Peer Community In Ecology

From cognition to range dynamics: advancing our understanding of macroecological patterns

Emanuel Alexis Fronhofer based on reviews by 2 anonymous reviewers

Open Access

A recommendation of:

Aaron Blaisdell, Zoe Johnson-Ulrich, Luisa Bergeron, Carolyn Rowney, Benjamin Seitz, Kelsey McCune, Corina Logan. **Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles** *(2019), In Principle Recommendation 2019. PCI Ecology.* http://corinalogan.com/Preregistrations/g_causal.html

Submitted: 20 August 2018, Recommended: 31 January 2019

Cite this recommendation as:

Emanuel Alexis Fronhofer (2019) From cognition to range dynamics: advancing our understanding of macroecological patterns. *Peer Community in Ecology, 100014.* 10.24072/pci.ecology.100014

Published: 31 January 2019

Understanding the distribution of species on earth is one of the fundamental challenges in ecology and evolution. For a long time, this challenge has mainly been addressed from a correlative point of view with a focus on abiotic factors determining a species abiotic niche (classical bioenvelope models; [1]). It is only recently that researchers have realized that behaviour and especially

plasticity in behaviour may play a central role in determining species ranges and their dynamics [e.g., 2-5]. Blaisdell et al. propose to take this even one step further and to analyse how behavioural flexibility and possibly associated causal cognition impacts range dynamics. The current preregistration is integrated in an ambitious long-term research plan that aims at addressing the above outlined question and focuses specifically on investigating whether more behaviourally flexible individuals are better at deriving causal inferences. The model system the authors plan on using are Great-tailed Grackles which have expanded their range into North America during the last century. The preregistration by Blaisdell et al. is a great example of the future of scientific research: it includes conceptual models, alternative hypotheses and testable predictions along with a sound sampling and analysis plan and embraces the principles of Open Science. Overall, the research the authors propose is fascinating and of highest relevance, as it aims at bridging scales from the microscopic mechanisms that underlie animal behaviour to macroscopic, macroecological consequences (see also [3]). I am very much looking forward to the results the authors will report.

References [1] Elith, J. & Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40: 677-697. doi: 10.1146/annurev.ecolsys.110308.120159 [2] Kubisch, A.; Degen, T.; Hovestadt, T. & Poethke, H. J. (2013) Predicting range shifts under global change: the balance between local adaptation and dispersal. Ecography 36: 873-882. doi: 10.1111/j.1600-0587.2012.00062.x [3] Keith, S. A. & Bull, J. W. (2017) Animal culture impacts species' capacity to realise climate-driven range shifts. Ecography, 40: 296-304. doi: 10.1111/ecog.02481 [4] Sullivan, L. L.; Li, B.; Miller, T. E.; Neubert, M. G. & Shaw, A. K. (2017) Density dependence in demography and dispersal generates fluctuating invasion speeds. Proc. Natl. Acad. Sci. USA, 114: 5053-5058. doi: 10.1073/pnas.1618744114 [5] Fronhofer, E. A.; Nitsche, N. & Altermatt, F. (2017) Information use shapes the dynamics of range expansions into environmental gradients. Glob. Ecol. Biogeogr. 26: 400-411. doi: 10.1111/geb.12547

Revision round #2

2018-11-06

Dear Dr. Blaisdell,

let me start by apologizing for the delay since the last round of revisions, unfortunately, referees were very hard to find. As a consequence only one of the original referees read your answers.

The referee remains unconvinced by some of your answers, specifically with regards to data analysis. I would like to encourage you to once more revisit your plans based on the referee's points. I am persuaded that clarifying these points at this state is the strength of preregistrations. I am looking forward to receiving a revised version of your preregistration.

Sincerely yours, Emanuel A. Fronhofer

Preprint DOI: 10.17605/OSF.IO/GCA5V

Reviewed by anonymous reviewer, 2018-10-11 14:38

Dear authors,

Thank you for your replies. Overall, I feel you cannot change much. I understand. However, there are still several points that I think you should still consider again.

1) I do not get your argument about adding an ID effect as a fixed and a random effect. I am reading, reviewing and editing many manuscripts about behavioral consistency and I never saw a single one with ID as a fixed and a random effect. I am not saying it cannot be done, but my statistician colleagues are also extremely surprised by doing this and I don't think the model is going to converge. From my reading, the PDF you sent only explains that a predictor can be either fixed or random depending on the context, not that you can add this predictor as fixed and random simultaneously, but I might have missed this section. If you have ID as fixed, what is the point of adding it as random because you already explain most variance due to ID with the fixed effect? The only model I know with a same predictor as fixed and random are model in which the predictor is coded as

continuous as a fixed effect and as categorical as a random effect (i.e. years) and it is debated whether it is correct to do so.

The “usual” way to study “interactions” between ID and other covariates is to add a random slope on top of a random intercept. It allows testing for individual variation around the mean slope which is what you want to test here. I would say that Dingemanse and Dochtermann 2013 (JAE) or van de Pol and Wright 2009 (Animal Behaviour) are great introductions about random slopes.

2) I will have to disagree again on the model with predictors and interactions between all of them. The issue is not having 40 or 64. The issues are about A) making meaningful conclusion when have triple or quadruple (or it seems even higher ranked interactions) interactions between continuous variable. You cannot explain a pattern involving many covariates interacting each others. B) The probability of obtaining a significant interaction by chance. For instance, I just ran a simulation with artificial dataset made of 1 variable to explain and 4 continuous predictors, made fully randomly. I added all interactions in a model and did it ten times. On the ten analyses, there was only 3 dataset for which there was no significant effect (interaction or simple effect). The 7 others had at least one significant effect (5 with at least two) and sometimes p-values were extremely low. You will have much more than 4 predictors. C) With 10 predictors, there are around 1000 potential interactions for a dataset of 64. In your power analysis, you indicated df = 10 which does not include interactions as added in your model.

3) Comment/Response 5: I was actually explaining that you can measure a metric explaining consistency and use it to study the relationship between consistency in exploration and flexibility. You actually explain more flexible individuals to be less consistent. However, you often need more than two repeats per individual. For ways to measure intra-individual variation, you can check Cleasby et al. 2015 (MEE) or Biro & Andriaenssens 2013 (Am Nat).

4) For the sample size, my point was more that the sample size within each population is important. It is difficult to have a subsample large enough to estimate behavioral variation but it depends on the population size (e.g. 16

individuals on a population made of 1000 individuals) and on the method to capture individuals which can bias the type of individuals captured.

5) For the number of populations used, I understand the difficulties. I repeat my advice as it is a main reason to reject a manuscript (as I saw as a reviewer and an editor). With three populations, you cannot conclude on the position on a range expansion. Let's imagine you have a predator around the core population. It will likely explain a large part of behavioral variation. I understand that you cannot increase sample size. You should at least sample individuals at different at different locations at the core/edge of the distributions, without increasing the number of individuals or at the very least you should choose populations the most similar (but it would require a complete knowledge of local ecological conditions).

Author's reply:

Dear Dr. Fronhofer and reviewer, We sincerely apologize for the delay in our revision. Due to some 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. Additionally, the grackles are making their way through the test battery more slowly than expected, which is good news in that we have not yet collected any data on on the causal cognition experiment.

We greatly appreciate the time you have taken for another round of revision! We have revised our preregistration (available at https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_causal.md) and we responded to your comments below (our responses are preceded by "> Response X").

Thank you very much for the opportunity to revise and resubmit!

All our best, Corina, Aaron, Zoe, Luisa, Carolyn, Benjamin, and Kelsey

Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles Aaron Blaisdell, Zoe Johnson-Ulrich, Luisa Bergeron, Carolyn Rowney, Benjamin Seitz, Corina Logan 10.17605/OSF.IO/GCA5V version v1.2 Submitted by Corina Logan 2018-08-20 11:09

Abstract This PREREGISTRATION has undergone one round of peer reviews. We have now revised the preregistration and addressed reviewer comments. The DOI was issued by OSF and refers to the whole GitHub repository, which contains multiple files. The specific file we are submitting is gcausal.Rmd, which is easily accessible at GitHub at https://github.com/corinalogan/grackles/blob/master/gcausal.Rmd. Note that viewing this file at OSF will result in not being able to see the figures as part of the .Rmd file. All changes are tracked at GitHub and are visible by clicking on the file's History button. Photo credit = Corina Logan (CC-BY-SA 4.0). We will start data collection in mid-October 2018 so it would be ideal if we could get through the review process before then. Keywords: Behavioral flexibility, causal cognition, comparative cognition, avian cognition

Round #2 Decision by Emanuel Alexis Fronhofer, 2018-11-06 10:15 Manuscript: 10.17605/OSF.IO/GCA5V version v1.2 Blaisdell et al. PCI Ecology preregistration - revisions

Dear Dr. Blaisdell, let me start by apologizing for the delay since the last round of revisions, unfortunately, referees were very hard to find. As a consequence only one of the original referees read your answers. The referee remains unconvinced by some of your answers, specifically with regards to data analysis. I would like to encourage you to once more revisit your plans based on the referee’s points. I am persuaded that clarifying these points at this state is the strength of preregistrations. I am looking forward to receiving a revised version of your preregistration. Sincerely yours, Emanuel A. Fronhofer

Reviews Reviewed by anonymous reviewer, 2018-10-11 14:38 Dear authors, Thank you for your replies. Overall, I feel you cannot change much. I understand. However, there are still several points that I think you should still consider again.

1) I do not get your argument about adding an ID effect as a fixed and a random effect. I am reading, reviewing and editing many manuscripts about behavioral consistency and I never saw a single one with ID as a fixed and a random effect. I am not saying it cannot be done, but my statistician colleagues are also extremely surprised by doing this and I don't think the model is going to converge. From my reading, the PDF you sent only explains that a predictor can be either fixed or random depending on the context, not that you can add this predictor as fixed and random simultaneously, but I might have missed this section. If you have ID as fixed, what is the point of adding it as random because you already explain most variance due to ID with the fixed effect? The only model I know with a same predictor as fixed and random are model in which the predictor is coded as continuous as a fixed effect and as categorical as a random effect (i.e. years) and it is debated whether it is correct to do so. The "usual" way to study "interactions" between ID and other covariates is to add a random slope on top of a random intercept. It allows testing for individual variation around the mean slope which is what you want to test here. I would say that Dingemanse and Dochtermann 2013 (JAE) or van de Pol and Wright 2009 (Animal Behaviour) are great introductions about random slopes.

Response 1. We were originally thinking that we would be able to use more than one number per variable per test per bird, which is why we wanted to use ID as both a fixed and a random effect in the same model. However, we revisited the data sheet and the model and realized that it is more feasible to just use one number per variable per test per bird, which means that we removed ID as a fixed effect and as a random effect in the models expl1 and expl2 (see the Exploration preregistration where these models appear https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_exploration.md). We also changed the response

variable to total number of trials to reverse on the most recent reversal for expl1.

For reference, here is the model we are discussing (examining the relationship between flexibility and exploration/boldness/neophilia/persistence): expl1 <- MCMCglmm(AvgTrialsToReverse ~ Condition + TimeOutsideNovelEnv + LatencyNovelEnv + AverageTimePerSectionNovelEnv + TotalNumberSectionsNovelEnv + LatencyTableObjectNeophilia + MultiaccessTouchesPerTime + LatencyObjectNeophobia + NoMotorActions + ID, random=~ID+Batch, family="poisson", data=explore, verbose=F, prior=prior, nitt=13000, thin=10, burnin=3000)

For reference, it is a recognized practice that a variable can be a fixed and a random effect in the same model. See the references below:

From Snijders (2005; http://www.stats.ox.ac.uk/~snijders/FixedRandomEffects.pdf), page 2: "The vantage point of multilevel analysis is that the effect of job level on work satisfaction (i.e., the regression coefficient of job level), could well be different across organisations. The fixed effect of this variable is the average effect in the entire population of organisations, expressed by the regression coefficient. Since mostly it is not assumed that the average effect of an interesting explanatory variable is exactly zero, almost always the model will include the fixed effect of all explanatory variables under consideration. When the researcher wishes to investigate differences between organisations in the effect of job level on work satisfaction, it will be necessary to specify also a random effect of this variable, meaning that it is assumed that the effect varies randomly within the population of organisations, and the researcher is interested to test and estimate the variance of these random effects across this population. Such an effect is also called a random slope."

Barr et al. (2013 J Mem Lang 68(3))

We were able to find several papers that include a variable as a fixed and a random effect in the same model, including papers by ecological statisticians including Loeske Kruuk (where Logan learned of this practice in the first place),

Jarrod Hadfield, and Andrew Bateman. Here are a few references with examples of where this occurs:

Garant, D., Kruuk, L. E., Wilkin, T. A., McCleery, R. H., & Sheldon, B. C. (2005). Evolution driven by differential dispersal within a wild bird population. Nature, 433(7021), 60. - page 64, selection analysis

English, S., Bateman, A. W., & Clutton-Brock, T. H. (2012). Lifetime growth in wild meerkats: incorporating life history and environmental factors into a standard growth model. Oecologia, 169(1), 143-153. - table 3, page 150

Phillimore, A. B., Leech, D. I., Pearce-Higgins, J. W., & Hadfield, J. D. (2016). Passerines may be sufficiently plastic to track temperature-mediated shifts in optimum lay date. Global change biology, 22(10), 3259-3272.

Here are some tutorials discussing this practice: https://stats.stackexchange.com/questions/173159/can-a-variable-be-both-random-and-fixed-effect-at-the-same-time-in-a-mixed-effec “The reason for this is that random effects are restrained to ∑γ=0∑γ=0, or always centered around 0. Thus, the random effect is the individual's estimated deviation from the group average for that individual. By leaving out the fixed effect, you would imply that the average effect of time must be 0.”

https://stats.stackexchange.com/questions/263194/does-it-make-sense-to-include-a-factor-as-both-fixed-and-random-factor-in-a-line “yes, it can make sense to include a factorial variable as fixed and a random effect. Depending on the data structure/ experimental design this may even be necessary to do so to arrive at valid conclusions”

2) I will have to disagree again on the model with predictors and interactions between all of them. The issue is not having 40 or 64. The issues are about A) making meaningful conclusion when have triple or quadruple (or it seems even higher ranked interactions) interactions between continuous variable. You cannot explain a pattern involving many covariates interacting each others. B) The probability of obtaining a significant interaction by chance. For instance, I just ran a simulation with artificial dataset made of 1 variable to explain and 4 continuous

predictors, made fully randomly. I added all interactions in a model and did it ten times. On the ten analyses, there was only 3 dataset for which there was no significant effect (interaction or simple effect). The 7 others had at least one significant effect (5 with at least two) and sometimes p-values were extremely low. You will have much more than 4 predictors. C) With 10 predictors, there are around 1000 potential interactions for a dataset of 64. In your power analysis, you indicated df = 10 which does not include interactions as added in your model.

Response 2. That's a fair point and we are not tied to examining interactions in most cases (actually, the interactions were likely a carry over from copying and pasting code from previous models and not something we specifically introduced on purpose). For models that had more than a couple of fixed effects, we replaced * with + to remove the interactions in the Exploration preregistration (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g*exploration.md), Flexibility preregistration (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g*flexmanip.md), and Inhibition preregistration (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_inhibition.md).

3) Comment/Response 5: I was actually explaining that you can measure a metric explaining consistency and use it to study the relationship between consistency in exploration and flexibility. You actually explain more flexible individuals to be less consistent. However, you often need more than two repeats per individual. For ways to measure intra-individual variation, you can check Cleasby et al. 2015 (MEE) or Biro & Andriaenssens 2013 (Am Nat).

Response 3. Because we are measuring individual differences in exploration and boldness using assays that incorporate novel and

threatening items, it is difficult to conduct more than two trials because habituation and learning begin to occur and cause change in the subject's responses. We are measuring within-test individual consistency with only two measurements (at time 1 and time 2) because we cannot conduct more measurements due to habituation to the objects. In terms of measuring whether there is something intrinsic to an individual that makes them behave the same way across different tests, we use several tests that have previously been shown to measure boldness or exploration, and this is recommended by various researchers in the field (e.g., Carter et al. 2013 Biol Rev).

We appreciate the suggestion to incorporate individual-level measures of consistency! It is an interesting idea. We want to validate that individuals are on average responding consistently to our tests because we first need to ensure that our methods have the capability of capturing an inherent trait. If results show that individuals have repeatable responses within a test, we will calculate whether individuals show consistent differences across the different tests. With two data points per test per individual we can calculate a difference score such that those individuals with scores closer to 0 are more consistent/predictable across tests. We will include this in our next revision of the Exploration preregistration, which is currently in review at PCI Ecology.

4) For the sample size, my point was more that the sample size within each population is important. It is difficult to have a subsample large enough to estimate behavioral variation but it depends on the population size (e.g. 16 individuals on a population made of 1000 individuals) and on the method to capture individuals which can bias the type of individuals captured.

Response 4. The total population size will vary among sites. There are probably ~400 great-tailed grackles in Tempe, AZ on the Arizona State University campus where our current field site is

centered. A population on the edge may have only 30 individuals, and we are not certain of population sizes in the center of their original range. Additionally, we don't have much control over our sample size because this species is highly unpredictable from site to site. For example, at Logan's original site in Santa Barbara, the grackles were easily trapped using large walk-in traps. These do not work on the Arizona grackles who have been extremely difficult to catch. Indeed, we have only caught and banded 42 grackles in 1 year in Arizona (and only a subset of these come into the aviaries). We would love to give some certainty around our sample size and we would absolutely love to increase the number of individuals we are able to test in the aviaries, however this is not looking promising, particularly given that the females are not very willing to participate in experiments so we have already had to replace 4 females in the aviaries because they never completed an experiment.

We realize that the trapping method can bias the type of individuals we might be catching (e.g., the more bold individuals go into the traps, therefore we are sampling only bold individuals, which is a subset of the population) and we are trying every trapping method we can think of that works in an urban environment (e.g., drop nets, mist nets, bownets, walk-in traps, carpet nooses), but our basic problem always comes down to none of these methods work very well so we have to take whatever individuals we can get. However, if we are sampling only the bold individuals, this will be the same across sites so sites will be comparable. Additionally, we are able to validate whether the aviary grackles are particularly bold by conducting boldness assays of the wild banded grackles, which will give us an indication of how generalizable the aviary results are to the broader population.

In terms of reducing the number of traits we sample, we are validating a variety of experiments using the Arizona field site as a testing ground and we will reduce the number of tests to the minimum number needed to answer our questions at subsequent field sites. Our Arizona findings will result in new preregistrations that will apply to the next field sites.

We are in the process of writing a new preregistration where we will sample population-level measures of behavior (i.e., flexibility, exploration, neophobia, persistence) using one behavioral test in the wild on unbanded grackles. It will allow us to compare the banded vs unbanded individuals at each of our three main field sites. For example, if the unbanded birds respond in the same way as the banded birds to this test, we will likely have representative sample of the population with regards to the other measures we are investigating.

5) For the number of populations used, I understand the difficulties. I repeat my advice as it is a main reason to reject a manuscript (as I saw as a reviewer and an editor). With three populations, you cannot conclude on the position on a range expansion. Let's imagine you have a predator around the core population. It will likely explain a large part of behavioral variation. I understand that you cannot increase sample size. You should at least sample individuals at different at different locations at the core/edge of the distributions, without increasing the number of individuals or at the very least you should choose populations the most similar (but it would require a complete knowledge of local ecological conditions).

Response 5. Please see our response 4 regarding the new preregistration. We are planning to use this new test on unbanded birds at several new sites beyond the three sites where we collect in-depth data on banded individuals.

Given the massive time and money investment in banding just the 42 grackles we have caught in one year on the Arizona State University campus, and keeping the field site and experiments running at the bare minimum level, we cannot spare the resources to establish additional trapping sites. Additionally, it appears impossible at this point to be able to determine in advance how similar great-

tailed grackle populations are to each other. The Santa Barbara and Tempe sites are very similar to each other in that both are urban environments where the grackles go from outdoor cafe to outdoor cafe at lunch time, raid garbage cans, and forage in grassy areas (e.g., sports fields and golf courses), however these two populations vary in behavior and in other traits (e.g., the Arizona grackles are smaller) in so many ways. We would not feel confident that even if we were able to choose two sites that are ecologically similar, that we would be choosing two grackle populations that are similar to each other.

We realize that in the causal cognition preregistration, the one involved in this review process, we cannot say anything about cross-population differences in causal abilities. This is why we were careful to phrase this preregistration as an investigation into whether this species possesses such abilities, and if they do, plan future experiments to investigate these abilities further.

Revision round #1

2018-09-23
Dear Dr. Blaisdell,

thank you very much for submitting your preregistration "Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles" to PCI Ecology. I would like to start by congratulating you and your co-authors for submitting this preregistration, you are pioneers and the ecology and evolution community should follow your example.

The science you propose is fascinating and of highest relevance, as it aims at bridging scales from the microscopic mechanisms that underlie animal behaviour to macroscopic, macroecological consequences (see also Keith & Bull, 2016, Ecography). I have now received two reviews of your preregistration and, while mirroring my enthusiasm, the two referees both mention a number of points that merit some thought. Some of these points certainly stem from them, at least in

part, not being familiar with approaches and ideas you mention. Nevertheless, these points should be clarified. Both reviews should help you clarify the experimental design and also very importantly the statistical analyses. Finally, I would like to point out that most of the links in the g_causal.Rmd file do not work, as they reference .html files and not .Rmd files. It would be very helpful if you could fix these issues for the future.

I suggest revising your preregistration in light of the referees' comments, accompanied by a detailed response to their criticism. I am looking forward to receiving a revised version of your preregistration.

Sincerely yours, Emanuel A. Fronhofer

Preprint DOI: 10.17605/OSF.IO/GCA5V

Reviewed by anonymous reviewer, 2018-09-07 16:12

Overall, the project is very interesting, quite ambitious, but doable to my opinion. I am really looking forward to see the results. All together, the different pieces will provide understand of range expansion of this species and in general. Below are few comments, which I hope will be useful. As the documents are full of information, I might have missed some. It can explain some of my comments.

I miss a document to know the time frame of experiment and which individuals are used in which experiment. I also miss some methods (I think). For instance, I could not find the timing for glucocorticoids measurements. It seems that you are measuring stress-response increase in glucorticoid, which means that you keep birds in hand more than 3 minutes. It is a good option, but baseline glucocorticoids levels are important to know in general and also to estimate the stress-induced increase in glucocorticoid (the difference between levels after stress and baseline).

Another example is that I am not sure how many times each individual will be measured in most behavioral assays. For exploration/neophillia/neophobia assays, I could not find the number of trials per individuals. It is important to know what possibilities you will have for statistical analyses. For example, if you have enough

trials, you can calculate a metric of consistency at the individual level, which can be add in models to test for relationships with flexibility.

Models often appear over-parameterized. For example, your model for flexibility-behavior relationships is: expl1 <- MCMCglmm(AvgTrialsToReverse ~ Condition * TimeOutsideNovelEnv * LatencyNovelEnv * AverageTimePerSectionNovelEnv * TotalNumberSectionsNovelEnv * LatencyTableObjectNeophilia * MultiaccessTouchesPerTime * LatencyObjectNeophobia * NoMotorActions * ID, random=~ID+Batch, family="poisson", data=explore, verbose=F, prior=prior, nitt=13000, thin=10, burnin=3000)

I see two issues. First, you cannot have an interaction with ID as a fixed effect as it is a random intercept. Second, in your model, you have interactions among up to 10 variables. You cannot run that, in particular with your sample size of 40.

This comment is true for many analyses (e.g. TrialsToReverseLast ~ Cort + NumberHeterophil + NumberLymphocytes + InflammatoryGene + NumberLeucocytes + AntiInflammatoryGene + Th2Gene + NumberParasites + NumberParasiteSpecies). Many variables will likely be correlated and you will have an issue of collinearity. I think it is better to test for correlations among variables first and run a PCA on them (for each part).

Sample sizes for behavioral assays are often too small (16 for causal cognition and 40 for 3 populations for other behaviors). Your project is mainly on individual variations and consistency. It requires a large sample size to capture the entire range of behavioral variation, in particular within each population. I can see that the project is huge, but you might want to consider reducing the number of traits measured to increase your sample size. It will also be important given your statistical analyses with so many traits.

For the expansion part, I think it is always better to have several sampling location at the edge and the core of range distribution instead of three sites. The age since expansion is not the only factor varying along site and local conditions (prey and predators among many things) may largely differ among sites, making the population age effect disappear. Given the scope and interest of the project, it would be a shame. Is it possible to have two recent and two old sites? You miss

the intermediate, but I think it is safer than one site per population age. You could even have 6 sites with the same total sample size.

For the flexibility, I wonder whether you would need another control group in which individuals learn a given color provide food (instead no color-food association)? Right now the manipulated and control groups are different for both having to make a right decision and the fact the right decision is changing over time.

Reviewed by anonymous reviewer, 2018-09-02 14:30

Dear Emanuel Fronhofer, dear Aaron Blaisdell and coauthors,

I have now read the protocol included in the preregistration entitled “Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles”. I think this project has the potential to produce interesting and potentially original results by making the link between areas of cognition biology that have rarely been considered together (causal reasoning and reversal learning). The species on which this protocol should be developed is also especially relevant, together with the context of the expansion of its range. However, lots of details are lacking in the protocol, making it very difficult to evaluate for me. Especially, the authors do not explain how they will control for inter-individual differences due to sex, age, size, or body condition, or to temperament. These different variables are more than likely to cause type 2 errors when testing for correlation between performances at cognitive tasks. In addition, the protocol lacks clear explanations on how the different performances will be measured, and mostly relies on other (published and unpublished) documents. Without reading another pre-registration and an article from 2006, I was thus not able to properly evaluate the experimental protocol. Finally, the theoretical framework provided here is still very preliminary (sometimes a bit misleading) and could be largely improved by considering the literature into more details. Note that my understanding of this protocol is affected by the fact that causal reasoning is out of my expertise. Despite these limitations, I tried to provide a feedback that might be useful to improve this protocol, by being as critical as possible. I provide more detailed comments below.

Abstract: I find the definition of behavioral flexibility misleading as it tends to include causal reasoning, while the aim of the project is to test whether causal reasoning is associated with behavioral flexibility or not. Maybe avoid the expression “behavioral flexibility” (for a discussion on the use and misuse of this expression, see also Audet and Lefebvre 2017. What’s flexible in behavioral flexibility? Behavioral Ecology, Volume 28, Issue 4, 1 August 2017, Pages 943–947), and instead directly mention which performance was measured to estimate it (e.g. “reversal learning”). The sentence “is rarely directly tested in species in a way that would allow us to determine how it works” is largely unclear. Especially what “how it works” refers to? Neuronal mechanisms? I guess not, but this is what the expression suggests. The long-tailed grackle is an Icteridae, family with species displaying high innovation rates in the wild, and large brains relative to their body size, making it easier to study their cognitive ability. The expansion history of the species in the area of investigation also adds an interesting aspect to this project, allowing to look at how cognition may favor range expansion. Differences in populations age/expansion histories however need to be better detailed to better understand why and how cognitive differences might be expected between them. Where will the birds tested here come from?

B. Partitioning the results: This is a very personal opinion, but it is unclear to me why separating the results into two manuscripts would be more relevant. Though it will largely depend on the birds’ ability to use causal inference, writing a single article would probably have more impact, be more interesting for the reader, and less time consuming for the authors.

C. Hypothesis: The opposite hypothesis can also be expected, e.g. in the speed-accuracy syndrome framework, where some individuals may be very fast at solving new problems because they are fast at interacting with new tasks and providing a range of trials and errors. The outcome will likely largely vary according to the difficulty of the problem, and whether just trial and error allows a fast success, or whether causal reasoning is needed. Note that the authors seem to associate “behavioral flexibility” with “problem solving” here, whereas they then seem to consider “behavioral flexibility” as a latent variable relying on performance at reversal learning. Building on a stronger conceptual framework

describing how cognitive performance are expected to covary, and which ones are expected to affect expansion abilities (but this seems to be a key aspect of the study in the abstract that then disappears in the actual project description), should help fixing this issue. See especially all the work by Andrea S. Griffin's lab (Newcastle, Australia), on problem solving, learning etc in the context of invasion and urbanisation in common mynas and miners, and more generally on how cognition is affecting the ability to solve new problems.

Predictions: -The authors seem to expect performance at serial reversal learning and at solving new tasks after previously solved tasks become unavailable to be highly correlated. The literature does not always provide strong evidence for this, and positive, negative or absent correlations could be expected: unless such a correlation has already been demonstrated in this species, this is a first step to develop, which will also inform on how cognitive abilities are associated within this grackle. If the two tests are not associated, how will the authors decide of which birds are "more flexible"? -"successfully solve new tasks after previously solved tasks become unavailable" needs more details: are these completely new tasks? Or do the birds have to inhibit a previous behaviour before finding a new solution? -the P1 and P2 are largely unclear to me here -how will the authors control for habituation to captivity? Behavioral differences (i.e. temperament traits, such as boldness, neophobia or exploratory behavior)? This is a fundamental aspect that cannot be ignored: you may obtain spurious associations between cognitive traits just because of temperament differences, or differences in sensitivity to captivity. -the interpretation provided in alternative 2 (negative association between measures of causal inference and behavioural flexibility) is largely unclear (why should relying on current cues rather than previous ones necessarily have a negative effect on causal inference?). The figures are largely unclear without any explanation.

Objectives: This is not my area of expertise, but at this stage, I still do not understand what exactly the authors mean by "causal models", and "causal models from contingency learning". The objectives are mostly a summary of a protocol used in a previous article (which I have not read), but do not explain

what this protocol will exactly allow to measure and why it is relevant. I guess more information on causal cognition before the objectives would also help.

D. Methods Experiments: The description of these experiments considers that the reader already knows this protocol, and has read the Blaisdell 2006 paper, which I haven't. As a result, I am not able to evaluate this protocol properly. A detailed description of the protocol would be very helpful here. E.g. what exactly are "the Tone and the Noise"? What are the "keys" the authors are referring to? Symbols on the screen? Unclear also what the role of "Light" is here. Detailed explanation of the aims of the protocol, with a wording directly understandable by a reader unaware of the Blaisdell 2006 paper would be useful. Note that just referring to the models presented in the figures was not very helpful as the authors do not explain these models, which are not self-understandable (to me at least).

The Apparatus: Has the species been tested before with this kind of apparatus? Are the authors confident that they will interact with it the way they are hoping for? Will the birds be trained to use this screen before the experiments start?

Dependent variables : How will the birds know about the food delivery symbol? I imagine that a training session is involved, but this is not detailed here. What is the cost of pecking on the food dispenser if no food is provided? If there is no cost at all, why would the birds necessarily decrease the number of pecks, even if they use causal inference? A negative result (e.g. no difference in the number of pecks) will be very difficult to interpret, an important aspect to acknowledge in this protocol.

Independent variables and predictions:

There is no information on the protocols to measure reversal learning and multi-access box. I understand that these may be included in a different preprint, but it makes it very hard to evaluate their relevance here.

E. Analysis plan: Unclear why you expect missing data. "The contribution of each independent variable will be evaluated using the Estimate in the full model". This will only be meaningful if all your explanatory variables are scaled to the same mean and variance (e.g. mean of 0 and variance of 1). Note that the power analyses provided are only meaningful if the authors neglect a range of

potentially confounding factors. Sex, body mass/size or body condition, individual variation in temperament however need to be considered in these analyses, as else, irrelevant but potentially strong correlations could be obtained.

F. Planned sample: For how long will the birds be kept in captivity? How will the birds be fed? Details on the habituation period are necessary.

Author's reply:

Dear Dr. Fronhofer and reviewers,

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 have revised our preregistration (available at https://github.com/corinalogan/grackles/blob/master/g_causal.Rmd) 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, Aaron, Zoe, Luisa, Carolyn, Benjamin, and Kelsey

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

Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles

Aaron Blaisdell, Zoe Johnson-Ulrich, Luisa Bergeron, Carolyn Rowney, Benjamin Seitz, Corina Logan

10.5281/zenodo.1346006 version 1.1

Submitted by Corina Logan 2018-08-20 11:09

Abstract

The DOI refers to the whole GitHub repository, which contains multiple files. The specific file we are submitting for peer review is gcausal.Rmd, which is easily accessible in this repo at https://github.com/corinalogan/grackles/blob/master/g_causal.Rmd. Photo credit = Corina Logan (CC-BY-SA 4.0). We will start data collection in mid-October 2018 so it would be ideal if we could get through the review process before then. Also, expect a follow up email with draft emails for reviewers regarding preregistrations, as well as instructions on how the prereg process works.

Keywords: Behavioral flexibility, causal cognition, comparative cognition, avian cognition

Round #1

Comment 1: Decision

by Emanuel Alexis Fronhofer, 2018-09-10 13:23

Manuscript: 10.5281/zenodo.1346006

Dear Dr. Blaisdell,

thank you very much for submitting your preregistration "Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles" to PCI Ecology. I would like to start by congratulating you and your co-authors for submitting this preregistration, you are pioneers and the ecology and evolution community should follow your example. The science you propose is fascinating and of highest relevance, as it aims at bridging scales from the microscopic mechanisms that underlie animal behaviour to macroscopic, macroecological consequences (see also Keith & Bull, 2016, Ecography).

Response 1: Thank you so much for your enthusiasm about the peer review of preregistrations! We're excited to be part of this new venture! Thank you also for the positive feedback and for pointing out a relevant paper that we were not aware of. Keith & Bull (2016) will help us structure our analyses when we refine our longer-term cross population preregistrations to prepare them for peer review.

Comment 2: I have now received two reviews of your preregistration and, while mirroring my enthusiasm, the two referees both mention a number of points that merit some thought. Some of these points certainly stem from them, at least in part, not being familiar with approaches and ideas you mention. Nevertheless, these points should be clarified. Both reviews should help you clarify the experimental design and also very importantly the statistical analyses. Finally, I would like to point out that most of the links in the g_causal.Rmd file do not work, as they reference .html files and not .Rmd files. It would be very helpful if you could fix these issues for the future.

Response 2: We agree that the comments are very useful for helping us clarify our work. Thank you for letting us know about the broken links! We were not aware of this issue and it appears to have caused problems during the review process because the reviewers couldn't get access to the other preregistration we referred to. We apologize! We went through the preregistration, changed all of the links from .html to .Rmd, and then tested the links to ensure they now work properly.

Reviewed by anonymous reviewer, 2018-09-07 16:12

Comment 3: Overall, the project is very interesting, quite ambitious, but doable to my opinion. I am really looking forward to see the results. All together, the different pieces will provide understand of range expansion of this species and in general. Below are few comments, which I hope will be useful. As the documents

are full of information, I might have missed some. It can explain some of my comments.

Response 3: Wow, we are so impressed that you read all of the preregistrations at our GitHub Grackle repository! Your feedback on the bigger picture of the project and how the pieces fit together is very valuable and welcome. We had imagined that each preregistration would be peer-reviewed as a separate piece, but now that we see your comments, perhaps this is a new way that the peer review of preregistrations for larger projects should move forward: by having at least one reviewer review all of the work and comment more broadly. Thank you! Since we had not planned on this, we completely understand why it is not very clear how all of the preregistrations fit together and what the sample sizes are for the cross-population preregistrations. Your comments have inspired us to work on integrating the preregistrations so it is clearer how they fit together, and also to make an overview figure of the whole project to show visually what the plan is. The figure is available at: https://github.com/corinalogan/grackles (scroll down a bit; or see it directly here: https://github.com/corinalogan/grackles/blob/master/GracklePr ojectTimeline.png). We also added a link to this figure in the abstract of the causal cognition preregistration.

Comment 4: I miss a document to know the time frame of experiment and which individuals are used in which experiment. I also miss some methods (I think). For instance, I could not find the timing for glucocorticoids measurements. It seems that you are measuring stress-response increase in glucorticoid, which means that you keep birds in hand more than 3 minutes. It is a good option, but baseline

glucocorticoids levels are important to know in general and also to estimate the stress-induced increase in glucocorticoid (the difference between levels after stress and baseline).

Response 4: For the time frame and for clarification about which individuals experience which tests, please see our new overview figure of the whole project at: https://github.com/corinalogan/grackles (scroll down to 5-year Project Overview).

Several of the preregistrations do not have the protocols listed yet - these protocols will have the more in depth methods for each experiment. We are adding these as we submit preregistrations for peer review at PCI Ecology (g*flexmanip.Rmd just went through its first peer review, and g*inhibition.Rmd and g_exploration.Rmd are just about ready to submit).

You are correct in that we will be measuring cort when the grackles have been in hand for more than 3 minutes. This is because we are not able to catch the bird, remove it from the trap, bring it to the processing tent, and get its blood within 3 minutes. We do not plan to obtain baseline cort levels for grackles because this would be too time intensive. However, we think this won't be too much of a problem because we are only interested in between-individual variation of the same cort measurement (e.g., after 3 min) to relate this to individual variation in behavioral measures. Additionally, for those birds we bring into the aviaries, we collect their blood before we release them. These blood collections are conducted the same number of

hours after sunrise across all birds to get a more standardized cort measure for the individuals with more data (i.e., those that have undergone the behavioral tests in the aviary).

Comment 5: Another example is that I am not sure how many times each individual will be measured in most behavioral assays. For exploration/neophillia/neophobia assays, I could not find the number of trials per individuals. It is important to know what possibilities you will have for statistical analyses. For example, if you have enough trials, you can calculate a metric of consistency at the individual level, which can be add in models to test for relationships with flexibility.

Response 5: We had not added the protocols for this preregistration yet, so this is why the methods weren't visible. We have now added them (available at https://github.com/corinalogan/grackles/blob/master/g_exploration.Rmd under Open Materials, or click this link to see them directly: Testing protocols).

We assess the consistency of individual differences in exploration of novel environments and novel objects, and boldness. First and second trials of each assay will be conducted approximately 6 weeks apart for each subject. For exploration we have 2 treatment conditions, novel environment and novel object exploration. In both conditions we will compare behavior (latency to approach, number of contacts) in the novel environment and with the novel object to the behavior of the subject in response to the familiar environment and a familiar object. For boldness we will measure behavioral responses to 3 treatment conditions: novel object, novel predator, known predator.

To determine individual consistency in response to boldness and exploration treatments we will use poisson mixed effect models where the dependent variable is the latency to approach (under Analysis Plan https://github.com/corinalogan/grackles/blob/master/g_exploration.Rmd). Fixed effects will include Day and treatment condition, and we will include a random effect for individual ID. Subjects show consistent individual differences in exploration and/or boldness if the addition of the random ID effect significantly improves the model over one without it, indicating more variance between than within individuals. If behavior is consistent, we will test for a relationship between flexibility and exploration/boldness by incorporating the average latency to approach, or the average latency to contact the novel object as a covariate in those models.

If we do not find that the subjects respond consistently across the first and second trials for exploration or boldness, we will not include these variables in further analyses.

Comment 6: Models often appear over-parameterized. For example, your model for flexibility-behavior relationships is: expl1 <- MCMCglmm(AvgTrialsToReverse ~ Condition * TimeOutsideNovelEnv * LatencyNovelEnv * AverageTimePerSectionNovelEnv * TotalNumberSectionsNovelEnv * LatencyTableObjectNeophilia * MultiaccessTouchesPerTime * LatencyObjectNeophobia * NoMotorActions * ID, random=~ID+Batch, family="poisson", data=explore, verbose=F, prior=prior, nitt=13000, thin=10, burnin=3000) I see two issues. First, you cannot have an interaction with ID as a fixed effect as it is a random intercept. Second, in your model, you have interactions among up to 10 variables. You cannot run that, in particular with your

sample size of 40. This comment is true for many analyses (e.g. TrialsToReverseLast ~ Cort + NumberHeterophil + NumberLymphocytes + InflammatoryGene + NumberLeucocytes + AntiInflammatoryGene + Th2Gene + NumberParasites + NumberParasiteSpecies). Many variables will likely be correlated and you will have an issue of collinearity. I think it is better to test for correlations among variables first and run a PCA on them (for each part).

Response 6: Including ID as both a fixed and random effect in the same model is a common practice to look at the effect of individuals on the response variable and to control for the effect of individual variation in general (e.g., Snijders 2005 http://www.stats.ox.ac.uk/~snijders/FixedRandomEffects.pdf). ID as a fixed effect gives information about whether particular individuals (the average of all responses for that individual) relative to the other individuals impact the response variable. Whereas, the random effect gives information about how clumped the measurements within an individual, which allows us to determine whether there are individual differences more generally (it also generally accounts for the non-randomness of the data due to repeated measures on the same individuals).

Regarding the sample size, we can see where this got a bit confusing because we weren't very clear. The sample size for the 5-year studies, which the model in your comment above is from, will be more like a minimum of 64 (see the new overview figure at: https://github.com/corinalogan/grackles). We revised the Sample Size Rationale to make this clearer in the following preregistrations:
https://github.com/corinalogan/grackles/blob/master/g*withinpo*

p.Rmd,
*https://github.com/corinalogan/grackles/blob/master/g*flexgenes.Rmd,
https://github.com/corinalogan/grackles/blob/master/g*flexforaging.Rmd,*
*https://github.com/corinalogan/grackles/blob/master/g*expansion.Rmd

We used a power analysis (performed in G*Power http://www.gpower.hhu.de/en.html) to determine what effect size we could estimate from a sample size of 64 (for flexibility measures) at a power of 0.70 with a model containing 10 predictors. Here is the output from the power analysis: F tests - Linear multiple regression: Fixed model, R^2 deviation from zero Analysis: A priori: Compute required sample size Input: Effect size $f^2 = 0.25$ α err prob = 0.05 Power (1-β err prob) = 0.7 Number of predictors = 10 Output: Noncentrality parameter $\lambda = 16.0000000$ Critical F = 2.0147024 Numerator df = 10 Denominator df = 53 Total sample size = 64 Actual power = 0.7070973

This means that, with our sample size of 64, we have a 71% chance of detecting a medium (approximated at f^2=0.15 by @cohen1988statistical) to large effect (approximated at f^2=0.35 by @cohen1988statistical). Note: the reason that some of the other preregistrations don't have power analyses yet is because we are not finished preparing them for peer review. We prefer to avoid PCA analyses because we are interested in the relationship between each explanatory variable and the response variable,

and we prefer to keep our data in as raw a format as possible to preserve as much information about individual variation as we can. We plan to examine the relationship between each explanatory variable with the response variable (noted in the Data Checking section), which will give us an idea of which relationships are more strongly correlated. To those models with several explanatory variables (e.g., g*expansion and g*exploration preregistrations), we can add an analysis that examines the relative influence of each variable on the response variable (using the dredge function in the MuMIn package in R). We have now added to the Analysis Plan in g_causal: “We realize that there are other variables that are not included in the analyses below that may have an influence in our models if they were included (e.g., individual differences in body size, sex, exploration, boldness, etc.). Many of these variables we will have measured on these particular individuals. We have chosen to keep the models as simple as possible because the sample sizes for each experiment are small. These experiments were designed to determine whether grackles attend to causal cues or not. If results show that they do, then we will conduct further tests to investigate the extent of these abilities. The combination of conducting multiple experiments on the same cognitive ability on different individuals at different times and locations will not only increase our overall sample size, but it will show that we were able to detect the trait we we were measuring.” Also, we are planning to run analyses in Stan after Logan learns more about the program (noted in the Alternative Analysis section). These Bayesian analyses are much

better at handling explanatory variables that are correlated with each other and at handling latent variables. We will add these new analyses to the preregistrations before conducting the analyses that are already outlined. References Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2E. Hillsdale, NJ: Lawrence Elrbaum Associates, Inc. J. Cohen1988Statistical Power Analysis for the Behavioral Sciences, 2EHillsdale, NJLawrence Elrbaum Associates. Snijders, Tom A.B. 'Fixed and Random Effects'. In: B.S. Everitt and D.C. Howell (eds.), Encyclopedia of Statistics in Behavioral Science. Volume 2, 664-665. Chicester (etc.): Wiley, 2005.

Comment 7: Sample sizes for behavioral assays are often too small (16 for causal cognition and 40 for 3 populations for other behaviors). Your project is mainly on individual variations and consistency. It requires a large sample size to capture the entire range of behavioral variation, in particular within each population. I can see that the project is huge, but you might want to consider reducing the number of traits measured to increase your sample size. It will also be important given your statistical analyses with so many traits.

Response 7: Sorry for the confusion around sample sizes! Please see Response 6 for clarification around the sample sizes across all of the populations, which is predicted to be more like 64. The sample size for the the two planned causal cognition experiments is 16 for each experiment, totalling 32 across both experiments for the Arizona population. If the grackles attend to causal cues at all, we are going to continue with the causal cognition experiments at the other populations as well (through year 5 of the project). So the sample size across all causal reasoning experiments will be much larger (~64), but this will depend on

whether the grackles are able to use these kinds of causal inferences at all. If not, then we will likely discontinue causal cognition experiments. If so, then we will explore their causal cognition abilities through a variety of experiments. The same experiment will likely not be used across all individuals as we explore their causal cognition, but there will be individual variation within each experiment, which will allow us to obtain a measure that should apply at the cross-experiment level as well.

Comment 8: For the expansion part, I think it is always better to have several sampling location at the edge and the core of range distribution instead of three sites. The age since expansion is not the only factor varying along site and local conditions (prey and predators among many things) may largely differ among sites, making the population age effect disappear. Given the scope and interest of the project, it would be a shame. Is it possible to have two recent and two old sites? You miss the intermediate, but I think it is safer than one site per population age. You could even have 6 sites with the same total sample size.

Response 8: Thank you for your insight on this! We have 5 years of funding for this project and we think that the maximum number of field sites that we will be able to set up and run in this period of time is 3 because it takes so long to band a population and set up aviaries and conduct the aviary tests. It is too late to skip the intermediate site in Arizona because we are just started data collection there (on the flexibility experiment) a few days ago. However, we have discussed running a flexibility/individual differences test on groups of unmarked grackles at additional sites, which could really help with the issue that you bring up. The extremely time consuming part of the project is catching grackles and also setting up aviaries and testing them there, so

eliminating these components could make additional sites feasible and perhaps we could catch a few birds at these other sites so we would have some indication of their blood and DNA measures at the group level. The test we are thinking of running at these additional sites is the multiaccess box. It would occur in the wild and not in visual isolation of other individuals because that wouldn't be feasible. The individuals would not be marked, but perhaps we could make it so the box marks an individual with paint as it interacts with the 4 options, thus we would know who has already solved versus who hasn't. We will write a preregistration for this later in the project because we expect we will want to try gathering this additional data on the edge when we are at the primary edge population field site and when we are at the primary population in the original part of their range.

Comment 9: For the flexibility, I wonder whether you would need another control group in which individuals learn a given color provide food (instead no color-food association)? Right now the manipulated and control groups are different for both having to make a right decision and the fact the right decision is changing over time.

Response 9: Interesting idea! We don't have a large enough sample size to create an additional group, and we thought about replacing the yellow tubes in the control group with just sticking with the last color they learned to prefer (for the control group this would be the rewarded color in their first reversal, which is the only reversal they receive). However, this would be training them to very strongly prefer one color over the other because they would get hundreds of trials of being rewarded for choosing

this color. This could pose a problem for the subsequent reversal learning experiment on the touch screen where there will be a light purple and a dark purple color to choose from - they might transfer to having a strong initial preference for the lighter or the darker color depending on which color (lighter gray or darker gray) was their rewarded color in the control group in the color tube reversal learning experiment. Also, using the yellow tubes where both tubes have rewards means that their environment is not giving them any information about how it might vary, whereas continuing to reward one color for the control group gives them some information about the environment. In our current set up with the yellow tubes, birds will still need to make a choice (which side to go to) because they will only get to look into one tube on each trial, just like the manipulated group. In the control group's case, the choice is simply about which side to choose because both tubes contain a reward.

Reviewed by anonymous reviewer, 2018-09-02 14:30

Comment 10: Dear Emanuel Fronhofer, dear Aaron Blaisdell and coauthors, I have now read the protocol included in the preregistration entitled "Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles". I think this project has the potential to produce interesting and potentially original results by making the link between areas of cognition biology that have rarely been considered together (causal reasoning and reversal learning). The species on which this protocol should be developed is also especially relevant, together with the context of the expansion of its range.

Response 10: thank you for your encouragement!

Comment 11: However, lots of details are lacking in the protocol, making it very difficult to evaluate for me. Especially, the authors do not explain how they will control for inter-individual differences due to sex, age, size, or body condition, or to temperament. These different variables are more than likely to cause type 2 errors when testing for correlation between performances at cognitive tasks.

Response 11: Sorry for the lack of protocol detail! We hope our responses above and below and our revisions help clarify things. We measure temperament, body condition and size, sex, and age for these individuals in other preregistrations (see https://github.com/corinalogan/grackles) and we examine the relationship of these other variables with flexibility. We do not plan to include these other variables in the causal cognition analyses because we will only have 16 individuals per experiment and we do not want to subset the data further in analyses. In terms of age, we will only measure adults in the causal cognition experiments. We are not able to age the grackles beyond juvenile (less than 1 year of age) or adult (more than 1 year of age). As well, each experiment is balanced for sex (50% female in each experiment allocated evenly across treatment conditions). We added this to the Assignment to Conditions section:

"*Sex* is balanced across each experiment (50% female in each experiment) and allocated evenly across treatment conditions."

We will be able to speculate about whether temperament could potentially play a role in the causal cognition experiments because we are investigating whether there are consistent individual differences in traits such as exploration and boldness

(see a separate preregistration: https://github.com/corinalogan/grackles/blob/master/g_exploration.Rmd). If we do not find consistent individual differences in these traits, then temperament will not be a potential confound because it won't exist for the traits we measured. Regardless, we are only measuring a small number of temperamental traits and there are any number of other variables that we are not measuring that could potentially influence grackle performance. This will always be a problem because it is not feasible to measure everything. That said, each trial on the touch screen is designed to begin only after the bird pecks a shape on the screen, which indicates it is motivated to participate. This will help eliminate temperamental confounds because if a bird participates at all, then this controls for differences in exploration, boldness, and potentially persistence. All individuals will be trained to use the touch screen so it won't just be the more exploratory or bold ones who know how to participate because they dared to explore and figure it out. In Logan's previous experience in working with this species, almost all individuals complete every experiment. Therefore, we are encouraged that this should be the case with future grackles as well.

Additionally, this first set of causal cognition experiments are designed to test whether this species attends to causal cues at all. If so, we will design subsequent experiments to explore this ability further, which will result in a much larger total sample size across all of these experiments over the 5-year project (see 5-

year project overview at https://github.com/corinalogan/grackles). This will help us to determine how robust our results are because if the grackles in each experiment keep showing that they attend to causal cues, then each study will build a larger set of evidence that we are measuring what we think we are.

Comment 12: In addition, the protocol lacks clear explanations on how the different performances will be measured, and mostly relies on other (published and unpublished) documents. Without reading another pre-registration and an article from 2006, I was thus not able to properly evaluate the experimental protocol.

Response 12: We apologize for the lack of clarity. We have added considerably more detail to the Methods section, and the Objective section now better summarizes the procedure used in the Blaisdell et al. (2006) paper. Namely, during test trials we have elaborated on the different response keys (4 cm blue square and 4 cm green triangle) that will be made available to subjects. The list of dependent variables has also been elaborated along with the procedure for both experiments 1 and 2. We hope these additions make the protocol easier to follow.

Comment 13: Finally, the theoretical framework provided here is still very preliminary (sometimes a bit misleading) and could be largely improved by considering the literature into more details. Note that my understanding of this protocol is affected by the fact that causal reasoning is out of my expertise. Despite these limitations, I tried to provide a feedback that might be useful to improve this protocol, by being as critical as possible. I provide more detailed comments below.

Response 13: Thank you for your detailed comments. We address them below.

Comment 14: Abstract: I find the definition of behavioral flexibility misleading as it tends to include causal reasoning, while the aim of the project is to test whether causal reasoning is associated with behavioral flexibility or not. Maybe avoid the expression "behavioral flexibility" (for a discussion on the use and misuse of this expression, see also Audet and Lefebvre 2017. What's flexible in behavioral flexibility? Behavioral Ecology, Volume 28, Issue 4, 1 August 2017, Pages 943–947), and instead directly mention which performance was measured to estimate it (e.g. "reversal learning"). The sentence "is rarely directly tested in species in a way that would allow us to determine how it works" is largely unclear. Especially what "how it works" refers to? Neuronal mechanisms? I guess not, but this is what the expression suggests. The long-tailed grackle is an Icteridae, family with species displaying high innovation rates in the wild, and large brains relative to their body size, making it easier to study their cognitive ability. The expansion history of the species in the area of investigation also adds an interesting aspect to this project, allowing to look at how cognition may favor range expansion. Differences in populations age/expansion histories however need to be better detailed to better understand why and how cognitive differences might be expected between them. Where will the birds tested here come from?

Response 14: In the last sentence of the abstract, where we discuss how flexibility relates to the causal cognition experiments, we replaced "behavioral flexibility" to more specifically refer to the ways in which we tested flexibility. It now reads: "Results will allow us to determine whether causal cognition is linked with our measures of flexibility (reversal learning and solution switching)"

Regarding the flexibility definition, we attempted to clarify the definition we use by revising it to read: "the ability to change behavior when circumstances change based on learning from

previous experience" We agree with Audet and Lefebvre (2017) that the term "behavioral flexibility" gets confusing because of its variable use in the literature. Therefore, just after our definition, we added a citation to the publication in which Logan and two philosophers of science developed the theory behind our particular definition (Mikhalevich et al. 2017; available at http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2016.0121). We also agree with Audet and Lefebvre (2017) that it is better to say what you measured rather than to give it an ambiguous term. In the first part of the abstract, we wanted to keep behavioral flexibility as a broader category because later in this 5-year project, we may want to measure other aspects beyond reversal learning and task switching on the multiaccess box. To clarify the distinction between our larger investigation of flexibility across populations and the two measures of flexibility that we compare here with causal cognition, we clarified in the abstract that the first part is about the "Project background" and the second part is about "This investigation".

We agree that some types of causal cognition experiments could test flexibility. However, according to our definition of flexibility, the causal cognition experiments we are using here do not. This is because here we are measuring whether they attend to causal cues at all. To turn this into a flexibility test, we would need to modify the experiment such that they have to use what they have learned previously about causal cues in a new situation in a way that shows that their behavior is functionally matching how

the environment has changed. We are very interested in designing such experiments, but first we want to see whether they can attend to causal cues at all before planning further experiments.

We apologize for the broken link to the preregistration on behavioral flexibility! We have now fixed the link, which is available here: https://github.com/corinalogan/grackles/blob/master/g_flexmanip.Rmd. In the flexibility preregistration we discuss how we investigate “how it works”. Specifically, we investigate whether it is manipulatable, whether flexibility generalizes to new contexts, whether it is repeatable within individuals and across contexts, and what learning strategies they might be using. For the causal cognition preregistration, we replaced the part about “how it works” with: “However, behavioral flexibility is rarely directly tested at the individual level, thus limiting our ability to determine how it relates to other traits, which limits the power of predictions about a species' ability to adapt behavior to new environments”

We plan to understand more about differences between population ages and expansion histories, and how flexibility relates, in two other preregistrations, which are available at: https://github.com/corinalogan/grackles/blob/master/g*flexgenes.Rmd* *and* *https://github.com/corinalogan/grackles/blob/master/g*expansion.Rmd.

The birds in this preregistration are from Tempe, Arizona. Because of your comment here and Reviewer 1's comments, we realized we needed to more explicitly address how the various pieces of this project come together across time and space. You can see how we have addressed this in Response 3 above.

Comment 15: B. Partitioning the results: This is a very personal opinion, but it is unclear to me why separating the results into two manuscripts would be more relevant. Though it will largely depend on the birds' ability to use causal inference, writing a single article would probably have more impact, be more interesting for the reader, and less time consuming for the authors.

Response 15: We agree with you and we are planning to have only one resulting manuscript from this preregistration. The reason we have this note is because usually one preregistration results in one manuscript (i.e., registered report formats at several journals) and we really like the fact that PCI Ecology allows us to submit one preregistration that could result in more than one manuscript. This allows us to preregister our broader project hypotheses. The reason for the note is to have the ability to keep our options open, but you are right: for this preregistration it will make sense to keep the preregistration as one paper. We revised the text to say: "We may decide to present these results in two separate papers".

Comment 16: C. Hypothesis: The opposite hypothesis can also be expected, e.g. in the speed-accuracy syndrome framework, where some individuals may be very fast at solving new problems because they are fast at interacting with new tasks and providing a range of trials and errors. The outcome will likely largely vary according to the difficulty of the problem, and whether just trial and error allows a fast success, or whether causal reasoning is needed.

Response 16: This is an interesting alternative hypothesis. Our tasks vary in difficulty level and also in how much trial and error learning could help in solving the problem faster. In the case of reversal learning on the color tubes, trial and error learning is the only way to solve this problem (which is an easy problem to solve), whereas with the multiaccess box, they could use causal information to learn how to solve before interacting with the 4 options (there is variation among options in their level of solving difficulty but all are feasible for a grackle [based on pilot data collected in April 2018]). With the causal cognition experiments, only causal cognition will allow them to solve these tests. Incorporating the speed-accuracy syndrome as an alternative hypothesis would suggest that problems that are easier (e.g., color tubes) should be able to be solved faster and with higher accuracy than problems that are more difficult (e.g., causal cognition experiments). The multiaccess box is likely our only test where individuals could use both trial and error and causal cognition or a combination of the two. However, the multiaccess box test will not allow us to determine whether individuals that are faster to switch between solving options are relying more on causal cognition than trial and error learning. We are not able to discern why some solve it faster because we are only measuring the latency to solve. If you have an idea about how we could discern this, we would love to hear it.

Comment 17: Note that the authors seem to associate “behavioral flexibility” with “problem solving” here, whereas they then seem to consider “behavioral flexibility” as a latent variable relying on performance at reversal learning.

Building on a stronger conceptual framework describing how cognitive performance are expected to covary, and which ones are expected to affect expansion abilities (but this seems to be a key aspect of the study in the abstract that then disappears in the actual project description), should help fixing this issue. See especially all the work by Andrea S. Griffin's lab (Newcastle, Australia), on problem solving, learning etc in the context of invasion and urbanisation in common mynas and miners, and more generally on how cognition is affecting the ability to solve new problems.

Response 17: We consider behavioral flexibility and problem solving distinct from each other so thank you for catching this! Logan has written a paper on the conceptual framework between flexibility and other cognitive abilities (Mikhalevich et al. 2017) and we have now modified the text to account for this. We added to C: Hypothesis in two ways:

1) We added "(faster at functionally changing their behavior when circumstances change)" and we added a citation to provide theoretical context to the sentence: "Individuals that are more behaviorally flexible (faster at functionally changing their behavior when circumstances change), as measured by reversal learning and switching between options on a multi-access box, are better able to derive accurate causal inferences (see @mikhalevich*is*2017 for theoretical background about the distinction between flexibility and complex cognition)".

2) We changed the sentence following point 1 to: "This is because causal cognition may facilitate flexibility: an individual could be faster at switching to new solutions that are more functional if it makes causal inferences about how the problem works, rather

than relying solely on trial and error learning to indiscriminately switch to new solutions."

Our theoretical framework differs from work for example by Griffin's lab in that we are interested in individual level rather than species level variation, we do not incorporate brain size into our framework (see Logan et al. 2018), and this species of grackle is a native species that does not necessarily need to start from scratch about learning how to survive in a new environment as an introduced species might need to. Additionally, our prediction about how flexibility and causal cognition relate differs from Griffin et al. (2016) who state: "Neither operant learning, nor any other cognitive process (e.g. causal inference) is implicated in discovering a solution" (p. 35). Thanks to your comment, we think we have made a clearer link with the theoretical framework we have previously established, as well as clarified the distinction between flexibility, problem solving, and causal cognition.

The abstract is composed of two elements: a broader project background, and the specific piece this preregistration is tackling. We can see where this would be confusing so we have labeled these parts within the abstract "Project background" and "This investigation" so it is clear that this preregistration is one piece of a larger puzzle. We did not get into the details about which cognitive abilities (or other traits if cognitive abilities are not the primary influential variable) are expected to covary with range expansion in this preregistration because our broader hypothesis examines how flexibility, not other cognitive abilities, varies

across the range. We outline how we expect flexibility to vary across the range in a separate preregistration: https://github.com/corinalogan/grackles/blob/master/g_expansion.Rmd. However, because of your comment, it is clear that we need to more effectively elucidate how the causal cognition experiments plug into the bigger picture. We have revised the Abstract to say: “Results will indicate whether causal cognition might play a role in switching to functionally relevant solutions based on how it correlates with measures of flexibility (reversal learning and solution switching). This will improve our understanding of which variables are involved in flexibility and how they are related, thus putting us in an excellent position to further investigate the mechanisms behind these links in future research.”

References Logan, C. J., Avin, S., Boogert, N., Buskell, A., Cross, F. R., Currie, A., ... & Shigeno, S. (2018). Beyond brain size: uncovering the neural correlates of behavioral and cognitive specialization. Comparative Cognition and Behavior Reviews 13:55-90. doi: 10.3819/CCBR.2018.130008.

Comment 18: Predictions: -The authors seem to expect performance at serial reversal learning and at solving new tasks after previously solved tasks become unavailable to be highly correlated. The literature does not always provide strong evidence for this, and positive, negative or absent correlations could be expected: unless such a correlation has already been demonstrated in this species, this is a first step to develop, which will also inform on how cognitive abilities are associated within this grackle. If the two tests are not associated, how will the authors decide of which birds are “more flexible”?

Response 18: We are exploring the relationship between reversal learning and solving new tasks after previously solved tasks become unavailable in a separate preregistration on behavioral flexibility (available at: https://github.com/corinalogan/grackles/blob/master/g_flexmanip.Rmd - sorry this link was broken in the original preregistration!). In the flexibility preregistration, we outline our predictions for each potential outcome: a positive, negative, or no relationship between these variables. However, you bring up a great point: we should clarify which measure we will use for flexibility in case these variables are not positively correlated. We updated the Predictions section with the following: "Alternative 3: If the flexibility measures do not positively correlate with each other (P2 alternative 2 in the flexibility preregistration), this indicates they measure different traits. In this case, we are interested in how each flexibility measure relates to performance on causal inference tasks: the reversal learning measure as an indication of flexibility, and task switching latency on the multiaccess box as a measure of a combination of flexibility and innovation."

Additionally, we updated the Independent Variables section with a new measure we are developing: "4) Flexibility comprehensive: This measure is currently being developed and is intended be a more accurate representation of all of the choices an individual made, as well as accounting for the degree of uncertainty exhibited by individuals as preferences change. If this measure

more effectively represents flexibility (determined using a modeled dataset and not the actual data), we may decide to solely rely on this measure and not use independent variables 2 and 3. If this ends up being the case, we will modify the code in the analysis plan below to reflect this change."

Comment 19: -"successfully solve new tasks after previously solved tasks become unavailable" needs more details: are these completely new tasks? Or do the birds have to inhibit a previous behaviour before finding a new solution? -the P1 and P2 are largely unclear to me here

Response 19: We changed the sentence to read: "successfully solve new tasks after previously solved tasks become unavailable on a multiaccess box". And we added a (working) link to the preregistration in which the details can be found.

P1 has been clarified to now read: "form causal models from contingency learning (i.e., observational learning). Contingency information could be represented in one of two ways. On the one hand, relations between events could be encoded as associations. On the other, they could be represented as causal. For example, if the sound of a bell is followed by delivery of food, one could represent the bell as associated with the food, and thus the sounding of the bell calls to mind an expectancy of food. Or, the subject could represent the bell as a cause of food. Blaisdell et al., 2006 (see also Leising et al., 2008) report evidence that rats can represent statistical relationships between events as causal." Thus, we predict the more flexible individuals will better learn the causal maps between all pairwise events (visual and auditory cues and food delivered from a food dispenser), and integrate

these individual maps into larger causal map structures, including a common-cause, two-effect map (if observing T, L caused it, thus F is present), and a direct cause-effect map (if N is present, it will cause F)."

P2 has been clarified to now read: "behave as if intervention can influence the type of causal inference made at test, depending on which causal model is being tested: dissociate between seeing and doing as evidenced by a lower rate of pecking a key to release food when they had the opportunity to intervene in a common cause condition (I caused T, thus it wasn't caused by L, thus there is no F), while intervening on a direct cause (I caused N, but N causes F, thus look for food) or a causal chain (I caused T, but T causes L which causes F, so look for F) will have no effect on key pecks."

Comment 20: -how will the authors control for habituation to captivity? Behavioral differences (i.e. temperament traits, such as boldness, neophobia or exploratory behavior)? This is a fundamental aspect that cannot be ignored: you may obtain spurious associations between cognitive traits just because of temperament differences, or differences in sensitivity to captivity.

Response 20: We are not quite sure how to address habituation to captivity because we have never seen it controlled for before. All individuals undergo a 3-4 day habituation period to the aviaries and then their experiments begin so in this sense they all have the same experience which is a form of control. We expect individual variation in causal cognition task performances, however, they will need to already be habituated to the touch screen before we will be able to collect any data from the

individuals in these experiments. Thus habituation should not be an issue once they pass touch screen training. Thanks to your comments, we realized that readers needed more information, including details about the habituation period and touch screen training, which we have now added to the document (see Responses 26 and 29). Regarding the associations between cognitive traits and temperament differences, please see our Response 11 above.

Comment 21: -the interpretation provided in alternative 2 (negative association between measures of causal inference and behavioural flexibility) is largely unclear (why should relying on current cues rather than previous ones necessarily have a negative effect on causal inference?).

Response 21: The grackles will need to rely on previous information about cues for what predicts what in order to correctly solve the causal cognition tests, therefore if they are relying more on current cues (just the presence or absence of a light that is not tied to the information about what the presence or absence means in terms of a food reward) they will not be able to solve these tasks consistently. We revised this alternative to say: “For example, relying solely on current cues (i.e., the immediate stimulus (e.g., tone, noise) or lack thereof) in the causal cognition test will not give them enough information to consistently solve the task. They will need to draw on their memory of what the presence or absence of the current stimulus means about the food reward based on their experience in previous trials to perform well on this task.”

Comment 22: The figures are largely unclear without any explanation.

Response 22: Excellent point! We added figure legends.

Comment 23: Objectives: This is not my area of expertise, but at this stage, I still do not understand what exactly the authors mean by “causal models”, and “causal models from contingency learning”. The objectives are mostly a summary of a protocol used in a previous article (which I have not read), but do not explain what this protocol will exactly allow to measure and why it is relevant. I guess more information on causal cognition before the objectives would also help.

Response 23: We apologize for not better introducing causal models in the introduction. We have made several additions to address this which are listed below.

In the abstract we write, “We aim to determine whether the more behaviorally flexible (measured in a separate preregistration) grackles are better able to make causal inferences (understanding relationships beyond their statistical covariations) in two experiments using a touch screen apparatus.”

In C: Hypothesis we write, “This is because causal cognition may facilitate flexibility: an individual could be faster at solving new problems if it makes causal inferences about how the problem works, rather than relying solely on trial and error learning. In this procedure, we assess whether grackles are able to derive correct predictions about causal interventions after observational learning, a core component of causal reasoning that can not be reduced to associative learning (@waldmann2005seeing).”

In Objective we write, “Blaisdell and colleagues (@blaisdell2006causal) taught rats that a light was a common cause of tone and food by presenting the light followed by the

tone on some trials and by the food on other trials during training. Rats also learned that a noise was a direct cause of food by presenting noise and food simultaneously during training. At test, some rats observed the tone or the noise. When they did, they looked for food. This shows that rats had formed the causal models of noise causes food and that tone is caused by light, which itself is a cause of food. Other rats were given the opportunity to intervene to make the tone or noise occur at test. This was done by giving the rats a novel lever that they had never seen before or been trained on. When the pressed the lever, this caused the tone (or noise) to turn on. When the noise was caused by a lever press, rats looked for food in the food hopper, but when lever pressing caused the tone to turn on, rats did not look for food. This shows that rats understood that, by intervening on the lever to cause the noise to occur, since the noise was a cause of food, they then expected food. But by intervening on the lever to cause the tone to occur, the rats realized that they had caused the tone, and not the light (which was an alternative cause of tone). As a result of attributing the tone to their own action rather than the light, they did not expect there to be any food in the food hopper."

In Objective we write, "this dissociation between seeing and doing suggests that subjects represent associated relationships as causal, and derive rational inference regarding an intervention on a cause versus an effect."

Comment 24: D. Methods Experiments: The description of these experiments considers that the reader already knows this protocol, and has read the Blaisdell

2006 paper, which I haven't. As a result, I am not able to evaluate this protocol properly. A detailed description of the protocol would be very helpful here. E.g. what exactly are "the Tone and the Noise"? What are the "keys" the authors are referring to? Symbols on the screen? Unclear also what the role of "Light" is here. Detailed explanation of the aims of the protocol, with a wording directly understandable by a reader unaware of the Blaisdell 2006 paper would be useful. Note that just referring to the models presented in the figures was not very helpful as the authors do not explain these models, which are not self-understandable (to me at least).

Response 24: We have now added protocol information. Please see our Responses 12 and 22 above.

Comment 25: The Apparatus: Has the species been tested before with this kind of apparatus? Are the authors confident that they will interact with it the way they are hoping for? Will the birds be trained to use this screen before the experiments start?

Response 25: This species has not been tested before with a touch screen apparatus, but we were able to pilot their ability to learn to use the food hopper in April 2018 on a captive great-tailed grackle. The grackle was quickly able to learn to use the hopper. We expect they will be able to use the touch screen in the way we have planned, but we will not know for sure until we start their touch screen training. They will undergo touch screen training and we have now added this information to the preregistration as two new sections under Methods: Apparatus, and Touch Screen Training

Comment 26: Dependent variables : How will the birds know about the food delivery symbol? I imagine that a training session is involved, but this is not detailed here. What is the cost of pecking on the food dispenser if no food is provided? If there is no cost at all, why would the birds necessarily decrease the

number of pecks, even if they use causal inference? A negative result (e.g. no difference in the number of pecks) will be very difficult to interpret, an important aspect to acknowledge in this protocol.

Response 26: Thank you for pointing out our absence of mentioning food delivery symbool training. The following sentence has now been added in the methods of experiment 1, “Before contingency training, subjects completed response key autoshaping and instrumental conditioning. Subjects were trained to peck at the response key to activate the food hopper using a mixed autoshaping/instrumental training procedure.”

You are correct in that checking for food is cheap, which could work against us in testing our predictions. But, this is generally the case with operant or Pavlovian behaviors, and even in the study of causal inferences in rats, and yet differences between test conditions were found. This bolsters our confidence that the procedure will be sensitive to causal inference behaviors in grackles. Independent variables and predictions:

Comment 27: There is no information on the protocols to measure reversal learning and multi-access box. I understand that these may be included in a different preprint, but it makes it very hard to evaluate their relevance here.

Response 27: We apologize for the broken link to the flexibility preregistration! It is available here: https://github.com/corinalogan/grackles/blob/master/g_flexmanip.Rmd. The flexibility preregistration is undergoing its own peer review process at PCI Ecology, therefore we did not include details about these experiments here.

Comment 28: E. Analysis plan: Unclear why you expect missing data. “The contribution of each independent variable will be evaluated using the Estimate in the full model”. This will only be meaningful if all your explanatory variables are scaled to the same mean and variance (e.g. mean of 0 and variance of 1). Note that the power analyses provided are only meaningful if the authors neglect a range of potentially confounding factors. Sex, body mass/size or body condition, individual variation in temperament however need to be considered in these analyses, as else, irrelevant but potentially strong correlations could be obtained.

Response 28: We don’t expect missing data, but we wanted to include an explanation of how we would handle if if it occurs. Given that we are testing individuals that must volunteer to participate in the tests, there is always the possibility that some might not finish an experiment, which would produce missing data for that individual in that particular experiment.

Regarding the need to scale explanatory variables to the same mean/variance, I think this question arose because we were unclear about what we are interested in here. We simply want to determine whether a variable has an effect or not, not to determine how much of the variance is being explained by each variable. We clarified this in the text: “We will determine whether an independent variable had an effect or not using the Estimate in the full model.”

Please see Response 11 for an explanation about whether to include additional variables in the analyses.

Comment 29: F. Planned sample: For how long will the birds be kept in captivity? How will the birds be fed? Details on the habituation period are necessary.

Response 29: We revised the text in Planned Sample to provide more details: “Grackles are individually housed in an aviary (each 244cm long by 122cm wide by 213cm tall) at Arizona State University for a maximum of three months where they have ad lib access to water at all times and are fed Mazuri Small Bird maintenance diet ad lib during non-testing hours (minimum 20h per day), and various other food items (e.g., peanuts, grapes, bread) during testing (up to 3h per day per bird). Individuals are given three to four days to habituate to the aviaries and then their test battery begins on the fourth or fifth day (birds are usually tested six days per week, therefore if their fourth day in the aviaries occurs on a day off, then they are tested on the fifth day instead).”