Dear Dr.'s Coulon, Dahirel, and Griffin,

We sincerely apologize for the delay in our revision. When the reviews arrived, we were in the middle of submitting four other preregistrations for peer review at PCI Ecology before data collection on those experiments began, which worked, but it required a large time investment and delayed this revision. Additionally, and with a long-term effect, due to staffing changes that occurred in the past few months, all of us were overcommitted just by trying to keep the experiments and field site running. Logan was in the field collecting data to help offset the setbacks, which meant that she was unable to lead the revision process until now.

Luckily, we received your reviews a few days before data collection on the color tube reversal learning experiment began (the first flexibility experiment) so we were able to determine whether any methodological changes were needed before collecting any data.

One general note: as a result of several other setbacks, we had to break the counterbalanced order of experiments for four birds in batch 1. For details, please see Table 1 in the protocol: https://docs.google.com/document/d/18D80XZV_XCG9urVzR9WzbfOKFprDV62v3P74upu01xU/edit?usp=sharing (exceptions highlighted in yellow).

We greatly appreciate the time you have taken to give us such useful feedback! We are very thankful for your willingness to participate in the peer review of preregistrations. We revised our preregistration (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_flexmanip.md) and protocol (see link in previous paragraph), and we responded to your comments (which we numbered for clarity) below (our responses are preceded by "> Response X").

We think the revised version is much improved due to your generous feedback!

All our best,

Corina, Carolyn, Luisa, Benjamin, Aaron, Zoe, and Kelsey

(Note the addition of a new co-author, Kelsey McCune, who recently joined the grackle team)

Decision

by Aurélie Coulon, 2018-09-13 09:05

Manuscript: 10.5281/zenodo.1303263

Decision on pre-registration "Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context?"

Dear authors,

I have now received two reviews of your pre-registration. It took longer time than we wished to get reviews, because of the (summer) timing and also because of the novelty of the pre-registration process. I apologize about this.

Both reviewers have interesting comments and suggestions about your hypotheses/predictions and protocol. I suggest revising your pre-registration accordingly, alongside writing a detailed response to each reviewer's comment.

I look forward to receiving a revised version of your pre-registration.

Yours sincerely,

Aurélie Coulon.

Reviews

Reviewed by Maxime Dahirel, 2018-07-29 13:56 Download the review (PDF file)

1. I've read with interest this preregistration submitted by Logan and colleagues to PCI Ecology. First, I commend the authors for submitting a pre-registration of their research project, something that is rare in ecology and evolution (I should probably get around it myself!). Although it is beyond the remit of this review, I want to say the broader project seems especially interesting, and that the current pre-registration seems like a well thought-out part of it.

> Response 1: Thank you so much for your support and positive feedback! It was very kind of you to also consider the broader project.

2. Regarding the pre-registration itself: As far as I can tell, the broad methods used to answer the questions at hand are sound and appropriate, the hypotheses (and alternatives!) are well-defined, and the broad details of the statistical analyses (which dependent and independent variables?) correct. There are however several comments regarding the sample size and statistical analyses themselves that need to be addressed before recommendation: (Remark: I address some issues in the first test where they are encountered, but they may also be present elsewhere in the pre-registration).

> Response 2: Thank you for the feedback! We look forward to addressing your detailed comments below (and making sure we carry the advice forward to apply it to other relevant parts of the preregistration).

3. 1- In "data checking", please note that data normality can only be assessed on residuals relative to a model, not on actual data (well, it can, but it doesn't mean anything with respect to analysis validity).Please also note that normality diagnostics plots can be extremely misleading and difficult to interpret for non-Gaussian GLM(M)s. I advise to look at the DHARMa R package (https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html) by Florian Hartig, which contains several diagnostic plots and tests that don't have this problem.

> Response 3: We really appreciate you sharing your R and modeling wisdom with us! There

have clearly been improvements since we learned these techniques - thank you for pointing them out to us and for teaching us how to use them. Your comment makes a lot of sense - the data must be checked in the context of the particular model and not separately. We made the following changes:

Analysis Plan > Data Checking: replaced the text with
"The data will be checked for overdispersion, underdispersion, zero-inflation, and heteroscedasticity with the DHARMa R package [@Hartig2019dharma] following methods by [Hartig](https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html). Note: DHARMa doesn't support MCMCglmm, therefore we will use the closest supported model: glmer from the R package lme4 [@lme4]."

We updated the code for each model (in that model's section of the Analysis Plan) using the DHARMa package and we choose a variety of tests to address issues that might come up with our particular data set (Analysis Plan > P1, P2, P2alternative2, P3a, P3alternative1, P3b, and P4).

4. 2- At several points throughout the registration, you plan to use generalized linear models (not mixed) when your list of independent variables clearly indicate a random effect of ID (see P1, but not only). If there are repeated individual measures, a mixed model should be used, even if the random effect is not what is tested in the current hypothesis.

> Response 4: Thank you for catching this! This was an oversight on our part. We changed the following models from GLMs to GLMMs: P1 (and also included batch as a random effect) and P2 (and also moved batch from a fixed effect to a random effect). The following models do not have repeated individual measures and so we kept them as GLMs: P2 alternative 2 and P3b.

5. 3- I have doubts about the ability of the chosen power analyses to correctly capture the true power of your tests, especially in cases involving random effects, where data points won't be independent. Isuggest the authors use simulations to get a better handle of how their models will actually behave. I acknowledge that simulating various effect sizes, for different data structures, can be time- consuming and difficult. I believe the package SQuID may be useful here (Allegue et al., 2017)(https://cran.r-project.org/web/packages/squid/index.html) (disclaimer: I have not used it yet,so can't vouch for its actual usefulness here).

> Response 5: We completely agree that the power analyses used were the wrong tool for the job. However, we didn't know of a better option and we wanted to have some representation of our ability to detect effects, which is why we used them. We had tried a few R packages that could have been more effective, but we were not able to get any of them to work with our mock data. Thank you for the recommendation regarding SQuID! We started using it and it is a great package with amazingly clear documentation that makes it easy to use.

After working with SQuID to simulate our model in P2, we were unable to find a way to represent the complexity of our model (e.g., compare population means between control vs. manipulated conditions) while manipulating effect sizes. Additionally, we have no prior information about what values to provide in the input for the simulation without looking at the data we are currently collecting (e.g., the multi-access box has never been presented to

grackles and we were unsure of how many options they would be able to solve). We are interested in the effect of the manipulation vs. everything else we are controlling for and, because of the complexity of the model, the effect is going to depend on the factors we control for as well as the boundaries of the dependent and independent variables. We currently don't have any estimates for any variables because these tests have never been done in grackles and we have not encountered previous research that has manipulated flexibility in this way. However, we will be able to estimate these boundaries from our data after it is collected. Then we can perform informed simulations which will allow us to understand what sample size we need to detect the effect of interest. Once we **have the data** (and before conducting the analyses in the preregistration), we can set priors by entering the boundaries for the variables in the analysis (while remaining blind to the effect - the relationship between the variables). We can do this by running the **null model** (dependent variable ~ 1 + random effects), which will allow us to understand what the effect can actually operate on, and will inform us about what a weak vs. a strong effect is for these models. From here (and also before conducting any of the analyses in the preregistration), we can run the **simulations** based on the null model and then we can explore the boundaries of influences (such as sample size) on our ability to detect the effects of interest of varying strengths. We will run these simulations using the principles in McElreath (2015, Statistical Rethinking; starting on page 249) as a starting point and in consultation with McElreath.

In terms of changes to the study design that would be possible to make as a result of simulation outcomes, pretty much the only thing we have some element of control over is the sample size. We have run into several unexpected complications at the Arizona field site (where we are currently collecting data), which is already indicating that we will not meet our projected sample size during our two years at this site (e.g., the grackles there are extremely difficult to catch and most of the females refuse to participate in tests). What we will do, before conducting the analyses in the preregistration, is run the simulations using the Arizona data to inform the simulation inputs and determine the lower sample size bounds for the analyses in this preregistration. If it turns out that our Arizona sample size is not larger than the lower boundary, we will change our experimental stopping criterion (which is currently to stop these experiments after two full aviary seasons in Arizona) and continue these experiments at our next field site until we meet the minimum sample size.

We updated the preregistration to lay out this new plan:

Analysis Plan > Ability to detect actual effects: "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."

Methods > 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](#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."

During this process, we also realized that for P1 we don't actually need a power analysis because it is implicit in the stopping criterion for the experiment. The prediction is that individuals will get faster to reverse with each reversal, but this has to be the case because our stopping criterion is that they must pass two consecutive reversals in 50 trials or less (our pilot data showed that 70 trials was the fastest first reversal, thus 50 trials would be an improvement). Therefore, we don't need a power analysis here to detect this effect since getting faster was already part of the flexibility manipulation. We will still need to run the MCMCglmm to determine the size of the effects. As a result, we made the following change:

Analysis Plan > P1: we removed the power analysis and model validation and replaced it with: "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."

6. 4- In P1 (and elsewhere), since there are only two models ("base" and "test"), there is no need to useAkaike weights. Independently of debates around their usefulness, they are meant/most useful to compare more than 2 models. With two models, you can simply use the AIC/ BIC/... and see whichmodel is best supported. If I remember my Burnham and Anderson correctly, a deltaAIC>4 when there are two models is equivalent to a weight of >90% given to the best model anyway.

> Response 6: Thank you so much for the advice! That makes sense because the R package is actually called multimodel inference (MuMIn), which implies the comparison of at least a few models. We were only able to find a table of deltaAIC in Burnham & Anderson (2002, p.70) showing that models with deltaAIC 0-2 should be considered substantially similar to each other, 4-7 considerably less similar to each other, and >10 no similarity to each other. We will use the deltaAIC 4 threshold where if one model's AIC is >3 in relation to the other model, the model with the smaller AIC will be considered the best fitting model. We changed the Analysis Plan > P2alternative2 model and text accordingly. (Note that the model validation was deleted from P1 - see Response 5)

7. 5- In P2, you use the average response as your dependent variable. Why not include all values and an individual random effect?

> Response 7: This is a good point - it is much better to make use of all of the data (especially because we have it!). We think we overlooked this because at one point we had the models set up with flexibility (number of trials to reverse) as the dependent variable, which would then result in only one measure per bird (because the control group only gets one reversal). Now that the average latency is the dependent variable, we are free from this restriction. P2 uses repeated individual measures and a GLMM, but the P2 alternative 2 model uses one number per test per bird and so applies to your comment and this is the model we changed as follows: changed from a GLM to a GLMM, changed the dependent variable from an average latency to solve new loci to the latency to solve new loci and from the average latency to attempt to solve new loci to the latency to attempt to solve new loci, and we added ID as a random effect because of the repeated individual measures. Note: we changed the dependent variable to number of trials to solve/attempt to solve a new locus - see our Response 20 for more details - and this dependent

variable includes repeated individual measures so it was a simple variable switch.

8. 6- In P2 also, please be aware that (in my experience) latencies will probably need transformation (I expect log) to fit within the statistical model.

> Response 8: It is good to know that latency data tends to be skewed, thus transforming them normalizes them for analyses. We ended up changing these response variables (in P2 alternative 2) from latencies to number of trials (see Response 20), which means we will likely not need to log transform the response variable because it will likely have a more normal distribution.

9. 7- In P3a, you use a linear mixed model (lmer) to estimate repeatability when a generalized linear mixed model (glmer) should be used, as the dependent variable is a count. This would lead to potentially wrong estimates of repeatability, first because the individual variance will probably be wrongly estimated, and then because the residual/overdispersion variance is not the same for a Gaussian vs a Poisson model (Nakagawa & Schielzeth, 2010)

> Response 9: We were trying to follow the instructions on Tom Houslay's (2016, http://rpubs.com/tomhouslay/200150) blog as closely as possible to make sure the models worked the way he explained them, which is why we used lmer. But we see your point and so we changed the P3a analysis from lmer to glmer. (But see Response 10, which describes why we ended up deleting this analysis.)

10. If you want to extract correct repeatabilities from a lme4 (G)LMM (and their 95% CI) on both the response and latent scales, you can use the rptR package (Stoffel, Nakagawa, & Schielzeth, 2017). However, since you plan to fit your GLMM using MCMCglmm you actually don't need that at all,since the latent scale adjusted repeatability and its credible interval can simply be obtained by mod$VCV[,ID]/(mod$VCV[,ID]+mod$VCV[,units]+mod$VCV[,any other included random effect]). For the raw repeatability, simply fit a model with no covariates. For the repeatability on the response scale, see (P. de Villemereuil, Morrissey, Nakagawa, & Schielzeth, 2018; Pierre de Villemereuil, Schielzeth, Nakagawa, & Morrissey, 2016) and the QGglmm R package.

> Response 10: This is amazing news - it really simplifies the analysis, thank you! We now include code to obtain the repeatability (posterior.mode of your code above) and 95% credible intervals (HPDinterval for your code above), and we made a raw repeatability model with no covariates for comparison (see revised code in Analysis Plan > P3a). We also removed the repeatability analysis based on Houslay's blog because your solution replaces it.

We are not entirely clear what the response scale is and how it relates to the terms used in the papers you suggested. For example, Villemereuil et al. (2018) and Villemereuil et al. (2016) refer to a "latent scale" and a "data scale", but not a "response scale", and the QGglmm package documentation additionally talks about an "observed scale" (https://cran.r-project.org/web/packages/QGglmm/QGglmm.pdf). We presume the "response scale" is the "data scale" and the "observed scale"? Going on this assumption, we added the variance from

the fixed effects to the code you provided for us and calculated the repeatability and its credible interval (see revised code in Analysis Plan > P3a).

11. 8- In P3b, I'd advise to use a multivariate mixed model rather than a univariate correlation between averages. See (Houslay & Wilson, 2018) and https://tomhouslay.com/tutorials/ for the rationale behind my suggestion and the potential problems with your approaches.

> Response 11: Yes, of course this makes sense - to use more of the data we collect and to account for repeated individual measures. We changed the model in P3b from a GLM to a GLMM by using the Number of trials to solve new loci on the multiaccess box as the dependent variable instead of the number of trials to reverse because the control group would only have one number for latter and thus no repeated measures. We included ID as a random effect.

12. 9- The paragraph F on sampling size is not clear (you mention 16 birds per treatment, then 8 birdsper experiment). Also please be precise what "many more" entails in "we expect to be able to test many more"; if only for ethical reasons, and then to design a proper maximum sample size stopping rule (even an approximate one). It may also be useful to run your power analyses for different potential sample sizes, not only the minimal one.

> Response 12: Sorry for the confusion - we think this was a typo carried over from a different preregistration. We deleted the text mentioning 8 birds and we also deleted "we expect to be able to test many more". For one thing, it is becoming quite clear that we aren't going to reach our target sample size of 32 at the Arizona field site because the birds are extremely difficult to catch, most of the females are refusing to participate in aviary tests, and it is taking the birds longer to get through the test battery than previously expected, which means we will likely not be able to test two full batches of 8 birds per batch per aviary season. We agree that it would be useful to run the power analyses at various sample sizes, but given how things are working out in the field, we don't think we will have the luxury of choosing a range of sample sizes at this field site. See Response 5 for the discussion of power analyses and simulations.

Reviewed by Andrea Griffin, 2018-09-13 06:03

The overall aim of this project is stated to be to "determine how behavioural flexibility works and how we can make predictions about a species' ability to adapt their behavior to new environments". Its aims are "to manipulate grackle behavioral flexibility … to determine whether their flexibility is generalizable across contexts, whether it is repeatable within individuals and across contexts, and what learning strategies they employ. Although I find these specific questions divorced from the background of invasive spread, and hence, do not particularly buy into the advantage of using an invasive species undergoing a current range expansion (in fact, I explain below how I find some predictions inherently contradictory with this background), the questions are worthwhile asking. Detailed comments are as follows:

13. Please clarify: “they generally get faster rather than getting faster with each reversal” and how the learning criterion was altered to accommodate this

> Response 13: Good point. We made the following changes:

State of the Data: we added “(each subsequent reversal may not be faster than the previous, however their average reversal speed decreases)” and we linked this text such that when the link is clicked, it will bring the reader to the section that describes all of the details involved (“*Determining the threshold: How many reversals are enough?*”)

Analysis Plan > In October, we also added a new paragraph:
“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's 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.”

14. P1: Although I find it interesting to compare the extent to which rule-learning generalizes across contexts as a function of experience with rule learning, the question seems to contradict the notion that behavioural flexibility (measured here as rule learning) can be a target of selection during a range expansion, meaning that individuals would need to differ in some inherent way in their capacity to rule learn. Hence, for behavioural flexibility to be a target of selection, behavioural flexibility should not be flexible, if that makes sense. Put differently, invasive birds should not become broader-context rule-learners as a consequence of experience with rule learning. In fact, learning this capacity would be predicted to delay selection on the trait. Hence, H3a seems the more logical prediction and test, given the larger context of the project.

> Response 14: Thank you for evaluating the big picture question for our study!
This discussion brought up interesting ideas that are worth considering. The problem is that we don’t think we will be able to fully address this comment with the variables we are measuring.

What we have changed as a result is the following (changes in bold):

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

P2: [...] 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](https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_flexgenes.md) (R1) and [expansion](https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_expansion.md) (H1))**.

We did not make changes to address the complicated issue of the processes that generate variation in behavioral flexibility and how selection might operate on them because it is beyond the scope of the current study. Consequently, we are not able to distinguish between hypotheses that suggest that variation in flexibility is due to a strong heritable component and/or early life canalization of flexibility versus hypotheses that imply a plastic/environmental component to behavioral flexibility. Indeed, both are likely so it is not contradictory to expect that flexibility both varies between individuals and can be manipulated. Though if the latter is true it may make it more difficult for selection to act on behavioral flexibility. In addition, the ability to learn a rule quickly does not preclude the ability to learn other rules and/or switch between rules. Therefore, we do not see the ability to quickly learn rules as an inhibitor of selection on flexibility, if it is under selection. Indeed, the tests we give them should be testing the ability to learn new rules quickly in new contexts since the apparatuses we give them are usually novel. We aren't only asking them to learn rules, we are asking them to make the correct choice in a particular situation, which is the functional choice (which varies from experiment to experiment and sometimes there is more than one functional choice). The constant attention to the functional option(s) would allow individuals to not only use the rule they previously learned, but search for other rules when the current rule no longer applies. Therefore, **what could be a target of selection is the ability to learn new rules, the ability to update rules, or the ability to switch between multiple rules** (what might be called "broader-context rule learning").

Additionally, rule learning may not be the only way they can get faster at serial reversals. They

might start from scratch with trial and error learning with each new reversal, but get faster because they decide to 1) persist less on the previously rewarded option (which could be a rule they learn: “give up more quickly in the face of non-reward”) or 2) explore because they are curious (though this shouldn’t apply to the birds in the manipulated condition because they receive several reversals which should reduce their curiosity about the task). We don’t know how to experimentally test these predictions, but theoretically these options, including pure curiosity, are possible.

You could be right in that if individuals are plastic in their expression of flexibility (e.g., no within-individual consistency) then there could be a delay in selection on flexibility in a new environment because the population would be faster to adapt if all individuals were at their biologically maximum level of flexibility, and because the plasticity in flexibility would limit the extent to which flexibility is directly passed on to offspring. However, we think the picture is more complicated because we would only expect all individuals to be at their maximum if all individuals constantly experience new environments. In established populations (i.e., populations that have been at that location for a longer period of time), they aren’t necessarily experiencing new environments. This allows for **uncertainty in how often an individual will experience new environments**, which means we need a more complex model.

In addition, there is a difference between having a capacity and expressing the capacity. The environment influences the expression of the capacity (e.g., canalization: even though an individual had a high capacity for a trait earlier in life, we would not be able to measure it if we measure the individual after canalization has occurred), so if, as argued above, individuals differ in how often they will experience new environments, we expect them to show different expressions of the capacity (which is what matters and what we can measure). For example, an individual that goes to a new environment might experience positive feedback to increase this capacity, and selection can operate on the capacity because it is important in adult life. Once the capacity is expressed in individuals in a population, there could be a reduction in the extent to which the environment can influence the expression of the trait. Therefore, 1) there will be within-individual consistency, and 2) how much flexibility is expressed might vary at the population level because it would be tuned to the local environment. This means that we would expect to see differences in the amount of flexibility expressed between populations because each population will be responding to its local environment and the individuals in each population will be a unique combination of how much flexibility is expressed. If an individual moves to a new population, the previously used rules might not work and the restrictions imposed on the expression of flexibility by the previous environment will be lifted. Now this individual could change its capacity according to its new experiences so there will no longer be within-individual consistency. This will lead to differences among individuals upon which selection can now operate. Note that this latter case could be operating when we give them new tests in the wild and when we bring them into the aviaries for testing: we are changing their environment, which might lift their normal environmental “restrictions” on flexibility, and could make flexibility more manipulatable.

We will keep these issues in mind as we continue to revise our cross-population preregistrations:

1) The genetics of behavioral flexibility across the range of a rapidly expanding species (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_flexgenes.md). Our research question 1 is:
   “R1 heritability: To what extent is variation in behavioral flexibility explained by genetic similarity? Based on flexibility testing in aviaries, we will develop a

flexibility test for individuals in the wild to increase our sample size. Genetic relationships among individuals will be determined using ddRADseq to then investigate this question using the Animal Model. If flexibility is not heritable, this is potentially because behavior is not usually very heritable, and/or we might fail to detect heritability because the A) sample size might not be large enough, and/or B) individuals may not vary enough in flexibility. If we find that flexibility is heritable, this could indicate that there is a stable polymorphism where some individuals have high and others low flexibility, which could give different benefits in a given environment."

2) How and why does behavioral flexibility vary across the range of a rapidly expanding species? (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_expansion.md). Our hypothesis 1 is:
"H1: Individuals are more behaviorally flexible (measured by reversal learning and switching between options on a multi-access box) near the edge of their geographic range (in the US relative to Honduras) if flexibility is required for range expansion." Which has the following predictions:
"Prediction 1: There will be a positive association between flexibility (individuals that are faster to reverse preferences on a reversal learning task and who also have lower latencies to switch to solving new loci after previously solved loci become unavailable) a population's proximity to the edge of the geographic range. If this is supported, other variables should also be involved in how such flexibility is achieved. These other variables will be investigated in the hypotheses below.
P1 alternative 1: If there no difference in flexibility between the edge and non-edge populations, this could indicate that most individuals are already highly flexible (ceiling effect).
P1 alternative 2: If there is a negative correlation or no difference in flexibility between the edge and non-edge populations, this suggests that other variables might play a larger role than flexibility in how this species is able to rapidly expand its geographic range (see hypotheses below)."

References

Logan CJ. 2016. Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ* 4:e2215 https://doi.org/10.7717/peerj.2215

15. P1 alternative: *"If the number of trials to reverse a preference does not correlate with or positively correlates with reversal number"*; this contradicts the general pattern of finding of an SDR (number of trials to criterion decreases with increased reversal number) and also the authors' own statement under P1: *"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". Is this a typo?

> Response 15: It wasn't a typo, we were trying to explain how we would interpret all potential correlations (positive, negative, and no correlation). P1 is our main prediction of what will happen with our results and is about a negative correlation: the number of trials to reverse a preference decreases with increasing reversals. In P1 alternative, we were investigating the alternatives to P1, which would be if there was no correlation or a positive correlation where the number of trials to reverse a preference increases with increasing reversals**.** We wanted to state up front how we would interpret the results through the presentation of these alternatives. We clarified this by adding to P1: "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,"

Methods:

16. Learning criterion: is the sequence re-started at each testing session? Ie what is planned if a bird is willing to do 10 choices on one day and then satiates. Is the trial number started again at each testing session? Or do the, for example, correct trials yesterday count towards today's? ie does the sliding window for scoring carry across testing sessions?

> Response 16: Good question, and one that we should clarify. It is a sliding window so we look at the most recent 10 trials for a bird, regardless of when the testing sessions occurred. We clarified this by adding to Methods > 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."

17. IVs: What is 'batch'? test cohort? How many birds are tested simultaneously?

> Response 17: You are correct, a batch is a test cohort and consists of 8 birds being tested simultaneously. Though batches can be mixed in the aviaries because as soon as a bird from batch 1 completes its tests, it is released and another bird is brought in to fill that aviary and becomes part of batch 2. We clarified this with the following change: Methods > Independent variables: "2) Batch (random effect because multiple batches included in the analysis). Note: batch is a test cohort, consisting of 8 birds being tested simultaneously"

18. "we should expect due to the lack of data on this species for these experiments": ok,

but there are several published comparative data sets on SDR that you could use (e.g. Tebbich et al.).

> Response 18: In terms of looking to the literature to estimate what kinds of effect sizes we might expect, we looked to the Darwin's finches who had the most similar reversal speeds to grackles. However, effect sizes were only available from cross-species analyses (Tebbich et al. 2010) and so they don't apply to our within-species analysis. Looking beyond the most similar species to grackles, we found that either the hypotheses were too different from ours to be able to draw parallels (e.g., Lissek & Gunturkun (2002) compared two groups of pigeons one with and one without NMDA receptor blockages), or they didn't conduct an analysis for the data that are most relevant to our question (e.g., Warren (1966) in their Figures 1 and 2).

There are effect sizes reported in Bond et al. (2007) for three species of corvids that parallel our P1 in that they are testing similar hypotheses: "Error rates declined significantly for all species across successive reversal blocks: 1st sessions, $F(4, 48) = 76.70$, $p < .001$; later sessions $F(4, 48) = 75.8$, $p < .001$" (Bond et al. 2007). We ran a power analysis on their study as if the study had not yet been conducted, so we could directly compare it with the power analyses we ran on our planned analyses. As in the power analyses in our preregistration, we used G*Power and set the power to 0.70. We then increased the effect size until the output displayed a total sample size of 15 which was the total number of birds used in their study (five individuals from three different species). Here is the output from G*Power:

F tests - Linear multiple regression: Fixed model, R^2 deviation from zero

Analysis:	A priori: Compute required sample size		
Input:	Effect size f^2	=	0,50
	α err prob	=	0,05
	Power (1-β err prob)	=	0,7
	Number of predictors	=	1
Output:	Noncentrality parameter λ	=	7,5000000
	Critical F	=	4,6671927
	Numerator df	=	1
	Denominator df	=	13
	Total sample size	=	15
	Actual power	=	0,7165040

This means that, for their study, they had a 72% chance of detecting a large effect (f^2=0.50). For prediction 1 in our study, we have a 71% chance of detecting a medium effect (f^2=0.21). (Effect size classification as medium or large from Cohen (1988)). This comparison is perhaps more useful than converting the Bond et al. (2007) effect size (F from an ANOVA) to the effect size used in MCMCglmms (our planned analysis method) to get an idea of what kind of an effect size we might expect because knowing what ability they had to detect the effect they obtained gives us an idea of whether we can rely on the obtained F statistic. Their F values were 76 and 77, which are both above the critical F value of 5 in the G*Power analysis. Thus, they detected a large effect and it looks like they had the power to do so.

In terms of what this exercise means for our study, it appears that, based on this one paper, we might expect a large effect size when examining whether error rates decline in reversal experiments, in which case we should be able to detect this large effect according to the power

analysis we performed for prediction 1. We were going to add this information to the preregistration, but then we finished addressing Comment 5, which made this obsolete (see Response 5 for details).

In terms of looking to the literature for a criterion that would tell us when we had manipulated an individual's flexibility (in case this was part of your comment), we were unable to find individual-level data for serial reversals, and there were no associated published data sets that we could find (e.g., Bond et al. 2007; Lissek et al. 2002, Tebbich et al. 2010, Warren 1966), which is why we had to rely entirely on our pilot data.

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

Lissek, S., B. Diekamp, and O. Güntürkün. "Impaired learning of a color reversal task after NMDA receptor blockade in the pigeon (Columbia livia) associative forebrain (Neostriatum Caudolaterale)." *Behavioral Neuroscience* 116, no. 4 (2002): 523.

Tebbich, S., Sterelny, K., & Teschke, I. (2010). The tale of the finch: adaptive radiation and behavioural flexibility. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *365*(1543), 1099-1109.

Warren, J. M. (1966). Reversal learning and the formation of learning sets by cats and rhesus monkeys. *Journal of Comparative and Physiological Psychology*, *61*(3), 421.

19. "Analysis": N=32: does that mean 16 experimental and 16 controls?

> Response 19: You are correct that N=32 means that 16 birds will be in the manipulated group and 16 will be in the control group. We clarified this by moving the Planned Sample text from the end of the preregistration to the first paragraph under Methods so it comes before readers get to the power analyses. We also clarified the text in Planned Sample to:
"Some individuals (~32: ~16 in the control group (they receive 1 reversal) and ~16 in the flexibility manipulation (they receive multiple reversals))"

20. In measuring the latencies to switch to new options on the multi-access box: how can 'work-time' be disentangled from 'non-work time' in a latency? I.e. a bird that tries lots of options and interacts with the box might have the same latency as a bird that does mostly other things but then tries once and by chance solves the box immediately. How can these two seemingly very different birds be differentiated in a latency? I guess this is equivalent to asking whether multi-access box testing yields any motivation variables, similar to number of beak-to-task contacts one can obtain from an innovative problem-solving task. I would predict that motivation variables, which end up providing some measure of persistence, will be associated with exploitation vs. exploration strategies. I could not find an answer to this question in the protocol.

> Response 20: This is an excellent point, thank you for bringing it up. We changed the multi-access box flexibility measure from "latency to attempt or solve a new locus" to the "number of trials to attempt a new locus" and the "number of trials to solve (pass criterion) a new locus". The number of trials should be a cleaner measure, particularly the number of trials to pass criterion, because the criterion is designed around the number of trials. A bird must successfully solve (on first touch) a particular locus in two consecutive trials to pass criterion, therefore now that we are paying attention only to trial number, we eliminate the noise that comes from the latency data where we would need to sum latencies between successful solves of different loci across sessions.

The number of trials measure still has an element of work-time involved, but we address this in a different preregistration called Exploration, which is currently in review at PCI Ecology (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_exploration.md). In the exploration preregistration, we measure persistence on the multi-access box using the number of touches to the multi-access box per time (P6 dependent variable 2) and we distinguish between the number of touches to the functional versus non-functional parts of the box (independent variables, P3). We then analyze these variables in models examining their relationship with flexibility.

21. "Learning strategies": The authors might find the following research and analysis strategy useful as it disentangles exploration from exploitation phases within each reversal. Federspiel, I. G., Garland, A., Guez, D., Bugnyar, T., Healy, S. D., Güntürkün, O., & Griffin, A. S. (2017). Adjusting foraging strategies: a comparison of rural and urban common mynas (Acridotheres tristis). Animal Cognition, 20, 65–74.

> Response 21: Thank you very much for pointing this out! In Federspiel et al. (2017), the number of 20-trial blocks where birds were choosing "randomly" (6-14 correct choices; called sampling blocks; akin to the *exploration phase* in P4 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 P4 in our preregistration). These ratios, calculated for each bird for their initial discrimination and reversal phases, quantitatively discerns the exploration phase from the exploitation phase. This analysis is useful for our hypothesis H4 to determine whether individuals rely more on an exploitative strategy as they progress through serial reversals, and we updated the Analysis Plan for P4 to include it.

Protocol_flexibility:
22. Why will birds only undergo 5 reversals on the touch screen? Is this enough to establish a significant slope in learning speed that can be compared across tasks?

> Response 22: We limited the touch screen serial reversals to five due to time constraints. We knew from pilot data that it was going to take the birds months to meet the criterion of passing two consecutive reversals at or under 50 trials, and besides the three experiments in this preregistration, they have five additional experiments to complete in their limited time in the aviary. We also know from the pilot data that the individual-level data is much more messy than we expected, which reduces the chances of obtaining a clear slope even after many, many reversals (38 reversals in the case of the pilot bird, see Figure 1). In this comparison, we are interested in whether the birds in the manipulated group (who receive 5 touch screen reversals) are on average faster than birds in the control group (who receive 1 touch screen reversal) and because the pilot bird's reversal speeds decreased in the first 5 trials (Figure 1), we think this should be enough to show a difference between the two groups if there is a difference to be found.

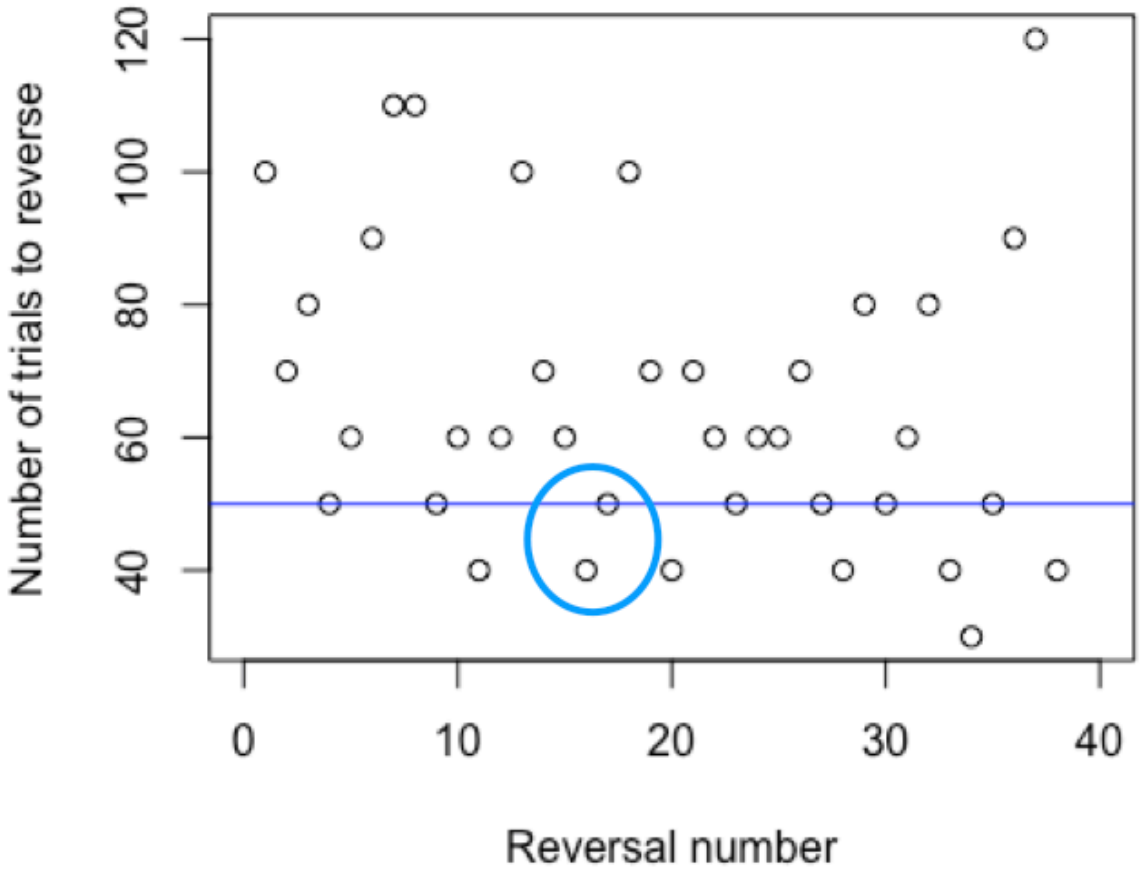

Figure 1. Pilot data from the color tube reversal learning experiment from Fajita (January-April 2018): almost the full range of variation in the number of trials to reverse a preference was encompassed in her first five reversals. The blue circle identifies the two consecutive reversals that were passed in 50 trials or less - it is at this point that she would have passed our new criterion.

We made the following change to the flexibility protocol:
Page 16: "*Criterion to pass touch screen serial reversals:* Individuals will meet criterion when they have passed criterion in 5 reversals (note this criterion is different from the serial reversal criterion in the colored tube reversal learning experiment due to time constraints and because here we are interested in whether there is an average difference between the manipulated (receives 5 reversals) and control group (receives 1 reversal), which should be detectable in the first 5 reversals based on pilot data)"
https://docs.google.com/document/d/18D80XZV_XCG9urVzR9WzbfOKFprDV62v3P74upu01xU/edit?usp=sharing

23. I am also confused by not testing the control group. Was the suggestion not to train birds on the tube task (experimental and control) and then to compare the transfer of both these groups to each of the two additional tasks (multi-access and touch screen)? Why would the controls not participate in the reversals?

> Response 23:
Controls don't participate in serial reversals because the serial reversals are an attempt to actually manipulate flexibility. In case the manipulation works, we need a non-manipulated control group for comparison. After their first reversal (so there is reversal data to compare with the manipulated group), the control individuals receive trials with identically colored tubes (two yellow tubes) that are both rewarded. This ensures they receive equal experience at retrieving food from tubes as in the manipulated group. The control group experiences the multi-access box and the touch screen just like the manipulated group, except the control group only receives 1 reversal on the touch screen and not 5 like the manipulated birds do. This is to avoid manipulating the control group's flexibility through touch screen serial reversals. We clarified this by adding further details about the control group and how they are compared with the manipulated group:

Hypotheses > 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"