Animal Behavior and Cognition
ABC 2022, 9(1):14-36
DOI: https://doi.org/10.26451/abc.09.01.03.2022

Are the More Flexible Great-Tailed Grackles Also Better at Behavioral Inhibition?

Corina J. Logan[1,*], Kelsey B. McCune[2], Maggie MacPherson[2], Zoe Johnson-Ulrich[2], Carolyn Rowney[1], Benjamin Seitz[3], Aaron P. Blaisdell[3], Dominik Deffner[1], and Claudia A. F. Wascher[4]

[1]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology
[2]Institute for Social, Behavioral and Economic Research, University of California Santa Barbara
[3]Department of Psychology, University of California Los Angeles
[4]School of Life Sciences, Anglia Ruskin University

*Corresponding author (Email: corina_logan@eva.mpg.de)

Citation – Logan, C. J., McCune, K. B., MacPherson, M., Johnson-Ulrich, Z., Rowney, C., Seitz, B., Blaisdell, A. P., Deffner, D., Wascher, C. A. F. (2022). Are the more flexible individuals also better at inhibition? *Animal Behavior and Cognition, 9*(1), 14-36. https://doi.org/10.26451/abc.09.01.03.2022

Abstract – Behavioral flexibility should, theoretically, be positively related to behavioral inhibition because one should need to inhibit a previously learned behavior to change their behavior when the task changes (flexibility). However, several investigations show no or mixed support of this hypothesis, which challenges the assumption that inhibition is involved in making flexible decisions. We tested the hypothesis that flexibility (reversal learning and solution switching on a multi-access box by Logan et al., 2022) is associated with inhibition (go/no go on a touchscreen and detour) by measuring all variables in the same individuals. Because touchscreen experiments had never been conducted in this species, we validated that they are functional for wild-caught grackles who learned to use it and completed go/no go on it. Performance on go/no go and detour tasks did not correlate, indicating they did not measure the same trait. Individuals who were faster to reverse took more time to attempt a new option on the multi-access box and were either faster or slower at go/no go depending on whether one individual, Taquito (accidentally tested beyond 200 trial cap), was included in the GLM. While the relationship between trials to reverse and trials to pass go/no go was strongly influenced by Taquito, the more comprehensive Bayesian flexibility model supported the positive relationship irrespective of whether Taquito was included. Performance on detour did not correlate with either flexibility measure, suggesting that they may measure separate traits. We conclude that flexibility is associated with certain types of inhibition, but not others, in great-tailed grackles.

Keywords – Behavioral flexibility, Behavioral inhibition, Go/no go, Detour, Birds, Grackles

Individuals who are more behaviorally flexible (the ability to change behaviors in response to a changing environment, Mikhalevich et al., 2017) are assumed to also be better at inhibiting a prepotent response (Ghahremani et al., 2009; Griffin & Guez, 2014; Liu et al., 2016; Manrique et al., 2013). This is because one should need to inhibit a previously learned behavior to change their behavior when the task changes. However, there is mixed support for the hypothesis that behavioral flexibility (hereafter, flexibility) and behavioral inhibition (hereafter, inhibition) are linked. Many investigations found no correlation between reversal learning (a measure of flexibility) and detour performance (a measure of inhibition) (Boogert et al., 2011; Brucks et al., 2017; Damerius et al., 2017; DuBois et al., 2018; Ducatez et al., 2019; Shaw et al., 2015), while others found mixed support that varied by species and experimental design (Deaner et al., 2006). Investigations using other measures of flexibility and inhibition have also failed to find a connection between the two (Johnson-Ulrich et al., 2018), and even between different

measures of inhibition (e.g., Bray et al., 2014; Fagnani et al., 2016). Further, causal evidence directly challenges the assumption that flexibility requires inhibition. For example, Homberg et al. (2007) showed that rats with improved inhibition (due to gene knockouts) did not perform better in a reversal learning experiment than non-knockout rats. Additionally, Ghahremani et al. (2009) found in humans that brain regions that are active during reversal learning are different from those that are active when someone inhibits a prepotent learned association. These results indicate that inhibition and flexibility are separate traits. The mixed support for a relationship between detour performance and reversal learning makes it difficult to determine whether inhibition is unrelated to flexibility or whether the detour or reversal learning tasks are instead inappropriate for some species.

It is important to use multiple experimental assays to validate that performance on a task reflects an inherent trait (Carter et al., 2013). We aimed to determine whether great-tailed grackles (*Quiscalus mexicanus*) that are better at inhibiting behavioral responses in three experiments (go/no go, detour, delay of gratification) are also more flexible (measured as reversal learning of a color preference, and the latency to attempt a new solution on a multi-access puzzle box by Logan et al., 2022). The go/no go experiment consisted of two different shapes sequentially presented on a touchscreen where one shape must be pecked to receive a food reward (automatically provided by a food hopper under the screen) and the other shape must not be pecked (indicating more inhibitory control) or there will be a penalty of a longer intertrial interval (indicating less inhibitory control). In the detour task, individuals are assessed on their ability to inhibit the motor impulse to try to reach a reward through the long side of a transparent cylinder (indicating less inhibitory control), and instead to detour and take the reward from an open end (indicating more inhibitory control) (Kabadayi et al., 2018; methods as in MacLean et al., 2014, who call it the "cylinder task"). We originally planned to conduct a delay of gratification task, where grackles must wait longer (indicating more inhibitory control) for higher quality (more preferred) food or for higher quantities (methods as in Hillemann et al., 2014); however, the grackles never habituated to the apparatuses and we were not able to conduct this experiment (we omit the details of our plans for the delay of gratification experiment from the rest of this article). The reversal learning of a color preference task involved one reversal (half the birds) or serial reversals (to increase flexibility; half the birds) of a light gray and a dark gray colored tube, one of which contained a food reward (the experiments and data are in Logan et al., 2022). Those grackles that were faster to reverse were more flexible. The multi-access box experimental paradigm was modeled after Auersperg et al. (2011) and consisted of four different access options to obtain food where each option required a different type of action to solve it (the experiments and data are in Logan et al., 2022). Once a grackle passed criterion for demonstrating proficiency in solving an option, that option became non-functional in all future trials. The measure of flexibility was the latency to switch to attempting a new option after a proficient option becomes non-functional, with shorter latencies indicating more flexibility. Employing several experimental assays to measure flexibility and inhibition supports a rigorous approach to testing whether the two traits are linked.

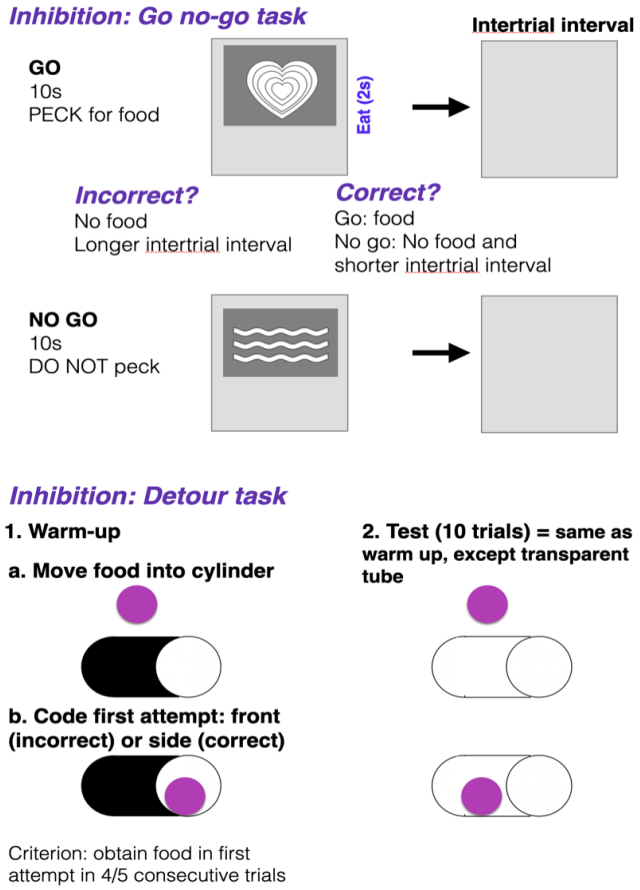
This investigation adds to current knowledge of inhibition and flexibility in several ways. First, our results indicate whether inhibition and flexibility are related and whether tests of inhibition measure the same trait in great-tailed grackles. In addition, touchscreen experiments had never been conducted in this species before, and it was one of our goals to validate whether this setup is viable for running an inhibition task on wild-caught adult grackles. Furthermore, when experimenters test subjects on a series of behavioral tasks, learning from previous tasks can carry over to affect performance on the focal task. Indeed, van Horik et al. (2018) found that previous experience with transparent materials influenced detour performance, while Isaksson et al. (2018) found no effect. Therefore, we also aimed to examine whether the extensive experience of obtaining food from tubes in the reversal learning experiment had an influence on a subject's detour performance, which also involves a tube with food in it.

We hypothesized that, if flexibility requires inhibition, then those individuals that were faster at reversing and switching to a different option after one became non-functional on a multi-access box, would also be better at inhibiting their responses in go/no go and detour tasks (Figure 1). We made the specific prediction (P1) that individuals that were faster to reverse preferences on a reversal learning task, and who also had lower latencies to successfully solve new loci after previously solved loci become unavailable

(multi-access puzzle box) (see Logan et al., 2022), would perform better in the go/no go task (methods similar to Harding et al., 2004) and in the detour task (methods as in MacLean et al., 2014). If there is no correlation between flexibility measures and performance on the inhibition tasks, we predict (P1 alternative 1) that this may indicate that the flexibility tasks may not require much inhibition (particularly if the inhibition results are reliable - see P1 alternative 2). If there is no correlation between flexibility measures and performance on the inhibition tasks, we predict (P1 alternative 2) that this may indicate that the inhibition tasks had low reliability and were therefore too noisy to correlate with flexibility.

Figure 1

The Experimental Designs of the Inhibition Tasks: Go/No Go and Detour



Note. In the go/no go task, the birds were presented with either wavy lines or a heart (the rewarded shape was counterbalanced across birds). Pecking the rewarded shape (the heart in the figure) resulted in the food hopper rising so the bird could eat for approximately 2-3 s. After a trial ended, the screen went blank for a variable number of seconds, depending on whether the individual gave a correct (~3 s) or incorrect (~10 s) response, before starting over again. If the bird failed to refrain from pecking the stimulus during the no go trials, then it was given a longer intertrial interval. In the detour task, individuals first received a warm up with an opaque tube where they learn that the experimenter will show them a piece of food and then move that piece of food into the tube. Subjects then had the opportunity to approach the tube and eat the food. A correct response was when their first approach was to go to the side of the tube to the opening to obtain the food and an incorrect response is when they try to access the food by pecking at the front of the tube (which has no opening). Once they passed the warm up, by solving correctly in 4 out of 5 consecutive trials, they moved on to the test, which used the same setup of tube and food except the tube was transparent. The idea was that being able to see the food through the tube wall might entice them to try to go through the wall rather than refrain from a direct approach to the food and instead go around the side through the tube opening.

If there is no correlation in performance across inhibition tasks, we predicted (P2) that it may indicate that one or more of these tasks does not measure inhibition, or that they measure different types of

inhibition (see Friedman & Miyake, 2004). If go/no go task performance strongly correlates with performance on the delayed gratification task, we predicted (P2 alternative) that this indicates these two tasks measure the same trait, which therefore validates a inhibition task using a touchscreen (the go/no go task).

If individuals perform well on the detour task and with little individual variation, we predicted (P3) that this is potentially because they will have had extensive experience looking into the sides of opaque tubes during reversal learning. To determine whether prior experience with opaque tubes in reversal learning contributed to their detour performance, a subset of individuals experienced the detour task before any reversal learning tests. If this subset performed the same as the others, then previous experience with tubes does not influence detour task performance. If the subset performed worse than the others, this indicates that detour task performance depends on the previous experiences of the individuals tested.

Method

The testing protocols are available for the inhibition and flexibility experiments, the analysis code is available at https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_inhibition.Rmd, and the data are available at: doi:10.5063/M043S3.

Ethics Statement

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

Planned Sample

Great-tailed grackles were caught in the wild in Tempe, Arizona, USA, for individual identification (colored leg bands in unique combinations). Some individuals (~32) were brought temporarily into aviaries for testing, and then were released back to the wild. Grackles were individually housed in an aviary (each 244 cm long by 122 cm wide by 213 cm tall) at Arizona State University for a maximum of three months where they had ad lib access to water at all times and were fed Mazuri Small Bird maintenance diet *ad lib* during non-testing hours (minimum 20 hrs per day), and various other food items (e.g., peanuts, grapes, bread) during testing (up to 3 hrs per day per bird). Individuals were given three to four days to habituate to the aviaries and then their test battery began on the fourth or fifth day (birds were usually tested six days per week, therefore if their fourth day in the aviaries occurred on a day off, then they were tested on the fifth day instead).

Sample Size Rationale

We tested as many birds as we could in the approximately three years at this field site given that the birds only participated in tests in aviaries during the non-breeding season (approximately September through March). The minimum sample size was set at 16 (8 per experiment). We tested 8 grackles in experiment 1 and did not conduct experiment 2 because they did not show evidence of causal cognition (as preplanned in the preregistration).

Randomization and Counterbalancing

Two individuals from each batch experienced the detour task before participating in the reversal learning experiment. These individuals were randomly selected using the random number generator at https://www.random.org. For the rest of the individuals (n = 6 per batch), the order of the inhibition tasks was counterbalanced across birds (using https://www.random.org). Half of the individuals experienced go/no go and then detour, and the other half the opposite order. Go and no go trials were presented randomly with the restriction that no more than four of the same trial type occurred in a row. The rewarded shape was counterbalanced across birds. In the detour experiment, the side from which the apparatus was baited was consistent within subjects, but counterbalanced across subjects. Data collected for interobserver reliability analyses were collected by hypothesis-blind video coders.

Ability to Detect Actual Effects

To begin to understand what kinds of effect sizes we can 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 see 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 Generalized Linear Models (GLM) (though the chosen test in G*Power appears to align with GLMs) or Generalized Linear Mixed Models (GLMM) 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.

Analyses

Analyses were conducted in R (R Core Team, 2017) and one model was run per dependent variable. The data were visually checked to determine whether they were normally distributed via two methods: 1) normality is indicated when the histograms of actual data match those with simulated data, and 2) normality is indicated when the residuals closely fit the dotted line in the Normal Q-Q plot (Zuur et al., 2009). If the data did not appear normally distributed, we visually checked the residuals. If they were patternless, then we assumed a normal distribution (Zuur et al., 2009). The detour data looked normal, the go/no go data were questionable, and both had patternless residuals; therefore, we presumed normality for both variables.

Go/no go experiment. We measured the number of trials to reach two passing criteria (85% correct, and 100% correct before trial 150 or 85% correct after trial 150) where correct responses involved pecking when the rewarded stimulus was displayed and not pecking when the unrewarded stimulus was displayed. Incorrect responses involved pecking when the unrewarded stimulus was displayed and not pecking when the rewarded stimulus was displayed. We ran one GLM (glm function, stats package) per criterion, with the explanatory variable being the number of trials to past first or last reversal, or the average latency to attempt a new option on the plastic multi-access box experiment or the log multi-access box experiment. We used a Poisson distribution and a log link. To determine our ability to detect actual effects, we ran a power analysis in G*Power with the following settings: test family = F tests, linear multiple regression: Fixed model (R^2 deviation from zero), *a priori* power analysis, 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). We found that with a sample size of 32, we would 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).

In addition to the number of trials it took birds to reverse a preference, we also used a more mechanistic multilevel Bayesian reinforcement learning model that takes into account all choices in the reversal learning experiment (see Blaisdell et al. (2021) for details and model validation). From trial to trial, the model updates the latent values of different options and uses those *attractions* to explain observed choices. For each bird j, we estimated a set of two different parameters. The *learning or updating rate* ϕ_j describes the weight of recent experience, the higher the value of ϕ_j, the faster the bird updates their attraction. This corresponds to the first and third connotation of behavioral flexibility as defined by (Bond et al., 2007), the ability to rapidly and adaptively change behavior in light of new experiences. The *random choice rate* λ_j controls how sensitive choices are to differences in attraction scores. As λ_j gets larger, choices become more deterministic, as it gets smaller, choices become more exploratory (random choice if $\lambda_j = 0$). This closely corresponds to the second connotation of internally generated behavioral variation, exploration or creativity (Bond et al., 2007). To account for potential differences between experimenters, we also included experimenter ID as a random effect.

This analysis yields posterior distributions for ϕ_j and λ_j for each individual bird. To use these estimates in a GLM that predicts their inhibition score, we propagated the full *uncertainty* from the reinforcement learning model by directly passing the variables to the linear model within a single large *stan* model. We included both parameters (ϕ_j and λ_j) as predictors and estimate their respective independent effect on the number of trials to pass criterion in go/no go as well as an interaction term. To model the number of trials to pass criterion, we used a Poisson likelihood and a standard log link function as appropriate for count data with an unknown maximum.

In the go/no go experiment, we also analyzed whether the latency (seconds) to first peck on the touchscreen (response variable) was associated with a correct response across trials (explanatory variables), and using bird ID as a random variable. A GLMM (MCMCglmm function, MCMCglmm package; (Hadfield, 2010)) was used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V = 1, nu = 0) (Hadfield, 2014). We ensured the GLMM showed acceptable convergence (lag time autocorrelation values < 0.01 after lag 0; (Hadfield, 2010)), and adjusted parameters if necessary. After collecting the data, we changed the distribution to Gaussian (with an identity link) because MCMCglmm would not run on a Poisson (it kept saying there were negative integers even after we removed them). A Gaussian distribution also works for this kind of data because the response variable is a latency in seconds. We conducted a power analysis as above and found that with a sample size of 32, we would have a 71% chance of detecting a large effect (approximated at $f^2 = 0.35$ by Cohen, 1988).

Detour experiment. We measured the first approach (physical contact with bill) where a correct response was through the tube's side opening and an incorrect response was to the front closed area of the tube (methods as in MacLean et al., 2014). There were two passing criteria: standard (first touch to apparatus) and grackle specific (first touch to apparatus minus frustration bites to the tube rim). We ran one GLM (glm function, stats package) per criterion, with the explanatory variable being the number of trials to past first or last reversal, or the average latency to attempt a new option on the plastic multi-access box experiment or the log multi-access box experiment. We used a binomial distribution and a logit link. The power analysis is the same as in the previous paragraph.

We ran the Bayesian analyses for the detour task with the more comprehensive computational measure of flexibility that takes into account all choices in the reversal learning experiment. We included both parameters (ϕ_j and λ_j) as well as their interaction to predict whether birds make correct choices in each trial of the detour task. We used a binomial likelihood as the outcome distribution and a logit link function (see Model 2a in the Results for full data preparation and analysis script).

To determine whether training improved detour performance, we conducted a GLM (glm function, stats package) with a binomial distribution and a logit link. The response variable was correct response (yes, no) and the explanatory variable was whether they received the detour experiment pre- or post-reversal

experiment. The power analysis showed that, with a sample size of 32, we would have a 71% chance of detecting a medium effect (approximated at $f^2 = 0.15$ by Cohen, 1988).

Unregistered Analysis: Interobserver Reliability of Dependent Variables

To determine whether experimenters coded the dependent variables in a repeatable way, hypothesis-blind video coders, Sophie Kaube (detour) and Brynna Hood (go/no go), were first trained in video coding the dependent variables (detour and go/no go: whether the bird made the correct choice or not), requiring a Cohen's unweighted kappa of 0.90 or above to pass training (using the psych package in R, Revelle (2017)). This threshold indicates that the two coders (the experimenter and the video coder) agreed with each other to a high degree (Landis & Koch, 1977). After passing training, the video coders coded 24% (detour) and 33% (go/no go) of the videos for each experiment and the unweighted Cohen's kappa was calculated to determine how objective and repeatable scoring was for this variable, while noting that the experimenter had the advantage over the video coder because watching the videos was not as clear as watching the bird participate in the trial from the aisle of the aviaries. The unweighted kappa was used because this is a categorical variable where the distances between the numbers are meaningless (0 = incorrect choice, 1 = correct choice, -1 = did not participate).

Detour: correct choice. We randomly chose four (Diablo, Queso, Chalupa, and Habanero) of the 11 birds that had participated in this experiment by November 2019 using https://www.random.org. First, S. Kaube analyzed all videos from Habanero and Diablo, and we analyzed the data using an intraclass correlation coefficient, which is not an appropriate test for categorical data. After learning this, we switched to use the Cohen's unweighted kappa and replaced Habanero and Diablo with two new randomly chosen grackles (Mole and Chilaquile). S. Kaube then analyzed all videos from Queso and Chalupa for training and passed (Cohen's unweighted kappa = 0.91, confidence boundary = 0.75-1.00, $n = 24$ data points). After passing training, S. Kaube analyzed all videos from Queso, Chalupa, Mole, and Chilaquile, and highly agreed with the experimenter's data (Cohen's unweighted kappa = 0.91, confidence boundary = 0.78-1.00, $n = 44$ data points).

Go/no go: correct choice. We randomly chose three (Diablo, Burrito, and Chilaquile) of the 12 birds that were estimated to complete this experiment using https://www.random.org. B. Hood then analyzed all videos from Diablo for training and passed (Cohen's unweighted kappa = 0.91, confidence boundary = 0.80-1.00, $n = 40$ data points). B. Hood then coded the rest of the videos and had substantial amounts of agreement with the experimenters (Cohen's unweighted kappa = 0.82, confidence boundary = 0.78-0.85, $n = 611$ data points).

We think the reason for the lower (but still acceptable) interobserver agreement for this variable is due to the fact that the correct choice data were not as objective to code as we had hoped due to the touchscreen malfunctioning (not registering touches to the screen), and to the subjective criterion that the bird had to be within a certain distance of the screen to be considered paying attention and thus be in position to make a choice or not. This indicates that our touchscreen set up could be greatly improved such that it is actually automated, rather than needing experimenter intervention for every trial.

Go/no go: latency to respond (peck the screen). Interobserver reliability was not conducted on this variable because we obtained this data from the automatically generated PsychoPy data sheets. However, we must note that when entering the latency to first screen peck into the main data sheet that the experimenter used to determine whether they made a correct choice or not, the two data sheets did not always match. This is because: 1) if a session started or ended with the bird not participating such that a trial was not triggered, this receives a -1 in the experimenter's data sheet and is not recorded by the PsychoPy data sheet; and 2) the touchscreen regularly failed to register screen pecks, which could result in an NA for the PsychoPy data sheet whereas the experimenter's data sheet recorded a choice.

Results

A total of 18 grackles participated to varying degrees in the test battery between September 2018 and May 2020 (Table 1). Sample sizes varied between the tests due to the extensive amount of time it took most birds to get through the test battery, in which case several had to be released before they were finished because, for example, they reached the end of the maximum amount of time we were allowed to temporarily hold them in the aviaries (see protocol[1] for details). Data are publicly available[2] at the Knowledge Network for Biocomplexity (Logan et al., 2020). Details on how the grackles were trained to use the touchscreen are in Seitz et al. (2021).

Table 1

Summarized Results per Bird in the Go/no go and Detour Inhibition Experiments, and the Reversal and Multi-access Box (MAB) Flexibility Experiments (flexibility data from Logan et al., 2022)

Bird	Go/no go trials to 85% correct after 150 trials	Go/no go trials to 85% correct	Detour proportion correct	Detour proportion correct modified	Detour pre- or post-reversal	Trials to reverse in first reversal	Trials to reverse in last reversal	Average latency to attempt new solution (MAB plastic)	Average latency to attempt new solution (MAB log)
Diablo	170	170	0.7	0.7	Post	80	40	25	NA
Burrito	190	190	0.5	0.9	Post	60	23	76	391
Adobo	160	160	0.4	0.6	Pre	100	50	31	79
Chilaquile	170	140	0.6	1.0	Post	40	30	44	170
Yuca	170	60	0.2	0.6	Post	80	80	132	77
Mofongo	201	60	0.8	1.0	Pre	40	40	502	630
Pizza	170	100	NA	NA	Post	60	60	NA	1482
Taquito	201	290	0.8	1.0	Post	160	160	NA	100
Queso	NA	NA	0.9	0.9	Pre	70	70	88	NA
Mole	170	170	0.8	0.9	Post	70	50	356	1173
Tomatillo	NA	NA	0.8	0.8	Post	50	50	317	NA
Tapa	NA	NA	1.0	1.0	Pre	100	100	685	NA
Chalupa	NA	NA	0.9	1.0	Post	90	50	NA	NA
Habanero	NA	NA	1.0	1.0	Post	80	40	28	NA
Pollito	NA	NA	0.9	0.9	Post	60	40	NA	668
Taco	NA	NA	0.2	1.0	Post	80	80	NA	117
Huachinango	NA	NA	0.7	0.7	Post	NA	NA	NA	NA
Pavo	NA	NA	0.8	0.8	Pre	NA	NA	NA	NA

Note. We used data from the MAB plastic experiment and the MAB wooden experiment because the wooden and plastic scores did not correlate with each other (Logan et al., 2022). *Go/no go trials to 85% correct after 150 trials* required the bird must achieve 100% correct before trial 150 and if they did not, then they passed after they achieved 85% correct. *Go/no go trials to 85% correct* is simply the number of trials to reach this criterion without the 150-trial threshold of needing to get 100% correct. A value of 201 for go/no go indicates that the bird did not pass criterion within the 200 trial maximum (but note the exception of Taquito who was tested beyond trial 200 until he passed due to experimenter error). *Detour proportion correct modified* accounts for the grackle-specific behavior of standing at the opening of the tube where they are about to reach their head inside the tube to get the food, but they appear frustrated and bite the edge of the plastic tube. These bites do not count as first touch to the plastic when the bird obtains the food immediately after the bite (see Results for the Detour task for justification of this coding).

There was no correlation between the two flexibility experiments: the number of trials to reverse a preference in the last reversal and the average number of seconds (latency) to attempt a new option on the multi-access box after a different locus has become non-functional because they passed criterion on it (Pearson's $r(9) = .52$, 95% CI = -0.12-0.85, $p = .10$, $t = 1.83$). The lack of a correlation between the two flexibility experiments could have arisen for a variety of reasons: 1) Perhaps comparing different types of data, number of trials to pass a criterion versus the number of seconds to switch to attempting a new option,

[1] https://docs.google.com/document/d/1oEQ66yLrkMFr4UJTXfPBRAEXqoUuOgRwcKOB_KcT7HE/edit
[2] https://knb.ecoinformatics.org/view/doi:10.5063/M043S3

distorts this relationship. Future experiments could obtain switch latencies from reversal learning to make the measures more directly comparable. 2) Perhaps one or both flexibility measures are not repeatable within individuals, in which case, it would be unlikely that a stable correlation would be found. 3) The multi-access box experimental design allows for unknown amounts of learning within a trial, whereas the reversal learning design allows only one learning opportunity per trial; perhaps this difference in experimental design introduces noise into the multi-access box experiment, thus making the comparison of their results ambiguous. Additionally, the average latency to attempt a new option did not correlate between the multi-access plastic and multi-access wooden experiments (Logan et al., 2022). Therefore, we conducted separate analyses for each flexibility experiment (reversal and multi-access) as well as separate analyses for the multi-access box and multi-access wooden apparatuses.

Prediction 1: The More Flexible Individuals are Also Better at Inhibition

Model 2a: Number of Trials to Pass Criterion in Go/No Go

Relationship Between Go/No Go (Inhibition) and Reversal Learning (Flexibility). There was a positive correlation between the number of trials to pass criterion in the go/no go experiment and the number of trials to reverse a preference (M =59 trials, SD = 41, range 23-160 trials, n = 9 grackles) in the colored tube reversal experiment (in their last reversal, thus for the control grackles, this was their first and only reversal, while for the manipulated grackles, this was their last reversal in the serial reversal manipulation) when using one of the two go/no go passing criteria: the number of trials to reach 85% correct (measured in the most recent 20 trial block; (M = 149 trials, SD =71, range 60-290 trials, n = 9 grackles; Table 2, Figure 2). The other passing criterion of achieving 100% correct performance by trial 150, and if this is not met then they pass when they reach 85% correct after trial 150 (measured in the most recent 20-trial block; M =178 trials, SD =15, range 160-200 trials) did not correlate with reversal performance.

Table 2

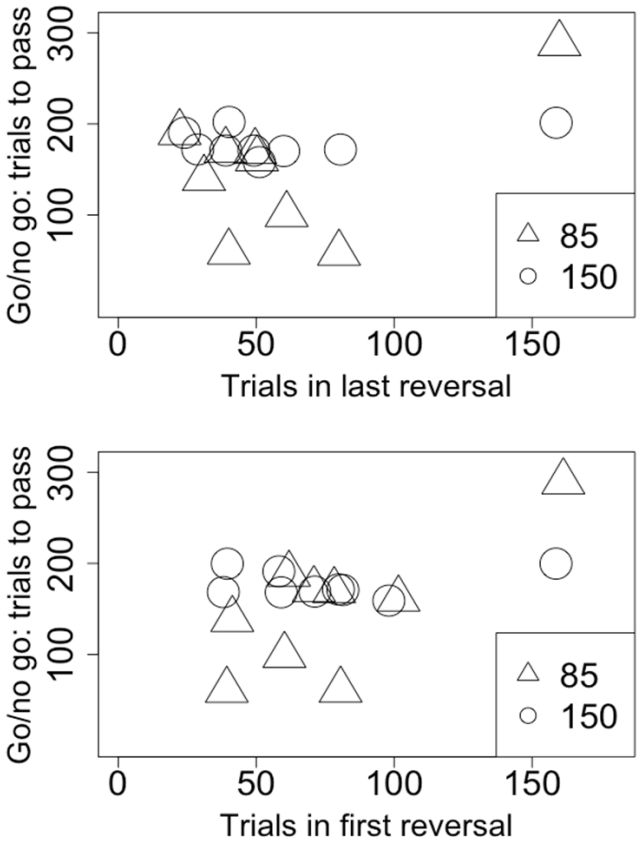
Results from the Go/no go and Reversal Learning GLMs with Taquito

Item	m1: 150 last reversal	m2: 85 last reversal	m3: 150 first reversal	m4: 85 first reversal
(Intercept)	5.14***	4.68***	5.15***	4.34***
	(0.05)	(0.05)	(0.06)	(0.07)
TrialsLast	0.00	0.01***		
	(0.00)	(0.00)		
TrialsFirst			0.00	0.01***
			(0.00)	(0.00)
N	9	9	9	9
AIC	75.91	278.00	76.96	211.92
BIC	76.30	278.40	77.36	212.31
Pseudo R2	0.15	1.00	0.04	1.00

Note. m1 and m2 show GLM outputs for the last reversal, while m3 and m4 show GLM outputs for the first reversal. m1 and m3 show results from the GLM using the number of trials to reach 85% correct if 100% correct was not achieved within the first 150 trials in go/no go, while m2 and m4 use the number of trials to reach 85% correct without the 150-trial threshold. The estimate is presented above the standard error, which is in parentheses; asterisks refer to p-value significance.
*p < .05 **p < .01 ***p < .001

Figure 2

The Number of Go/No Go Trials to Pass Criterion per Bird in their Last and First Reversals



Note. The number of go/no go trials to pass criterion per bird (n = 9 grackles) using the 85% correct (triangles) or 85% correct after 150 trials (circles) criteria and the number of trials to reverse a color preference in their last reversal (top panel) and first reversal (bottom panel).

Regardless of the criterion used, we capped the number of trials for the go/no go experiment at 200, with the exception of two individuals who were tested past trial 200 due to experimenter error (Mofongo continued to trial 249 and did not pass the 85% criterion; and Taquito continued to trial 290 and passed the 85% criterion). We repeated the above analyses for the 85% criterion using a data set without Taquito because this would make the individuals more comparable as not all grackles were given the chance to pass criterion after trial 200.

Results for the analyses without Taquito showed that, instead of a positive correlation, there was a negative correlation between the number of trials to pass criterion in the go/no go experiment and the number of trials to reverse a preference in the colored tube reversal experiment (in their last reversal; M = 47, SD = 17, range 23-80, n = 8 grackles) using the 85% criterion (M = 131 trials, SD = 51, range 60-190 trials, n = 8 grackles; Table 3, Figure 2).

The two results from the data set that included Taquito were confirmed using a more comprehensive computational measure of reversal learning that accounts for all of the choices an individual made as well as the degree of uncertainty exhibited as preferences change (flexibility 4 in the Methods). We used multilevel Bayesian reinforcement learning models to investigate a bird's learning rate and random choice rate per reversal (see Methods for more details; results presented as posterior means and 89% highest posterior density intervals (HPDI)). With the 85% correct criterion, we found a negative relationship

between reversal learning rate and the number of go/no go trials to pass criterion. This means that birds who were faster to update their behavior in the reversal experiment were also faster to reach criterion in the go/no go task (β_{ϕ} = -0.37, HPDI = -0.54 to -0.16). This confirms the positive relationship between numbers of trials to reverse a preference and trials to reach criterion in the go/no go task, because fewer trials to reverse preferences tended to be reflected in higher learning rates in the computational model. Moreover, birds that exhibited a higher random choice rate in the reversal experiment took longer to reach the 85% correct criterion compared to birds that were less random in their choices (β_{λ} = -0.34, HPDI = -0.52 to -0.12). We also found some evidence for a positive interaction between both learning parameters (reversal learning rate and random choice rate; $\beta_{\phi\chi\lambda}$ = 0.27, HPDI = 0.02 – 0.58), suggesting a buffering effect among parameters such that the influence of random choice rate is weaker for individuals that are fast learners.

Table 3

Results from the Go/No Go and Reversal Learning GLMs without Taquito

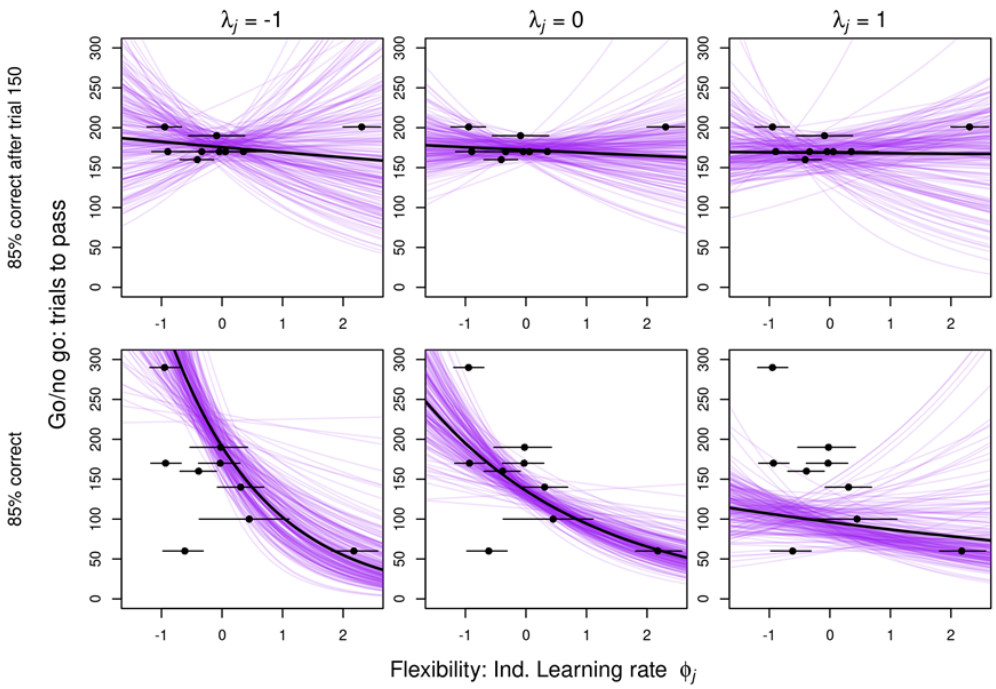
Item	m1: 85 last reversal	m2: 85 first reversal
(Intercept)	5.51 ***	4.51 ***
	(0.09)	(0.11)
TrialsLast	-0.01 ***	
	(0.00)	
TrialsFirst		0.01 ***
		(0.00)
N	8	8
AIC	158.86	201.12
BIC	159.02	201.28
Pseudo R2	1.00	0.77

Note. m1 shows GLM outputs for the last reversal, while m2 shows GLM outputs for the first reversal. Both models show results from the GLM using the number of trials to reach 85% correct without the 150-trial threshold. The estimate is presented above the standard error, which is in parentheses; asterisks refer to *p*-value significance.
*$p < .05$ **$p < .01$ ***$p < .001$

Figure 3 plots posterior predictions for the effect of learning rate ϕ_j on the number of trials to pass criterion for three different levels of the random choice rate λ_j. Focusing on the bottom row (85% correct criterion), the model, in general, predicts that fast learners in the reversal learning experiment also reach the criterion in the go/no go experiment in fewer trials. There appears to be a trade-off between learning parameters, such that fast learners who are somewhat exploratory are predicted to perform better than fast learners who show very limited randomness in their choices. Lastly, overall individuals who show fewer random choices in the reversal learning experiment are predicted to perform better in the go/no go inhibition experiment.

Figure 3

The Number of Go/No Go Trials to Pass Criterion per Bird in Relation to their Reversal Learning Rate and Random Choice Rate



Note. Results from the computational learning model (flexibility 4; n = 9). Posterior predicted number of trials to pass go/no go using the 85% correct after 150 trials (top row) or 85% correct (bottom row) criteria, based on estimates for the individual-level learning rates from the reinforcement learning model (ϕ_j; black dots show posterior means, black horizontal lines indicate 89% highest posterior density intervals). Curves are plotted for high (left; λ_j = -1), average (middle; λ_j= 0) and low (right; λ_j = 1) random choice rates. Purple lines represent 200 independent draws from the posterior, the black lines show posterior means. Both predictors (λ_j and ϕ_j) were standardized before calculations.

As with the other analysis, there was no robust association between either learning rate (β_ϕ = - 0.02, HPDI = -0.15 – 0.12) or random choice rate (β_λ = -0.02, HPDI = -0.12 – 0.07) and the number of trials to pass the other go/no go criterion (100% correct by trial 150). There was no interaction between the learning parameters ($\beta_{\phi\chi\lambda}$ = 0.01, HPDI = -0.23 – 0.19).

Reassuringly, excluding Taquito did not change the overall patterns. There was still a negative relationship between reversal learning rate and the number of go/no go trials to pass the 85% correct criterion (β_ϕ = -0.26, HPDI = -0.47 to -0.01), a positive relationship between random choice rate and go/no go trials (β_λ = -0.34, HPDI = -0.53 to -0.06) and a positive interaction between both learning parameters ($\beta_{\phi\chi\lambda}$ = 0.27, HPDI = -0.13 – 0.53). The results for the other go/no go criterion also did not change for the data set that included Taquito.

Overall, these results indicate that those individuals that have more inhibition are also faster at changing their preferences when circumstances change. While the relationship between trials to reverse preference and trials to reach the go/no go criterion was strongly influenced by Taquito, who was very slow in both experiments, the more comprehensive model of flexibility that takes all trials into account and does not rely on an arbitrary passing criterion provided support for the relationship irrespective of whether Taquito was included or not. Still, we would need a larger sample size to determine to what degree the relationship is perturbed by individual variation.

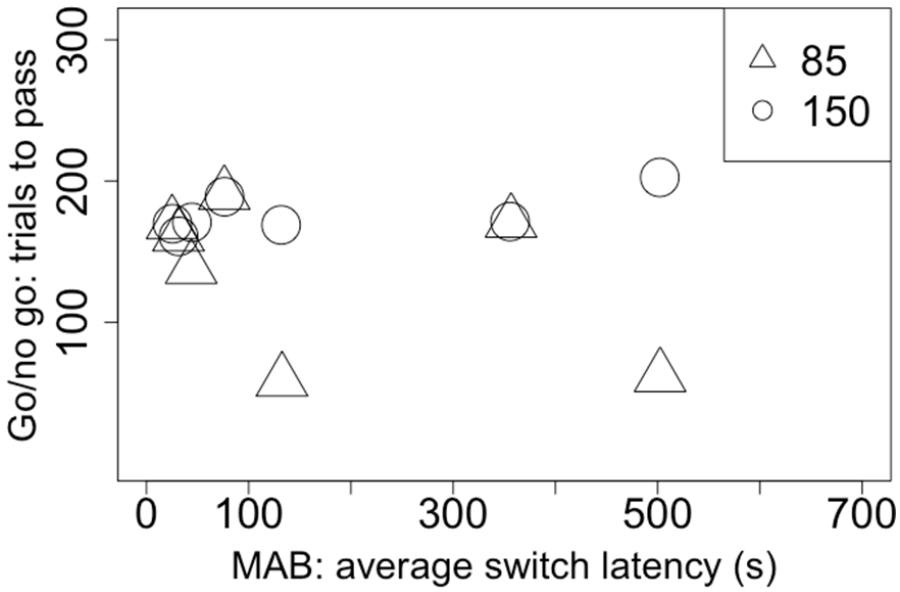
Unregistered Analyses. We additionally analyzed the relationship between go/no go performance and the number of trials to reverse a color preference (M = 76, SD = 37, range: 40–160, n = 8 grackles) in

the first reversal to make our results comparable across more species. This is because most studies do not conduct serial reversals, but only one reversal. The results that included Taquito (Table 2) were the same as the results that excluded Taquito (Table 3): there was a positive correlation between go/no go and reversal learning performance when using the 85% go/no go criterion, and no relationship when using the 100% by 150-trial criterion. In comparison with the results for the last reversal, these results are the same as those that included Taquito (positive relationship; Table 2), and the opposite of those that excluded Taquito (negative relationship; Table 3).

Relationship Between Go/No Go (Inhibition) and Multi-Access Box (Flexibility). The average latency to attempt a new option on both multi-access box (MAB) experiments (plastic and log) negatively correlated with go/no go performance when using the 85% go/no go criterion (plastic sample: M = 136, SD = 54, range: 60-190, n = 7 grackles, does not include Taquito; log sample: M = 146, SD = 76, range: 60–290, n = 8 grackles, includes Taquito). There was no correlation when using the 150-trial threshold (M = 176, SD = 14, range 160-201, n = 7 grackles; Table 4, Figure 4). Results from the log MAB that exclude Taquito show no relationship between the average latency to attempt a new option (M = 572, SD = 559, range: 77-1482, n = 7 grackles) and go/no go performance using the 85% criterion (M =125, SD = 53, range 60–190, n = 7 grackles). On the plastic MAB, the average of the average latency per bird to attempt a new solution was 167 s (SD = 188, range: 25–502, n = 7 grackles). On the log MAB, the average of the average latency per bird to attempt a new solution was 513 s (SD = 544, range 77–1482, n = 8 grackles).

Figure 4

The Number of Trials to Pass Criterion on the Go/no go Experiment in Relation to the Average Number of Seconds Taken to Switch Options on the Multi-access Box (MAB)



Note. The number of go/no go trials to pass criterion per bird (n = 7) using the 85% correct (triangles) or 85% correct after 150 trials (circles) criteria and the average latency to attempt a new locus on the multi-access box (MAB) plastic.

Table 4

Results from the Go/No Go and Multi-access Box GLMs with Taquito

Item	m1: 150 plastic	m2: 85 plastic	m3: 150 log	m4: 85 log
(Intercept)	5.13***	5.09***	5.20***	5.10***
	(0.04)	(0.04)	(0.04)	(0.04)
AvgLatencyPlastic	0.00	-0.00 ***		
	(0.00)	(0.00)		
AvgLatencyLog			-0.00	-0.00***
			(0.00)	(0.00)
N	7	7	8	8
AIC	57.23	163.75	69.83	315.01
BIC	57.12	163.65	69.99	315.17
Pseudo R2	0.31	0.99	0.02	0.88

Note. m1 and m3 show results from the GLM using the number of trials to reach 85% correct if 100% correct was not achieved within the first 150 trials in go/no go, while m2 and m4 use the number of trials to reach 85% correct without the 150-trial threshold. m1 and m2 show results from the plastic multi-access box, while m3 and m4 show results from the log multi-access box. The estimate is presented above the standard error, which is in parentheses; asterisks refer to *p*-value significance. Note that an estimate of -0.00 simply means that rounding to two decimal places obscured additional digits that show this is a slightly negative number.
*$p < .05$ **$p < .01$ ***$p < .001$

Model 2b: Latency to Peck Screen in Go/No Go

The model that examined whether the latency of the first peck to the screen per trial (response variable) was associated with the outcome of the trial (correct/incorrect) did not converge. This is probably because the correct choice on the no go trials was not to peck the screen and so this level of the categorical choice variable has much less data than the other two levels (incorrect choice and correct choice on the go trials; Figure 5). Therefore, we cannot include the analysis here or make conclusions based on it. Additionally, there was a problem matching the latency data across data sheets. Latency data was brought in from the PsychoPy data sheets, however, the number of trials reported by the experimenter and by PsychoPy sometimes differed for reasons that are unclear. Therefore, the first latency to peck the screen is not completely accurately matched between the two data sheets.

Table 5

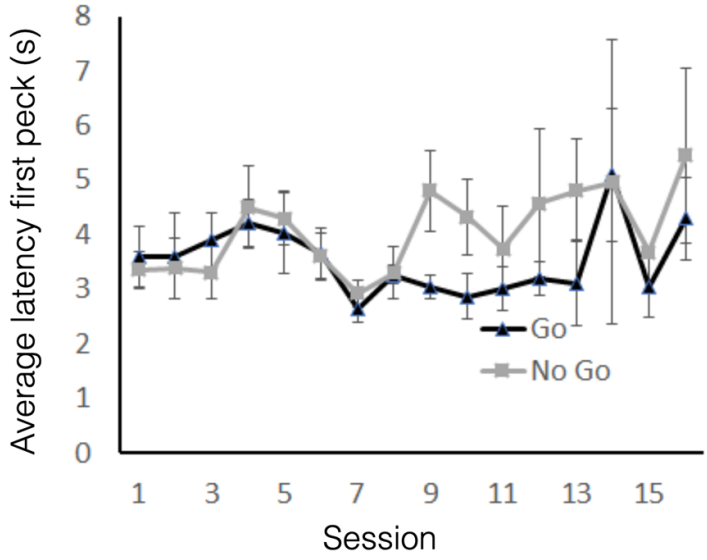
Results from the Go/No Go and Log Multi-access Box GLM without Taquito

Item	85 log
(Intercept)	4.84 ***
	(0.05)
AvgLatencyLog	-0.00
	(0.00)
N	7
AIC	193.62
BIC	193.51
Pseudo R2	0.00

Note. GLM using the number of trials to reach 85% correct without the 150-trial threshold. The estimate is presented above the standard error, which is in parentheses; asterisks refer to *p*-value significance. Note that an estimate of -0.00 simply means that rounding to two decimal places obscured additional digits that show this is a slightly negative number.
*$p < .05$ **$p < .01$ ***$p < .001$

Figure 5

The Average Latency to First Peck on the Touchscreen across Sessions for Go and No Go Trials



Note. The average latency (seconds) across all birds ($n = 9$) to first peck the screen in a trial per session according to whether it was a go trial (when they should peck; black triangles and black regression line) or a no go trial (when they should not peck; gray squares and gray regression line) (error bars = standard error of the mean).

Relationship Between Detour (Inhibition) and Reversal Learning (Flexibility)

There was no correlation between the proportion correct on the detour experiment ($M = 0.71$, $SD = 0.25$, range 0.20–1.00, $n = 18$ grackles) and the number of trials to reverse their last preference in the reversal learning experiment (Table 6, Figure 6). The same result was found using the more comprehensive flexibility measure with the Bayesian reinforcement model: we found no relationship between the learning rate ($\beta_{\phi} = 0.12$, HPDI = -0.13 to 0.38) or random choice rate ($\beta_{\lambda} = -0.07$, HPDI = -0.55 to 0.46) and the proportion of correct choices in the detour experiment. There was also no interaction among parameters (learning rate and random choice rate; $\beta_{\phi\chi\lambda} = 0.01$, HPDI = -0.39 to 0.38).

Table 6

Results from the Detour and Reversal Learning GLMs

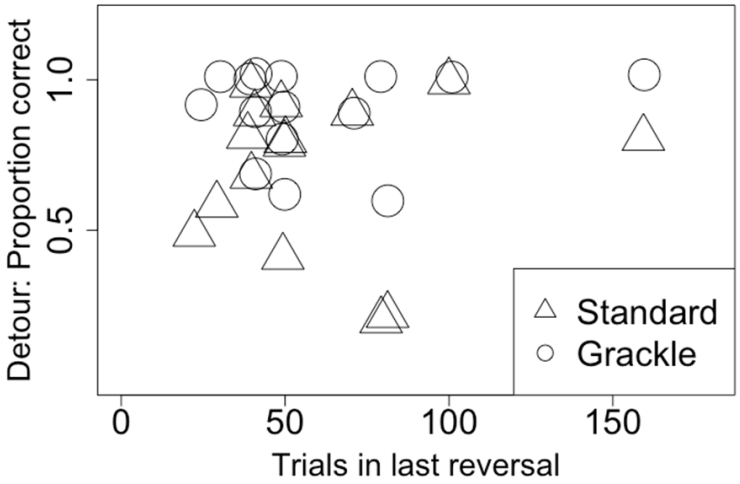
Item	m1: std & last rev	m2: std & 1st rev	m3: grackle & last rev	m4: grackle & 1st rev
(Intercept)	0.82	0.73	1.66	0.73
	(1.16)	(1.63)	(1.78)	(1.63)
TrialsLast	0.00		0.01	
	(0.02)		(0.03)	
TrialsFirst		0.00		0.00
		(0.02)		(0.02)
N	15	15	15	15
AIC	21.47	21.52	7.62	21.52
BIC	22.89	22.93	9.03	22.93
Pseudo R2	0.00	-0.00	-0.00	-0.00

Note. m1 and m2 show GLM outputs using the standard MacLean et al. (2014) method of scoring (std), while m3 and m4 show GLM outputs using the grackle-specific scoring method (grackle). m1 and m3 show results using the last reversal (last rev), while m2 and m4 use the first reversal (1st rev).
*$p < .05$ **$p < .01$ ***$p < .001$

Unregistered Analyses. We additionally analyzed the relationship between detour performance and the number of trials to reverse a color preference in the first reversal to make our results comparable across more species. This is because most studies do not conduct serial reversals, but only one reversal. The results remained the same regardless of whether the first or last reversal were analyzed: there was no relationship between detour and reversal learning performance (Tables 4 and 5).

Figure 6

The Proportion of Trials Correct per Bird in the Detour Experiment in Relation to the Number of Trials each Bird Required to Pass Criterion in their Last Reversal



Note. The proportion of detour trials correct per bird (n = 15) using the standard calculation method (triangles) or the grackle-specific calculation method (circles) and the number of trials to reverse a color preference in their last reversal.

As we conducted this experiment, we discovered that scoring whether the grackle made a correct or incorrect first choice is more complicated than the scoring method used in MacLean et al. (2014). In MacLean et al. (2014), and most other studies using a detour task, to our knowledge, if the plastic is touched first, then it is an incorrect choice, whereas if the food is touched first, it is a correct choice. If the plastic is touched first, it is assumed that the individual touched the plastic on the long side of the tube and not on the rim side where the opening is because they were trying to reach the food through plastic (which is non-functional). We found that many grackles have a habit of standing at the tube opening, biting the rim of the tube and then immediately afterwards putting their head in to obtain the food, possibly due to reluctance to put their heads into the tube. This behavior did not appear to be an attempt to reach the food through the plastic because: 1) it was always followed by immediate food retrieval, and 2) it was distinct from other pecks to plastic on the long side. For these reasons, we coded an additional variable, the "grackle-specific correct choice." In this variable, a bite to the plastic rim does not count as an incorrect choice if they then obtained the food without having touched the front (non-edge) of the plastic tubing between their bite to the rim and their obtaining the food. Instead, this counts as a correct choice. We therefore conducted *post hoc* analyses of the proportion correct on the detour task in relation to their reversal performance (Tables 4 and 5). The results were the same as above: there is no correlation between detour performance (using the grackle-specific correct choice) and the number of trials to reverse their last or first preference. With this scoring method, grackles averaged 87% correct (*SD* 25%, range: 60–100%). Results were also identical to above for the more comprehensive flexibility measure using the Bayesian model: there was no relationship between detour performance (using the grackle-specific method) and learning rate (β_{ϕ} = 0.17, HPDI = -0.11 to 0.44) or random choice rate (β_{λ} = -0.13, HPDI = -0.44 to 0.21) and no interaction ($\beta_{\phi\chi\lambda}$ = 0.06, HPDI = -0.28 to 0.38).

Relationship Between Detour (Inhibition) and Multi-Access Box (Flexibility)

We conducted a separate analysis to determine whether the proportion of correct responses in the detour experiment was related to the average latency to attempt a new option on the multi-access boxes (plastic and log) and found no relationship (using the MacLean et al. (2014) method of scoring; Table 7).

Table 7

Results from the Detour and Multi-access Box (MAB) GLMs

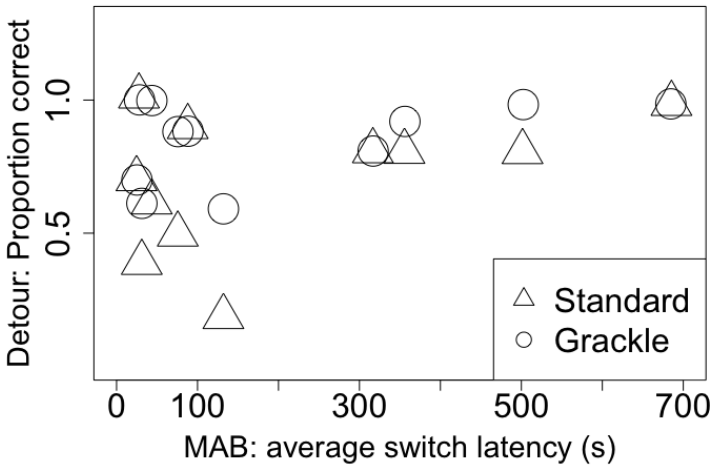
Item	m1: std & plastic	m2: grackle & plastic	m3: std & log	m4: grackle & log
(Intercept)	0.33	-0.47	1.27	1.55
	(0.90)	(1.03)	(1.12)	(1.42)
AvgLatencyPlastic	0.00		0.00	
	(0.00)		(0.01)	
AvgLatencyLog		0.00		0.00
		(0.00)		(0.00)
N	11	9	11	9
AIC	15.45	13.84	7.51	6.37
BIC	16.25	14.23	8.31	6.76
Pseudo R2	0.18	0.33	-0.02	-0.01

Note. m1 and m3 show GLM outputs using the standard MacLean et al. (2014) method of scoring (std), while m2 and m4 show GLM outputs using the grackle-specific scoring method (grackle). m1 and m2 show results from the MAB plastic experiment, while m3 and m4 show results from the MAB log experiment.
*$p < .05$ **$p < .01$ ***$p < .001$

Unregistered Analyses. There was no correlation between the proportion of correct responses in the detour experiment using the grackle-specific scoring method and the average latency to attempt a new option on either of the multi-access boxes (plastic or log; Table 7, Figure 7).

Figure 7

The Proportion of Trials Correct per Bird in the Detour Experiment in Relation to the Average Number of Seconds it Took them to Switch Options on the Multi-access Box (MAB)



Note. The proportion of detour trials correct per bird (n = 11) using the standard calculation method (triangles) or the grackle-specific calculation method (circles) and the average latency to attempt a new locus on the MAB plastic.

Prediction 2: No Correlation Between Inhibition Tasks

There was no correlation between the inhibition tasks go/no go and detour. Cronbach's alpha showed low reliability equal to zero for all comparisons (go/no go 150 threshold and detour standard = 0.03, go/no go 150 and detour grackle specific=0.03, go/no go 85 and detour standard = 0.005, go/no go 85 and detour grackle specific = 0.003).

Prediction 3: Does Training Improve Detour Performance?

There was no difference in the proportion correct on the detour task and whether the individual received the detour experiment before or after their reversal learning experiment (which also involved obtaining food from tubes; Table 8). Seventeen grackles participated in the detour experiment with 5 in the pre-reversal condition and 12 in the post-reversal condition.

Unregistered Analysis

We conducted a post-hoc analysis using the detour grackle-specific proportion of correct responses (see full explanation in P1: detour > Unregistered analyses) and found that the result is the same as above: there is no difference in detour performance relative to their experience with reversal tubes (Table 8).

Table 8

Results from the Detour GLMs to Determine Whether Experience with Reversal Tubes Improves Detour Performance

Item	Detour standard	Detour grackle-specific
(Intercept)	0.73 (0.62)	1.95 * (0.87)
DetourprepostPre	0.53 (1.24)	-0.13 (1.56)
N	17	17
AIC	22.83	8.71
BIC	24.50	10.38
Pseudo R2	0.00	-0.00

Note. Detour standard shows GLM outputs using the MacLean et al. (2014) method of scoring, Detour grackle-specific shows GLM outputs using the grackle-specific scoring method, Condition refers to whether they received the detour test before (pre) or after (post) their reversal experiment.
*$p < .05$ **$p < .01$ ***$p < .001$

Discussion

We found mixed support for the hypothesis that inhibition and flexibility are associated with each other. Inhibition measured using the go/no go task was associated with flexibility (reversal task and multi-access box tasks), but inhibition measured using the detour task was not associated with either flexibility measure. While the relationship between the number of trials to reverse a preference and the number of trials to reach go/no go criterion depended on the inclusion or exclusion of one individual, flexibility measured through our more mechanistic computational model showed a consistent association with go/no go performance, such that the more flexible learners were also better at inhibition. This shows the need to move beyond rather arbitrary thresholds towards more theoretically grounded measures of cognitive traits, based on, for example, cognitive modeling of behavior. Regardless, the change of direction of the relationship given the addition or removal of one individual from the data set indicates that individuals should be tested beyond an arbitrary threshold in the go/no go test to better understand individual variation at the high end of the spectrum. The negative correlation between performance on go/no go and the multi-access boxes could indicate that solution switching on the multi-access box is hindered by self-control.

Performance on the multi-access box improves when one explores the other options faster. Perhaps inhibition hinders such exploration, resulting in slower switching times.

Our results confirm previous findings where detour performance was not associated with flexibility as measured by the multi-access box locus switching performance (Johnson-Ulrich et al., 2018) or by reversal learning (Boogert et al., 2011; Brucks et al., 2017; Damerius et al., 2017; DuBois et al., 2018; Ducatez et al., 2019; Shaw et al., 2015). This mixed support could be because the two inhibition tests, go/no go and detour, did not correlate with each other, indicating that they did not measure the same trait in great-tailed grackles.

There is controversy around how to best assess inhibition given the several experimental paradigms that are available. Inhibitory control is a multi-level construct and an integral part of executive functioning. One aspect of inhibition is motor self-regulation (i.e., stopping a prepotent but counterproductive movement; Diamond, 2013), which is usually assessed with the detour task in non-human animals. Another aspect of inhibitory control is self-control (i.e., the ability to withhold an immediate response towards a present stimulus in favor of a later stimulus; Nigg, 2017). To assess self-control in non-human animals, a task must crucially involve a component of decision making, such as deciding between obtaining a less preferred reward now or tolerating a delay for a more valuable outcome in the future (Beran, 2015). In non-human animals, self-control is typically assessed using experimental paradigms, such as the accumulation paradigm, exchange paradigm, hybrid delay, and intertemporal choice task (for an overview see: Beran, 2018; Miller et al., 2019). A major concern associated with the comparison of performance on inhibition tasks is that measures are not always consistent when different experimental paradigms are used (Addessi et al., 2013; Brucks et al., 2017; van Horik et al., 2018), which is further confirmed by our findings. This indicates that it is crucial to compare inhibition paradigms with each other on the same individuals to understand whether and how they relate to each other and in which contexts. In addition, it may be best to refer to the different inhibition paradigms with distinct terms to differentiate them (e.g., "motor inhibition" for detour-like tasks and "self-control" for delay of gratification tasks).

In the go/no go experiment, the 85% correct passing criterion was more relevant to the grackles, and the one we recommend using in the future. Setting an arbitrary threshold of needing 100% correct in the first 150 trials to pass criterion, which is not generally used in go/no go inhibition tasks, was not ecologically relevant for grackles. In reversal learning tests, which are similar to the go/no go experimental design in that they learn to discriminate between two shapes, grackles almost always continue to explore their options regardless of whether they already have a color preference (e.g., Logan, 2016). There was also more individual variation using the 85% passing criterion, which makes it a more useful measure for comparison.

Although great-tailed grackles had never experienced touchscreen experiments before, we found that the grackles were able to learn to use the touchscreen and to complete the go/no go experiment on it. This validates the use of this setup for future experiments in this species and shows that it could be a viable option for wild-caught birds from other species as well. However, there are several caveats to the feasibility of touchscreen tasks for behavioral testing (see Seitz et al., 2021 for details). First, touchscreen hardware and software can be prone to error. We recommend future studies ensure that the touchscreens accurately record the target behaviors prior to intensive experimentation. Second, touchscreen experimentation should be as fully automated as possible; it can be difficult for observers to objectively code bird behaviors as the birds interact with a touchscreen. Our interobserver reliability was not as reliable as we had hoped, although it was still acceptable for data analysis, due to some of these issues (see details in Methods).

Performance on the detour inhibition test was not affected by extensive experience obtaining hidden food from tubes in the reversal learning test. Grackles who received the detour experiment before reversal training did not perform differently from those who received the detour experiment after reversal training. These two contexts appear to be different enough to solicit independent responses without interference due to a grackle's previous test history. The development of our grackle-relevant detour scoring method resulted in improved performance for 9 out of the 16 grackles we tested. This indicates that cross-species comparisons on this test that are not attuned to the species under study could underestimate inhibitory ability. This finding could partially explain why so many of the 36 species in MacLean et al. (2014)

performed so poorly on this task, aside from actually having poor motor inhibition.

Our developments and modifications to these inhibition tests confirm that it is necessary to accommodate species-relevant behavioral differences in apparatus design and when scoring choices to measure the actual potential of a given species (e.g., Thornton & Lukas, 2012). Such developments are required to determine what inherent trait inhibition tests measure, whether it is appropriate to categorize different tests as measuring the same ability, and how inhibition relates to other traits.

Conclusions

Our results support the idea that flexibility used in reversal learning and in task switching on multi-access boxes may only be associated with the "self-control" type of inhibition (as measured by the go/no go task) and not motor inhibition (as measured by the detour task) in great-tailed grackles. We confirm previous findings that suggest inhibition is multiple constructs that are potentially independent, as has been suggested for humans and dogs (Brucks et al., 2017; Friedman & Miyake, 2004). It is possible that inhibition represents a set of cognitive pathways that is evolutionarily ancient (such that birds and mammals share types of inhibition from a common ancestor) or that there has been convergent evolution of these abilities in multiple lineages.

Note: This paper is the final report based on an accepted peer-reviewed pre-registered submission that can be found here: https://doi.org/10.31234/osf.io/vpc39. The pre-study peer reviews can be found here: https://doi.org/10.24072/pci.ecology.100016, and the post-study peer reviews can be found here: https://doi.org/10.24072/pci.ecology.100081

Acknowledgements

The authors thank Dieter Lukas for help polishing the predictions; Ben Trumble for providing us with a wet lab at Arizona State University and Angela Bond for lab support; Melissa Wilson for sponsoring our affiliations at Arizona State University; Kevin Langergraber for serving as the local PI on the ASU IACUC; 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; 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; Erin Vogel, our preregistration Recommender at PCI Ecology, and Simon Gingins and two anonymous reviewers for their wonderful feedback; Aliza le Roux, our post-study Recommender at PCI Ecology, and reviewers Pizza Ka Yee Chow and Alex DeCasian for their useful feedback; Debbie Kelly for advice on how to modify the go/no go experiment; Melissa Folsom, Sawyer Lung, and Luisa Bergeron for field and aviary support; Brynna Hood and Sophie Kaube for interobserver reliability video coding; 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. This research was funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology, and by a Leverhulme Early Career Research Fellowship to Logan in 2017-2018.

Author Contributions

Logan: Hypothesis development, experimental design (go/no go task), data collection, data analysis and interpretation, write up, revising/editing, materials/funding. **McCune:** Data collection, data interpretation, revising/editing. **MacPherson:** Data collection, data interpretation, revising/editing. **Johnson-Ulrich:** Touchscreen programming for go/no go task, data interpretation, revising/editing. **Rowney:** Data

collection, data interpretation, revising/editing. **Seitz:** Experimental design (go/no go task), touchscreen programming (go/no go task), data interpretation, revising/editing. **Blaisdell:** Experimental design (go/no go task), data interpretation, revising/editing. **Deffner:** Data analysis (Flexibility 4 model), revising/editing. **Wascher:** Hypothesis development, experimental design (delayed gratification and detour tasks), data analysis and interpretation, write up, revising/editing. All authors gave final approval for publication.

Conflict of Interest: The authors declare no competing interests.

Data Availability: The data are available at: doi:10.5063/M043S3.

References

Addessi, E., Paglieri, F., Beran, M. J., Evans, T. A., Macchitella, L., De Petrillo, F., & Focaroli, V. (2013). Delay choice versus delay maintenance: Different measures of delayed gratification in capuchin monkeys (*Cebus apella*). *Journal of Comparative Psychology*, *127*(4), 392-398.

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.

Beran, M. (2018). *Self-control in animals and people*. Academic Press.

Beran, M. J. (2015). The comparative science of "self-control": What are we talking about? *Frontiers in Psychology*, *6*, 51.

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. *PsyArXiv,* https://doi.org/10.31234/osf.io/z4p6s.

Bond, A. B., Kamil, A. C., & Balda, R. P. (2007). Serial reversal learning and the evolution of behavioral flexibility in three species of North American corvids (*Gymnorhinus cyanocephalus*, *Nucifraga columbiana*, *Aphelocoma californica*). *Journal of Comparative Psychology*, *121*(4), 372–379.

Boogert, N. J., Anderson, R. C., Peters, S., Searcy, W. A., & Nowicki, S. (2011). Song repertoire size in male song sparrows correlates with detour reaching, but not with other cognitive measures. *Animal Behaviour*, *81*(6), 1209–1216.

Bray, E. E., MacLean, E. L., & Hare, B. A. (2014). Context specificity of inhibitory control in dogs. *Animal Cognition*, *17*(1), 15–31.

Brucks, D., Marshall-Pescini, S., Wallis, L. J., Huber, L., & Range, F. (2017). Measures of dogs' inhibitory control abilities do not correlate across tasks. *Frontiers in Psychology*, *8*, 849.

Carter, A. J., Feeney, W. E., Marshall, H. H., Cowlishaw, G., & Heinsohn, R. (2013). Animal personality: What are behavioural ecologists measuring? *Biological Reviews*, *88*(2), 465–475.

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

Damerius, L. A., Graber, S. M., Willems, E. P., & van Schaik, C. P. (2017). Curiosity boosts orang-utan problem-solving ability. *Animal Behaviour*, *134*, 57–70.

Deaner, R. O., van Schaik, C. P., & Johnson, V. (2006). Do some taxa have better domain-general cognition than others? A meta-analysis of nonhuman primate studies. *Evolutionary Psychology*, *4*(1), 147470490600400114.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168.

DuBois, A. L., Nowicki, S., Peters, S., Rivera-Cáceres, K. D., & Searcy, W. A. (2018). Song is not a reliable signal of general cognitive ability in a songbird. *Animal Behaviour*, *137*, 205–213.

Ducatez, S., Audet, J.-N., & Lefebvre, L. (2019). Speed–accuracy trade-off, detour reaching and response to PHA in carib grackles. *Animal Cognition*, *22*(5), 625–633.

Fagnani, J., Barrera, G., Carballo, F., & Bentosela, M. (2016). Is previous experience important for inhibitory control? A comparison between shelter and pet dogs in a-not-b and cylinder tasks. *Animal Cognition*, *19*(6), 1165–1172.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.

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.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135.

Ghahremani, D. G., Monterosso, J., Jentsch, J. D., Bilder, R. M., & Poldrack, R. A. (2009). Neural components underlying behavioral flexibility in human reversal learning. *Cerebral Cortex*, *20*(8), 1843–1852.

Griffin, A. S., & Guez, D. (2014). Innovation and problem solving: A review of common mechanisms. *Behavioural Processes*, *109*, 121–134.

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/

Harding, E. J., Paul, E. S., & Mendl, M. (2004). Animal behaviour: Cognitive bias and affective state. *Nature*, *427*(6972), 312–312.

Hillemann, F., Bugnyar, T., Kotrschal, K., & Wascher, C. A. (2014). Waiting for better, not for more: Corvids respond to quality in two delay maintenance tasks. *Animal Behaviour*, *90*, 1–10.

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.

Isaksson, E., Urhan, A. U., & Brodin, A. (2018). High level of self-control ability in a small passerine bird. *Behavioral Ecology and Sociobiology*, *72*(7), 118.

Johnson-Ulrich, L., Johnson-Ulrich, Z., & Holekamp, K. (2018). Proactive behavior, but not inhibitory control, predicts repeated innovation by spotted hyenas tested with a multi-access box. *Animal Cognition*, *21*(3), 379–392.

Kabadayi, C., Bobrowicz, K., & Osvath, M. (2018). The detour paradigm in animal cognition. *Animal Cognition*, *21*(1), 21–35.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

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 and problem solving in an invasive bird. *PeerJ*, *4*, e1975.

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

Logan, C. J., McCune, K. B., MacPherson, M., Johnson-Ulrich, Z., Rowney, C., Seitz, B., Blaisdell, A., Deffner, D., & Wascher, C. (2020). Great-tailed grackle inhibition data [Data set]. Knowledge Network for Biocomplexity. https://doi.org/10.5063/M043S3

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N. Burkart, J. M., Call, J., Cantlon, J. F., Cheke, L. G., Clayton, N. S., Delgado, M. M., DiVincenti, L. J., Fujita, K., Herrmann, E., Hiramatsu, C., Jacobs, L. F., Jordan, K. E., Laude, J. R., Leimgruber, K. L., Messer, E. J. E., de A. Moura, . C., Ostoji, L., Picard, A., Platt, M. L., Plotnik, J. M., Range, F., Reader, S. M., Reddy, R. B., Sandel, A. A., Santos, L. R., Schumann, K., Seed, A. M., Sewall, K. B., Shaw, R. C., Slocombe, K. E., Su, Y., Takimoto, A., Tan, J., Tao, R., van Schaik, C. P., Virányi, Z., Visalberghi, E., Wade, J. C., Watanabe, A., Widness, J., Young, J. K., Zentall, T. R., & Zhao, Y. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140–E2148.

Manrique, H. M., Völter, C. J., & Call, J. (2013). Repeated innovation in great apes. *Animal Behaviour*, *85*(1), 195–202.

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.

Miller, R., Boeckle, M., Jelbert, S. A., Frohnwieser, A., Wascher, C. A., & Clayton, N. S. (2019). Self-control in crows, parrots and nonhuman primates. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1504.

Nigg, J. T. (2017). Annual research review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *58*(4), 361–383.

R Core Team. (2017). *R: A language and environment for statistical computing* (version 4.0.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

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

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.

Thornton, A., & Lukas, D. (2012). Individual variation in cognitive performance: Developmental and evolutionary perspectives. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1603), 2773–2783.

van Horik, J. O., Langley, E. J., Whiteside, M. A., Laker, P. R., Beardsworth, C. E., & Madden, J. R. (2018). Do detour tasks provide accurate assays of inhibitory control? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875), 20180150.

Zuur, A. F., Ieno, E. N., & Saveliev, A. A. (2009). *Mixed effects models and extensions in ecology with R*. Springer.