Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context?

Dr. Corina Logan (Max Planck Institute for Evolutionary Anthropology, corina_logan@eva.mpg.de), Ca

2020-06-15

```
#Make code wrap text so it doesn't go off the page when Knitting to PDF
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

###ABSTRACT

This is one of the first studies planned for our long-term research on the role of behavioral flexibility in rapid geographic range expansions. Behavioral flexibility, the ability to adapt behavior to new circumstances, is thought to play an important role in a species' ability to successfully adapt to new environments and expand its geographic range (e.g., (Lefebvre et al. 1997), (Griffin and Guez 2014), (Chow, Lea, and Leaver 2016), (Sol and Lefebvre 2000), (Sol, Timmermans, and Lefebvre 2002), (Sol et al. 2005), (Sol et al. 2007)). However, behavioral flexibility is rarely directly tested in species in a way that would allow us to determine how it works and how we can make predictions about a species' ability to adapt their behavior to new environments. We use great-tailed grackles (a bird species) as a model to investigate this question because they have rapidly expanded their range into North America over the past 140 years ((Wehtje 2003), (Peer 2011)). We aim to manipulate grackle behavioral flexibility (color tube reversal learning) to determine whether their flexibility is generalizable across contexts (touch screen reversal learning and multi-access box), whether it is repeatable within individuals and across contexts, and what learning strategies they employ. Results will allow us to understand more about what flexibility is and how it works, and validate whether a touch screen measures the same ability as the color tubes (thus facilitating faster testing that can be conducted in the wild).

###A. STATE OF THE DATA

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

This preregistration was submitted to PCI Ecology for peer review (July 2018), we received the first round of peer reviews a few days before data collection began (Sep 2018), we revised and resubmitted after data collection had started (Feb 2019) and it passed peer review (Mar 2019) before any of the planned analyses had been conducted. See the peer review history at PCI Ecology.

###B. PARTITIONING THE RESULTS

We may present the different hypotheses in separate papers.

###C. HYPOTHESES

H1: Behavioral flexibility, as measured by reversal learning using colored tubes, is manipulatable.

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

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

H2: Manipulating behavioral flexibility (improving reversal learning speed through serial reversals using colored tubers) improves flexibility (rule learning and/or switching) and problem solving in a new context (multi-access box and serial reversals on a touch screen).

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

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

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

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

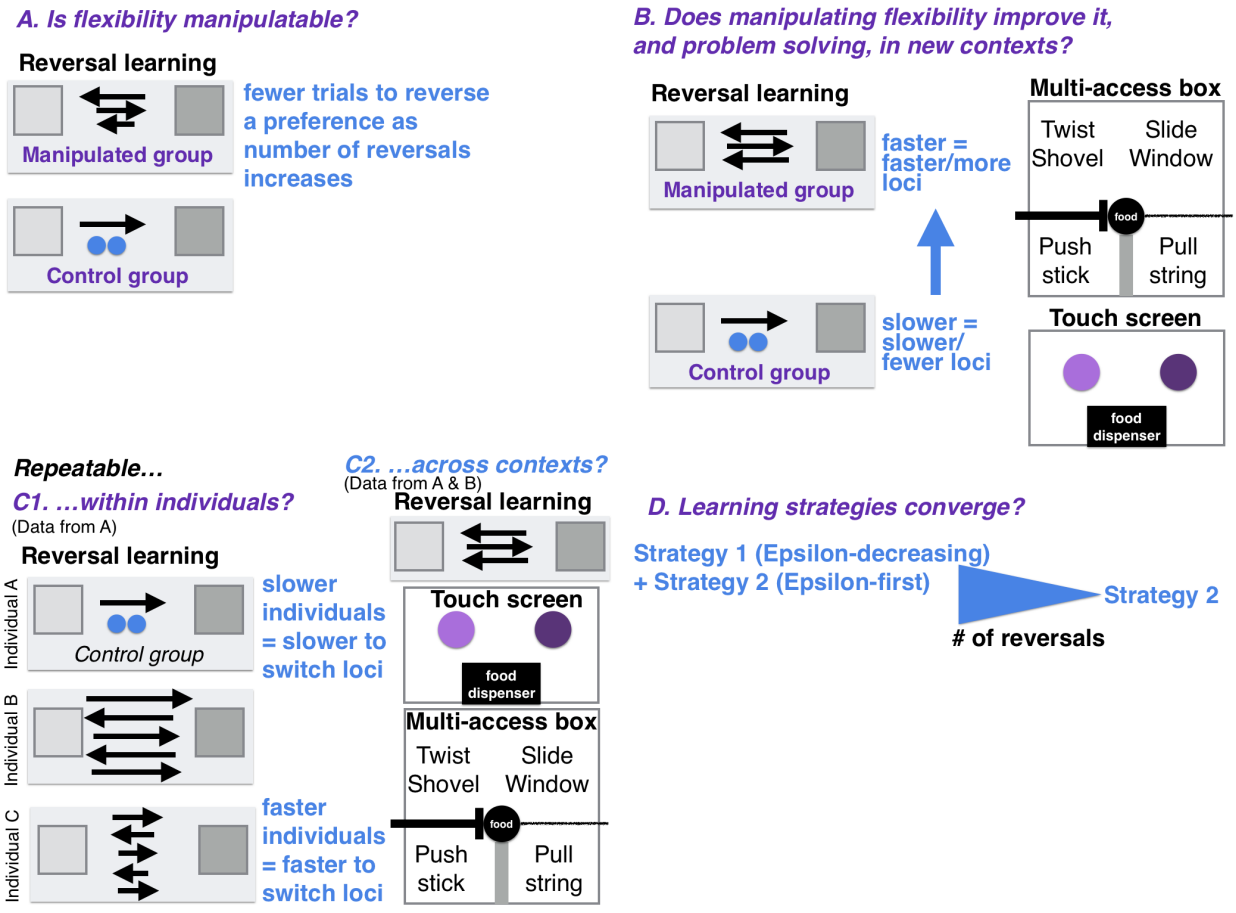

Figure 1: Figure 1. A visual illustration of Hypothesis 1 (A), Hypothesis 2 (B), Hypothesis 3 (C1 and C2), and Hypothesis 4 (D).

H3a: Behavioral flexibility within a context is repeatable within individuals.

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

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

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

H3b: The consistency of behavioral flexibility in individuals across contexts (context 1=reversal learning on colored tubes, context 2=multi-access box, context 3=reversal learning on touch screen) indicates their ability to generalize across contexts.

Individual consistency of behavioral flexibility is defined as the number of trials to reverse a color preference being strongly positively correlated within individuals with the latency to solve new loci on the multi-access box and with the number of trials to reverse a color preference on a touch screen (total number of touch screen reversals = 5 per bird).

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

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

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

H4: Individuals should converge on an epsilon-first learning strategy (learn the correct choice after one trial) as they progress through serial reversals.

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

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

###D. METHODS

Planned Sample

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

Sample size rationale

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

Data collection stopping rule

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

####**Open materials**

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

Testing protocols for all three experiments: color tube reversal learning, multi-access box, and touch screen reversal learning

####**Randomization and counterbalancing**

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

####**Blinding of conditions during analysis**

No blinding is involved in this study.

####**Dependent variables**

P1-P3

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

P2 alternative 2: additional analysis: latency and motor diversity

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

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

Number of trials to solve a new loci on the multi-access box

P4: learning strategies

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

####**Independent variables**

####*P1: reversal speed gets faster with serial reversals*

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

####*P2: serial reversals improve rule switching & problem solving*

1) Average latency to attempt to solve a new locus after solving a different locus
2) Average latency to solve a new locus after solving a different locus
3) Total number of loci solved

4) Experimental group (manipulated=multiple reversals with color stimuli; control=one reversal plus equalized experience making choices where both are the same color and both contain a reward)

5) Batch (random effect because multiple batches included in the analysis). Note: batch is a test cohort, consisting of 8 birds being tested simultaneously

####P2 alternative 2: additional analysis: latency and motor diversity

1) Number of trials to reverse a preference in the last reversal that individual participated in

2) Motor diversity: the number of different motor actions used when attempting to solve the multi-access box

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

####P3a: repeatable within individuals within a context (reversal learning)

1) Reversal number

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

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

1) Trial number

2) Latency from the beginning of the trial to when they make a choice

3) Minutes since maintenance diet was removed from the aviary

4) Cumulative number of rewards from previous trials on that day

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

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

####P3b: repeatable across contexts

1) Reversal number

2) Condition (color tubes, multi-access box, touch screen)

3) Latency to solve a new locus

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

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

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

####P4: serial reversal learning strategy

1) Trial number

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

###E. ANALYSIS PLAN

We do not plan to **exclude** any data. When **missing data** occur, the existing data for that individual will be included in the analyses for the tests they completed. Analyses will be conducted in R (current version 3.6.3; 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.

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

####Data checking

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

####Determining the threshold: How many reversals are enough?

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

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

####Revising the choice criterion and the criterion to pass the control condition

Choice criterion: At the beginning of the second bird's initial discrimination in the reversal learning color tube experiment (October 2018), we revised the criterion for what counts as a choice from A) the bird's head needs to pass an invisible line on the table that ran perpendicular to the the tube opening to B) the bird needs to bend its body or head down to look in the tube. Criterion A resulted in birds making more choices than the number of learning opportunities they were exposed to (because they could not see whether there was food in the tube unless they bent their head down to look in the tube) and appeared

to result in slower learning. It is important that one choice equals one learning opportunity, therefore we revised the choice criterion to the latter. Anecdotally, this choice matters because the first three birds in the experiment (Tomatillo, Chalupa, and Queso) learned faster than the pilot birds (Empanada and Fajita) in their initial discriminations and first reversals. Thus, it was an important change to make at the beginning of the experiment.

Criterion to pass the control condition: Before collecting experimental data, we set the number of trials experienced by the birds in the control group as 1100 because this is how many trials it would have taken the pilot bird in the manipulated group, Fajita, to pass serial reversals 2-17 according to our revised serial reversal passing criterion. However, after 25 and 17 days (after Tomatillo and Queso's first reversals, respectively) of testing the first two individuals in the control group it became apparent that 1100 trials is impractical given the time constraints for how long we are permitted to keep each bird temporarily in captivity and would prevent birds from completing the test battery before their release. Additionally, after revising the choice criterion, it was going to be likely that birds in the manipulated group would require fewer than 1100 trials to meet the serial reversal passing criterion. Therefore, reducing the number of trials control birds experience would result in a better match of experience with birds in the manipulated group. On 2 November 2018 we set the number of trials control birds experience after their first (and only) reversal to the number of trials it requires the first bird in the manipulated group to pass (the first bird has not passed yet, therefore we do not yet know what this number is). After more individuals in the manipulated group pass, we will update this number to the average number of trials to pass.

####P1: negative relationship between the number of trials to reverse a preference and the number of reversals?

Analysis: A Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) will be used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01; (Hadfield 2010)), and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

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

```
seriald <- read.csv("/Users/corina/GTGR/data/data_reverse.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutput <- simulateResiduals(fittedModel = glmer(TrialsToReverse ~
    ReverseNumber + (1 | ID) + (1 | Batch), family = poisson,
    data = seriald), n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(ReverseNumber, simulationOutput$scaledResiduals)  #plot the residuals against other predi

# GLMM
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0)), G = list(G1 = list(V = 1,
```

```
    nu = 0), G2 = list(V = 1, nu = 0)))
serial <- MCMCglmm(TrialsToReverse ~ ReverseNumber, random = ~ID +
    Batch, family = "poisson", data = seriald, verbose = F, prior = prior,
    nitt = 13000, thin = 10, burnin = 3000)
summary(serial)
# autocorr(serial$Sol) #Did fixed effects converge?
# autocorr(serial$VCV) #Did random effects converge?

# AIC calculation
library(MuMIn)
options(na.action = "na.fail")
base1 <- dredge(MCMCglmm(TrialsToReverse ~ ReverseNumber, random = ~ID +
    Batch, family = "poisson", data = seriald, verbose = F, prior = prior,
    nitt = 13000, thin = 10, burnin = 3000))
library(knitr)
kable(base1, caption = "Table 2: Model selection output.")
```

####P2: serial reversal improves rule switching & problem solving

Analysis: Because the independent variables could influence each other, we will analyze them in a single model. A Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) will be used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01; (Hadfield 2010)), and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

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

Input:

Effect size $f^2 = 0.41$

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)).

```
improve <- read.csv("/Users/corina/GTGR/data/data_reversemulti.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOut <- simulateResiduals(fittedModel = glmer(TrialsToReverse ~
    Condition + AvgLatencySolveNewLoci + AvgLatencyAttemptNewLoci +
        TotalLoci + (1 | Batch), family = poisson, data = improve),
    n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOut$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity in
testDispersion(simulationOut)  #if under- or over-dispersed, then p-value<0.05, but then check the disp
testZeroInflation(simulationOut)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOut)  #check for heteroscedasticity ('a systematic dependency of the dispersio
plot(simulationOut)  #...there should be no pattern in the data points in the right panel
plotResiduals(Condition, simulationOut$scaledResiduals)  #plot the residuals against other predictors (

# GLMM
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0)), G = list(G1 = list(V = 1,
    nu = 0), G2 = list(V = 1, nu = 0)))
imp <- MCMCglmm(TrialsToReverse ~ Condition + AvgLatencySolveNewLoci +
    AvgLatencyAttemptNewLoci + TotalLoci, random = ~Batch, family = "poisson",
    data = improve, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000)
summary(imp)
# autocorr(imp$Sol) #Did fixed effects converge?
# autocorr(imp$VCV) #Did random effects converge?
```

P2 alternative 2: additional analysis: latency and motor diversity

A Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) will be used with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01; (Hadfield 2010)), and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

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

Input:

Effect size $f^2 = 0.27$

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)).

```
# Latency to attempt to solve a new locus
diversity <- read.csv("/Users/corina/GTGR/data/data_reversemulti.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutp <- simulateResiduals(fittedModel = glmer(TrialsToSolveNewLoci ~
    TrialsToReverseLast + NumberMotorActionsMulti + (1 | ID),
    family = poisson, data = diversity), n = 250)  #250 simulations, but if want higher precision chang
simulationOutp$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity in
testDispersion(simulationOutp)  #if under- or over-dispersed, then p-value<0.05, but then check the dis
testZeroInflation(simulationOutp)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutp)  #check for heteroscedasticity ('a systematic dependency of the dispersi
plot(simulationOutp)  #...there should be no pattern in the data points in the right panel
plotResiduals(NumberMotorActionsMulti, simulationOutp$scaledResiduals)  #plot the residuals against oth
plotResiduals(TrialsToReverseLast, simulationOutp$scaledResiduals)

# GLMM
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0)), G = list(G1 = list(V = 1, nu = 0)))
div <- MCMCglmm(TrialsToSolveNewLoci ~ TrialsToReverseLast +
    NumberMotorActionsMulti, random = ~ID, family = "poisson",
    data = diversity, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000)
summary(div)
# autocorr(div$Sol) #Did fixed effects converge?
# autocorr(div$VCV) #Did random effects converge?

# AIC calculation
library(MuMIn)
options(na.action = "na.fail")
base1 <- dredge(MCMCglmm(TrialsToSolveNewLoci ~ TrialsToReverseLast +
    NumberMotorActionsMulti, random = ~ID, family = "poisson",
    data = diversity, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000))
library(knitr)
kable(base1, caption = "Table 5: Model selection output.")
```

```
# Latency to solve a new locus
diversity <- read.csv("/Users/corina/GTGR/data/data_reversemulti.csv",
    header = T, sep = ",", stringsAsFactors = F)
```

```
# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutpu <- simulateResiduals(fittedModel = glmer(TrialsToAttemptNewLoci ~
    TrialsToReverseLast + NumberMotorActionsMulti + (1 | ID),
    family = poisson, data = diversity), n = 250)  #250 simulations, but if want higher precision chang
simulationOutpu$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity i
testDispersion(simulationOutpu)  #if under- or over-dispersed, then p-value<0.05, but then check the di
testZeroInflation(simulationOutpu)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutpu)  #check for heteroscedasticity ('a systematic dependency of the dispers
plot(simulationOutpu)  ##...there should be no pattern in the data points in the right panel
plotResiduals(NumberMotorActionsMulti, simulationOutpu$scaledResiduals)  #plot the residuals against ot
plotResiduals(TrialsToReverseLast, simulationOutpu$scaledResiduals)

# GLMM
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0)), G = list(G1 = list(V = 1, nu = 0)))
div <- MCMCglmm(TrialsToAttemptNewLoci ~ TrialsToReverseLast +
    NumberMotorActionsMulti, random = ~ID, family = "poisson",
    data = diversity, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000)
summary(div)
# autocorr(div$Sol) #Did fixed effects converge?
# autocorr(div$VCV) #Did random effects converge?

# AIC calculation
library(MuMIn)
options(na.action = "na.fail")
base1 <- dredge(MCMCglmm(TrialsToAttemptNewLoci ~ TrialsToReverseLast +
    NumberMotorActionsMulti, random = ~ID, family = "poisson",
    data = diversity, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000))
library(knitr)
kable(base1, caption = "Table 5: Model selection output.")
```

####P3a: repeatable within individuals within a context (reversal learning)

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

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

(n=32). The protocol of the power analysis is here:

Input:

Effect size $f^2 = 0.21$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 1

Output:

Noncentrality parameter = 6.7200000

Critical F = 4.1708768

Numerator df = 1

Denominator df = 30

Total sample size = 32

Actual power = 0.7083763

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

```
serial2 <- read.csv("/Users/corina/GTGR/data/data_reverse.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutput <- simulateResiduals(fittedModel = glmer(TrialsToReverse ~
    ReverseNumber + (1 | ID), family = poisson, data = serial2),
    n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(ReverseNumber, simulationOutput$scaledResiduals)  #plot the residuals against other predi

# GLMM
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0)), G = list(G1 = list(V = 1,
    nu = 0)))
serial <- MCMCglmm(TrialsToReverse ~ ReverseNumber, random = ~ID,
    family = "poisson", data = serial2, verbose = F, prior = prior,
    nitt = 13000, thin = 10, burnin = 3000)
summary(serial)
# autocorr(serial$Sol) #Did fixed effects converge?
# autocorr(serial$VCV) #Did random effects converge?

# REPEATABILITY In MCMCglmm, the latent scale adjusted
# repeatability and its credible interval can simply be
# obtained by:
# serial$VCV[,ID]/(serial$VCV[,ID]+serial$VCV[,units]) -
# advice from Maxime Dahirel
```

```
repeata <- serial$VCV[, "ID"]/(serial$VCV[, "ID"] + serial$VCV[,
    "units"])  #latent scale adjusted repeatability and its credible interval
mean(repeata)  #0.79 variance
var(repeata)  #0.15 variance
posterior.mode(repeata)  #0.99879
HPDinterval(repeata, 0.95)  #2.6e-14 to 0.99999

# Repeatability on the data/observed scale (accounting for
# fixed effects) code from Supplementary Material S2 from
# Villemereuil et al. 2018 J Evol Biol
vf <- sapply(1:nrow(serial[["Sol"]]), function(i) {
    var(predict(serial, it = i))
})  #estimates for each iteration of the MCMC

repeataF <- (vf + serial$VCV[, "ID"])/(vf + serial$VCV[, "ID"] +
    serial$VCV[, "units"])  #latent scale adjusted + data scale
posterior.mode(repeataF)  #1.0
HPDinterval(repeataF, 0.95)  #0.9999 to 1.0

# Now compare with the raw repeatability: null model
serialraw <- MCMCglmm(TrialsToReverse ~ 1, random = ~ID, family = "poisson",
    data = serial2, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000)
summary(serialraw)

repeataraw <- serialraw$VCV[, "ID"]/(serialraw$VCV[, "ID"] +
    serialraw$VCV[, "units"])  #latent scale adjusted repeatability and its credible interval
posterior.mode(repeata)  #0.99879
HPDinterval(repeata, 0.95)  #2.6e-14 to 0.99999
```

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

Analysis: Because the independent variables could influence each other or measure the same variable, I will analyze them in a single model: Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) with a binomial distribution (called categorical in MCMCglmm) and logit link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01; (Hadfield 2010)), and adjust parameters if necessary. The contribution of each independent variable will be evaluated using the Estimate in the full model.

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

Input:

Effect size $f^2 = 0.31$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 4

Output:

Noncentrality parameter = 11.4700000

Critical F = 2.6684369

Numerator df = 4

Denominator df = 32

Total sample size = 37

Actual power = 0.7113216

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

```
rr <- read.csv("/Users/corina/GTGR/data/data_reverseraw.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutput <- simulateResiduals(fittedModel = glmer(CorrectChoice ~
    Trial + LatencyToChoose + MinSinceFoodRemoved + NumberRewardsFromPrevTrials +
        (1 | ID) + (1 | Batch), family = binomial, data = rr),
    n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(LatencyToChoose, simulationOutput$scaledResiduals)  #plot the residuals against other pre

# GLMM - Is trial the main independent variable associated
# with learning performance (CorrectChoice) or are other
# variables associated with performance, including motivation
# and hunger?
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0), R3 = list(V = 1, nu = 0), R4 = list(V = 1, nu = 0)),
    G = list(G1 = list(V = 1, nu = 0), G2 = list(V = 1, nu = 0)))

rr1 <- MCMCglmm(CorrectChoice ~ Trial + LatencyToChoose + MinSinceFoodRemoved +
    NumberRewardsFromPrevTrials, random = ~ID + Batch, family = "categorical",
    data = rr, verbose = F, prior = prior, nitt = 13000, thin = 10,
    burnin = 3000)
summary(rr1)
autocorr(rr1$Sol)  #Did fixed effects converge?
autocorr(rr1$VCV)  #Did random effects converge?
```

####P3b: individual consistency across contexts

Analysis: Do those individuals that are faster to reverse a color preference also have lower latencies to switch to new options on the multi-access box? Do those individuals that are faster to reverse a color preference also have lower latencies to switch to new options on the multi-access box? A Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield 2010)) will be used with a

Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield 2014). We will ensure the GLMM shows acceptable convergence (lag time autocorrelation values <0.01; (Hadfield 2010)), and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

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

Input:

Effect size $f^2 = 0.21$

err prob = 0.05

Power (1- err prob) = 0.7

Number of predictors = 1

Output:

Noncentrality parameter = 6.7200000

Critical F = 4.1708768

Numerator df = 1

Denominator df = 30

Total sample size = 32

Actual power = 0.7083763

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

```
improve1 <- read.csv("/Users/corina/GTGR/data/data_reversemulti.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutput <- simulateResiduals(fittedModel = glmer(TrialsToSolveNewLoci ~
    Condition + ReversalNumber + TrialsToReverseT + TrialsToReverse +
        (1 | ID), family = poisson, data = improve1), n = 250)  #250 simulations, but if want higher pr
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(ReverseNumber, simulationOutput$scaledResiduals)  #plot the residuals against other predi

# GLMM color reversal tubes compared with multi-access box
# and reversal on the touchscreen
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0)), G = list(G1 = list(V = 1, nu = 0)))
```

```
rm <- MCMCglmm(TrialsToSolveNewLoci ~ Condition * ReversalNumber *
    TrialsToReverseT * TrialsToReverse, random = ~ID, family = "poisson",
    data = improve1, verbose = F, prior = prior, nitt = 13000,
    thin = 10, burnin = 3000)
summary(rm)
# autocorr(rm$Sol) #Did fixed effects converge?
# autocorr(rm$VCV) #Did random effects converge?
```

####P4: learning strategies (for birds in the manipulated group only)

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

From Logan (2016) (emphasis added):

> The following equations refer to the different phases involved in each strategy:
>
> Equation 1 (exploration phase):
>
> $$\epsilon N$$
>
> Equation 2 (exploitation phase):
>
> $$(1-\epsilon)N$$
>
> N is the number of trials given, and epsilon,
>
> $$\epsilon$$
>
> , represents the subject's uncertainty about the location of the reward, starting at complete uncertainty ($\epsilon = 1$) at the beginning of the experiment and decreasing rapidly as individuals gain experience with the task (exploration phase where the rewarded [option] is chosen below or at chance levels) and switch to the exploitative phase (the rewarded [option] is chosen significantly above chance levels). Because the [subjects] needed to learn the rules of the task, they necessarily had an exploration phase. The **epsilon-first strategy** involves an exploration phase followed by an entirely exploitative phase. The optimal strategy overall would be to explore one color in the first trial and the other color in the second trial, and then switch to an exploitative strategy (choose the rewarded [option] significantly above chance levels). In this case there would be no pattern [in the learning curve] in the choices [during] the exploration phase because it would consist of sampling each [option] only once. In the **epsilon-decreasing strategy**, subjects would start by making some incorrect choices and then increase their choice of the rewarded [option] gradually as their uncertainty decreases until they choose the rewarded [option] significantly above chance levels. In this case, a linear pattern emerges [in the learning curve] during the exploration phase.

We will then quantitatively determine to what degree each bird used the exploration versus exploitation strategy using methods in [federspiel2017adjusting] by calculating the number of 20-trial blocks where birds were choosing "randomly" (6-14 correct choices; called sampling blocks; akin to the exploration phase in our preregistration) was divided by the total number of blocks to reach criterion per bird. This ratio was also calculated for "acquisition" blocks where birds made primarily correct choices (15-20 correct choices; akin to the exploitation phase in our preregistration). These ratios, calculated for each bird for their serial reversals, quantitatively discern the exploration from the exploitation phases.

```
strat <- read.csv("/Users/corina/GTGR/data/data_strategy.csv",
    header = T, sep = ",", stringsAsFactors = F)

# DATA CHECKING
library(DHARMa)
library(lme4)
```

```
simulationOutput <- simulateResiduals(fittedModel = glmer(ratioExplore ~
    ReversalNumber + (1 | ID), family = poisson, data = strat),
    n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(ReverseNumber, simulationOutput$scaledResiduals)  #plot the residuals against other predi

# GLMM explore strategy (6-14 correct choices/20 trial block)
# ratio
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0)), G = list(G1 = list(V = 1, nu = 0)))
st <- MCMCglmm(ratioExplore ~ ReversalNumber, random = ~ID, family = "poisson",
    data = strat, verbose = F, prior = prior, nitt = 13000, thin = 10,
    burnin = 3000)
summary(st)
# autocorr(st$Sol) #Did fixed effects converge?
# autocorr(st$VCV) #Did random effects converge?

# DATA CHECKING
library(DHARMa)
library(lme4)
simulationOutput <- simulateResiduals(fittedModel = glmer(ratioExploit ~
    ReversalNumber + (1 | ID), family = poisson, data = strat),
    n = 250)  #250 simulations, but if want higher precision change n>1000
simulationOutput$scaledResiduals  #Expect a flat distribution of the overall residuals, and uniformity
testDispersion(simulationOutput)  #if under- or over-dispersed, then p-value<0.05, but then check the d
testZeroInflation(simulationOutput)  #compare expected vs observed zeros, not zero-inflated if p<0.05
testUniformity(simulationOutput)  #check for heteroscedasticity ('a systematic dependency of the disper
plot(simulationOutput)  #...there should be no pattern in the data points in the right panel
plotResiduals(ReverseNumber, simulationOutput$scaledResiduals)  #plot the residuals against other predi

# GLMM exploit strategy (6-14 correct choices/20 trial block)
# ratio
library(MCMCglmm)
prior = list(R = list(R1 = list(V = 1, nu = 0), R2 = list(V = 1,
    nu = 0)), G = list(G1 = list(V = 1, nu = 0)))
et <- MCMCglmm(ratioExploit ~ ReversalNumber, random = ~ID, family = "poisson",
    data = strat, verbose = F, prior = prior, nitt = 13000, thin = 10,
    burnin = 3000)
summary(et)
# autocorr(et$Sol) #Did fixed effects converge?
# autocorr(et$VCV) #Did random effects converge?
```

####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.

###F. ETHICS

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

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

###G. AUTHOR CONTRIBUTIONS

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

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

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

Seitz: Prediction revision, programmed the reversal learning touch screen experiment, data interpretation, revising/editing.

Blaisdell: Prediction revision, assisted with programming the reversal learning touch screen experiment, data interpretation, revising/editing.

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

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

###H. FUNDING

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

###I. ACKNOWLEDGEMENTS

We thank our PCI Ecology recommender, Aurelie Coulon, and reviewers, Maxime Dahirel and Andrea Griffin, for their feedback on this preregistration; Dieter Lukas for help polishing the hypotheses, developing an additional dependent variable (flexibility ratio), and assistance in responding to reviewer comments; Ben Trumble for hosting the grackle project via office and lab space; Angela Bond for daily logistical support in the lab; Melissa Wilson Sayres for sponsoring our affiliations at Arizona State University and lending lab equipment; Kristine Johnson for technical advice on great-tailed grackles; Jay Taylor for grackle scouting and connecting us with facilities at Arizona State University; 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; and Richard McElreath for project support.

###J. REFERENCES

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Bergstrom, Carl T, and Michael Lachmann. 2004. "Shannon Information and Biological Fitness." In *Information Theory Workshop, 2004. IEEE*, 50–54. IEEE.

Chow, Pizza Ka Yee, Stephen EG Lea, and Lisa A Leaver. 2016. "How Practice Makes Perfect: The Role of Persistence, Flexibility and Learning in Problem-Solving Efficiency." *Animal Behaviour* 112. Elsevier: 273–83.

Cohen, Jacob. 1988. "Statistical Power Analysis for the Behavioral Sciences 2nd Edn." Erlbaum Associates, Hillsdale.

Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. "Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses." *Behavior Research Methods* 41 (4). Springer: 1149–60.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2). Springer: 175–91.

Griffin, Andrea S, and David Guez. 2014. "Innovation and Problem Solving: A Review of Common Mechanisms." *Behavioural Processes* 109. Elsevier: 121–34.

Hadfield, Jarrod D. 2010. "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33 (2): 1–22. http://www.jstatsoft.org/v33/i02/.

Hadfield, JD. 2014. "MCMCglmm Course Notes." http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf.

Hartig, Florian. 2019. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. http://florianhartig.github.io/DHARMa/.

Homberg, Judith R, Tommy Pattij, Mieke CW Janssen, Eric Ronken, Sietse F De Boer, Anton NM Schoffelmeer, and Edwin Cuppen. 2007. "Serotonin Transporter Deficiency in Rats Improves Inhibitory Control but Not Behavioural Flexibility." *European Journal of Neuroscience* 26 (7). Wiley Online Library: 2066–73.

Lefebvre, Louis, Patrick Whittle, Evan Lascaris, and Adam Finkelstein. 1997. "Feeding Innovations and Forebrain Size in Birds." *Animal Behaviour* 53 (3). Elsevier: 549–60.

Liu, Yuxiang, Lainy B Day, Kyle Summers, and Sabrina S Burmeister. 2016. "Learning to Learn: Advanced Behavioural Flexibility in a Poison Frog." *Animal Behaviour* 111. Elsevier: 167–72.

Logan, C J. 2016. "Behavioral Flexibility in an Invasive Bird Is Independent of Other Behaviors." *PeerJ* 4. PeerJ Inc.: e2215.

Manrique, Héctor Marín, Christoph J. Völter, and Josep Call. 2013. "Repeated Innovation in Great Apes." *Animal Behaviour* 85 (1): 195–202. https://doi.org/10.1016/j.anbehav.2012.10.026.

McElreath, Richard. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press. http://xcelab.net/rm/statistical-rethinking/.

McInerney, Rob E. 2010. "Multi-Armed Bandit Bayesian Decision Making." *Univ. Oxford, Oxford, Tech. Rep*.

O'Hara, Mark, Ludwig Huber, and Gyula Kopanny Gajdon. 2015. "The Advantage of Objects over Images in Discrimination and Reversal Learning by Kea, Nestor Notabilis." *Animal Behaviour* 101. Elsevier: 51–60.

Peer, Brian D. 2011. "Invasion of the Emperor's Grackle." *Ardeola* 58 (2). BioOne: 405–9.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Sol, Daniel, Richard P Duncan, Tim M Blackburn, Phillip Cassey, and Louis Lefebvre. 2005. "Big Brains, Enhanced Cognition, and Response of Birds to Novel Environments." *Proceedings of the National Academy of Sciences of the United States of America* 102 (15). National Acad Sciences: 5460–5.

Sol, Daniel, and Louis Lefebvre. 2000. "Behavioural Flexibility Predicts Invasion Success in Birds Introduced to New Zealand." *Oikos* 90 (3). Wiley Online Library: 599–605.

Sol, Daniel, Tamas Székely, Andras Liker, and Louis Lefebvre. 2007. "Big-Brained Birds Survive Better in Nature." *Proceedings of the Royal Society of London B: Biological Sciences* 274 (1611). The Royal Society: 763–69.

Sol, Daniel, Sarah Timmermans, and Louis Lefebvre. 2002. "Behavioural Flexibility and Invasion Success in Birds." *Animal Behaviour* 63 (3). Elsevier: 495–502.

Wehtje, Walter. 2003. "The Range Expansion of the Great-Tailed Grackle (Quiscalus Mexicanus Gmelin) in North America Since 1880." *Journal of Biogeography* 30 (10). Wiley Online Library: 1593–1607.