Peer Community In Ecology

Are the more flexible great-tailed grackles also better at behavioral inhibition?

Logan CJ[1] McCune KB[2] MacPherson M[2] Johnson-Ulrich Z[2] Rowney C[1] Seitz B[3] Blaisdell AP[3] Deffner D[1] Wascher CAF[4]

2021-05-04

Affiliations: 1) Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; 2) University of California Santa Barbara, Santa Barbara, California, USA; 3) University of California Los Angeles, Los Angeles, California, USA; 4) Anglia Ruskin University, Cambridge, UK. *Corresponding author: corina_logan@eva.mpg.de

This research article has been peer reviewed and recommended by:

Aliza le Roux. 2021. Great-tailed grackle research reveals need for researchers to consider their own flexibility and test limitations in cognitive test batteries. *Peer Community in Ecology*, 100081. 10.24072/pci.ecology.100081. Reviewers: Pizza Ka Yee Chow and Alex DeCasian.

Cite as: Logan CJ, McCune KB, MacPherson M, Johnson-Ulrich Z, Rowney C, Seitz B, Blaisdell AP, Deffner D, Wascher CAF. 2021. Are the more flexible individuals also better at inhibition? PsyArXiv, version 7, peer-reviewed and recommended by *Peer Community in Ecology*. doi: https://doi.org/10.31234/osf.io/vpc39

See the easy-to-read HTML version and the reproducible manuscript (Rmd) version for the code

This article began as a preregistration, which was pre-study peer reviewed and received an In Principle Recommendation of the version on 6 March 2019 by:

Erin Vogel. 2019. Adapting to a changing environment: advancing our understanding of the mechanisms that lead to behavioral flexibility. *Peer Community in Ecology*, 100016. 10.24072/pci.ecology.100016. Reviewers: Simon Gingins and two anonymous reviewers.

ABSTRACT

Behavioral flexibility (hereafter, flexibility) should theoretically be positively related to behavioral inhibition (hereafter, inhibition) because one should need to inhibit a previously learned behavior to change their behavior when the task changes [the flexibility component; Manrique et al. (2013); Griffin and Guez (2014); Liu et al. (2016)]. However, several investigations show no or mixed support of this hypothesis, which challenges the assumption that inhibition is involved in making flexible decisions. We aimed to test the hypothesis that flexibility (measured as reversal learning and solution switching on a multi-access box by

Logan et al. 2019) is associated with inhibition by measuring both variables in the same individuals and three inhibition tests (a go/no go task on a touchscreen, a detour task, and a delay of gratification experiment). We set out to measure grackle inhibition to determine whether those individuals that are more flexible are also better at inhibition. Because touchscreen experiments had never been conducted in this species, we additionally validated that a touchscreen setup is functional for wild-caught grackles who learned to use the touchscreen and completed the go/no go inhibition task on it. Performance on the go/no go and detour inhibition tests did not correlate with each other, indicating that they did not measure the same trait. Individuals who were faster to update their behavior in the reversal experiment took more time to attempt a new option in the multi-access box experiments, and they were either faster or slower to reach criterion in the go/no go task depending on whether the one bird, Taquito, who was accidentally tested beyond the 200 trial cap was included in the GLM analysis. While the relationship between the number of trials to reverse a preference and the number of trials to reach the go/no go criterion was strongly influenced by Taquito, who was very slow in both experiments, the more comprehensive Bayesian model of flexibility that takes all trials into account and does not rely on an arbitrary passing criterion provided support for the positive relationship irrespective of whether Taquito was included. Performance on the detour inhibition task did not correlate with either measure of flexibility, suggesting that detour performance and the flexibility experiments may measure separate traits. We were not able to run the delay of gratification experiment because the grackles never habituated to the apparatuses. We conclude that flexibility is associated with certain types of inhibition, but not others, in great-tailed grackles.

Video summary https://youtu.be/TXFOYqZztf4

INTRODUCTION

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; Manrique et al. 2013; Griffin and Guez 2014; Liu et al. 2016). 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; Shaw et al. 2015; Brucks et al. 2017; Damerius et al. 2017; DuBois et al. 2018; Ducatez et al. 2019), 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 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 puzzle box (multi-access) by Logan et al. 2019). 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'). In the delay of gratification task, 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). 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. 2019). Those grackles that are faster to reverse are more flexible. The multi-access box experimental paradigm is modeled after Auersperg et al. (2011) and consists of four different access options to obtain food where each option requires a different type of action to solve it (the experiments and data are in Logan et al. 2019). Once a grackle passes criterion for demonstrating proficiency in solving an option, that option becomes non-functional in all future trials. The measure of flexibility is 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, 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.

ASSOCIATED PREREGISTRATION

Our hypotheses, methods, and analysis plans are described in the peer-reviewed preregistration of this article, which is included below as the Methods.

DEVIATIONS FROM THE PREREGISTRATION

After data collection began and before data analysis:

1) Jan 2020 go/no go performance: the preregistration listed the passing criterion as the number of trials to reach 85% correct, while the protocol associated with the preregistration that the experimenters used when testing the birds listed the passing criterion as 100% correct within 150 trials or 85% correct between 150-200 trials. Therefore, we tested birds according to the latter criterion and conducted all analyses for both criteria. It was previously not specified over what number of trials 85% accuracy was calculated, therefore we decided to calculate it at the level of the most recent sliding 10 trial block (i.e., the most recent 10 trials, regardless of whether it is an even 20, 30, 40 trials).

2) Jul 2020: in the section 'Independent variables > P1 go/no go > Model 2b', removed the variable "flexibility condition" because, by definition, the birds in the manipulated condition were faster to reverse.

3) Sep 2020: Prediction 1 alternative 2 analysis - when we tried to run the code, we discovered that the Cronbach's alpha is not the appropriate test to run on our experimental design to test the internal validity of the experiment (e.g., does this test actually measure what we think it does). To test internal

validity, we would need to change the experimental design, which was not the goal of our current study. Therefore, we did not conduct this analysis.

RESULTS

A total of 18 grackles participated to varying degrees in the test battery between Sep 2018 and May 2020 (Table 1). Sample sizes vary 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 for details). Data are publicly available 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)

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=0.52, 95% confidence interval=-0.12-0.85, t=1.83, df=9, p=0.10). 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, 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. 2019). 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.

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. 2019). 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. 2019). **Go/no go trials to 85% correct after 150 trials** requires the bird must achieve 100% correct before trial 150 and if they did not, then they pass after they achieve 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).

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

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 (average=59 trials, standard deviation=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; average=149 trials, standard deviation=71, range=60-290 trials, n=9 grackles; Table 2a, Figure 1). 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; average=178 trials, standard deviation=15, range=160-200 trials) did not correlate with reversal performance.

Regardless of the criterion used, we capped the number of trials for the go/no go experiment at 200, with the exception of 2 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**; average=47, standard deviation=17, range=23-80, n=8 grackles) using the 85% criterion (average=131 trials, standard deviation=51, range=60-190 trials, n=8 grackles; Table 2b, Figure 1).

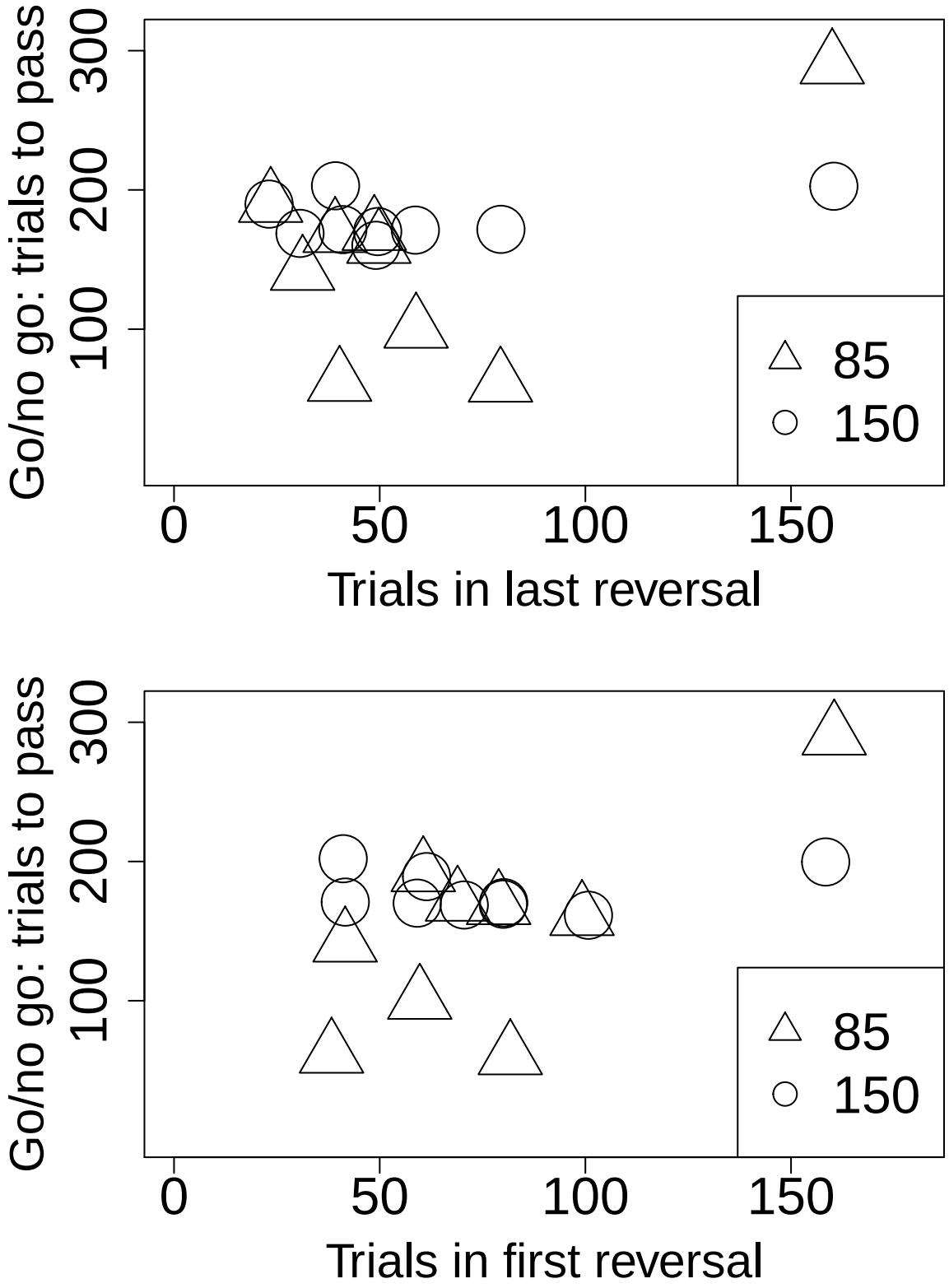

Figure 1. The number of go/no go trials to pass criterion per bird (n=9 grackles) using the 85% correct (triangles) or 85% correct after 150 trials (circles) criteria and the number of trials to reverse a color preference in their last reversal (top panel) and first reversal (bottom panel).

The two results from the data set that included Taquito were confirmed using a more comprehensive com-

putational 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 use 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 are 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 tend 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.

Figure 2 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.

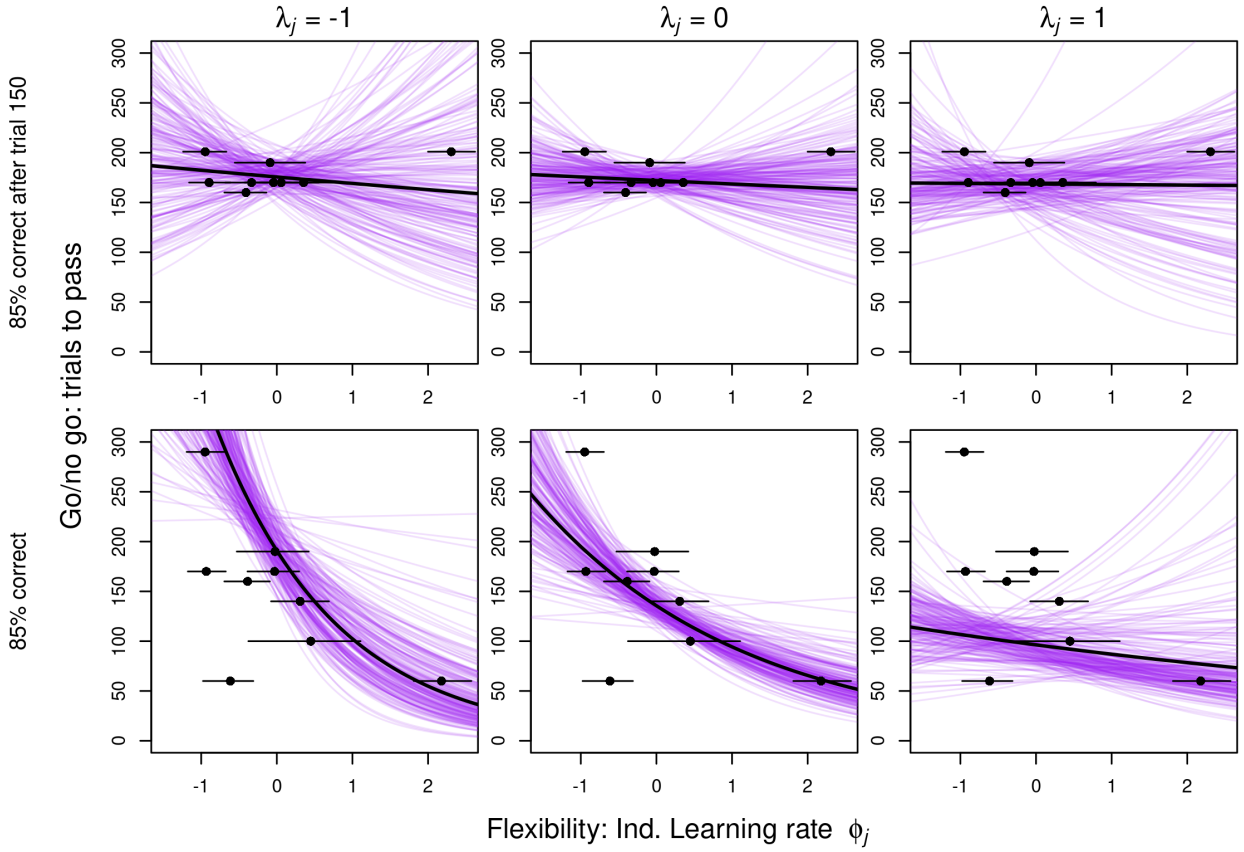

Figure 2. Results from the computational learning model (flexibility 4; n=9). Posterior predicted number of trials to pass go/no go using the 85% correct after 150 trials (top row) or 85% correct (bottom row) criteria, based on estimates for the individual-level learning rates from the reinforcement 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 **Unregistered analysis 1:** We additionally analyzed the relationship between go/no go performance and the number of trials to reverse a color preference (average=76, standard deviation=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 2a) were the same as the results that excluded Taquito (Table 2b): 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 2a), and the opposite of those that excluded Taquito (negative relationship; Table 2b).

Table 2a. Results from the go/no go and reversal learning GLMs (WITH Taquito): **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.

	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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 2b. Results from the go/no go and reversal learning GLMs (WITHOUT Taquito): **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.

	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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Relationship between go/no go (inhibition) and multi-access box (flexibility) The average latency to attempt a new option on both MAB experiments (plastic and log) negatively correlated with go/no go performance when using the 85% go/no go criterion (plastic sample: average=136, standard deviation=54, range=60-190, n=7 grackles, does not include Taquito; log sample: average=146, standard deviation=76, range=60-290, n=8 grackles, includes Taquito). There was no correlation when using the 150 trial threshold (average=176, standard deviation=14, range=160-201, n=7 grackles; Table 3a, Figure 3). Results from the log MAB that exclude Taquito show no relationship between the average latency to attempt a new option (average=572, standard deviation=559, range=77-1482, n=7 grackles) and go/no go performance using the 85% criterion (average=125, standard deviation=53, range=60-190, n=7 grackles). On the plastic multi-access box, the average of the average latency per bird to attempt a new solution was 167 seconds (standard deviation=188, range=25-502, n=7 grackles). On the log multi-access box, the average of the average latency per bird to attempt a new solution was 513 seconds (standard deviation=544, range=77-1482, n=8 grackles).

Table 3a. Results from the go/no go and multi-access box GLMs (WITH Taquito): **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.

	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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 3b. Results from the go/no go and log multi-access box GLM (WITHOUT Taquito) 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.

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

*** p < 0.001; ** p < 0.01; * p < 0.05.

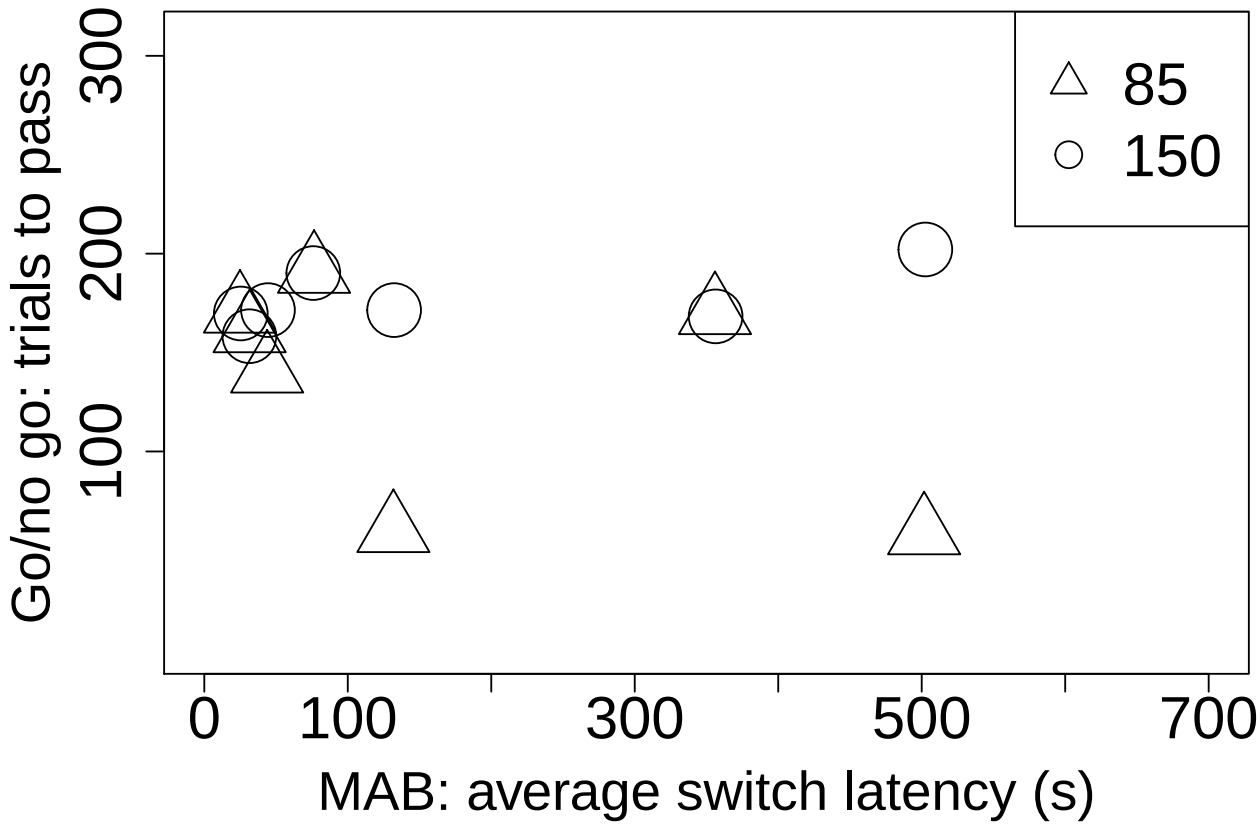

Figure 3. The number of go/no go trials to pass criterion per bird (n=7) using the 85% correct (triangles) or 85% correct after 150 trials (circles) criteria and the average latency to attempt a new locus on the multi-access box (MAB) plastic.

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 4). 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.

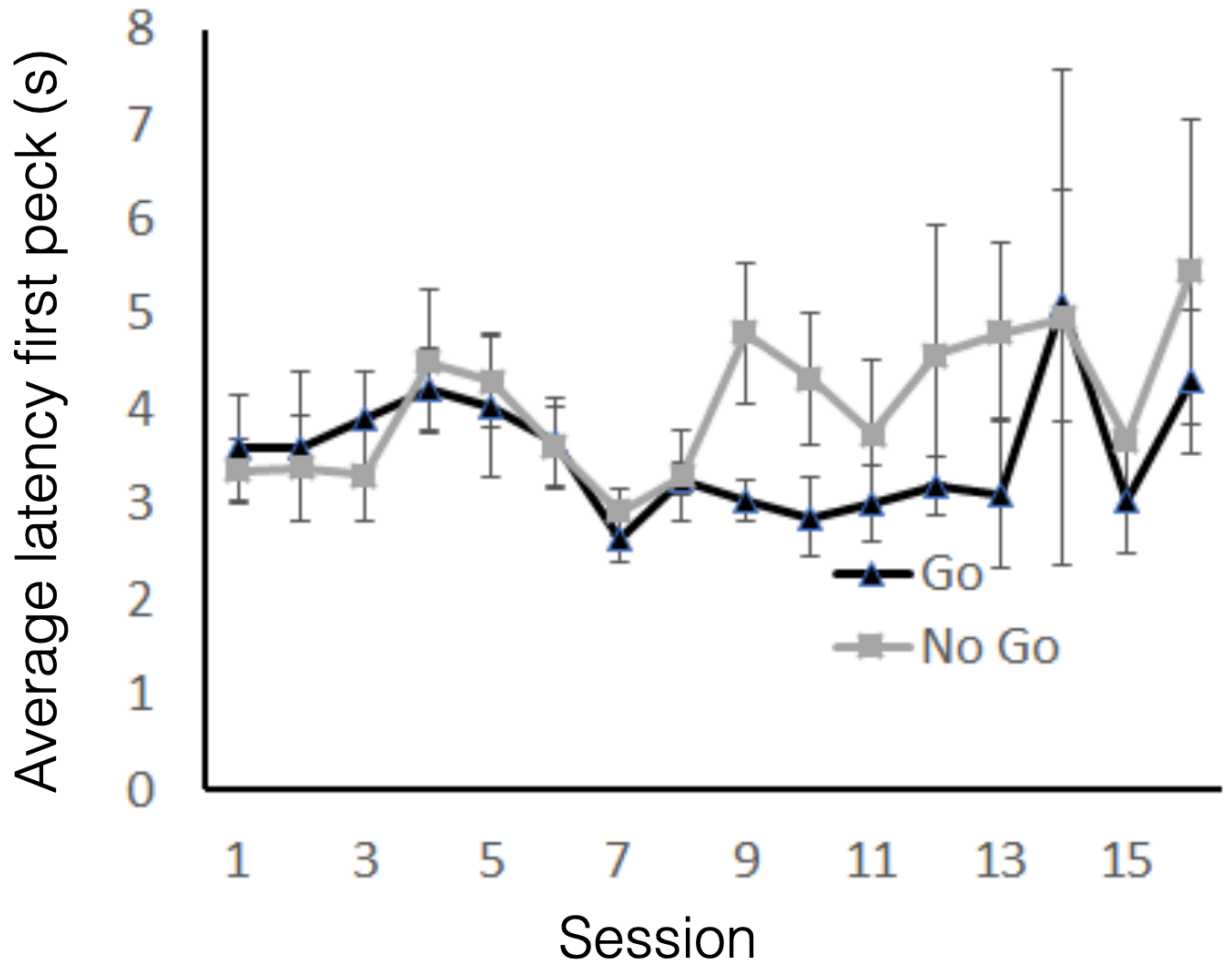

Figure 4. The average latency (seconds) across all birds (n=9) to first peck the screen in a trial per session according to whether it was a go trial (when they should peck; black triangles and black regression line) or a no go trial (when they should not peck; gray squares and gray regression line) (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 (average=0.71, standard deviation=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 3, Figure 5). The same result was found using the more comprehensive flexibility measure with the Bayesian reinforcement model: we found no relationship between the learning rate (β_ϕ = 0.12, HPDI = -0.13 to 0.38) or random choice rate (β_λ = -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).

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 (Table 3).

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 (Table 3). 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 (standard deviation: 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).

Table 4. Results from the detour and reversal learning GLMs: **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).

	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

*** p < 0.001; ** p < 0.01; * p < 0.05.

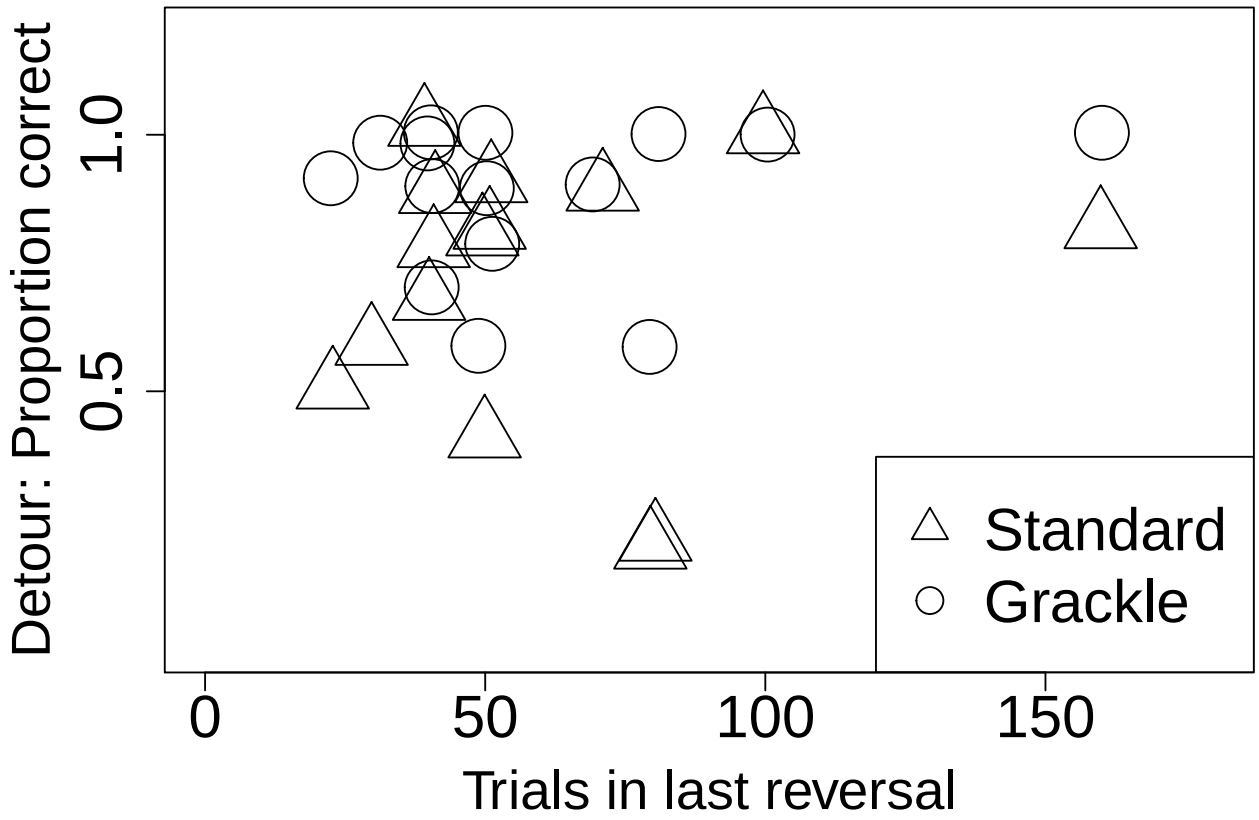

Figure 5. The proportion of detour trials correct per bird (n=15) using the standard calculation method (triangles) or the grackle-specific calculation method (circles) and the number of trials to reverse a color preference in their last reversal.

Relationship between detour (inhibition) and multi-access box (flexibility)

We conducted a separate analysis to determine whether the proportion correct 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 5].

Unregistered analyses

There was no correlation between the proportion correct 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 5, Figure 6).

Table 5. Results from the detour and multi-access box GLMs: **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.

	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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

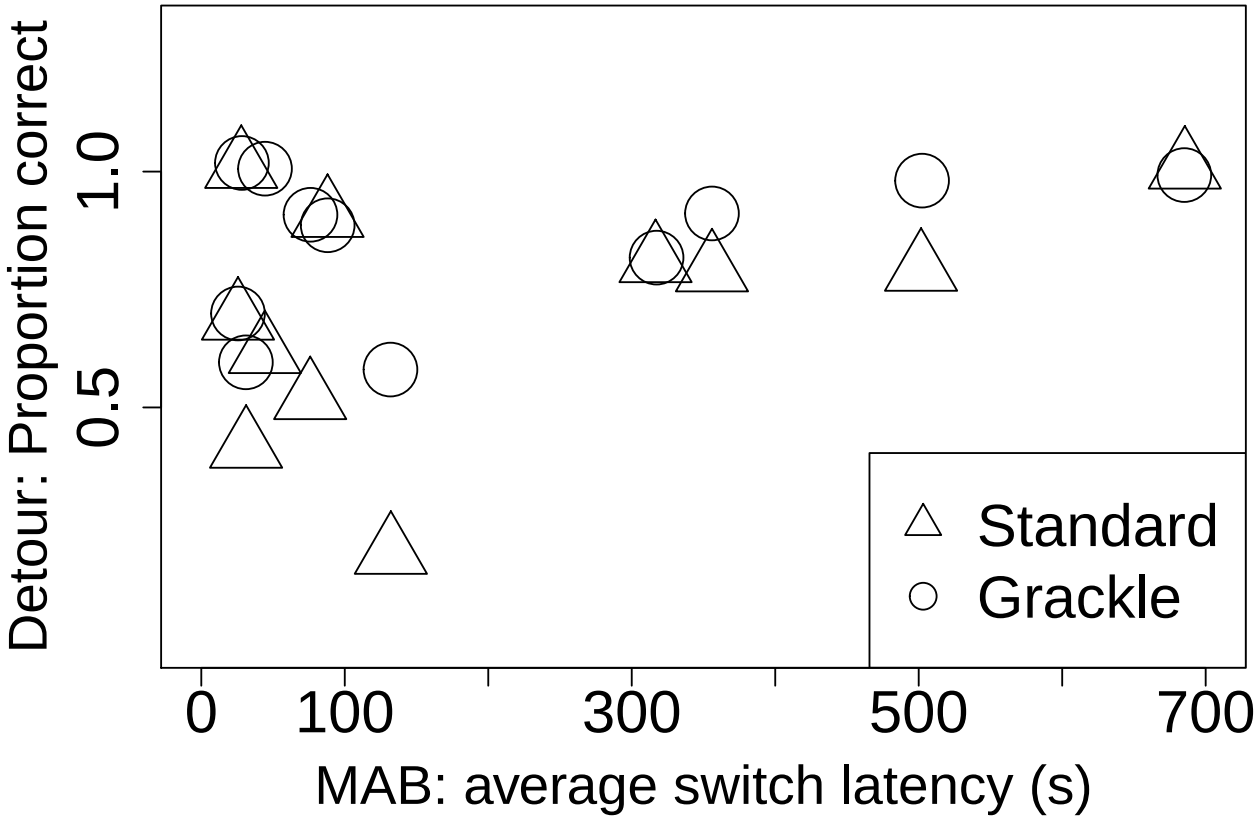

Figure 6. The proportion of detour trials correct per bird (n=11) using the standard calculation method (triangles) or the grackle-specific calculation method (circles) and the average latency to attempt a new locus on the multi-access box (MAB) plastic.

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 4). 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 6).

Table 6. Results from the detour GLMs to determine whether experience with reversal tubes improves detour performance: **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.

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

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

We were not able to conduct the delay of gratification experiment because the grackles never habituated to the apparatuses, therefore the inhibition results come only from the go/no go and detour experiments.

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; Shaw et al. 2015; Brucks et al. 2017; Damerius et al. 2017; DuBois et al. 2018; Ducatez et al. 2019). 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. While 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; 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 and 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.

In conclusion, 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 (Friedman and Miyake 2004; Brucks et al. 2017). 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.

METHODS

A. STATE OF THE DATA

Prior to collecting any data: This preregistration was written.

After data collection had begun (and before any data analysis): This preregistration was submitted to PCI Ecology (Oct 2018) for peer review after starting data collection on the detour task for the pre-reversal subcategory of subjects (for which there was data from one bird). Reviews were received, the preregistration was revised and resubmitted to PCI Ecology (Jan 2019) at which point there was detour data for six birds, data on a few training trials for the delay of gratification task for one bird, and no data from the go/no go experiment. This preregistration passed peer review and was recommended by PCI Ecology in March 2019 (see the review history).

B. PARTITIONING THE RESULTS

We may decide to present the results from different tests in separate papers. NOTE: everything in the preregistration is included in this one manuscript.

C. HYPOTHESIS

If behavioral flexibility requires behavioral inhibition, then individuals that are more behaviorally flexible (indicated by individuals that are faster at functionally changing their behavior when circumstances change), as measured by reversal learning and switching to a different option after one becomes non-functional on a multi-access box, will also be better at inhibiting their responses in three tasks: delayed gratification, go/no go, and detour (Figure 7).

P1: Individuals that are faster to reverse preferences on a reversal learning task and who also have lower latencies to successfully solve new loci after previously solved loci become unavailable (multi-access box) (see flexibility preregistration) will 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) who call it the "cylinder task"), and they will wait longer for higher quality (more preferred) food, but not for higher quantities of food (methods as in Hillemann et al. (2014)). Waiting for higher quality food has been validated as a test of inhibition in birds, while waiting for a higher quantity of food does not appear to measure inhibition (Hillemann et al. (2014)).

P1 alternative 1: If there is no correlation between flexibility measures and performance on the inhibition tasks, this may indicate that the flexibility tasks may not require much inhibition (particularly if the inhibition results are reliable - see *P1 alternative 2*).

P1 alternative 2: If there is no correlation between flexibility measures and performance on the inhibition tasks, this may indicate that the inhibition tasks had low reliability and were therefore too noisy to correlate with flexibility.

P2: If there is no correlation in performance across inhibition tasks, it may indicate that that one or more of these tasks does not measure inhibition, or that they measure different types of inhibition (see Friedman and Miyake (2004)).

P2 alternative: If go/no go task performance strongly correlates with performance on the delayed gratification task, this indicates these two tasks measure the same trait, which therefore validates a inhibition task using a touchscreen (the go/no go task).

P3: If individuals perform well on the detour task and with little individual variation, 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 will experience the detour task before any reversal learning tests.

If this subset performs the same as the others, then previous experience with tubes does not influence detour task performance. If the subset performs worse than the others, this indicates that detour task performance depends on the previous experiences of the individuals tested.

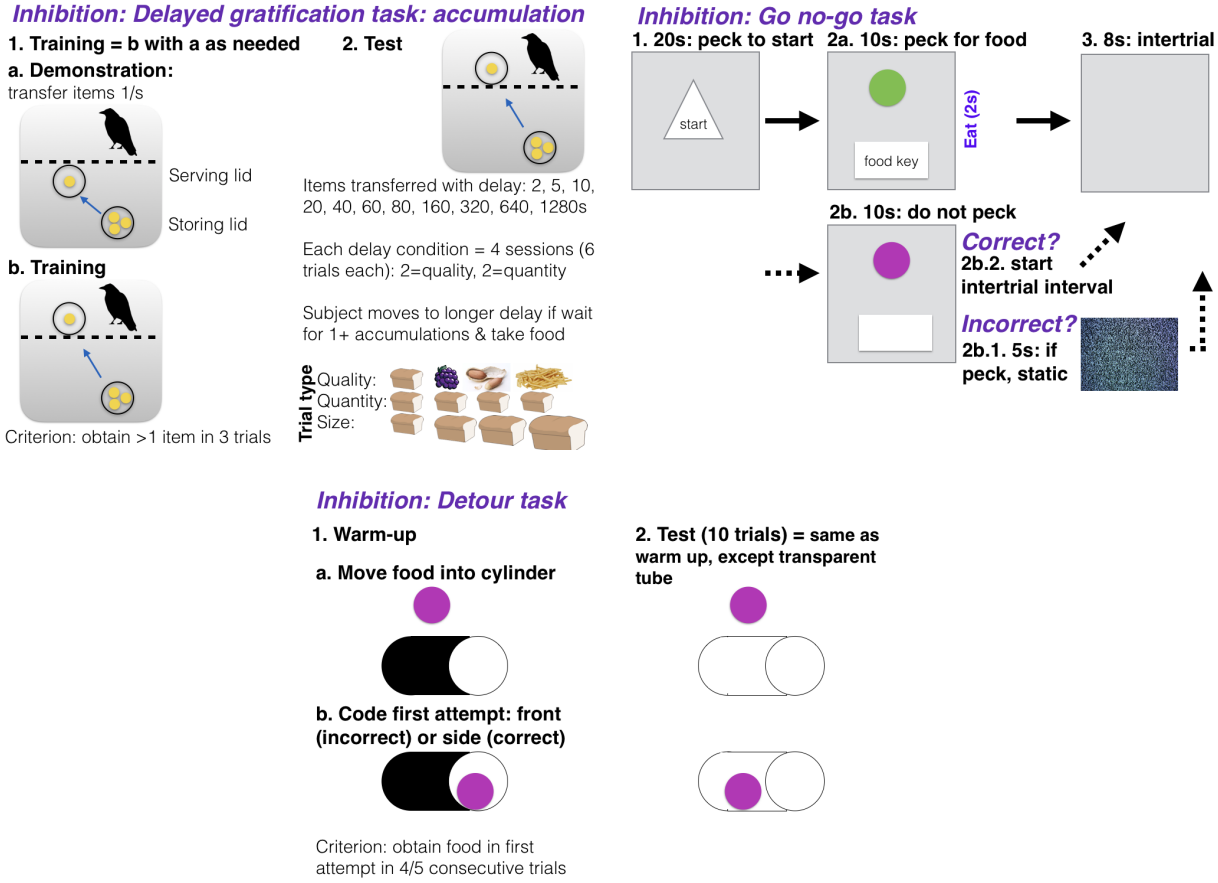

Figure 7a. PLANNED: The experimental designs of the three tasks: delayed gratification, go/no go, and detour (see protocol for details). In the **delay of gratification** task, individuals learn that food items will be transferred by the experimenter from a storing lid (near the experimenter) to a serving lid (near the bird) one at a time, and that they have access to the food in the serving lid from which they can eat at any time: they will have the opportunity to learn that they will have access to more food if they wait longer for the experimenter to transfer food items. Once they pass training (by waiting for more than one food item in three trials), they move on to the test where food items are transferred from the serving to the storing lid with delays ranging from 2-1280 seconds. Birds will be tested on whether will wait for food items that increase in quality (i.e., are more preferred) or increase in quantity (i.e., the same food type accumulates in the serving lid). In the **go/no go** task, after pecking a start key on the touchscreen to show they are attending to a trial, they will be presented with either a green circle or a purple circle (the rewarded circle color is counterbalanced across birds). Pecking the food key while the rewarded colored circle (green in the figure) is on the screen will result in the food hopper rising so the bird can eat food for 2 seconds, after which point the trial ends and the screen goes blank for 8 seconds before starting over again. If the non-rewarded colored circle (purple in the figure) appears on the screen after the start key is pecked, then the correct response is to refrain from pecking the food key for 10 seconds. If the bird succeeds in refraining, the next intertrial interval starts. If the bird fails and pecks the food key while the purple circle is on the screen, then it is given an aversive stimuli for 5 seconds (TV static screen). In the **detour** task, individuals first receive 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 have the opportunity to approach the tube and eat the food. A correct response is when their first approach is 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 pass the warm up, by solving correctly in 4 out of 5 consecutive trials, they move on to the test, which uses the same setup of tube and food except the tube is transparent. The idea is 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.

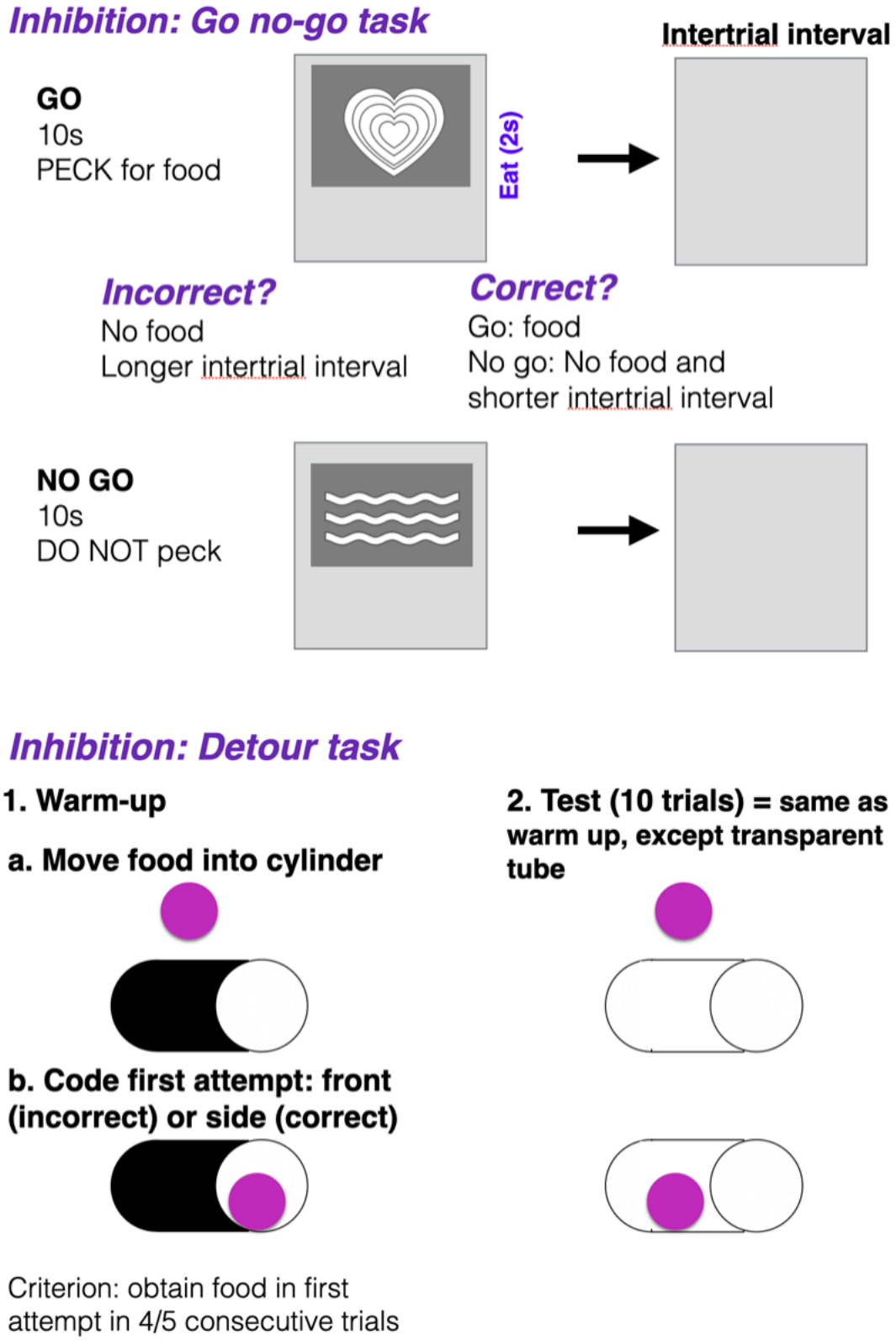

Figure 7b. ACTUAL: The experimental designs of the two tasks that were able to be conducted, go/no go and detour, and the modifications we made to go/no go (see protocol for details on why these changes were made). In the **go/no go** task, the green circle was replaced with a white heart, the purple circle was replaced with white wavy lines, the trial start key was removed, and the food key was removed to make it

so that pecking the go stimulus could result in a food reward. After a correct go response, the food hopper was available for approximately 2-3 s. After a trial ended, the screen went blank for for a variable number of seconds, depending on whether the individual gave a correct (~3 s) or incorrect (~10s) 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 rather than an aversive stimuli (TV static screen). No changes were made to the **detour** task.

D. METHODS

Open materials ADDED Sep 2020: Testing protocols: inhibition for the three inhibition experiments: go/no go, detour, and delay of gratification

Testing protocols: flexibility for the experiments: color tube reversal learning and multi-access box

Open data When the study is complete, the data will be published in the Knowledge Network for Biocomplexity's data repository.

Randomization and counterbalancing P3

Two individuals from each batch will experience the detour task before participating in the flexibility manipulation. These individuals will be randomly selected using the random number generator at https://www.random.org.

P1-P2

For the rest of the individuals (n=6 per batch), the order of the three inhibition tasks will be counterbalanced across birds (using https://www.random.org to randomly assign individuals to one of three experimental orders). 1/3 of the individuals will experience:

1. Delayed gratification task
2. Go/no go task
3. Detour

1/3 of the individuals will experience:

1. Go/no go task
2. Detour
3. Delayed gratification task

1/3 of the individuals will experience:

1. Detour
2. Delayed gratification task
3. Go/no go task

NOTE (Sep 2020): the delayed gratification task was not conducted because the grackles never habituated to the apparatuses. The following birds experienced go/no go first, then detour: Burrito, Chilaquile, Pizza, Yuca, and Pollito.

Delayed gratification

- Food preference test: food will be presented in random combinations over six sessions of 12-15 trials.
- Training trials: The type of demonstration and training trials varied randomly (with more demo trials near the beginning of training), incorporating trials in which food of the same sort accumulated (quantity), food of ascending quality accumulated (quality), and trials in which we added increasingly larger food pieces throughout the trial (size).
- Test: we will test each food quality (low, mid, high) twice in randomized order in each session.

Go/no go

Go and no go trials will be presented randomly with the restriction that no more than four of the same type will occur in a row. The rewarded color will be counterbalanced across birds.

Detour

The side from which the apparatus is baited will be consistent within subjects, but counterbalanced across subjects.

Blinding of conditions during analysis No blinding is involved in this study. NOTE (Sep 2020): interobserver reliability analyses were conducted by hypothesis-blind video coders.

Dependent variables

P1: the more flexible individuals are better at inhibition

1) **Delayed gratification:** Number of food pieces waited for (0-3). A successful wait is defined as waiting for at least one additional piece of food to be added to the serving lid of the three possible additional food items, and accepting at least one of the reward pieces.
2) **Go/no go:**
    a) The number of trials to reach criterion (85% correct) where correct responses involve pecking when the rewarded stimulus is displayed and not pecking when the unrewarded stimulus is displayed, and incorrect responses involve pecking when the unrewarded stimulus is displayed, and not pecking when the rewarded stimulus is displayed
    b) The latency to respond (peck the target key)
3) **Detour:** First approach (physical contact with bill): Correct (to the tube's side opening) or Incorrect (to the front closed area of the tube) (methods as in MacLean et al. (2014)).

One model will be run per dependent variable.

P3: does training improve detour performance?

1) First approach (physical contact): Correct (to the tube's side opening) or Incorrect (to the front closed area of the tube) (methods as in MacLean et al. (2014)).

Independent variables

P1: delayed gratification

1) Food quality or quantity (Quality: High, Med, Low; Quantity: Smaller, Medium, Larger)
2) Trial
3) Delay (2, 5, 10, 20, 40, 60, or 80 seconds)
4) Flexibility 1: **Number of trials to reverse** a preference in the last reversal an individual experienced (reversal learning; 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). See behavioral flexibility preregistration.
5) Flexibility 3: If the number of trials to reverse a preference does not positively correlate with the latency to attempt or solve new loci on the multi-access box (an additional measure of flexibility), then the **average latency to solve** and the **average latency to attempt** a new option on the multi-access box will be additional dependent variables. See behavioral flexibility preregistration.
6) Flexibility 4: This measure is currently being developed and is intended to be a more accurate representation of all of the choices an individual made, as well as accounting for the degree of uncertainty exhibited by individuals as preferences change. If this measure more effectively represents flexibility (determined using a modeled dataset and not the actual data), we may decide to solely rely on this measure and not use flexibility measures 1 through 3. If this ends up being the case, we will modify the code in the analysis plan below to reflect this change.

P1: go/no go Model 2a: number of trials to reach criterion

1) Flexibility 1: Number of trials to reverse a preference in the last reversal an individual experienced (reversal learning; as above)
2) Flexibility 3: If the number of trials to reverse a preference does not positively correlate with the latency to attempt or solve new loci on the multi-access box, then the **average latency to solve** and the **average latency to attempt** a new option on the multi-access box will be additional independent variables (as above).
3) Flexibility 4: This measure is currently being developed and is intended to be a more accurate representation of all the choices an individual made, as well as accounting for the degree of uncertainty exhibited by individuals as preferences change. If this measure more effectively represents flexibility (determined using a modeled dataset and not the actual data), we may decide to solely rely on this measure and not use flexibility measures 1 through 3. If this ends up being the case, we will modify the code in the analysis plan below to reflect this change.

Model 2b: latency to respond

1) Correct or incorrect response
2) Trial
3) Flexibility Condition: control, flexibility manipulation
4) ID (random effect because multiple measures per bird)

NOTE Jul 2020: remove flexibility condition as a variable because, by definition, the birds in the manipulated group were faster to reverse their preferences.

P1: detour

1) Trial

NOTE (Aug 2020): Because the data are analyzed in a GLM, meaning that there is only one row per bird, trial number is not able to be included because it would need to be conducted on multiple rows per bird. Therefore, we removed this independent variable from this analysis.

2) Flexibility 1: Number of trials to reverse a preference in the last reversal an individual experienced (reversal learning; as above)

3) Flexibility 3: If the number of trials to reverse a preference does not positively correlate with the latency to attempt or solve new loci on the multi-access box, then the **average latency to solve** and the **average latency to attempt** a new option on the multi-access box will be additional independent variables (as above).

4) Flexibility 4: This measure is currently being developed and is intended to be a more accurate representation of all of the choices an individual made, as well as accounting for the degree of uncertainty exhibited by individuals as preferences change. If this measure more effectively represents flexibility (determined using a modeled dataset and not the actual data), we may decide to solely rely on this measure and not use flexibility measures 1 through 3. If this ends up being the case, we will modify the code in the analysis plan below to reflect this change.

P3: does training improve detour performance?

1) Condition: pre- or post-reversal learning tests

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) agree with each other to a high degree (Landis and 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 Nov 2019 using random.org. First, 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 using the Cohen's unweighted kappa and replaced Habanero and Diablo with two new randomly chosen grackles (Mole and Chilaquile). 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, 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 random.org. 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). 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.

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)). 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.

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), Faul et al. (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.

Data checking The data will be visually checked to determine whether they are 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 do not appear normally distributed, visually check the residuals. If they are patternless, then assume a normal distribution (Zuur et al. 2009). Detour data look normal, go/no go data are questionable, and both have patternless residuals, therefore we presume normality for both variables.

P1: delayed gratification Assess food preferences: Conduct preference tests between pairs of different foods. Rank food preferences into three categories (High, Medium, Low) in the order of the percentage of times a food was chosen.

Analysis: Generalized Linear Model (GLM; glm function, stats package) with a Poisson distribution and log link, unless the only choices made were 0 (they didn't wait for food) and 1 (they waited for 1 piece of food but not for 2 or 3), in which case we will use a binomial distribution with a logit link. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

To determine our ability to detect actual effects, 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 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)).

These analyses were not conducted because the experiment failed due to the grackles never habituating to the test apparatuses.

P1: go/no go Analysis:

Model 2a: number of trials to reach criterion in the go/no go experiment Generalized Linear Model (GLM; glm function, stats package) with a Poisson distribution and a log link. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

To determine our ability to detect actual effects, 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 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)).

Flexibility comprehensive: 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 A 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 estimate 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 propagate the full *uncertainty* from the reinforcement learning model by directly passing the variables to the linear model within a single large *stan* model. We include 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.

Model 2b: latency to respond in the go/no go experiment A Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) will be used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01 after lag 0; (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.

NOTE (Sep 2020): 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.

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 $f^2 = 0{,}32$

err prob = 0,05

Power (1- err prob) = 0,7

Number of predictors = 3

Output:

Noncentrality parameter = 10,2400000

Critical F = 2,9466853

Numerator df = 3

Denominator df = 28

Total sample size = 32

Actual power = 0,7061592

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)).

P1: detour Analysis: Generalized Linear Model (GLM; glm function, stats package) with a binomial distribution and a logit link. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

See the protocol for the power analyses for Model 2b above for the rough estimation of our ability to detect actual effects with this model.

Flexibility comprehensive: We again repeat the 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 include 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 use a binomial likelihood as the outcome distribution and a logit link function (see section 2a for full data preparation and analysis script).

P1 alternative 2: are inhibition results reliable? The reliability of the inhibition tests will be calculated using Cronbach's Alpha (as in Friedman and Miyake (2004); R package: psy (Falissard 2012), function: cronbach), which is indicated by alpha in the output.

NOTE (Sep 2020): when we tried to run this code we discovered that this is not the appropriate test to run on our experimental design to test the internal validity of the experiment (e.g., does this test actually measure what we think it does). To test internal validity, we would need to change the experimental design, which was not the goal of our current study. Therefore, we did not conduct this analysis.

P2: correlation across inhibition tasks See analysis description for P1 alternative 2.

P3: does training improve detour performance? Analysis: Generalized Linear Model (GLM; glm function, stats package) with a binomial distribution and a logit link. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

To determine our ability to detect actual effects, 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 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)).

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. See the State of the Data for a description of the analysis changes we made.

F. PLANNED SAMPLE

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

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). The minimum sample size will be 16, however we expect to be able to test up to 32 grackles.

Data collection stopping rule

We will stop testing birds once we have completed two full aviary seasons (likely in March 2020). NOTE: the two full aviary seasons concluded in May 2020. NOTE (Sep 2020): data collection stopped after two full aviary seasons in May 2020.

G. 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)

H. 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.

I. FUNDING

This research is 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.

J. CONFLICT OF INTEREST DISCLOSURE

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

K. ACKNOWLEDGEMENTS

We 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.

L. REFERENCES

Addessi E, Paglieri F, Beran MJ, Evans TA, 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.

Auersperg AMI, Bayern AMP von, Gajdon GK, 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. doi:10.1371/journal.pone.0020231. [accessed 2017 May 15]. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020231.

Beran M. 2018. Self-control in animals and people. Academic Press.

Beran MJ. 2015. The comparative science of 'self-control': What are we talking about? Frontiers in Psychology. 6:51.

Blaisdell A, Johnson-Ulrich Z, Bergeron L, Rowney C, Seitz B, McCune KB, Folsom M, MacPherson M, Logan C. 2021. Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles. In principle acceptance by PCI Ecology of the version on 31 Jan 2019. doi:10.31234/osf.io/z4p6s. psyarxiv.com/z4p6s.

Bond AB, Kamil AC, Balda RP. 2007. Serial reversal learning and the evolution of behavioral flexibility in three species of North American corvids (*Gymnorhinus cyanocephalus*, *Nucifraga columbiana*, *Aphelocoma californica*). Journal of Comparative Psychology. 121(4):372–379. doi:10.1037/0735-7036.121.4.372.

Boogert NJ, Anderson RC, Peters S, Searcy WA, 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. doi:10.1016/j.anbehav.2011.03.004. [accessed 2017 May 15]. http://www.sciencedirect.com/science/article/pii/S0003347211001011.

Bray EE, MacLean EL, Hare BA. 2014. Context specificity of inhibitory control in dogs. Animal Cognition. 17(1):15–31.

Bray EE, MacLean EL, Hare BA. 2014. Context specificity of inhibitory control in dogs. Animal Cognition. 17(1):15–31.

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

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

Carter AJ, Feeney WE, Marshall HH, 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 edn.

Damerius LA, Graber SM, Willems EP, Schaik CP van. 2017. Curiosity boosts orang-utan problem-solving ability. Animal Behaviour. 134:57–70.

Deaner RO, Schaik CP van, 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. doi:10.1177/147470490600400114. [accessed 2017 May 15]. http://dx.doi.org/10.1177/147470490600400114.

Diamond A. 2013. Executive functions. Annual review of psychology. 64:135–168.

DuBois AL, Nowicki S, Peters S, Rivera-Cáceres KD, Searcy WA. 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.

Falissard B. 2012. Psy: Various procedures used in psychometry. https://CRAN.R-project.org/package=psy.

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 NP, Miyake A. 2004. The relations among inhibition and interference control functions: A latent-variable analysis. Journal of experimental psychology: General. 133(1):101.

Friedman NP, Miyake A. 2004. The relations among inhibition and interference control functions: A latent-variable analysis. Journal of experimental psychology: General. 133(1):101.

Ghahremani DG, Monterosso J, Jentsch JD, Bilder RM, Poldrack RA. 2009. Neural components underlying behavioral flexibility in human reversal learning. Cerebral cortex. 20(8):1843–1852.

Griffin AS, 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 JD. 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 EJ, Paul ES, Mendl M. 2004. Animal behaviour: Cognitive bias and affective state. Nature. 427(6972):312–312.

Hillemann F, Bugnyar T, Kotrschal K, Wascher CA. 2014. Waiting for better, not for more: Corvids respond to quality in two delay maintenance tasks. Animal behaviour. 90:1–10.

Homberg JR, Pattij T, Janssen MC, Ronken E, De Boer SF, Schoffelmeer AN, Cuppen E. 2007. Serotonin transporter deficiency in rats improves inhibitory control but not behavioural flexibility. European Journal of Neuroscience. 26(7):2066–2073.

Horik JO van, Langley EJ, Whiteside MA, Laker PR, Beardsworth CE, Madden JR. 2018. Do detour tasks provide accurate assays of inhibitory control? Proceedings of the Royal Society B: Biological Sciences. 285(1875):20180150.

Isaksson E, Urhan AU, 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 JR, Koch GG. 1977. The measurement of observer agreement for categorical data. biometrics.:159–174.

Liu Y, Day LB, Summers K, Burmeister SS. 2016. Learning to learn: Advanced behavioural flexibility in a poison frog. Animal Behaviour. 111:167–172.

Logan C. 2016. Behavioral flexibility and problem solving in an invasive bird. PeerJ. 4:e1975.

Logan CJ, MacPherson M, Rowney C, Bergeron L, Seitz B, Blaisdell A, Folsom M, Johnson-Ulrich Z, McCune K. 2019. Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context? In principle acceptance by PCI Ecology of the version on 26 Mar 2019. http://corinalogan.com/Preregistrations/g_flexmanip.html.

Logan C, McCune K, MacPherson M, Johnson-Ulrich Z, Rowney C, Seitz B, Blaisdell A, Deffner D, Wascher C. 2020. Great-tailed grackle inhibition data. Knowledge Network for Biocomplexity. Data package. doi:10.5063/M043S3.

MacLean EL, Hare B, Nunn CL, Addessi E, Amici F, Anderson RC, Aureli F, Baker JM, Bania AE, Barnard AM, et al. 2014. The evolution of self-control. Proceedings of the National Academy of Sciences. 111(20):E2140–E2148.

Manrique HM, Völter CJ, Call J. 2013. Repeated innovation in great apes. Animal Behaviour. 85(1):195–202. doi:10.1016/j.anbehav.2012.10.026. [accessed 2017 May 23]. http://www.sciencedirect.com/science/article/pii/S0003347212004861.

McElreath R. 2016. Statistical rethinking: A bayesian course with examples in r and stan. CRC Press. http://xcelab.net/rm/statistical-rethinking/.

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. doi:10.1098/rsfs.2016.0121. [accessed 2017 May 29]. http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2016.0121.

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

Nigg JT. 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.

Nigg JT. 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. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

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

Seitz BM, McCune K, MacPherson M, Bergeron L, Blaisdell AP, Logan CJ. 2021. Using touchscreen equipped operant chambers to study animal cognition. Benefits, limitations, and advice. PloS one. 16(2):e0246446.

Shaw RC, Boogert NJ, Clayton NS, Burns KC. 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. doi:10.1098/rstb.2012.0214. [accessed 2017 May 24]. http://rstb.royalsocietypublishing.org/content/367/1603/2773.

Zuur AF, Ieno EN, Saveliev AA. 2009. Mixed effects models and extensions in ecology with r. Springer.