds in the control group. Those birds who needed only few trials in their last reversal, irrespective of whether they were in the control or the manipulated group (first and only reversal for control birds, last reversal for manipulated birds) were also on average better at solving the multi-access boxes (see results above on rule switching and Figure 4).

Across both manipulated and control birds, ϕ was more consistently associated with the number of trials individuals needed to reverse, and ϕ changed more than λ across reversals for the manipulated birds (Figure 8). However, changes in ϕ and λ independently correlated with changes in the improvement in performance of the manipulated birds from the first to the last reversal (association of change in number of trials from first to last reversal with standardized change in ϕ: 11, 89% PI: 6-15 and with standardized λ: 6, 89% PI: 1-10; Table 7: Model 19).

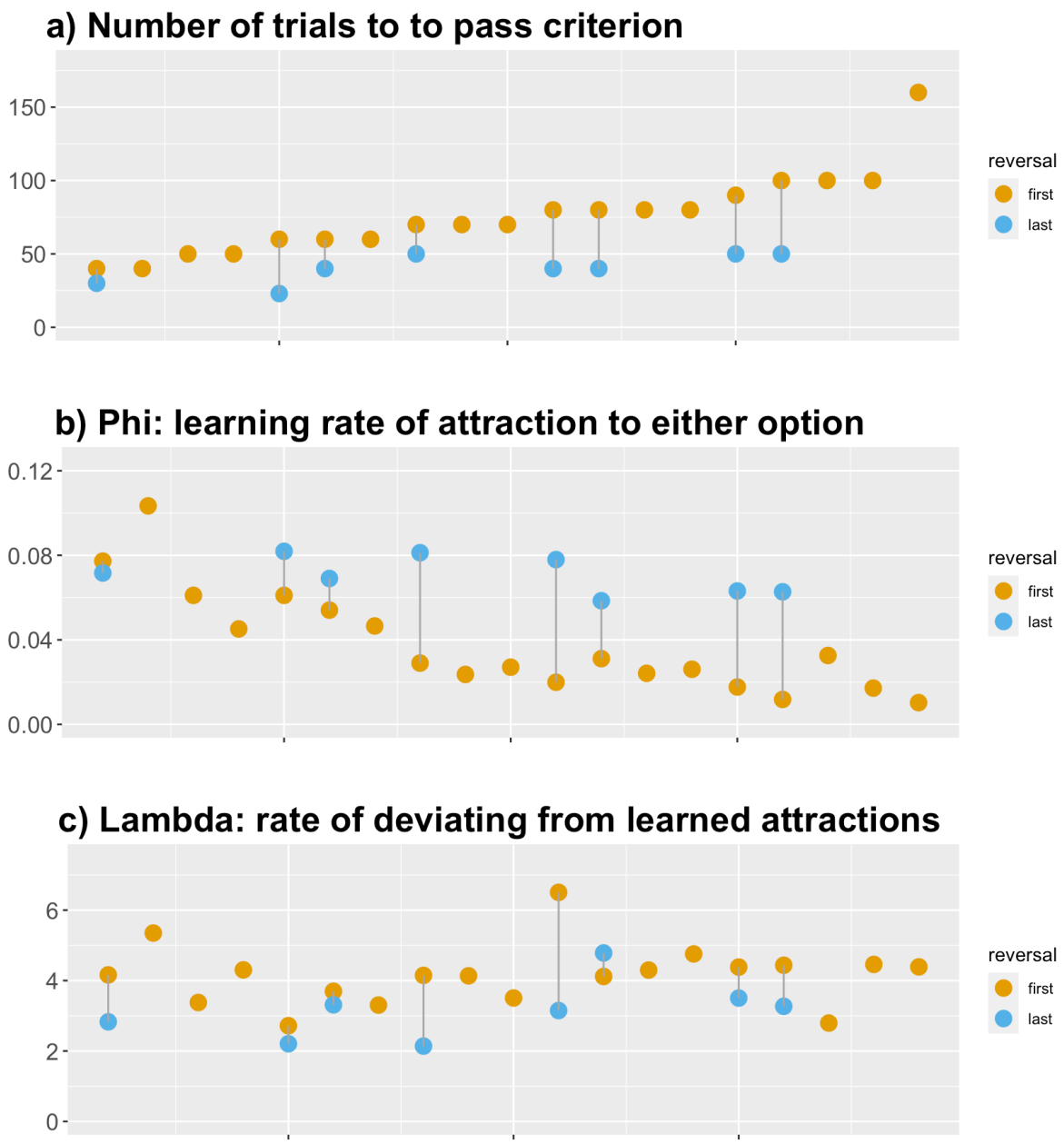


Figure 8. Comparisons of the different measures of performance in the reversal task for each of the 19 birds. The figure shows a) the number of trials to pass criterion for the first reversal (orange; all birds) and the last reversal (blue; only manipulated birds); b) the ϕ values reflecting the learning rate of attraction to the two options from the initial discrimination and first reversal (orange; all birds) and from the last two reversals (blue; manipulated birds); and c) the λ values reflecting the rate of deviating from the learned attractions to the two options from the initial discrimination and first reversal (orange; all birds) and from the last two reversals (blue; manipulated birds). Individual birds have the same position along the x-axis in all three panels. Birds that needed fewer trials to reverse their preference generally had higher ϕ values, whereas λ appeared to reflect whether any choices of the unrewarded color occurred throughout the trials or only at the beginning. For the manipulated birds, their ϕ values changed more consistently than their λ values, and the ϕ values of the manipulated individuals were generally higher than those observed in the control individuals, while their λ values remained within the range also observed in the control group.

The pairwise analyses above indicate that the number of trials in the last reversal is correlated with the number of trials in the first reversal, with ϕ, and with λ. The number of trials in the first reversal, ϕ, and λ are also correlated with each other (Figure 9). With the Bayesian approach, we can use one model to estimate all potential links simultaneously to identify the pathways through which the variables interact with each other (e.g., some variables might be correlated because both are influenced by a third variable). We therefore simultaneously estimated support for the following pathways:

- trials last reversal ~ trials first reversal + ϕ last reversal + λ last reversal
- trials first reversal ~ ϕ first reversal + λ first reversal;
- ϕ last reversal ~ ϕ first reversal
- λ last reversal ~ λ first reversal

Results from this simultaneous estimation of the potential pathways shows that the ϕ from the initial learning and first reversal determines the number of trials to pass the first reversal, which, in turn, explains how many trials they need to pass their last reversal. The ϕ for the last reversal does not appear to provide any additional information about the number of trials in the last reversal, and λ is not directly associated with the number of trials birds need to reverse (Table 7: Model 20) (Figure 9).

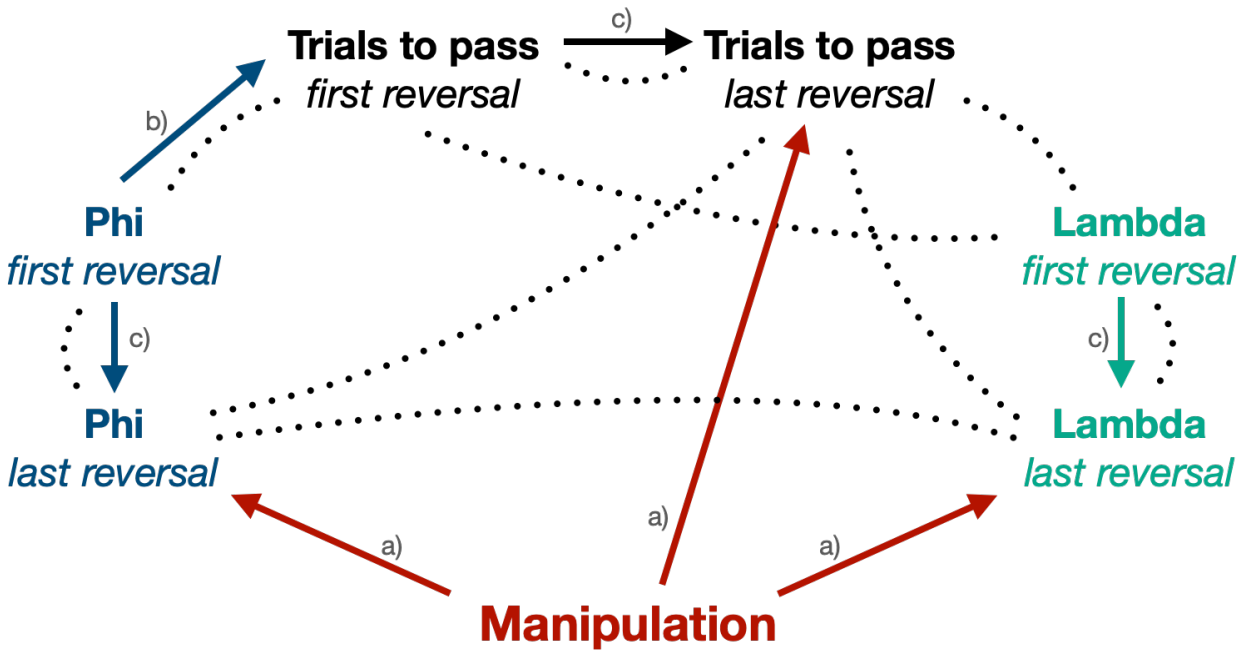


Figure 9. Causal graph showing the relationships between the number of trials to pass a reversal, ϕ, λ, and the flexibility manipulation. In the pairwise assessments (dotted lines), most of the variables are indicated as being associated with each other. The combined model identifies which of these associations are likely to be direct (solid lines with arrows). The results from the combined model indicate that a) the manipulation worked, b) ϕ has a more direct influence on performance in the reversals than λ does, and c) individuals have some consistency both in their abilities and in their performance.

Table 7. Model outputs for the pairwise comparisons (models 15-19) and for the combined model (model 20) explaining the changes during the manipulation. SD=standard deviation, the 89% prediction intervals are shown, n_eff=effective sample size, Rhat4=an indicator of model convergence (1 is ideal).

	Mean	SD	5.5%	94.5%	n_eff	Rhat4
MODEL 15 (improvement) trials ~ a[bird] + b[bird]*reversal						
b_bar	-30.30	3.51	-35.65	-24.65	109	1.00
sigma_bar	2.13	2.93	0.17	9.77	9	1.00
sigma	6.54	2.42	0.23	9.41	10	1.00
MODEL 16 (improvement): trials ~ a[reversal] + b[bird,reversal]						
rho	0.34	0.39	-0.40	0.85	2452	1.00
MODEL 17 (phi improvement): phi ~ a[bird] + b[bird]*reversal						
a	0.00	0.02	-0.02	0.03	620	1.00
b	0.03	0.01	0.02	0.05	207	1.01
rho	-0.29	0.46	-0.93	0.52	1492	1.00
sigma	0.02	0.01	0.01	0.03	184	1.01
MODEL 18 (lambda improvement): lambda ~ a[bird] + b[bird]*reversal						
a	5.36	0.35	4.57	6.18	255	1.01
b	-1.10	0.30	-1.57	-0.64	260	1.01
rho	-0.08	0.44	-0.77	0.64	566	1.01
sigma	0.85	0.20	0.58	1.19	648	1.00
MODEL 19 (improvement association): performanceimprovement ~ a + b*phiimprovement + c*lambdaimprovement						
a	32.74	2.52	28.76	36.79	1362	1.00
b	10.63	3.09	5.68	15.31	1155	1.00
c	5.58	3.03	0.73	10.20	1223	1.00
sigma	7.22	1.36	5.31	9.56	1322	1.00
MODEL 20 (combined)						
trials last ~ trials first	0.62	0.36	0.04	1.17	1166	1.00
trials last ~ phi last	-0.28	0.51	-1.07	0.54	1095	1.00
trials last ~ lambda last	-0.22	0.48	-0.98	0.55	1278	1.00
trials first ~ phi first	-1.04	0.15	-1.26	-0.80	1059	1.00
trials first ~ lambda first	0.18	0.16	-0.41	0.06	890	1.00
phi last ~ phi first	0.29	0.37	-0.31	0.86	1696	1.00
lambda last ~ lambda first	0.19	0.38	-0.41	0.79	1806	1.00

3) Association between ϕ and λ with performance on the multi-access boxes We modified the analyses from the preregistered analyses in the Results section that assessed potential links between reversal learning and performance on the multi-access boxes by replacing the number of trials it took individuals to reverse with ϕ (learning rate of attraction to either option) and λ (rate of deviating from learned attractions) estimated from the reversal performances. The modified analyses did not find matches with any of the three previously detected correlations between reversal learning and performance on the two multi-access boxes (latency to attempt a locus on the plastic multi-access box, number of loci solved on the plastic and wooden multi-access boxes) (Table 8). We detected a different correlation: the latency to attempt a new locus on the wooden multi-access box was positively correlated with ϕ in the last reversal (Table 8: Model 28). This correlation appears to arise not because of a linear increase of the latency with increasing ϕ values, but because there are several individuals who have both a long latency and a large ϕ. However, there are also some individuals who have a long latency with a low ϕ (see below for additional analyses). This indicates that individuals who were faster to update their associations in reversal learning (higher ϕ, therefore needed fewer trials in their last reversal) took more time to attempt a new locus. Even though ϕ was closely associated with the number of trials a bird needed to reach the reversal criterion, we presumably could not recover the previous correlations because of our small sample sizes. In addition, we estimated ϕ and λ across at least one

reversal (initial discrimination plus first reversal, or last two reversals for manipulated birds), whereas the previous analyses using the number of trials to reverse were based on a single reversal (first or last reversal).

For the manipulated birds, we found that during their last reversal there was a positive correlation between ϕ and λ, with individuals with higher ϕ values also showing higher λ values. This positive correlation could lead to worse performance on the multi-access boxes for birds with intermediate values. There could be two alternative routes to better performances on the multi-access boxes with some birds solving a new locus faster because they quickly update previously learned associations (higher ϕ) despite also deviating more from learned associations (higher lambda), while other birds might attempt a new locus faster because they are more likely to deviate from learned associations (lower λ) despite also not updating information as quickly (lower ϕ). Our data shows that, for the number of loci solved on both the plastic and the wooden multi-access boxes, there is a U-shaped association, particularly with λ values in the last reversal (Table 8: models 39 & 46) (Figure 10), with birds with intermediate values of λ solving fewer loci on both multi-access boxes (Figure 4). For the latency to attempt a new locus, there is also a U-shaped association, particularly with ϕ, with birds with intermediate values of ϕ showing shorter latencies to attempt a new locus (Table 8: models 25 & 32). Given that there is also a positive correlation between number of loci solved and the latency to attempt a new locus, there might be a trade off, where birds with extreme ϕ and λ values solve more loci, but need more time, whereas birds with intermediate values have shorter latencies, but solve fewer loci.

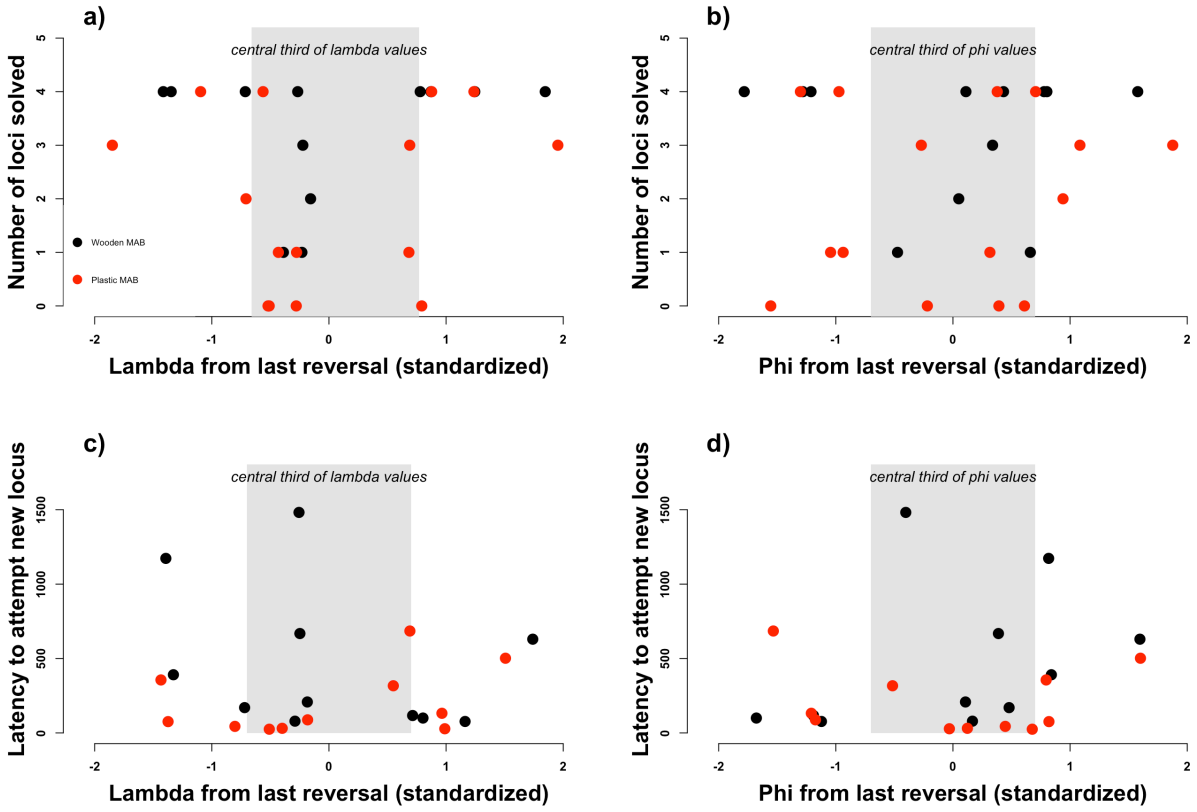


Figure 10. Relationships between phi and lambda from the last reversal and performance on the wooden (black dots) and plastic (red dots) multi-access boxes. Birds with intermediate λ values in their last reversal (a) were less likely to solve all four loci on the multi-access boxes than birds with either high or low λ values. Birds who solved two or fewer loci on either box all fall within the central third of the λ values observed for the last reversal, while 12 of the 14 birds who solved all four loci fall outside this central range. An individual's ϕ and λ values change 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. There are no clear relationships between (b) ϕ and the number of loci solved, (c) λ and the latency to attempt a locus, or (d) ϕ and the latency to attempt a new locus.

Table 8. Model outputs for the **latency** to switch loci after passing criterion on a different locus on the plastic (models 21-27) and wooden (models 28-34) multi-access boxes in relation to ϕ and λ. SD=standard deviation, the 89% prediction intervals are shown, n_eff=effective sample size, Rhat4=an indicator of model convergence (1 is ideal), b=the slope of the relationship between loci solved or average switch latency and ϕ or λ.

	Mean	SD	5.5%	94.5%	n_eff	Rhat4
MODEL 21 (plastic phi): latency ~ a + b*phi						
a	4.99	0.31	4.51	5.48	1354	1
b	-0.07	0.24	-0.45	0.31	1769	1
var	0.80	0.31	0.39	1.34	1527	1
MODEL 22 (plastic lambda): latency ~ a + b*lambda						
a	4.97	0.30	4.50	5.46	1547	1
b	0.32	0.27	-0.10	0.74	1260	1
var	0.87	0.34	0.40	1.46	1425	1
MODEL 23 (plastic both): latency ~ a+ b*phi + c*lambda						
a	4.99	0.31	4.52	5.46	1183	1
b	0.33	0.27	-0.09	0.76	1736	1
c	-0.01	0.25	-0.41	0.42	1556	1
var	0.83	0.32	0.39	1.42	1321	1
MODEL 24 (plastic interaction): latency ~ a + b*phi*lambda						
a	5.02	0.31	4.51	5.49	886	1
b	0.07	0.21	-0.25	0.42	1256	1
var	0.80	0.30	0.39	1.33	1493	
MODEL 25 (plastic U shaped): latency ~ a+ b*abs(lambda) + c*abs(phi)						
a	3.07	0.52	2.29	3.91	1210	1
b	0.82	0.53	-0.02	1.68	1353	1
c	1.49	0.47	0.76	2.27	1226	1
var	1.27	0.48	0.61	2.12	1456	1
MODEL 26 (plastic phi first): latency ~ a + b*phi						
a	4.97	0.30	4.49	5.44	1105	1
b	0.16	0.26	-0.24	0.60	1376	1
var	0.80	0.30	0.39	1.32	1218	1
MODEL 27 (plastic lambda first): latency ~ a + b*lambda						
a	4.95	0.34	4.40	5.47	1284	1
b	0.20	0.27	-0.53	0.88	1334	1
var	0.80	0.34	0.36	1.41	1614	1
MODEL 28 (wooden phi): latency ~ a + b*phi						
a	5.73	0.28	5.27	6.15	1064	1
b	0.47	0.30	0.00	0.94	1144	1
var	1.06	0.44	0.48	1.86	1364	1
MODEL 29 (wooden lambda): latency ~ a + b*lambda						
a	5.76	0.30	5.28	6.21	1373	1
b	-0.25	0.25	-0.63	0.15	1415	1
var	0.96	0.37	0.35	1.62	1532	1
MODEL 30 (wooden both): latency ~ a+ b*phi + c*lambda						
a	5.72	0.31	4.52	5.46	1183	1
b	-0.29	0.27	-0.09	0.76	1736	1
c	0.47	0.25	-0.41	0.42	1556	1
var	1.07	0.32	0.39	1.42	1321	1
MODEL 31 (wooden interaction): latency ~ a + b*phi*lambda						
a	5.80	0.30	5.31	6.23	1259	1
b	0.15	0.24	-0.22	0.56	1448	1
var	0.92	0.35	0.44	1.54	1342	1
MODEL 32 (wooden U shaped): latency ~ a+ b*abs(lambda) + c*abs(phi)						
a	5.07	0.53	4.20	5.90	739	1
b	0.68	0.59	-0.23	1.68	867	1
c	0.39	0.77	-0.81	1.62	931	1
var	0.78	0.34	0.34	1.42	932	1
MODEL 33 (wooden phi first): latency ~ a + b*phi						
a	5.75	0.30	5.27	6.22	1172	1
b	0.30	0.33	-0.22	0.82	1467	1
var	0.95	0.40	0.43	1.65	1216	1
MODEL 34 (wooden lambda first): latency ~ a + b*lambda						
a	5.76	0.30	5.28	6.21	1250	1
b	-0.21	0.25	-0.60	0.21	1233	1
var	0.94	0.37	0.45	1.59	1537	1

Table 9. Model outputs for the **number of loci solved** on the plastic (models 35-41) and wooden (models 42-48) multi-access boxes in relation to ϕ and λ. SD=standard deviation, the 89% prediction intervals are shown, n_eff=effective sample size, Rhat4=an indicator of model convergence (1 is ideal), b=the slope of the relationship between loci solved or average switch latency and ϕ or λ.

	Mean	SD	5.5%	94.5%	n_eff	Rhat4
MODEL 35 (plastic phi): loci solved ~ a + b*phi						
a	0.02	0.30	-0.45	0.50	1153	1
b	0.24	0.26	-0.16	0.65	1463	1
MODEL 36 (plastic lambda): loci solved ~ a + b*lambda						
a	0.00	0.25	-0.40	0.41	1369	1
b	0.14	0.22	-0.21	0.49	1200	1
MODEL 37 (plastic both): loci solved ~ a+ b*phi + c*lambda						
a	4.99	0.31	4.52	5.46	1183	1
b	0.33	0.27	-0.09	0.76	1736	1
c	-0.01	0.25	-0.41	0.42	1556	1
MODEL 38 (plastic interaction): loci solved ~ a + b*phi*lambda						
a	5.02	0.31	4.51	5.49	886	1
b	0.07	0.21	-0.25	0.42	1256	1
MODEL 39 (plastic U shaped): loci solved ~ a+ b*abs(lambda) + c*abs(phi)						
a	-0.66	0.50	-1.45	0.15	947	1
b	1.51	0.60	0.61	2.48	845	1
c	-0.55	0.58	-1.45	0.37	861	1
MODEL 40 (plastic phi first): loci solved ~ a + b*phi						
a	0.02	0.26	-0.41	0.42	1313	1
b	0.20	0.22	-0.17	0.54	1624	1
MODEL 41 (plastic lambda first): loci solved ~ a + b*lambda						
a	0.01	0.26	-0.41	0.42	1346	1
b	0.29	0.23	-0.08	0.66	1536	1
MODEL 42 (wooden phi): loci solved ~ a + b*phi						
a	1.35	0.34	0.83	1.90	1329	1
b	-0.08	0.27	-0.52	0.37	1268	1
MODEL 43 (wooden lambda): loci solved ~ a + b*lambda						
a	1.34	0.33	0.83	1.87	1566	1
b	0.20	0.27	-0.24	0.63	1444	1
MODEL 44 (wooden both): loci solved ~ a+ b*phi + c*lambda						
a	0.75	0.42	0.07	1.43	1186	1
b	0.37	0.34	-0.18	0.92	1354	1
c	0.56	0.36	-0.01	1.14	1131	1
MODEL 45 (wooden interaction): loci solved ~ a + b*phi*lambda						
a	0.92	0.38	0.34	1.53	966	1
b	0.67	0.32	0.17	1.19	952	1
MODEL 46 (wooden U shaped): loci solved ~ a+ b*abs(lambda) + c*abs(phi)						
a	0.40	0.50	-0.43	1.20	902	1
b	1.52	0.75	0.33	2.70	827	1
c	0.43	0.67	-0.60	1.52	1002	1
MODEL 47 (wooden phi first): loci solved ~ a + b*phi						
a	1.34	0.34	0.82	1.19	1259	1
b	0.05	0.28	-0.37	0.48	1434	1
MODEL 48 (wooden lambda first): loci solved ~ a + b*lambda						
a	1.34	0.33	0.82	1.88	1283	1
b	-0.11	0.27	-0.52	0.32	1111	1

DISCUSSION

The flexibility manipulation worked

Although animal behavior can affect conservation outcomes (Greggor et al., 2016), behavioral manipulations other than predator recognition training have rarely been attempted (Jolly et al., 2018; Moseby et al., 2012; Ross et al., 2019; West et al., 2018; see review in Tetzlaff et al., 2019). Here, we conducted a controlled experiment to evaluate whether serial reversal learning affected behavioral flexibility. We found that the number of trials to reverse decreased with increasing reversal number, and, when examining last reversals, there was a difference between the manipulated and control groups. This indicates that the flexibility manipulation was effective in that it manipulated reversal learning speeds. The post-hoc Bayesian analyses further showed that performance in the last reversal was not linked with how many reversals they needed to reach criterion. Most grackles performed worse in the middle of the manipulation (e.g., reversals 2 through their third to last reversal) before improving and reaching criterion. That we were able to manipulate flexibility is a novel and important contribution because manipulating flexibility, which is thought of as a generalizable cognitive ability, has the potential to change not only the trained behavior, but may also allow trained individuals to change other behaviors related to this general cognitive ability.

The post-hoc Bayesian analyses revealed that the primary component of flexibility that was manipulated was the learning rate (ϕ), which more than doubled between the first and last reversals. The increase in the learning rate might reflect that birds recognize that this is an environment where new information should be prioritized over previously learned associations. In contrast, the rate of deviating from learned preferences (λ) did not correlate with the number of trials to reverse. The decrease in the rate of deviation from the first to the last reversal might indicate that individuals learned a meta-rule about the serial reversal experiment, that this is an environment where information from the last few trials is highly predictive of the reward location and that they should deviate from their previous attractions as soon as the reward changes.

Serial reversals affected performance on both multi-access boxes

While performance differed between the two multi-access boxes, the serial reversal flexibility manipulation did affect flexibility in a new context as well as innovativeness. Grackles that were faster to reverse a preference in their first and last reversals, and those in the manipulated condition, were also faster to attempt to solve a new locus on the **plastic** multi-access box. Similarly, the flexibility manipulation affected innovativeness because grackles in the manipulated condition solved on average 1.2 more loci on the **wooden** multi-access box than those birds in the control condition and there was a positive correlation between the number of loci solved on the **plastic** multi-access box and the number of trials to reverse in the last reversal. That our results were not consistent across first reversal, last reversal, and condition (Figure 4) on the two different multi-access boxes could be due to the small sample sizes because even in the control group there were several individuals who solved their first and only reversal in very few trials. Furthermore, the lack of correlation between the number of trials to reverse in the first reversal and the number of loci solved on either multi-access box indicates that flexibility is not an inherently utilized tool, but one that is shaped by experience. If it was an inherently utilized tool, the variation in the number of trials to complete first reversals would likely have resulted in a correlation with the number of loci solved. The analyses linking ϕ and λ to the performance on the multi-access boxes suggest that birds might also use different strategies to solve a larger number of loci on the multi-access box, either being potentially quicker at discounting the no longer rewarded locus or alternatively being more likely to explore new loci. In addition, it is also possible that performance on the multi-access boxes relies on other cognitive abilities in which individuals may differ. For example, we previously found that grackles who are faster to complete go no-go, an inhibition task, were slower to switch loci on the multi-access boxes (Logan, McCune, et al., 2021). As such, variation in self control may affect performance on flexibility and innovation tasks by decreasing exploratory behaviors.

Repeatability of flexibility and reversal learning strategies

Examining only the manipulated grackles, there was **repeatability of flexibility performance within a context (serial reversal learning with colored tubes), but not across contexts (correlation of reversal learning and solution switching on the multi-access boxes)**. Individuals who were faster at reversing a color preference in reversal 1 were also generally faster at reversing in subsequent reversals. The post-hoc Bayesian analyses replicated this result because manipulated birds exhibited among-individual variation in performance across reversals. Consequently, it is possible to formulate a general rule for determining when the manipulation is complete by using individual performance in reversal 1: the number of trials in the last reversal equaling roughly (trials first reversal)^2 / 200.

While one third of the grackles switched from an exploratory **strategy** (epsilon-decreasing) to an exploitative strategy (epsilon-first) in their last reversal, there was no correlation between either strategy and reversal number, indicating that the grackles did not use a particular strategy earlier or later in their serial reversals. This could suggest that the grackles did not learn the overarching rule that once food is not present in the preferred color's tube, they must switch to preferring the other color. Instead, they may learn each preference change as if it was new.

Why did performance on a touchscreen vary so drastically from a traditional approach?

We assumed that reversal learning performance using **shape on the touchscreen** would directly compare to and be interchangeable with reversal learning performance using colored tubes. However, it quickly became clear that the touchscreen experiment may have been asking a different question compared with the traditional reversal learning approach using physical objects. Unfortunately, we did not have the time to explore what might have caused the differences between the two tests, but we speculate below. We conclude that these two methods, the traditional physical object and the touchscreen, do not measure the same construct in this species and with this reversal learning experiment.

One possible explanation for the difference between the two experiments is that grackles might require more trials to learn to discriminate between shapes than between colors. Shapes are known to require a few more trials for a preference to develop (e.g., Shaw et al., 2015: mean=40 trials color, mean=55 trials shape in toutouwai; Isden et al., 2013: mean=6 trials color, mean=10 trials shape in spotted bowerbirds), however grackles required hundreds more trials to learn shapes, therefore this explanation seems unlikely. Moreover, grackles may not have understood how the touchscreen worked and therefore it was the apparatus that interfered with their performance, yet grackles successfully completed a go no-go inhibition task using the same touchscreen apparatus (Logan, McCune, et al., 2021). The go no-go task similarly used two different white shapes (wavy lines or a heart), but the shapes were presented sequentially rather than simultaneously (as in the reversal touchscreen experiment). Given this difference between the two touchscreen experiments, it is possible that the grackles found touching the screen in the reversal experiment rewarding in and of itself because something happened whenever they made a response. That is, if they touched the correct stimulus, they received food; if they touched the incorrect stimulus, the screen went blank immediately. This is in contrast with the go no-go experiment where the stimulus stayed on the screen for a set amount of time after an incorrect choice. Another potential reason for the difference between performances on the two touchscreen experiments was that making the incorrect choice in the reversal experiment was not costly enough. In the reversal touchscreen experiment, they could get through many trials, receiving some rewards, in a short amount of time. Consequently, there was potentially not enough incentive to learn quickly, thus explaining the differences in learning speeds between the two reversal experiments.

We are not the first group to attempt to transfer a traditional lab or field task to a touchscreen apparatus (e.g., Drayton & Santos, 2014). Despite some of the challenges associated with touchscreen apparatuses, other attempts to transfer tasks to a touchscreen have been more successful (e.g., A. P. Blaisdell & Cook, 2005; Kangas & Bergman, 2017; Sawa et al., 2005). We maintain that touchscreens have the potential to be an incredibly useful tool for studying comparative cognition in some systems (for reviews and methods, see Bussey et al., 2008; Cook et al., 2004; Kangas & Bergman, 2017; Logan, McCune, et al., 2021; Seitz et al., 2021; Wolf et al., 2014).

Conclusion

We demonstrate that it is possible to manipulate flexibility using a paradigm such as reversal learning. This opens up many opportunities to better understand what flexibility is and whether and how it is causally related to other behaviors or forms of cognition. Understanding how flexibility causally relates to other traits will allow researchers to develop robust theory about the mechanisms and functional impact of flexibility, and when to invoke it as a primary driver in a given context, such as a rapid geographic range expansion. Indeed, we are already in the process of testing the latter hypothesis by conducting cross-population research on great-tailed grackles to test whether a population on the range edge is more flexible (Logan CJ et al., 2020). That we were able to manipulate flexibility, which had causal effects on flexible behavior in a different context (multi-access box) as well as a different cognitive ability (innovativeness), demonstrates that flexibility manipulations could be useful in training individuals of other species in how to be more flexible. This could have important implications for threatened and endangered taxa (such as informing the choice of individuals for captive breeding or introduction programs where individuals or their offspring are released into novel areas), as well as for habituating zoo animals or other managed populations to novelty. If such a flexibility manipulation was successful, it could then change their behavior in this and other domains, giving them a better chance of succeeding in human modified environments.

METHODS

Below is our preregistration that received in principle acceptance at PCI Ecology (PDF version)

A. STATE OF THE DATA

This preregistration was written (2017) prior to collecting data. Pilot data on serial reversal learning (using colored tubes) in one grackle was collected January through April 2018, which informed the revision of 1) the criterion to pass serial reversal learning, 2) more accurate language for H1 P1 (each subsequent reversal may not be faster than the previous, however their average reversal speed decreases), 3) the removal of shape reversals from H3a and H3b (to reduce the amount of time each bird is tested), and 4) a new passing criterion for touchscreen serial reversals in H3b. Part way through data collection on reversal learning (using colored tubes) for the first two birds, the criterion for what counts as making a choice was revised (October 2018) and part way through data collection on the first four birds (October 2018; see below for details) the number of trials that birds in the control group receive was revised to make the test battery feasible in the time given.

This preregistration was submitted to PCI Ecology for peer review (July 2018), we received the first round of peer reviews a few days before data collection began (Sep 2018), we revised and resubmitted 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.

B. PARTITIONING THE RESULTS

We may present the different hypotheses in separate papers (Nov 2020: all hypotheses are included in this one post-study article).

D. METHODS

Planned Sample Great-tailed grackles will be caught in the wild in Tempe, Arizona, USA for individual identification (colored leg bands in unique combinations). Some individuals (~32: ~16 in the control group (they receive 1 reversal) and ~16 in the flexibility manipulation (they receive multiple reversals)) will be brought temporarily into aviaries for testing, and then they will be released back to the wild.

Sample size rationale We will test as many birds as we can in the approximately three years at this field site given that the birds only participate in tests in aviaries during the non-breeding season (approximately September through March).

Data collection stopping rule We will stop testing birds once we have completed two full aviary seasons (likely in March 2020) if the sample size is above the minimum suggested boundary based on model simulations (see section "Ability to detect actual effects" below). If the minimum sample size is not met by this point, we will continue testing birds at our next field site (which we move to in the summer of 2020) until we meet the minimum sample size.

Open materials Design files for the plastic multi-access box: 3D printer files and laser cutter files

Testing protocols for all three experiments: colored tube reversal learning, plastic multi-access box, wooden multi-access box, and touchscreen reversal learning

NOTE (Oct 2020): Touchscreen training data and a summary of the training process is detailed in Seitz et al. (2021)

Open data The data are available at the Knowledge Network for Biocomplexity's data repository: https://knb.ecoinformatics.org/view/corina_logan.84.42.

Randomization and counterbalancing H1: Subjects will be randomly assigned to the manipulated or control group. In the reversal learning trials, the rewarded option is pseudorandomized for side (and the option on the left is always placed first). Pseudorandomization consisted of alternating location for the first two trials of a session and then keeping the same color on the same side for at most two consecutive trials thereafter. A list of all 88 unique trial sequences for a 10-trial session, following the pseudorandomization rules, will be generated in advance for experimenters to use during testing (e.g., a randomized trial sequence might look like: LRLLRRLRLR, where L and R refer to the location, left or right, of the rewarded tube). Randomized trial sequences will be assigned randomly to any given 10-trial session using a random number generator (random.org) to generate a number from 1-88.

Blinding of conditions during analysis No blinding is involved in this study.

Dependent variables *P1-P3*

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.

P2 alternative 2: additional analysis: latency and motor diversity

1) Number of trials to attempt a new locus on the multi-access boxes
2) Number of trials to solve (meet criterion) a new locus on the multi-access boxes

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

P4: learning strategies

Proportion of correct choices in a non-overlapping sliding window of 4-trial bins across the total number of trials required to reach the criterion of 17/20 correct choices (as in P1-P3).

Independent variables

P1: reversal speed gets faster with serial reversals

1) Reversal number
2) Batch (random effect because multiple batches included in the analysis). Note: batch is a test cohort, consisting of 8 birds being tested simultaneously
3) ID (random effect because repeated measures on the same individuals)

P2: serial reversals improve rule switching & problem solving

1) Average latency to attempt to solve a new locus after solving a different locus
2) Average latency to solve a new locus after solving a different locus
3) Total number of loci solved
4) Experimental group (manipulated=multiple reversals with color stimuli; control=one reversal plus equalized experience making choices where both are the same color and both contain a reward)
5) Batch (random effect because multiple batches included in the analysis). Note: batch is a test cohort, consisting of 8 birds being tested simultaneously

Note April 2020: we realized that the average latency to solve a new locus after solving a different locus is confounded with the total number of loci solved because the measure of innovation is included in the definition. Therefore, we will remove this independent variable when conducting the analysis so that we are only examining pure measures of flexibility (average latency to attempt to solve) and innovation (total number of loci solved).

P2 alternative 2: additional analysis: latency and motor diversity

1) Number of trials to reverse a preference in the last reversal that individual participated in
2) Motor diversity: the number of different motor actions used when attempting to solve the multi-access boxes
3) ID (random effect because repeated measures on the same individuals)

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

1) Reversal number
2) Condition (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)

P4: serial reversal learning strategy

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

E. ANALYSIS PLAN

We do not plan to **exclude** any data. When **missing data** occur, the existing data for that individual will be included in the analyses for the tests they completed. Analyses will be conducted in R [current version 4.0.3; R Core Team (2017)], using several R packages: Zhu (2021), Hlavac (2018), J. D. Hadfield (2010), Bartoń (2020), McElreath (2020), Stan Development Team (2020), Xie (2019), Ushey et al. (2020), Eddelbuettel & François (2011), Wickham (2016), knitr (Xie, 2013, 2017, 2018), Wickham et al. (2021), Gabry & Češnovar (2021), posterior (Bürkner et al., 2020), cowplot (Wilke, n.d.), bayesplot (Gabry et al., 2019), irr (Gamer et al., 2012), psych (Revelle, 2014, 2017), Lin (2020), DHARMa (Hartig, 2019), lme4 (Bates et al., 2012; Bates et al., 2015). When there is more than one experimenter within a test, experimenter will be added as a random effect to account for potential differences between experimenters in conducting the tests. If there are no differences between models including or excluding experimenter as a random effect, then we will use the model without this random effect for simplicity.

Unregistered analysis: interobserver reliability of dependent variables To determine whether experimenters coded the dependent variables in a repeatable way, hypothesis-blind video coders were first trained in video coding the dependent variable, and then they coded at least 20% of the videos in the reversal (tubes) and multi-access box experiments. We randomly chose a subset of all of the birds who participated in each experiment using random.org:

- Reversal 6/20 grackles (30% with half from the control group): Chalupa, Avocada, Diablo, Fideo, Tomatillo, Adobo
- Multi-access box plastic 3/15 grackles (20%): Habanero, Queso, Chalupa
- Multi-access box log 3/12 grackles (25%): Diablo, Adobo, Yuca

Video coders then analyzed all videos from these birds. The experimenter's data was compared with the video coder data using the intra-class correlation coefficient (ICC) to determine the degree of bias in the regression slope (Hutcheon et al. (2010), using the irr package in R: Gamer et al. (2012)). Note that the data in columns from coders 1 and 2 in the data sheets were aligned based on similar numbers between coders to prevent disagreements near the top of the data sheet from misaligning all subsequent entries.

Interobserver reliability training To pass **interobserver reliability (IOR) training**, video coders needed an ICC score of 0.90 or greater to ensure the instructions were clear and that there was a high degree of agreement across coders (see R code comments for details).

Alexis Breen (compared with experimenter's live coding):

- Multi-access box: correct choice unweighted Cohen's Kappa=0.90 (confidence boundaries=0.77-1.00, n=33 data points)
- Multi-access box: locus solved unweighted Cohen's Kappa=0.90 (confidence boundaries=0.76-1.00, n=33 data points)

Note: Breen was not a hypothesis-blind video coder. She contributed to extensive video coding across the whole project, however, for interobserver reliability analyses, her data were always compared with a hypothesis-blind coder's data.

Anja Becker (compared with experimenter's live coding):

- Reversal: correct choice ICC=1.00 (confidence boundaries=1.00-1.00, n=25 data points)

Tiana Lam (compared with experimenter's live coding):

- Multi-access box: correct choice ICC=0.90 (confidence boundaries=0.77-1.00, n=33 data points)
- Multi-access box: locus solved unweighted Cohen's Kappa=0.95 (confidence boundaries=0.84-1.00, n=33 data points)

Brynna Hood (compared with experimenter's live coding):

- Multi-access log: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)
- Multi-access log: locus solved unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)

Interobserver reliability Interobserver reliability scores (minimum 20% of the videos) were as follows:

Brynna Hood (compared with experimenter's live coding):

- Multi-access log: correct choice unweighted Cohen's Kappa=0.91 (confidence boundaries=0.76-1.00, n=39 data points)
- Multi-access log: locus solved unweighted Cohen's Kappa=1.0 (confidence boundaries=1.0-1.00, n=39 data points)

Tiana Lam (compared with experimenter's live coding):

- Multi-access box: correct choice unweighted Cohen's Kappa=0.83 (confidence boundaries=0.73-0.92, n=102 data points)
- Multi-access box: locus solved unweighted Cohen's Kappa=0.90 (confidence boundaries=0.830-0.97, n=102 data points)

Anja Becker (compared with experimenter's live coding):

- Reversal: correct choice ICC=0.99 (confidence boundaries=0.98-0.99, n=3280 data points)

These scores indicate that the dependent variables are repeatable to a high or extremely high degree given our instructions and training.

Unregistered analyses: Bayesian Flexibility models In addition to the planned analyses, we conducted post-hoc exploratory analyses on the serial reversal learning data to better understand the effect the flexibility manipulation had on performance. We used the version of the Bayesian model that was developed by A. Blaisdell et al. (2021) and modified by Logan CJ et al. (2020see 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 learning rate of attraction to either option (ϕ) and the rate of deviating from learned attractions (λ). We wanted to address the following questions:

1) **What did the manipulation change? Can we determine what mechanisms of flexibility the birds in the manipulated group who were already fast at reversing rely on?** We predicted that birds that were already faster at reversing would have similar deviation rates from the learned attractions between the first and last reversals and lower learning rates than slower birds, which would allow them to change their preference more quickly because the attraction would be weaker and easier to reverse.
2) **Does the manipulation shift birds beyond what is naturally observed and does it make them more similar?** In the analyses in the Results section, it was unclear how there was an effect on innovation and flexibility in the multi-access box experiments when, in some cases, there was no difference between the control and manipulated conditions. Therefore, for both the control and manipulated groups, we investigated whether the learning rate and rate of deviating from learned attractions differed between a bird's first 10 trials of the first and last reversals and whether what we observe among the manipulated birds at the end might already naturally be present in some birds in the control group. In addition, we wanted to know whether the manipulations affected all birds equally or if we could still detect variation.
3) **Are ϕ or λ, the two components of flexibility in reversal learning, associated with performance on the multi-access boxes across control and manipulated birds?** In the analyses in the Results section, we detected some associations between a bird's performance in the reversal learning task and on the multi-access boxes. Examining the two parameters, ϕ and λ, separately might offer a more detailed understanding of potential abilities that might influence performance on the different tasks.

Using simulations to check models estimating potential factors underlying performance in the reversal experiment We first ran the Bayesian model on simulated data to better understand how the two parameters might lead to differences in performance and whether we could detect meaningful differences between control and manipulated birds. The settings for the simulations were based on the previous analysis of data from grackles in a different population (Santa Barbara, A. Blaisdell et al. (2021)). When we used only the choices simulated individuals made during their one reversal, the estimated ϕ and λ values did not match those the individuals had been assigned. We realized that ϕ and λ values were consistently shifted in a correlated way. When estimating these values from only a single reversal, there was equifinality: multiple combinations of the two parameters ϕ and λ could potentially explain the performance of birds during this reversal, and the estimation adjusts both learning parameters towards the mean. However, when we combined data from across at least one switch in the color of the rewarded option, combining initial discrimination learning with the first reversal, the model accurately recovered the ϕ and λ values that simulated what the individuals had been assigned.

In terms of the influence of the two parameters ϕ and λ on the number of trials birds needed to reverse a color preference, the ϕ values assigned to simulated individuals had a stronger influence than the λ values (estimated association of number of trials with standardized values of ϕ: -21, 89% prediction interval (PI):-22 to -19; with standardized values of λ -14, 89% PI: -16 to -13). In particular, low numbers of trials to reverse could be observed across the full range of λ values, though when λ was smaller than 8, simulated birds might need 150 or more trials to reverse a preference (Figure M1). In contrast, there was a more linear relationship between ϕ and the number of trials to reverse, with birds needing fewer trials the larger their ϕ.

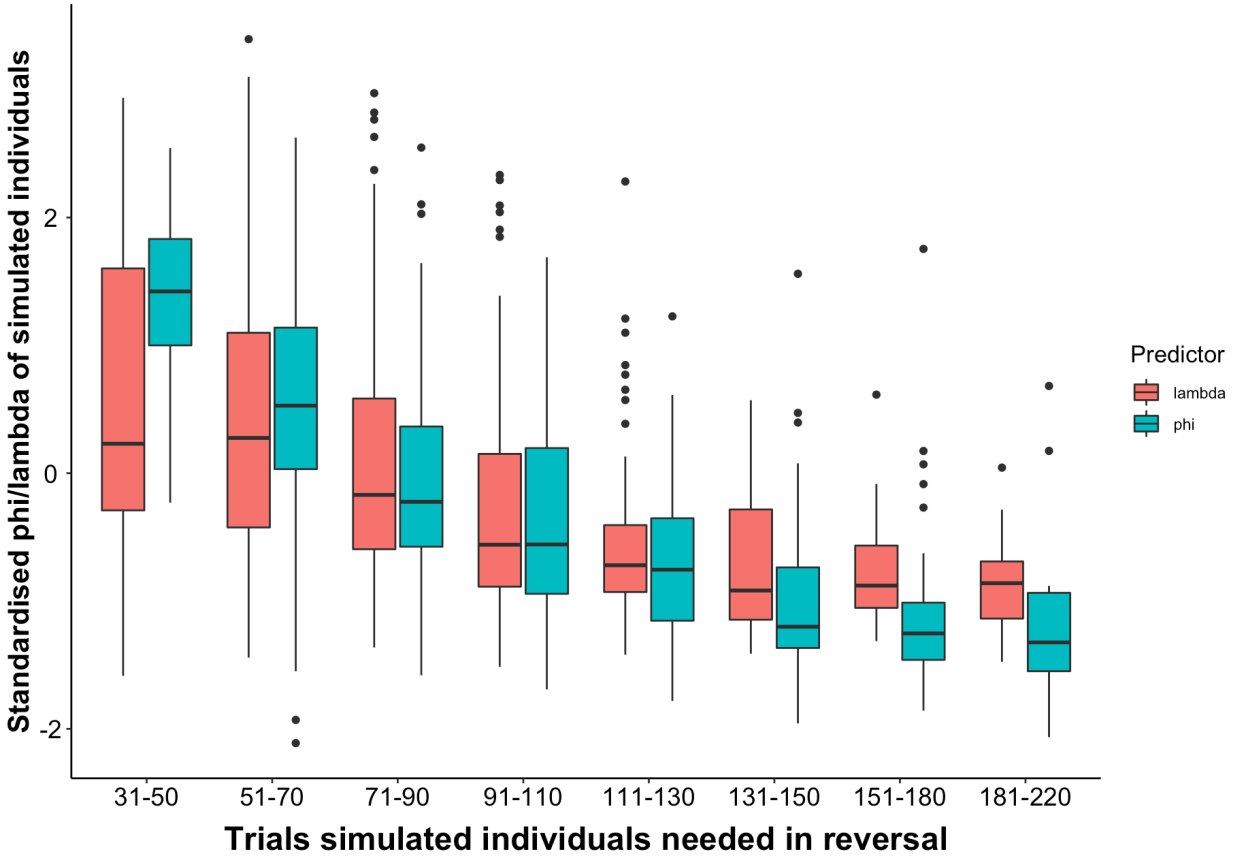


Figure M1. In the simulations, the ϕ values assigned to individuals (green) had a clearer influence on the number of trials these individuals needed to reverse than their λ values (red). ϕ and λ values were standardized for direct comparison. In general, individuals needed fewer trials to reverse if they had larger ϕ and λ values. However, relatively small λ values could be found across the range of reversal performances, whereas there was a more clear distinction with ϕ values.

Ability to detect actual effects To begin to understand what kinds of effect sizes we will be able to detect given our sample size limitations and our interest in decreasing noise by attempting to measure it, which increases the number of explanatory variables, we used G*Power (v.3.1, Faul et al., 2007, 2009) to conduct power analyses based on confidence intervals. G*Power uses pre-set drop down menus and we chose the options that were as close to our analysis methods as possible (listed in each analysis below). Note that there were no explicit options for GLMs (though the chosen test in G*Power appears to align with GLMs) or GLMMs or for the inclusion of the number of trials per bird (which are generally large in our investigation), thus the power analyses are only an approximation of the kinds of effect sizes we can detect. We realize that these power analyses are not fully aligned with our study design and that these kinds of analyses are not appropriate for Bayesian statistics (e.g., our MCMCglmm below), however we are unaware of better options at this time. Additionally, it is difficult to run power analyses because it is unclear what kinds of effect sizes we should expect due to the lack of data on this species for these experiments.

To address the power analysis issues, we will run simulations on our Arizona data set before conducting any analyses in this preregistration. We will first run null models (i.e., dependent variable ~ 1 + random effects), which will allow us to determine what a weak versus a strong effect is for each model. Then we will run simulations based on the null model to explore the boundaries of influences (e.g., sample size) on our ability to detect effects of interest of varying strengths. If simulation results indicate that our Arizona sample size is not larger than the lower boundary, we will continue these experiments at the next field site until we meet the minimum suggested sample size.

SIMULATIONS APRIL 2020 (pre-data analysis): following procedures in McElreath (2018), we first constructed a **hypothesis-appropriate mathematical model** that encompasses the relationship between the variables of interest for each analysis: 1) number of loci solved on the multi-access box ~ trials to reverse, and 2) latency to attempt a new locus on the multi-access box ~ trials to reverse.

Simulation and model: number of loci solved on the multi-access box ~ trials to reverse

The model takes the form of:

locisolved ~ Binomial(4, p) *[likelihood]*

logit(p) ~ α[batch] + βtrials *[model]*

locisolved is the number of loci solved on the multi-access box, 4 is the total number of loci on the multi-access box, p is the probability of solving any one locus across the whole experiment, α is the intercept and each batch gets its own, β is the expected amount of change in locisolved for every one unit change in trials, and trials is the number of trials to reverse a color preference.

Expected values for the number of loci solved on the multi-access box were set to either 2 or 0 (out of 4 loci maximum) because we were unsure of whether the grackles would be able to solve any loci on the multi-access box because this experiment had never been done on this species before. Expected values for reversal learning using colored tubes (mean, standard deviation, and range of number of trials to reverse a color preference) were based on previously published data on great-tailed grackles (Logan, 2016). This data indicates that the average number of trials to reverse a preference is 91 and the standard deviation is 21. In our model, the variation in the actual data is reflected by both the population standard deviation and the expected amount of change related to the explanatory variable. After running simulations, we identified the following distributions and priors to be the most likely for our expected data:

α ~ Normal(4,10) *[α prior]*

β ~ Normal(0,5) *[β prior]*

We used normal distributions for α and β because they are (or are based on) sums with large means (see Figure 10.6 in McElreath, 2018). For the β prior, we had no expectation about whether the relationship would be positive or negative, therefore we centered it on 0 (the mean).

Simulation and model: latency to attempt a new locus on the multi-access box ~ trials to reverse

For the average latency to attempt a new locus on the multi-access box as it relates to trials to reverse (both are measures of flexibility), we simulated data and set the model as follows:

latency ~ gamma-Poisson(λ_i, ϕ) *[likelihood]*

$\log(\lambda_i) \sim \alpha$[batch] + βtrials *[the model]*

latency is the average latency to attempt a new locus on the multi-access box, λ_i is the rate (probability of attempting a locus in each second) per bird (and we take the log of it to make sure it is always positive; birds with a higher rate have a smaller latency), ϕ is the dispersion of the rates across birds, α is the intercept for the rate per batch, β is the expected amount of change in the rate of attempting to solve in any given second for every one unit change in trials, and trials is the number of trials to reverse a color preference.

Expected values for the latency to attempt a new locus on the multi-access box was set to between 1-2700 sec because the experiment ends for a bird if they do not obtain the food in 3 consecutive trials, and each trial can last up to 15 min. Because we did not have prior data for this species on this test, we set the mean to 300 sec, which is half way through a usual 10 min trial because it seems likely that if a bird is going to attempt another locus, it will likely do so at the next opportunity, especially after being successful in the previous trial. Expected values for reversal learning using colored tubes are the same as above. After running simulations, we identified the following to be the most likely distributions and priors for our expected data:

$\phi \sim 1/(\text{Exponential}(1))$ *[ϕ prior]*

$\alpha \sim$ Normal(300,50) *[α prior]*

$\beta \sim$ Normal(0,5) *[β prior]*

We used a gamma-Poisson distribution for latency because it constrains the values to be positive and to primarily occur sooner rather than later, which is what we expect from the grackles (based on data from New Caledonian crows and kea in Auersperg et al., 2011). For ϕ, we used an exponential distribution because it is standard for this paramter. We used normal distributions for α and β because they are (or are based on) sums with large means (see Figure 10.6 in McElreath, 2018). For the β prior, we had no expectation about whether the relationship would be positive or negative, therefore we centered it on 0 (the mean).

We **translated the simulation output into effect sizes** and examined what kind of effect size these parameter values represent (Table M1). For each β, we calculated the effect size (Box 13.3 in Lajeunesse et al., 2013: linear regression):

r = β (SDx / SDy) = β (1.5 / 21)

Where r is the Pearson product moment correlation and SD is the standard deviation. For the standard deviation of x (number of loci solved on the multiacccess box), we estimated a possible value of 1.5. For the standard deviation of y (trials to reverse), we used 21 from the Santa Barbara grackle data (Logan, 2016). We then calculated the effect sizes and R^2 values for each value of β.

Table M1. The connection between β and effect sizes (SDx=standard deviation of x, which is the number of loci solved; SDy=standard deviation of y, which is the number of trials to reverse; R^2=R squared).

We then used the simulations to run **models** on simulated data to estimate the measurement error associated with varying sample size, β, and the range of multi-access box loci solved or latency to attempt a new locus (Table M2). Before running the models, we decided that a model would detect an effect if 89% of the posterior sample was on the same side of zero (following McElreath, 2018). We ran the simulation with β=3 (latency) because this was a high value at which an appropriate range of values were observed in the simulation testing phase, β=0 because this would be the scenario in which there is no relationship between the response variable and the trials to reverse, and β=-1 to determine how small of a difference we can detect and with what amount of associated noise (σ). **Sigma (σ) is the standard deviation in the trials to reverse if the trials to reverse is a normal distribution. In all simulations, the mean in the trials to reverse was set to 91. Therefore, a (σ) of 14 is 15% noise (14/91). We found that when (σ) is larger than 14, we cannot detect even the largest effect of trials to reverse on loci solved or latency because there are some simulations where the estimated regression coefficient crosses zero.** When β=0 we want all of the regression coefficients to cross zero (10 out of 10

random repetitions) and when $\beta \neq 0$ we want none of the regression coefficients to cross zero (0 out of 10 random repetitions). We ran the models several times with various parameters to determine at what point this was the case for each combination of parameters.

Table M2. Simulation outputs from varying β, sample size (n), σ, and whether the actual range of multi-access box [MAB] loci solved were 0-2 or 0-4 **(we did not know how many loci the grackles would be able to solve before we started collecting data so we ran two simulations. The grackles ended up being able to solve all four loci on both multi-access boxes, therefore we must use only those rows associated with "Range of MAB loci solved" = 0-4). This table is useful for the analyses involving the number of loci solved on the multi-access box, but not the latency to switch to attempting a new locus on the multi-access box, which uses a different (gamma poisson) model.**

This shows that we would have the power to detect a medium effect (-0.357 in Table M1) with a sample size of 15 if the noise (σ) is <15%. We would be unlikely to get a false negative because there were no false negatives in the simulations (i.e., the posterior sample range did not cross zero). With this sample size, when β=0, there are no false positives (i.e., the posterior sample range always included zero). However, we would not be able to detect a weak effect unless the noise (σ) was much smaller.

Data checking The data will be checked for overdispersion, underdispersion, zero-inflation, and heteroscedasticity with the DHARMa R package (Hartig, 2019) following methods by Hartig. Note: DHARMa doesn't support MCMCglmm, therefore we will use the closest supported model: glmer from the R package lme4 (Bates et al., 2015).

Determining the threshold: How many reversals are enough? We initially (in 2017) set as the passing criterion: During the data collection period, the number of trials required to reverse a preference will be documented per bird, and reversals will continue until the first batch of birds tested reaches an asymptote (i.e., there are negligible further decreases in the number of trials required to reverse a preference). The number of reversals to reach the asymptote will be the number of reversals that subsequent birds experience.

Due to delays in setting up the field site, we were only able to test two grackles in early 2018 (January through April) and, due to randomization, only one (Fajita) was in the experimental condition that involved undergoing the flexibility manipulation (Empanada was in the control condition). While Fajita's reversal speeds generally improved with increasing serial reversals, she never reached an asymptote (which we defined as passing three consecutive reversals in the same number of trials), even after 38 reversals. These 38 reversals took 2.5 months, which is an impractical amount of time if birds are to participate in the rest of the test battery after undergoing the reversal manipulation (we are permitted to keep them in aviaries for up to three months per bird). Because our objective in this experiment is to manipulate an individual's flexibility, we decided to revise our serial reversal passing criterion to something more species relevant based on Fajita's serial reversal performance and the performance of seven grackles in Santa Barbara who underwent only one reversal in 2014 and 2015 (Logan, 2016). **The revised serial reversal passing criterion is: passing two sessions in a row at or under 50 trials**. 50 trials is fewer trials than any of the nine grackles required to pass their first reversal (range 70-130), therefore it should reflect an improvement in flexibility.

Revising the choice criterion and the criterion to pass the control condition **Choice criterion:** At the beginning of the second bird's initial discrimination in the reversal learning colored tube experiment (October 2018), we revised the criterion for what counts as a choice from A) the bird's head needs to pass an invisible line on the table that ran perpendicular to the the tube opening to B) the bird needs to bend its body or head down to look in the tube. Criterion A resulted in birds making more choices than the number of learning opportunities they were exposed to (because they could not see whether there was food in the tube unless they bent their head down to look in the tube) and appeared to result in slower learning. It is important that one choice equals one learning opportunity, therefore we revised the choice criterion to the latter. Anecdotally, this choice matters because the first three birds in the experiment (Tomatillo, Chalupa,

and Queso) learned faster than the pilot birds (Empanada and Fajita) in their initial discriminations and first reversals. Thus, it was an important change to make at the beginning of the experiment.

Criterion to pass the control condition: Before collecting experimental data, we set the number of trials experienced by the birds in the control group as 1100 because this is how many trials it would have taken the pilot bird in the manipulated group, Fajita, to pass serial reversals 2-17 according to our revised serial reversal passing criterion. However, after 25 and 17 days (after Tomatillo and Queso's first reversals, respectively) of testing the first two individuals in the control group it became apparent that 1100 trials is impractical given the time constraints for how long we are permitted to keep each bird temporarily in captivity and would prevent birds from completing the test battery before their release. Additionally, after revising the choice criterion, it was going to be likely that birds in the manipulated group would require fewer than 1100 trials to meet the serial reversal passing criterion. Therefore, reducing the number of trials control birds experience would result in a better match of experience with birds in the manipulated group. On 2 November 2018 we set the number of trials control birds experience after their first (and only) reversal to the number of trials it requires the first bird in the manipulated group to pass (the first bird has not passed yet, therefore we do not yet know what this number is). After more individuals in the manipulated group pass, we will update this number to the average number of trials to pass. Note on 16 April 2020: this is what we did for all birds in the control condition, except Mofongo who was a slow participator and would not have finished his test battery by the time it got too hot to keep birds in the aviaries if we used the current average number of trials (420). Instead, we matched him with the fastest bird in the manipulated group (Habanero=290 trials) to make it more likely that Mofongo could get through the rest of the test battery in time.

P1: negative relationship between the number of trials to reverse a preference and the number of reversals? **Analysis:** A Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCMCglmm package; J. D. 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) (J. Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; J. D. 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.

We do not need a power analysis to estimate our ability to detect actual effects because, by definition, the individuals that complete this experiment must get faster at reversing in order to be able to pass the stopping criterion (two consecutive reversals in 50 trials or less). According to previous grackle data (from the pilot and from Santa Barbara), the fastest grackle passed their first reversal in 70 trials, which means that passing our serial reversal stopping criterion would require them to have improved their passing speed.

P2: serial reversal improves rule switching and problem solving **Note on 14 May 2020:** Please see our Alternative Analyses section where we describe that we will conduct this analysis as in the new models in the Ability to detect actual effects section, which will replace the analysis listed below.

Analysis: Because the independent variables could influence each other, we will analyze them in a single model. A Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCMCglmm package; J. D. 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) (J. Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; J. D. 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 $f^2 = 0.41$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 5

Output:

Noncentrality parameter = 13.1200000

Critical F = 2.5867901

Numerator df = 5

Denominator df = 26

Total sample size = 32

Actual power = 0.7103096

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).

We will first determine whether the total loci solved, the latency to solve or attempt at new loci are correlated across the two distinct multi-access boxes. If there is a positive correlation, then we will only use the variables for the plastic multi-access box (for which we will likely have more data), as presented below. If there is no correlation, we will incorporate the total loci solved, the latencies to solve and attempt at new loci for each of the multi-access boxes as independent variables in our model.

P2 alternative 2: additional analysis: latency and motor diversity A Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCMCglmm package; J. D. 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) (J. Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; J. D. 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 $f^2 = 0.27$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 2

Output:

Noncentrality parameter = 8.6400000

Critical F = 3.3276545

Numerator df = 2

Denominator df = 29

Total sample size = 32

Actual power = 0.7047420

This means that, with our sample size of 32, we have a 70% chance of detecting a medium (approximated at f^2=0.15 by Cohen, 1988) to large effect (approximated at f^2=0.35 by Cohen, 1988).

We will perform separate models for each multi-access box (plastic and wooden).

NOTE (Aug 2021): when attempting to run the below model, we realized the model has to be a GLM and not a GLMM because there is only one data point per bird, so we changed this accordingly.

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 (J. D. 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; J. Hadfield (2014)]. We will ensure the GLMM shows acceptable convergence [i.e., lag time autocorrelation values <0.01; J. D. 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; J. D. 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) (J. Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; J. D. 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? 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; (J. D. 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) (J. Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; J. D. 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).

P4: learning strategies (for birds in the manipulated group only) **Analysis:** Learning strategies will be identified by matching them to the two known approximate strategies of the contextual, binary multi-armed bandit: epsilon-first and epsilon-decreasing (McInerney, 2010; as in Logan, 2016).

From Logan (2016) (emphasis added):

> The following equations refer to the different phases involved in each strategy:
>
> Equation 1 (exploration phase):
>
> $$\epsilon N$$
>
> Equation 2 (exploitation phase):
>
> $$(1-\epsilon)N$$
>
> N is the number of trials given, and epsilon,
>
> $$\epsilon$$
>
> , represents the subject's uncertainty about the location of the reward, starting at complete uncertainty ($\epsilon = 1$) at the beginning of the experiment and decreasing rapidly as individuals gain experience with the task (exploration phase where the rewarded [option] is chosen below or at chance levels) and switch to the exploitative phase (the rewarded [option] is chosen significantly above chance levels). Because the [subjects] needed to learn the rules of the task, they necessarily had an exploration phase. The **epsilon-first strategy** involves an exploration phase followed by an entirely exploitative phase. The optimal strategy overall would be to explore one color in

the first trial and the other color in the second trial, and then switch to an exploitative strategy (choose the rewarded [option] significantly above chance levels). In this case there would be no pattern [in the learning curve] in the choices [during] the exploration phase because it would consist of sampling each [option] only once. In the **epsilon-decreasing strategy**, subjects would start by making some incorrect choices and then increase their choice of the rewarded [option] gradually as their uncertainty decreases until they choose the rewarded [option] significantly above chance levels. In this case, a linear pattern emerges [in the learning curve] during the exploration phase.

We will then quantitatively determine to what degree each bird used the exploration versus exploitation strategy using methods in (Federspiel et al., 2017) by calculating the number of 20-trial blocks where birds were choosing "randomly" (6-14 correct choices; called sampling blocks; akin to the exploration phase in our preregistration) was divided by the total number of blocks to reach criterion per bird. This ratio was also calculated for "acquisition" blocks where birds made primarily correct choices (15-20 correct choices; akin to the exploitation phase in our preregistration). These ratios, calculated for each bird for their serial reversals, quantitatively discern the exploration from the exploitation phases.

NOTE (Aug 2021): the grackles were tested in 10-trial blocks and not 20-trial blocks as in Federspiel et al. (2017), which would mean that if there were <20 trials in the last block of a reversal, they would be omitted from the analysis. Therefore, we changed the block size to 10 trials and adjusted the sampling blocks to 2-9 correct choices, and the acquisition blocks to 9-10 correct choices using significance levels in the binomial test as did Federspiel et al. (2017).

Alternative Analyses We anticipate that we will want to run additional/different analyses after reading McElreath (2016). We will revise this preregistration to include these new analyses before conducting the analyses above.

14 May 2020: After reading McElreath (2018) and taking McElreath's stats course, we changed a couple of things about the analysis plan in this preregistration (before we analyzed any of our data). These are the changes we made:

1) Ability to detect actual effects: We added two simulations and hypothesis-specific models for P2. One examines the relationship between the number of loci solved on the multi-access box and the number of trials to reverse a preference. The other examines the latency to attempt another locus on the multi-access box and the number of trials to reverse a preference.

2) P2: serial reversal improves rule switching and problem solving: In conducting point 1, we realized that we had misinterpreted which variable should be the response variable in this analysis. We originally set the number of trials to reverse as the response variable, however we should have instead set the number of loci solved as the response variable and then planned to conduct a second model with the latency to attempt a new locus as the response variable and number of trials as the explanatory variable. This is because a) we manipulated the number of trials to reverse, therefore it must be the explanatory variable; and b) they should be split into two models because of a and because these are two very different relationships that should be considered in their own models. We also realized that Condition (manipulated or control) does not need to be a variable in any of our models because the manipulated birds have, by definition, faster reversal speeds. For these reasons, when we conduct the P2 analysis in this preregistration, we will use the custom models we made in point 1 above rather than the planned MCMCglmm model.

F. 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])

G. AUTHOR CONTRIBUTIONS

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

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.

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

H. 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).

I. 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 on the Managing Board at PCI Ecology.

J. 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.

K. REFERENCES

Auersperg, A. M. I., Bayern, A. M. P. von, Gajdon, G. K., Huber, L., & Kacelnik, A. (2011). Flexibility in problem solving and tool use of kea and New Caledonian crows in a multi access box paradigm. *PLOS ONE*, *6*(6), e20231. https://doi.org/10.1371/journal.pone.0020231

Bartoń, K. (2020). *MuMIn: Multi-model inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes (2011). R package version 0.999375-42*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bergstrom, C. T., & Lachmann, M. (2004). Shannon information and biological fitness. *Information Theory Workshop, 2004. IEEE*, 50–54.

Blaisdell, A. P., & Cook, R. G. (2005). Integration of spatial maps in pigeons. *Animal Cognition*, *8*(1), 7–16.

Blaisdell, A., Seitz, B., Rowney, C., Folsom, M., MacPherson, M., Deffner, D., & Logan, C. J. (2021). *Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles (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

Bussey, T. J., Padain, T. L., Skillings, E. A., Winters, B. D., Morton, A. J., & Saksida, L. M. (2008). The touchscreen cognitive testing method for rodents: How to get the best out of your rat. *Learning & Memory*, *15*(7), 516–523.

Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2020). Posterior: Tools for working with posterior distributions. *Earthquake Spectra, R Package Version 0.1*, *3*.

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

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

Cook, R. G., Geller, A. I., Zhang, G.-R., & Gowda, R. (2004). Touchscreen-enhanced visual learning in rats. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 101–106.

Diquelou, M. C., Griffin, A. S., & Sol, D. (2015). *The role of motor diversity in foraging innovations: A cross-species comparison in urban birds*.

Drayton, L. A., & Santos, L. R. (2014). Insights into intraspecies variation in primate prosocial behavior: Capuchins (cebus apella) fail to show prosociality on a touchscreen task. *Behavioral Sciences*, *4*(2), 87–101.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'CmdStan'*.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package 'irr.' *Various Coefficients of Interrater Reliability and Agreement*.

Greggor, A. L., Berger-Tal, O., Blumstein, D. T., Angeloni, L., Bessa-Gomes, C., Blackwell, B. F., St Clair, C. C., Crooks, K., Silva, S. de, Fernández-Juricic, E., & others. (2016). Research priorities from animal behaviour for maximising conservation progress. *Trends in Ecology & Evolution*, *31*(12), 953–964.

Griffin, A. S., & Guez, D. (2014). Innovation and problem solving: A review of common mechanisms. *Behavioural Processes*, *109*, 121–134. https://doi.org/10.1016/j.beproc.2014.08.027

Griffin, A. S., Guez, D., Federspiel, I., Diquelou, M., & Lermite, F. (2016). Invading new environments: A mechanistic framework linking motor diversity and cognition to establishment success. *Biological Invasions and Animal Behaviour*, 26e46.

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

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

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22. http://www.jstatsoft.org/v33/i02/

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

Hlavac, M. (2018). *Stargazer: Well-formatted regression and summary statistics tables*. Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer

Homberg, J. R., Pattij, T., Janssen, M. C., Ronken, E., De Boer, S. F., Schoffelmeer, A. N., & Cuppen, E. (2007). Serotonin transporter deficiency in rats improves inhibitory control but not behavioural flexibility. *European Journal of Neuroscience*, *26*(7), 2066–2073.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *Bmj*, *340*, c2289. https://doi.org/10.1136/bmj.c2289

Isden, J., Panayi, C., Dingle, C., & Madden, J. (2013). Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Animal Behaviour*, *86*(4), 829–838.

Jolly, C. J., Kelly, E., Gillespie, G. R., Phillips, B., & Webb, J. K. (2018). Out of the frying pan: Reintroduction of toad-smart northern quolls to southern kakadu national park. *Austral Ecology*, *43*(2), 139–149.

Kangas, B. D., & Bergman, J. (2017). Touchscreen technology in the study of cognition-related behavior. *Behavioural Pharmacology*, *28*(8), 623. https://doi.org/10.1097/FBP.0000000000000356

Lajeunesse, M. J., Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). Recovering missing or partial data from studies: A survey of conversions and imputations for meta-analysis. *Handbook of Meta-Analysis in Ecology and Evolution*, 195–206.

Lefebvre, L., Whittle, P., Lascaris, E., & Finkelstein, A. (1997). Feeding innovations and forebrain size in birds. *Animal Behaviour*, *53*(3), 549–560. https://doi.org/10.1006/anbe.1996.0330

Lin, G. (2020). *Reactable: Interactive data tables based on 'react table'*. https://CRAN.R-project.org/package=reactable

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

Logan, C. J. (2016). Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ*, *4*, e2215.

Logan, C. J., Avin, S., Boogert, N., Buskell, A., Cross, F. R., Currie, A., Jelbert, S., Lukas, D., Mares, R., Navarrete, A. F., & others. (2018). Beyond brain size: Uncovering the neural correlates of behavioral and cognitive specialization. *Comparative Cognition & Behavior Reviews*.

Logan, C. J., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & McCune, K. B. (2021). Reversal learning and multi-access box data for great-tailed grackles. *Knowledge Network for Biocomplexity*, *Data package*.

Logan, C. J., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & McCune, K. B. (2021). Reversal learning and multi-access box data for great-tailed grackles. *Knowledge Network for Biocomplexity*, *Data package*.

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, 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

Manrique, H. M., Völter, C. J., & Call, J. (2013). Repeated innovation in great apes. *Animal Behaviour*, *85*(1), 195–202. https://doi.org/10.1016/j.anbehav.2012.10.026

McCune, KB, MacPherson, M, Rowney, C, Bergeron, L, Folsom, M, & Logan, C. (2019). Is behavioral flexibility linked with exploration, but not boldness, persistence, or motor diversity? *In Principle Acceptance by PCI Ecology of the Version on 27 Mar 2019*. http://corinalogan.com/Preregistrations/g_exploration.html

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC Press. https://doi.org/10.1201/9781315372495

McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

McElreath, R. (2020). *Rethinking: Statistical rethinking book package*.

McInerney, R. E. (2010). Multi-armed bandit bayesian decision making. *Univ. Oxford, Oxford, Tech. Rep.*

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

Moseby, K. E., Cameron, A., & Crisp, H. A. (2012). Can predator avoidance training improve reintroduction outcomes for the greater bilby in arid australia? *Animal Behaviour*, *83*(4), 1011–1021.

O'Hara, M., Huber, L., & Gajdon, G. K. (2015). The advantage of objects over images in discrimination and reversal learning by kea, nestor notabilis. *Animal Behaviour*, *101*, 51–60.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Revelle, W. (2014). Psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, Illinois*, *165*, 1–10.

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Ross, A. K., Letnic, M., Blumstein, D. T., & Moseby, K. E. (2019). Reversing the effects of evolutionary prey naiveté through controlled predator exposure. *Journal of Applied Ecology*, *56*(7), 1761–1769.

Sawa, K., Leising, K. J., & Blaisdell, A. P. (2005). Sensory preconditioning in spatial learning using a touch screen task in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(3), 368.

Seitz, B. M., McCune, K., MacPherson, M., Bergeron, L., Blaisdell, A. P., & Logan, C. J. (2021). Using touchscreen equipped operant chambers to study animal cognition. Benefits, limitations, and advice. *PloS One*, *16*(2), e0246446.

Shaw, R. C., Boogert, N. J., Clayton, N. S., & Burns, K. C. (2015). Wild psychometrics: Evidence for 'general'cognitive performance in wild new zealand robins, petroica longipes. *Animal Behaviour*, *109*, 101–111.

Sol, D., Duncan, R. P., Blackburn, T. M., Cassey, P., & Lefebvre, L. (2005). Big brains, enhanced cognition, and response of birds to novel environments. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5460–5465. https://doi.org/10.1073/pnas.0408145102

Sol, D., & Lefebvre, L. (2000). Behavioural flexibility predicts invasion success in birds introduced to new zealand. *Oikos*, *90*(3), 599–605. https://doi.org/10.1034/j.1600-0706.2000.900317.x

Sol, D., Székely, T., Liker, A., & Lefebvre, L. (2007). Big-brained birds survive better in nature. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1611), 763–769.

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

Stan Development Team. (2020). *RStan: The R interface to Stan*. http://mc-stan.org/

Tetzlaff, S. J., Sperry, J. H., & DeGregorio, B. A. (2019). Effects of antipredator training, environmental enrichment, and soft release on wildlife translocations: A review and meta-analysis. *Biological Conservation*, *236*, 324–331.

Ushey, K., Allaire, J., Wickham, H., & Ritchie, G. (2020). *Rstudioapi: Safely access the RStudio API*. https://CRAN.R-project.org/package=rstudioapi

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

West, R., Letnic, M., Blumstein, D. T., & Moseby, K. E. (2018). Predator exposure improves anti-predator responses in a threatened mammal. *Journal of Applied Ecology*, *55*(1), 147–156.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wilke, C. (n.d.). Cowplot: Streamlined plot theme and plot annotations for 'ggplot2.' R package version 0.9. 2; 2017. *URL Https://CRAN. R-Project. Org/Package= Cowplot*.

Wolf, J. E., Urbano, C. M., Ruprecht, C. M., & Leising, K. J. (2014). Need to train your rat? There is an app for that: A touchscreen behavioral evaluation system. *Behavior Research Methods*, *46*(1), 206–214.

Xie, Y. (2013). Knitr: A general-purpose package for dynamic report generation in r. *R Package Version*, *1*(7).

Xie, Y. (2017). *Dynamic documents with r and knitr*. Chapman; Hall/CRC.

Xie, Y. (2018). Knitr: A comprehensive tool for reproducible research in r. In *Implementing reproducible research* (pp. 3–31). Chapman; Hall/CRC.

Xie, Y. (2019). *formatR: Format r code automatically*. https://CRAN.R-project.org/package=formatR

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax.* https://CRAN.R-project.org/package=kableExtra