Peer Community In Ecology

# Probing behaviors correlated with behavioral flexibility

Jeremy Van Cleve based on reviews by 2 anonymous reviewers

A recommendation of:

Kelsey McCune, Carolyn Rowney, Luisa Bergeron, Corina Logan. **Is behavioral flexibility linked with exploration, but not boldness, persistence, or motor diversity?** *(2019), In Principle Recommendation. PCI Ecology.* http://corinalogan.com/Preregistrations/g_exploration.html

Open Access

Published: 26 March 2019

Copyright: This work is licensed under the Creative Commons Attribution-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/4.0/

*Submitted: 27 September 2018, Recommended: 26 March 2019*

**Cite this recommendation as:**

Jeremy Van Cleve (2019) Probing behaviors correlated with behavioral flexibility. *Peer Community in Ecology, 100020.* 10.24072/pci.ecology.100020

Behavioral plasticity, which is a subset of phenotypic plasticity, is an important component of foraging, defense against predators, mating, and many other behaviors. More specifically, behavioral flexibility, in this study, captures how quickly individuals adapt to new circumstances. In cases where individuals disperse to new environments, which often occurs in range expansions, behavioral flexibility is likely crucial to the chance that individuals can establish in these environments. Thus, it is important to understand how best to measure behavioral flexibility and how measures of such flexibility might vary across individuals and behavioral contexts and with other measures of learning and problem solving. In this preregistration, Logan and colleagues

propose to use a long-term study of the great-tailed grackle to measure how much they can manipulate behavioral flexibility in a reversal learning task, how much behavioral flexibility in one task predicts flexibility in another task and in problem solving a new task, and how robust these patterns are within individuals and across tasks. Logan and colleagues lay out their hypotheses and predictions for each experiment in a clear and concise manner. They also are very clear about the details of their study system, such as how they determined the number of trials they use in their learning reversal experiments, and how those details have influenced their experimental design. Further, given that the preregistration uses RMarkdown and is stored on GitHub (as are other studies in the larger project), their statistical code and its history of modification are easily available. This is a crucial component of making research more reproducible, which is a recent emphasis in behavioral sciences more broadly. Reviewers of this preregistration found the study of substantial merit. The authors have responded to the reviewers' comments and their revisions have made the preregistration much clearer and cogent. I am happy to recommend this preregistration.

# Revision round #1

*2019-02-03*

Dear Dr. Logan,

Thank you for submitting your preregistration to PCI Ecology. Please forgive the delay in getting reviews and a decision to you.

Your study looks very interesting and both reviewers are enthusiastic about the potential results from the project. The reviewers suggest some areas for improvement in the preregistration and both focus on the need to better define terms within the "Hypotheses" section. They both also point out some confusion with the P6 alternatives. Finally, one reviewer points out some issues with the statistical analysis that should be addressed.

I encourage you to revise the preregistration according to the reviewer's comments and look forward to seeing the revision.

Please also note a couple of minor points of my own below.

Best wishes,

Jeremy Van Cleve

JVC Comments:

Figure 1. I am a bit confused reading this figure. Each experiment could be pictured by a box and then alongside it all the variables manipulated and measured listed. This might be more readable than the flowchart style.

Titles for predictions (P1-P8) are inconsistent and sometimes phrased as statements or questions (e.g., P1-P5). Please make these consistent.

Independent vs Dependent variables: Listing these for each prediction got me a little confused since some appeared multiple times in each category and some in both categories. I wonder if a simpler presentation would have a description of each variable and then a table listing whether the variable is independent or dependent for each prediction.

Phrase " If they are patternless, then assume a normal distribution". What does "patternless" mean?

"Power (1- err prob)". Specify which Type or error probability (i.e., Type II).

*Preprint DOI:* 10.17605/OSF.IO/GCA5V

## Reviewed by anonymous reviewer, 2018-12-24 15:48

First of all, I apologize for the delay. Overall the project sounds good. I have identified some points that could be improved either for data analysis but also for the clarity/reliability of the predictions.

H1: This first part mainly depends on whether or not behavioural flexibility is manipulatable, which made me uncomfortable to apprehend the predictions. I am also doubtful on the independence of reversal learning (which basically

depends on the exploration skill of an individual as it requires the focus individual to explore the different opportunities – right? – and its cognitive abilities too obviously) from novel environment exploration stricto sensus. I believe that these definitions should be clearly stated.

P1-P5 alternative: not sure that an absence of correlation is meaningful… It depends on the level of correlation the authors expect (?) and statistical robustness.

H2: To me behavioural consistency (also known as personality) is defined across time and/or context and is a property of the individual so I do not really understand P6-1 vs. P6-2 opposition. Personality measure is the highest value the heritability of the behavioural trait can reach. I am not sure we can separate the low and high repeatability score (additionally 95% CI are often large).

H3: Do the author consider trapping bias? It may be a more robust hypothesis than the effect of unfamiliar environment…

I did not review the R script but I have issues with data analyses: - Statistical power: 0.70 could be sufficient enough if the authors are expected large correlation coefficients but it is not usually the case with behavioural traits. - Poisson distribution: I suppose that this is because of count data. Negative binomial distribution is often more accurate for behavioural data.

## Reviewed by anonymous reviewer, 2018-11-26 21:54

Download the review (PDF file)

## Author's reply:

Dear Dr. Van Cleve 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 also appreciate the opportunity to submit a revision.

We revised our preregistration (https://github.com/corinalogan/grackles/blob/master/EasyToReadFiles/g_flexmanip.md) and protocol (https://docs.google.com/document/d/1sEMc5z2fw6S9C-

wVfc2zV331CRPpu3NuA7IhSFUZJpE/edit?usp=sharing), 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, Kelsey, Carolyn, Luisa, and Corina

Does manipulating behavioral flexibility affect exploration, but not boldness, persistence, or motor diversity? Kelsey McCune, Carolyn Rowney, Luisa Bergeron, Corina Logan 10.17605/OSF.IO/GCA5V version 1.3 Submitted by Corina Logan 2018-09-27 03:35 Abstract This is a PREREGISTRATION. The DOI was issued by OSF and refers to the whole GitHub repository, which contains multiple files. The specific file we are submitting is gexploration.Rmd, which is easily accessible at GitHub at https://github.com/corinalogan/grackles/blob/master/gexploration.Rmd (note: if the URL is broken, it's because the underscore was deleted by the system upon submission. There is an underscore between g and exploration). Note that viewing this file at OSF will result in not being able to see the figures as part of the .Rmd file. Photo credit: Zoe Johnson-Ulrich (CC-BY-SA 4.0). We will likely start data collection in late October or early November 2018 so it would be ideal if we could get through the review process before then. Keywords: Behavioral flexibility, exploration, boldness, persistence, motor diversity, comparative cognition, avian cognition Round #1

Your decision by Jeremy Van Cleve, 2019-02-03 04:17 Manuscript: 10.17605/OSF.IO/GCA5V Probing behaviors correlated with behavioral flexibility

Dear Dr. Logan, Thank you for submitting your preregistration to PCI Ecology. Please forgive the delay in getting reviews and a decision to you. Your study looks very interesting and both reviewers are enthusiastic about the potential results from the project. The reviewers suggest some areas for improvement in the preregistration and both focus on the need to better define terms within the "Hypotheses" section. They both also point out some confusion with the P6

alternatives. Finally, one reviewer points out some issues with the statistical analysis that should be addressed. I encourage you to revise the preregistration according to the reviewer's comments and look forward to seeing the revision. Please also note a couple of minor points of my own below. Best wishes, Jeremy Van Cleve

JVC Comments: 1. Figure 1. I am a bit confused reading this figure. Each experiment could be pictured by a box and then alongside it all the variables manipulated and measured listed. This might be more readable than the flowchart style.

Response 1. Thank you for this feedback. We took your advice and revised Figure 1 and its caption.

- Titles for predictions (P1-P8) are inconsistent and sometimes phrased as statements or questions (e.g., P1-P5). Please make these consistent.

Response 2. Thank you for pointing this out! We changed all of the titles for predictions in the Analysis Plan into statements.

- Independent vs Dependent variables: Listing these for each prediction got me a little confused since some appeared multiple times in each category and some in both categories. I wonder if a simpler presentation would have a description of each variable and then a table listing whether the variable is independent or dependent for each prediction.

Response 3. A table is a great idea! Unfortunately, we wrestled for hours with several packages in R to get a table to work in .Rmd and .md, but we were not successful. Therefore, we made a Google sheet with all of the variables in a table format and provided a link to this document at the top of Methods > Variables included in analyses 1-5 (https://docs.google.com/spreadsheets/d/1nhFkqTFWeAeWli8FU8n7mDiWGBuCeduzf8tWN3wPQeE/edit?usp=sharing). Additionally, within the .Rmd and md files, we re-organized the

variables according to which analysis they were in (1-5), which will hopefully make it easier to follow within the preregistration.

- Phrase " If they are patternless, then assume a normal distribution". What does "patternless" mean?

Response 4. Per reviewer comments we received on a separate preregistration (g_flexmanip.md), we ended up revising how we conduct data checking in that preregistration and also in this one. Please see the new data checking process in Analysis Plan > Data Checking, plus the new “data checking” code within each R analysis.

Additionally, this same reviewer pointed out that we don’t need to run an additional analysis to get repeatabilities when we are already running an MCMCglmm and can extract this information from the output. We applied their suggestion to this preregistration as well and made the following change to the text (plus we updated the R code):

Analysis Plan > REPEATABILITY: replaced the text about how we calculated repeatability with “We will obtain repeatability estimates that account for the observed and latent scales, and then compare them with the raw repeatability estimate from the null model. The repeatability estimate indicates how much of the total variance, after accounting for fixed and random effects, is explained by individual differences (ID). We will run this GLMM using the MCMCglmm function in the MCMCglmm package ([@hadfieldMCMCglmmpackage]) with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) [@hadfield2014coursenotes]. We will ensure the GLMM shows acceptable convergence (i.e., lag time autocorrelation values <0.01; [@hadfieldMCMCglmmpackage]), and adjust parameters if necessary.” 5. "Power (1- err prob)". Specify which Type or error probability (i.e., Type II). > Response 5. We added a note to clarify in Analysis Plan > P1-5 and P3: “Power (1-β err prob - note: β=probability of making a Type II error)”

Reviews Reviewed by anonymous reviewer, 2018-12-24 15:48 6. First of all, I apologize for the delay. Overall the project sounds good. I have identified some points that could be improved either for data analysis but also for the clarity/reliability of the predictions. H1: This first part mainly depends on whether or not behavioural flexibility is manipulatable, which made me uncomfortable to apprehend the predictions. I am also doubtful on the independence of reversal learning (which basically depends on the exploration skill of an individual as it requires the focus individual to explore the different opportunities – right? – and its cognitive abilities too obviously) from novel environment exploration stricto sensus. I believe that these definitions should be clearly stated.

Response 6. Please see our Response 22 for details about how we revised the preregistration to be clearer about how the flexibility manipulation relates to our expectations of the relationship between flexibility and the other measured variables. In short, the flexibility manipulation shouldn't have any effect on the nature of the relationship between flexibility and the other variables, it should just enhance our ability to detect a correlation if one exists. The predictions apply equally to both situations: whether the flexibility manipulation worked or not, which we now explain in the preregistration and detail below - please see the details in Response 13.

Regarding the independence of reversal learning from exploration, the two are not necessarily linked. For example, reversal learning is commonly thought to require inhibition to be able to inhibit choosing the previously rewarded option just after the reward has moved into the previously non-rewarded option. However, recent research summarized in Logan (2016 PeerJ https://peerj.com/articles/1975/) shows:

"Reversal learning speed is thought to positively correlate with inhibition: when the task changes subjects must inhibit the previously learned behavior to be able to learn the new behavior (Manrique, Völter & Call, 2013; Griffin & Guez, 2014;

Liu et al., 2016). However, this idea is challenged by an experiment in rats that were genetically modified to increase inhibition (Homberg et al., 2007). Knock out rats with improved inhibition showed no difference in their reversal learning speed from non-modified rats (Homberg et al., 2007). This suggests that behavioral flexibility may rely more on individuals continuing to sample their environment rather than simply inhibiting a response when a behavior is no longer rewarded."

This example highlights how little we know about how flexibility actually works and what other behaviors it is actually linked with. In a previous study on this species, no correlation was found between reversal learning and exploration (Logan 2016 PeerJ https://peerj.com/articles/2215/), which suggests that flexibility is independent from exploration. This doesn't mean that exploration could never be involved when birds are making choices in trials in the reversal learning paradigm. Rather, it means that consistent individual differences in exploration (if they exist) were not linked with reversal performance.

Additionally, the reversal learning colored tubes will not be novel objects for the birds because they undergo habituation to identical tubes (but with a different color - yellow) before beginning the reversal experiment with the light gray and dark gray tubes. They participate in an average of 374 trials of reversal learning so any novel effects that might occur would wear off quickly and most of the data would be free of this effect if one still exists after habituation (see details in the flexibility preregistration protocol for Experiment 1 https://docs.google.com/document/d/18D80XZV_XCG9urVzR9WzbfOKFprDV62v3P74upu01xU/edit?usp=sharing). We also include the flexibility ratio as a dependent variable, which accounts for exploration effects within the reversal experiment. In Methods > Dependent variables: 2) The **ratio of correct divided by incorrect trials** for the first 40 trials in their final reversal after the individual has seen the newly rewarded option once. These 40 trials include trials where individuals were offered the test and chose not to participate (i.e., make a choice). This accounts for flexibility that can occur when some individuals inhibit their previously rewarded preference (thus exhibiting flexibility because they changed their behavior when circumstances changed), but are not as exploratory as those

who have fewer 'no choice' trials. 'No choice' data is data that is otherwise excluded from standard reversal learning analyses. Including 'no choice' trials, controls for individual differences in exploration because those that refuse to choose are not exploring new options, which would allow them to learn the new food location.

- P1-P5 alternative: not sure that an absence of correlation is meaningful… It depends on the level of correlation the authors expect (?) and statistical robustness.

Response 7. We are not sure what to expect given that no correlations were found between a flexibility measure (reversal learning using colored tubes) and neophobia, exploration, risk aversion, motor diversity, or persistence in a previous study on this species (Logan 2016 PeerJ https://peerj.com/articles/2215/). We also wanted to account for all possible outcomes (positive, negative, or no correlation) before collecting the data to ensure we were making a priori explanations for whatever results we might find. This is why we have a “no correlation” option in the predictions for each hypothesis. However, perhaps we are misinterpreting your comment because in Hypotheses > “P1-P5 alternative”, it doesn’t discuss the direction of the correlation, but rather whether the flexibility manipulation worked or not. Perhaps this was not clear, so we made the following change to address this:

Hypothesis > P1-P5 alternative: “If the flexibility manipulation does not work in that those individuals in the experimental condition are not more flexible than control individuals, then we will analyze the individuals from both conditions as one group. In this case, we will assume that we were not able to influence their flexibility and that whatever level of flexibility they had coming into the experiment reflects the general individual variation in the population. This

experiment will then elucidate whether general individual variation in flexibility relates to exploratory behaviors. ”

In terms of our expectations regarding statistical robustness, please see our Response 10 for more details.

- H2: To me behavioural consistency (also known as personality) is defined across time and/or context and is a property of the individual so I do not really understand P6-1 vs. P6-2 opposition. Personality measure is the highest value the heritability of the behavioural trait can reach. I am not sure we can separate the low and high repeatability score (additionally 95% CI are often large).

Response 8. You are right that our sample size is likely not going to be large enough to distinguish low from high repeatability. We also now realize that H2 overlaps heavily with H1, and in particular, P3-P5. Therefore, we deleted H2, P6, P6 alternative 1, and P6 alternative 2 (the new H2 and P6 in the revised preregistration was the previous H3 and P7).

- H3: Do the author consider trapping bias? It may be a more robust hypothesis than the effect of unfamiliar environment…

Response 9. Thank you for this suggestion! Although we used multiple trapping methods, some of which should create less trapping bias (i.e., mist nets), it is likely a bias could still occur. Trapping bias would apply to all individuals in our study because we trapped and color-banded individuals to bring into captivity as well as those that remained in the wild before measuring their exploration and boldness. We have incorporated this prediction in Hypotheses > P6 alternative 2: “There is no difference in exploration and boldness between individuals in captivity and individuals in the wild (matched for season), potentially because

in both contexts our data is biased by sampling only the types of individuals that were most likely to get caught in traps."

1.I did not review the R script but I have issues with data analyses:

Statistical power: 0.70 could be sufficient enough if the authors are expected large correlation coefficients but it is not usually the case with behavioural traits.

Response 10. We expect that the power analyses we used underestimate our ability to detect actual effects because it is the wrong tool for the job and does not account for the particulars of our experiment. We addressed this concern in response to reviewer comments on a separate preregistration (g*flexmanip) and we copy and paste that conversation below. We also made the same changes to this preregistration as we did for g*flexmanip.

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

After working with SQuID to simulate our model in P2, we were unable to find a way to represent the complexity of our model (e.g., compare population means between control vs. manipulated conditions) while manipulating effect sizes. Additionally, we have no prior information about what values to provide in the input for the simulation without looking at the data we are currently collecting (e.g., the multi-access box has never been presented to grackles and we were unsure of how many options they would be able to solve). We are interested in the effect of the manipulation vs. everything else we are controlling for and, because of the complexity of the model, the effect is going to depend on the factors we control for as well as the boundaries of the dependent and

independent variables. We currently don't have any estimates for any variables because these tests have never been done in grackles and we have not encountered previous research that has manipulated flexibility in this way. However, we will be able to estimate these boundaries from our data after it is collected. Then we can perform informed simulations which will allow us to understand what sample size we need to detect the effect of interest. Once we have the data (and before conducting the analyses in the preregistration), we can set priors by entering the boundaries for the variables in the analysis (while remaining blind to the effect - the relationship between the variables). We can do this by running the null model (dependent variable ~ 1 + random effects), which will allow us to understand what the effect can actually operate on, and will inform us about what a weak vs. a strong effect is for these models. From here (and also before conducting any of the analyses in the preregistration), we can run the simulations based on the null model and then we can explore the boundaries of influences (such as sample size) on our ability to detect the effects of interest of varying strengths. We will run these simulations using the principles in McElreath (2015, Statistical Rethinking; starting on page 249) as a starting point and in consultation with McElreath. In terms of changes to the study design that would be possible to make as a result of simulation outcomes, pretty much the only thing we have some element of control over is the sample size. We have run into several unexpected complications at the Arizona field site (where we are currently collecting data), which is already indicating that we will not meet our projected sample size during our two years at this site (e.g., the grackles there are extremely difficult to catch and most of the females refuse to participate in tests). What we will do, before conducting the analyses in the preregistration, is run the simulations using the Arizona data to inform the simulation inputs and determine the lower sample size bounds for the analyses in this preregistration. If it turns out that our Arizona sample size is not larger than the lower boundary, we will change our experimental stopping criterion (which is currently to stop these experiments after two full aviary seasons in Arizona) and continue these experiments at our next field site until we meet the minimum sample size.

We updated the preregistration to lay out this new plan: Analysis Plan > Ability to detect actual effects: "To address the power analysis issues, we will run

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

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

- Poisson distribution: I suppose that this is because of count data. Negative binomial distribution is often more accurate for behavioural data.

Response 11. We will first check the dispersion in our count data (see our new data checking procedure in Response 4). If they are over-dispersed and thus violate the assumptions of a Poisson distribution, then we will use a negative binomial distribution (Zuur et al. 2009, p.383).

A.F. Zuur et al., Mixed Effects Models and Extensions in Ecology with R, 383 Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6 16,

Reviewed by anonymous reviewer, 2018-11-26 21:54 12. This preregistration describes a study aiming to measure how individual variation in behavioral flexibility relates to other behavioral traits such as exploration and boldness. It is argued that elucidating these behavioral patterns will help improve our understanding of how species' are able to adapt to new or changing environments. The authors plan to measure behavior in captive and wild great-

tailed grackles which seem like a good system for examining questions relating behavior to species expansion. The overall goal is to manipulate individual flexibility by utilizing different training protocols such as a serial reversal learning task in a subset of individuals and measuring if this manipulation impacts other behavioral traits. Group differences in these traits would suggest some connection between flexibility and that trait, while a lack of group differences would suggest independence with flexibility. Even if the manipulation does not work, the authors state that they would have the ability to examine individual-level behavioral patterns between flexibility, exploration, and boldness in this system. Overall, I find the topic to be of particular scientific merit as there is clearly growing interest in the animal cognition field for both measuring individual level variation in behavioral flexibility and tying that variation to other behavioral traits (e.g., coping styles, cognitive syndromes, etc). To then tie that variation to species expansion would be particularly exciting. While I am a little skeptical in terms of the manipulation working because it is unclear how well it will generalize across contexts, I think the approach is well thought out, and I agree with the authors that even if it doesn't work they will still have a worthwhile dataset in order to examine underlying behavioral patterns. These patterns along with measures of repeatability both involving captive and wild individuals would be a worthwhile dataset for publication. Below I have included some comments and questions.

## Response 12. Thank you so much for your supportive feedback! We look forward to addressing your comments and questions below.

1. Predictions: I understand that the section "P1-P5 alternative" is describing the alternative to the section "Predictions 1-5". I believe it is saying that even if the manipulation doesn't work you should still be able to examine underlying patterns of correlation between these traits among individuals. However, I am a little confused by sections following (e.g., P1 alternative 1). Are these predictions being made under the assumption that the manipulation failed?? I think this could be a little clearer.

Response 13. Yes, we could definitely be clearer here, thank you for pointing this out. To clarify, we added to the end of Hypotheses > P1-P5 alternative: “The following alternatives apply to both cases: if the manipulation works (in which case we would expect stronger effects for the manipulated group), and if the manipulation doesn't work (in which case we expect individuals to vary across all of the measured variables and for these variables to potentially interact).”

1. P1 alternative 1: Unclear what is meant by “... could indicate another trait is present, such as boldness.” Do you mean that this other trait could be explaining both individual variation in exploration and flexibility?

Response 14. Yes, you are correct. We revised the Hypothesis > P1 alternative 1 to say: “This suggests that flexibility is not independent of exploration and could indicate that another trait is present that could be explaining individual variation in flexibility as well as in exploration. This other trait or traits could be something such as boldness or persistence.”

1. P1 alternative 2: Would it matter which of the dependent outcomes it was correlated with (i.e., the one that accounts for exploration in reversal learning or the one that does not)? Also, how can flexibility be described as totally independent in this case if one of the two measures of flexibility are associated with exploration?

Response 15. Good point. We now split P1 alternative 2 into 2a and 2b to address this. We revised Hypothesis > H1 > P1 alternative 2 to : **P1 alternative 2a:** There is a positive correlation between exploration and the dependent variable that does not account for exploration (number of trials to reverse), but not the flexibility ratio, which suggests that performance overall in

reversal learning is partially explained by variation in exploration, but that flexibility and exploration are separate traits because using a measure that accounts for exploration still shows variation in flexibility.

**P1 alternative 2b:** There is a negative correlation between exploration and the flexibility ratio that accounts for exploration, but not with the number of trials to reverse. This could be an artifact of accounting for exploration in both variables.

1. Figure 1: Is time 1 before or after the manipulation?

Response 16. Yes, Time 1 is after the manipulation. We revised Figure 1 (see details in Response 1) and ended up taking out this part of the figure (the figure is now more descriptive rather than looking at interactions). To make sure this point is clear in the rest of the preregistration, we updated the Figure 1 caption and Analysis Plan > P1-P5 text.

1. H2: What if the manipulation itself manipulates these other traits independently. For example, the repeated trials of the manipulation habituate the animal to handling and other experimental stressors and therefore results in them showing more exploratory behavior because they are no longer shutting down behaviorally from these stressors.

Response 17. The control group is matched to the manipulated group in terms of how much experience they get at interacting with tubes and in the experiment. After their first reversal, the control group receives trials with two yellow tubes that both contain food so it doesn't matter which tube they choose, they just need to make a choice. With the yellow tubes, they receive the average number of trials it takes a bird in the manipulated group to pass their serial reversal criterion. Therefore, we expect

the manipulation to affect other traits in the same way as for control individuals.

1. P6: Alternative. Repeatability and changing behavior are not mutually exclusive as it is how behavior changes relative to other individuals. All individuals can change their behavior across time and still have high repeatability (e.g., those with the highest scores at time 1 still have highest scores at time 2 even though the exact scores may differ considerably). Also, even with lower repeatability you would still say that the traits are at least partially a property of the individual.

Response 18. Good point! We deleted H2, P6, P6 alternative 1, and P6 alternative 2 because we also realized that this hypothesis overlaps with H1 and Predictions 3-5 (note the H2 and P6 in the revised version were the previous H3 and P7).

1. Novel Environment: What is the rationale for having the familiar environment measure always first? Are you comparing main effects in terms of movement between familiar and novel environments, or just relative differences between individuals?

Response 19. One critique of the interpretation of behavior during a novel environment test is that the more active individuals might visit more areas of the environment, regardless of their actual exploratory tendency (Carter et al. 2013; Perals et al. 2017). By measuring the familiar environment first we can control for daily individual variation in activity levels, as well as inter-individual variation in activity. We incorporated familiar environment activity into our analyses by changing the following:

Methods > Independent variables: 2) Time spent in each of the different sections inside a novel environment or the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: activity in novel environment vs. activity in familiar environment

3) Time spent per section of a novel environment or in the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: time spent in novel environment vs. time spent in familiar environment

Analysis Plan > P1-P5: replaced AverageTimePerSectionNovelEnv and TotalNumberSectionsNovelEnv with AverageTimePerSection*EnvCondition and TotalNumberSections*EnvCondition, respectively. EnvCondition has two levels: familiar and novel.

References Carter, A. J., Feeney, W. E., Marshall, H. H., Cowlishaw, G. & Heinsohn, R. (2013). Animal personality: What are behavioural ecologists measuring? Biological Reviews 88(2), 465-475. Perals, D., Griffin, A., Bartomeus, I. & Sol, D. (2017). Revisiting the open-field test: What does it really tell us about animal personality? Animal Behaviour 123, 69-79.

- This protocol seems different from your reference in that the bird is really examining a large object in a familiar space vs. entering a whole new space as was tested in Mettke-Hoffman et al. 2009. Seems arguable if this is novel space or novel object. This might be important in terms of interpretation of results and distinguishing between predictions 1 and 2.

Response 20. By introducing the novel environment into the home cage of the individual we hoped to create a situation analogous to a free-entry open field test described in Mettke-Hofmann et al. (2009), which is more likely to measure exploration than boldness (Carter et al. 2013). However, we are not aware of another researcher that has conducted this exact manipulation. We believe if our version of the novel environment test does turn out to measure exploration (via external validation, see the next paragraph) that it will be a more logistically feasible experimental design for many researchers.

As in free-entry open field tests, the novel environment (a small tent) is placed away from the food and water in the home cage. So the subject can choose to enter or avoid this new space. However, because this manipulation involves novelty, you are correct that the behavioral responses of subjects could also be governed by boldness. Researchers of animal personality infrequently test exploration and boldness in multiple different ways with the same individuals, so the division between measurements of boldness and measurements of exploration is fuzzy. To address this we will conduct several variations on exploration and boldness tests to determine if individuals respond to the small tent in the same way that they respond to a novel object, and additionally relate behavior to that in response to a very obvious known threat (a taxidermied predator), which should only elicit a boldness response. In so doing, we can validate this method (or not) as an independent test of exploration.

We have clarified in Hypotheses > Predictions 1-5 that we are not exactly copying methods of Mettke-Hofmann et al. 2009: “will be more exploratory of new environments (P1; methods similar to free-entry open field test as in @mettke2009spatial)”

We also added that we will compare responses to our novel environment with novel object and boldness manipulations in Hypotheses > P1 alternative 4: “These measures of exploration both incorporate novelty and thus may measure boldness rather than exploration. This is supported by a positive correlation between behavioral responses to our exploration and boldness assays.” We added the corresponding analysis in Analysis Methods > P1 alternative 4.

1. Are you tracking unsuccessful wild assay attempts? I think it will be important to track overall participation in order to argue against possible critiques of self-selection bias due to personality differences.

Response 21. Yes, we will have data on all wild assay attempts. We have added this information as an independent variable for Methods > Independent Variables > P6: “2) Number of times we attempted to assay boldness or exploration but failed due to lack of participation”

1. Analysis: It is unclear why flexibility measures are being used as dependent variables when the research is being framed as examining how manipulation of flexibility may or may not affect other behavioral traits. This also seems particularly troublesome when condition is

Response 22. It appears that this comment somehow got cut off in the PDF, however we believe your main point was communicated and so we will address it here. Thank you for pointing this out - we were not clear about this in the Hypothesis section and we needed to be. We don't actually know, and we can't predict, whether there will be any causal relationships between flexibility and the other behaviors. If we find correlations between them, we won't know in what direction they go (i.e., which is the causal variable). We don't know now and we won't know after we conduct the analyses which variable might affect the other, so we are simply looking at whether flexibility is correlated with any of these other behaviors. In this case, it doesn't matter whether flexibility is a dependent or independent variable because we are just trying to detect a correlation at all.

The flexibility manipulation would also not causally change the nature of the relationship between flexibility and any of the other variables we measure in this preregistration. Instead, the flexibility manipulation would potentially enhance the individual variation, thus making it easier for us to detect a correlation with another behavior if one exists.

As a result of this discussion, we made the following changes (in bold):

Abstract: "In this piece of the long-term project, we aim to understand whether grackle behavioral flexibility (color tube reversal learning - described in a separate preregistration) correlates (or not) with individual differences in the exploration of new environments and novel objects, boldness, persistence, and motor

diversity (and whether the flexibility manipulation made such correlations more detectable). Results will indicate whether consistent individual differences in these traits might interact with measures of flexibility (reversal learning and solution switching). This will improve our understanding of which variables are linked with flexibility and how they are related, thus putting us in an excellent position to further investigate the mechanisms behind these links in future research."

Hypotheses > H1 > Predictions 1-5: we added: "We do not expect the flexibility manipulation to causally change the nature of the relationship between flexibility and any of the other measured variables. Instead, we expect the manipulation to potentially enhance individual variation, thus making it easier for us to detect a correlation if one exists."