Peer Community In Ecology

An experiment to improve our understanding of the link between behavioral flexibility and innovativeness

Aurélie Coulon *based on peer reviews by* ***Maxime Dahirel****,* ***Andrea Griffin****,* ***Aliza le Roux*** *and 1 anonymous reviewer*

Logan CJ, Lukas D, Blaisdell AP, Johnson-Ulrich Z, MacPherson M, Seitz BM, Sevchik A, McCune KB (2023) Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context. EcoEvoRiv, ver. 5, peer-reviewed and recommended by Peer Community in Ecology. https://doi.org/10.32942/osf.io/5z8xs

Submitted: 13 January 2022, Recommended: 15 May 2023

Cite this recommendation as:
Coulon, A. (2023) An experiment to improve our understanding of the link between behavioral flexibility and innovativeness. *Peer Community in Ecology*, 100407. 10.24072/pci.ecology.100407

Published: 15 May 2023
Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Whether individuals are able to cope with new environmental conditions, and whether this ability can be improved, is certainly of great interest in our changing world. One way to cope with new conditions is through behavioral flexibility, which can be defined as "the ability to adapt behavior to new circumstances through packaging information and making it available to other cognitive processes" (Logan et al. 2023). Flexibility is predicted to be positively correlated with innovativeness, the ability to create a new behavior or use an existing behavior in a few situations (Griffin & Guez 2014).

The post-study manuscript by Logan et al. (2023) proposes to test flexibility manipulability, and the relationship between flexibility and innovativeness. The authors did so with an experimental study on great-tailed grackles (*Quiscalus mexicanus*), an expanding species in the US, known to be flexible.

The authors used serial reversal learning to investigate (1) whether behavioral flexibility, as measured by reversal learning using tubes of different shades, is manipulable; (2) whether manipulating (improving/training) behavioral flexibility improves flexibility and innovativeness in new contexts; (3) the type of learning strategy used by the individuals throughout the serial reversals.

The study described in this manuscript was pre-registered in Logan et al. (2019) and received in-principle recommendation on 26 Mar 2019 (Coulon 2019). One hypothesis from this original preregistration will be treated in a separate manuscript.

Among several interesting results, what I found most striking is that flexibility, in this species, seems to be a trait that is acquired by experience (vs. inherent to the individual). This opens exciting interrogations on the

role of social learning, and on the impact of rapid environmental changes (which may force the individuals to experiment new ways to access to resources, for example), on individual flexibility and adaptability to new conditions. **REFERENCES**

Coulon A (2019) Can context changes improve behavioral flexibility? Towards a better understanding of species adaptability to environmental changes. Peer Community in Ecology, 100019. https://doi.org/10.24072/pci.ecology.100019

Griffin, A. S., & Guez, D. (2014). Innovation and problem solving: A review of common mechanisms. Behavioural Processes, 109, 121–134. https://doi.org/10.1016/j.beproc.2014.08.027

Logan C, Rowney C, Bergeron L, Seitz B, Blaisdell A, Johnson-Ulrich Z, McCune K (2019) Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context? In Principle Recommendation 2019. PCI Ecology. http://corinalogan.com/Preregistrations/g_flexmanip.html

Logan CJ, Lukas D, Blaisdell AP, Johnson-Ulrich Z, MacPherson M, Seitz B, Sevchik A, McCune KB (2023) Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context. EcoEcoRxiv, version 5 peer-reviewed and recommended by Peer Community in Ecology. https://doi.org/10.32942/osf.io/5z8xs

Reviews

Evaluation round #4

DOI or URL of the preprint: **https://doi.org/10.32942/osf.io/5z8xs**
Version of the preprint: 4

Authors' reply, 10 May 2023

Dear Aurélie Coulon,

Thank you so much for checking our manuscript again and for being ready to recommend it! We are very glad that you are happy with our revisions. Thank you also for your further two comments, which we address here.

COMMENT 1: p.5, p2 alternative 1: shouldn't it read "increase" instead of "decrease" in "If the manipulation does not work in that those individuals in the experimental condition do not decrease their reversal speeds more than control individuals"

RESPONSE 1: we can see where this language is confusing because it is a decrease in the number of trials to solve but an increase in the speed of solving. This is confusing so we changed the sentence to:

"If the manipulation does not work in that those individuals in the experimental condition do not **reverse faster** than control individuals"

COMMENT 2: Figure 2, legend: as displayed in the figure, the letters A to D should refer to the different experimental apparatuses, not the different ways of opening the wooden box. + The description of the apparatuses (in the legend) could be made in the same order as they are presented in the figure.

RESPONSE 2: good points, thank you. We changed the figure caption so that the letters A to D that refer to the doors on the multiaccess box are described as "locus A", "locus B", etc. And we put them in the same order as they are presented in the figure (counterclockwise starting with Swing). We added the figure panel labels to the caption. We also realized that Figure 3 now appears before Figure 2 so we switched their numbering.

The track changes are shown in this commit at GitHub: https://github.com/corinalogan/grackles/commit/fe31129b5a8eb76b81a83ca246773a74a0a06061

Let us know if you need anything else. We are excited to share this research and the recommendation with the world!

All our best,
Corina Logan (on behalf of all co-authors)

Decision by Aurélie Coulon, posted 02 May 2023, validated 02 May 2023

2 minor points on "Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context"

Dear Corina Logan and co-authors,

I have read your replies to the reviewers' comments. I am happy to tell you I am ready to write the recommandation for your paper. I only have 2 very minor points to raise:

- p.5, p2 alternative 1: shouldn't it read "increase" instead of "decrease" in "If the manipulation does not work in that those individuals in the experimental condition do not decrease their reversal speeds more than control individuals"

- Figure 2, legend: as displayed in the figure, the letters A to D should refer to the different experimental apparatuses, not the different ways of opening the wooden box. + The description of the apparatuses (in the legend) could be made in the same order as they are presented in the figure.

Best,
Aurélie Coulon.

Evaluation round #3

DOI or URL of the preprint: https://doi.org/10.32942/osf.io/5z8xs
Version of the preprint: 3

Authors' reply, 18 April 2023

Dear Aurélie Coulon, Maxime Dahirel, and anonymous reviewer,

Thank you all for checking this version of the manuscript and for providing your useful feedback! We really appreciate the time you have taken over the years for this piece of research that we are really excited about. We revised per your comments and provide detailed responses below.

Here are the links to the various versions of the article, depending on which you prefer:

-**PDF** at EcoEvoRxiv (version 4): https://doi.org/10.32942/osf.io/5z8xs

-**html** with a floating table of contents: http://corinalogan.com/Preregistrations/g_flexmanip2.html

-**Rmd** file with the text and the code (version-tracked): https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip.Rmd

For reference, here is the PDF of the **preregistration**: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanipPassedPreStudyPeerReview26Mar2019.pdf

Thank you so much for all of your help throughout this whole research process!

All our best,
Corina Logan (on behalf of all co-authors)

Author response
Round #3

by Aurélie Coulon, 04 Apr 2023 18:59

Manuscript: https://doi.org/10.32942/osf.io/5z8xs version 3

revision needed for the preprint "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context"

Dear Dr Logan and collaborators,

I have received two reviews of the revised version of your preprint called "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a
new context" (now renamed "Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context"). One of those reviewers, Maxime Dahirel, had reviewed the previous version of your ms; the second one, anonymous, had not. The latter is very enthusiastic about your work and M. Dahirel and myself acknowledge the work you have done on your ms and think it has improved its flow. M. Dahirel is also satisfied with all the answers you made to his other previous suggestions/comments, including statistical issues.
Both reviewers however still highlighted a few difficulties in the method section, and especially the way the deviations from the preregistration are included. They made 2 different suggestions on how to handle this problem, that both seem like good ideas to me.
They also have a few other rather minor comments.
Once these are addressed, your preprint should be ready for recommendation.

Best,

Aurélie Coulon.
Recommender, PCI Ecology. **Reviews**
Reviewed by Maxime Dahirel, 31 Mar 2023 13:12

COMMENT 1: I have now read the latest revision of the recommended-in-principle manuscript entitled "Behavioral flexibility is manipulable and it improves flexibility and problem solving in a new context" by Logan et al.

On the substance of the manuscript, I pretty much have nothing left to say. The authors have consistently taken into account my comments about the limitations of their data, and about conveying definitions to a broader audience, resulting in a significantly better manuscript.

I do have however some comments left about some stylistic, syntax and organisation choices and/or mistakes. Some are continuations of previous comments about how to integrate preregistered and post-registration material in a way that is stylistically consistent and easy-to follow by the reader. Some are more about internal inconsistencies in style/syntax within the manuscript itself, especially around equation syntax. I acknowledge that some of these stem directly from the methods used to layout the text of the manuscript (I reviewed, again, the EcoEvoRxiv pdf version); nonetheless, please be very careful and proofread the entire manuscript against the comments I detail below, not just the specific lines where I point these issues. In my opinion, once these stylistic issues are thoroughly ironed out, I think the manuscript is ready to recommend.

RESPONSE 1: We are so glad that you think we did a good job of addressing your comments and improving the manuscript! Indeed, it is tricky to figure out the format that works best for readers when trying to integrate the preregistration with the post-study article. It seems like no matter how many times we proofread the article and get multiple eyes on it, typos get through. Thanks for your patience with this. We respond to your detailed comments below.

Minor comments

COMMENT 2: Line 23: please remove “in species” to avoid unnecessary repetition with “species” later in the same sentence

RESPONSE 2: Thank you! We made the change.

COMMENT 3: Line 108: Maybe the header should be “Preregistered hypotheses” rather than simply “hypotheses” to really hammer the point home

RESPONSE 3: Good idea! We added this.

COMMENT 4: Line 148: should be “(some of which ARE tested in McCune et al 2023)”, not “will be”. Please be careful with both tense and year when referencing the companion preprints to this one.

RESPONSE 4: Thanks for spotting this! We made the change and we searched through the other references to our preprints to make sure the correct tense was used.

COMMENT 5: On the Methods: It should be made unambiguously clear whether everything under the “Methods” header (lines 181-342) is a summary of the preregistered methods or of the actually used methods. Current wording and styling obfuscates this; for instance, the header “deviations from preregistrations” line 343 is of the same style as the main “Methods” header, implying it is something different from it. I suggest bringing the deviations under the general methods header, and being extremely clear with subheaders or within the text what is a summary of the preregistered methods, and what is not.

RESPONSE 5: Thanks for helping us get clear about how to best present this! We updated the first paragraph in Methods to make our rationale clear for making changes and what changes we made and that the Methods describe the final approach we used. We moved the Deviations section up to the beginning under the general Methods header as you suggested and we removed the header for this section, which now also reflects the temporal order of going from preregistration through deviations to the final methods that are presented afterward.

COMMENT 6: Line 261: “following methods by Hartig” can be deleted, since the DHARMa package is already cited.

RESPONSE 6: We made the change.

COMMENT 7: Lines 264 vs 196, and others: in some places of the Methods (L196), text after a colon is treated as part of the same sentence with the text immediately before, and therefore not capitalised. In others (L264) text immediately after a colon is capitalized, indicating it is considered a separate sentence. This inconsistency make it difficult to read the text, and in some cases results in verb-less sentences where there should really be a verb. Please check thoroughly the manuscript, especially the methods, for this. Similarly, variables descriptions

are treated in many (but not all) places as verb-less lists, yet are including inline as normal sentences would be. E.g. Line 265 “Explanatory variable: reversal number. Random variables: batch (batch is a test cohort, consisting of 8 birds being tested simultaneously and there were multiple batches included in the analysis) and ID (random effect because there were repeated measures on the same individuals).”. (see also L284-285). This hurts readability; please convert that either to actual lists (e.g. bullet points) or rewrite to full sentences.

RESPONSE 7: Thank you for pointing this out. We went through the document and made these consistent and made sure they were complete sentences or a bulleted list (changes made in Summary of Testing Protocols, Analyses, and Results).

COMMENT 8: Across the methods and supplementary: There are several issues around the writing of equations. I suggest the authors use (Edwards & Auger-Méthé, 2019) as one example of recommended practices. Among these issues (again, I am only pointing representative examples, the authors should check carefully the entire manuscript):

• Equations are supposed to be parts of sentences. The phrase starting L287 should include a comma after the first equation, and a comma, semi-colon or a period after the second equation, depending on whether the authors consider line 290 to start a new sentence or not.

• More importantly, there are major inconsistencies in indexing – If a variable is indexed by individual in one equation (e.g. λi , the individual rate by bird in equation line 295), then all individual-level variables/elements should also be indexed similarly across the manuscript unless explicitly noted. For 1 instance, this should be at least latencyi and not latency line 295, and individual indexing should also be added in the equation lines 288-289, among others. Since there are multiple observation per individual, the notation should most likely be at least:

$\text{latency}_{i,j} \sim \text{gamma} - \text{Poisson}(\lambda_{i,j}, \square), \log(\lambda_{i,j}) \sim \alpha + \beta \text{trials}_{i,j}$,

where i is the individual and j the observation. Please be also careful about whether or not coefficients (α and β) should be individual-indexed (αi , βi) or not, depending on the presence or not of individual-level random effects, and what that says about the way priors should then be written.

RESPONSE 8: Thank you very much for sharing the Edwards & Auger-Méthé article with us. We are not mathematicians and appreciate the How To. We went through all of the equations in the article (Methods and Supplementary Material 1) and implemented the conventions suggested by you and by the article. Note that these changes are not marked at the tracked changes PDF, but only in the rmd at GitHub.

COMMENT 9: Line 451: just to check, when you say you removed the observer random effect because inter-observer reliability was virtually 1, can you confirm you also removed/merged together the different observer replicates? To avoid artificially inflate the sample size.

RESPONSE 9: You are correct, we only used one set of data in the analyses in the Results section and we did not include in our analyses the duplicate data that was used for the interobserver reliability analyses.

COMMENT 10: Line 507: I would suggest to start the caption by “In the manipulated birds, the number(. . .)” Figure 4: I am still not sure why the authors do not add the predicted trend line from table 2 to this plot. This would be helpful.

RESPONSE 10: Thanks for the caption suggestion - we implemented it. For Figure 4, we didn’t realize that you

had previously suggested this - sorry for missing it! We added a predicted trend line to Figure 4.

COMMENT 11: Figure 6: it might be interesting to also add the unmanipulated birds to that plot, on a second row for instance, to create a contrast. As far as I can tell the relevant posteriors exist _ from the same models table 1 & 3 used to extract the predictions from the manipulated birds.

RESPONSE 11: Good idea. We added one more box to the existing row that shows all control birds' observed and estimated values from their single reversal, and updated the Figure 6 caption.

COMMENT 12: Line 726: there is a missing period between "shapes" and "Shapes"

RESPONSE 12: Thank you! We fixed it.

COMMENT 13: Line 895: "parameter", not "paramter"

RESPONSE 13: Thank you for catching this! We fixed the typo.

COMMENT 14: Reference list: the main text and the supplementary material should have separate reference lists. Please ignore this comment if this was already on the to-do list for the final version, and the current state just reflects the way the document was knitted for submission

References

Edwards, A. M., & Auger-Méthé, M. (2019). Some guidance on using mathematical notation in ecology. Methods in Ecology and Evolution, 10(1). https://doi.org/10.1111/2041-210X.13105

RESPONSE 14: We combine all references into one list on purpose because, when the article eventually goes to a journal, the reference list for the main text gets indexed and we want all references to get citation credit. The references that go with the supplementary material, when listed separately, result in authors not receiving full citation credit (Seeber 2008, Weiss et al. 2010, Rafferty et al. 2015).

References

Rafferty, A. R., Wong, B. B., & Chapple, D. G. (2015). An increasing citation black hole in ecology and evolution. Ecology and Evolution, 5(1), 196-199.

Seeber, F. Citations in supplementary information are invisible. Nature 451, 887 (2008). https://doi.org/10.1038/451887d

Weiss, M. S., Einspahr, H., Baker, E. N., Dauter, Z., Kaysser-Pyzalla, A. R., Kostorz, G., & Larsen, S. (2010). Citations in supplementary material. Acta Crystallographica Section D: Biological Crystallography, 66(12), 1269-1270.

Reviewed by anonymous reviewer, 22 Mar 2023 14:31

Review of "Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context."

COMMENT 15: This manuscript investigates whether flexibility can be manipulated in one task and for this flexibility to then be generalized to a different task. This is a really interesting paper, and I think would be of interest to many readers. I also applaud the authors for their commitment to the preregistration process

as well as open data and code for their manuscript. However, the organization of the manuscript leaves it difficult to follow. I was lost in all the different models, especially when the authors would bring them up in the discussion, and I think a table explaining the models and what they were for would go a long way in fixing this.

RESPONSE 15: Thank you very much for your positive feedback! We realize that there is lots going on in this article and we thank you for your advice on how to make this clearer for readers. We had hoped that Figure 7 was a good summary of the various results all in one place, but it seems that we did not fully achieve this goal. We are not exactly sure what information you would like in a summary table. Figure 7 links the research questions with the respective analyses, and Table SM3 lists all of the model specifications. Therefore, we linked these two pieces by referring, in each cell of Figure 7, to the model number in Table SM3 rather than creating an additional table that duplicates information from both of these sources. We find that when we read the discussion, we refer to Figure 7 to help remember all of the various relationships. Therefore, we added more references to Figure 7 in the Discussion to help readers orient.

COMMENT 16: Additionally, although I think it is great how they demonstrate how their investigation differed from pre-registration, it takes away from the flow of the manuscript and makes it difficult to read. I'm fine with a small explanation of these issues and putting this in the supplemental materials. This way, the authors are fully transparent and the flow/conciseness of the manuscript is maintained.

RESPONSE 16: Please see Response 5 for how we addressed confusion around the methods we actually used versus those that were preregistered. We decided to leave the description of the changes from the preregistration in the main text because we wanted to explain the rationale for the methods we ended up using and we think it is helpful for readers to consider when planning similar studies.

COMMENT 17: I'm also struggling with the authors assertiveness in their results. The authors did a cross validation (which should probably be described a bit better as I first assumed they were doing a LOO/pareto smoothing), and found that by taking out one individual that had the highest number of trials to reversal changed some of their results to non-significance (again, I had to look up which models were which). When looking at the graphs, this individual 'looks' like it may be driving their results. I think that this may require them to temper their discussion.

RESPONSE 17: Thank you for bringing this up. We see now that how we worded this in the Discussion could be confusing. We revised the text to now say:

"Because of the variation in our small sample (Taquito was by far the slowest to reverse a preference), we conducted a validation check to determine whether removing a bird from the data set changed the model results. Removing either Taquito or a random bird from the data set changed the conclusions for one of the three models (Model 2, but not Models 6 or 12). This change in results after removing a data point indicates that we should be less confident in the conclusion that individuals who are faster to reverse a preference in their last reversal also solved more loci on the plastic multi-access box. However, it did not matter whether we removed Taquite, the slowest performer, or a random bird, indicating that this outlier did not drive the results but rather that the result is constrained by our small sample size."

Our discussion of the one model (out of three) whose results changed because of the validation check is tempered by phrases throughout this paragraph (e.g., "we should be less confident in the conclusion", "the result is constrained by our small sample size").

What we did was not strictly a cross validation or a LOO cross validation because we did not use an independent sample or training data for the check. Therefore we updated the text to say that we did a "validation check", which should cue readers to the fact that our check was unique for our particular purposes.

While it looks like this individual could drive the results, we found in round 2 revisions that the removal of Taquito (the outlier) or a random other individual has the same effect. Therefore, Taquito is not driving the

results. We copy and paste our response from round 2 here to fully explain.

From the round 2 author response:

We also conducted another set of cross validation checks by removing a randomly selected bird (not Taquito) from the data to determine whether removing any bird makes a difference or only Taquito (if his removal makes a difference).

We found that removing Taquito or the random bird **did not change the results** for:

Model 12: wooden MAB latency, Fig 8b - there remained no relationship

Model 6: wooden MAB loci solved, Fig 8d - there remained no relationship

We found that removing Taquito or the random bird **did change the result** for:

Model 2: plastic MAB loci solved, Fig 8c - there was no relationship (whereas there is a relationship when the whole dataset is analyzed)

Because the changes were consistent based on removing a random data point and not specific to Taquito, we conclude that his high reversal learning value was not difficult for the model to account for. Therefore, when a model result changes due to the removal of a random bird (whether it was Taquito or not), it means that we need to be less confident in the main model. We added an asterisk to Figure 7, which indicates that the result is not as confident due to the small sample size

COMMENT 18: The authors also present simulations for effect sizes given specific sample sizes. I find this great. Perhaps I am wrong about this, but their effect sizes are based on previous data and the standard deviation. Given their new data has much more noise with a larger standard deviation, I would think this would potentially mean that they need to increase their sample size.

RESPONSE 18: Sorry for the confusion around the sample size simulations! We realize that this section is difficult to figure out - we wrote the simulations a few years ago and have since developed better methods so it is also difficult for us to go back and figure out what we did. In fact, we noticed that some of the R code and description for Table SM1.2 was missing from the current version of the manuscript, so we re-added it back into Supplementary Material 1 (r code chunk "sim"). In our preregistration, we did not set out to find a prespecified effect size. The simulations were based on what other literature considered small or medium effects and we used them primarily to assess what kinds of effect sizes we would be able to detect with an assumed sample size of 15 individuals. You are correct that, because of the larger standard deviation (SD) of our actual data, our sample size restricts us to detecting effects that are larger than we had initially assumed based on simulations. This becomes an issue only for the cases where we concluded that we did not find evidence of a relationship because the interval crossed zero. Therefore, we added the following to the Discussion: "In the cases where there was no correlation between loci solved and reversal performance, it is possible that the effect size was too small for us to have the power to detect (Figure 7)."

In our more recent work on behavioral flexibility and innovativeness in grackles, we have gotten better at developing bespoke Bayesian power analyses. The newer analyses are much more efficient and we have consistently found that the minimum sample size for detecting small effects is 12-15 individuals, regardless of the specific hypothesis being tested (Logan et al. 2022, Logan et al. 2023). Therefore, it is likely that this general minimum sample size also applies to the Bayesian analyses we use in the Results section of the current article.

References

Logan CJ, Shaw R, Lukas D, McCune KB. 2022. How to succeed in human modified environments (http://corinalogan.com/ManyIndividuals/mi1.html) In principle acceptance by PCI Registered Reports of the version on 25 Aug 2022

Logan CJ, McCune KB, Rolls C, Marfori Z, Hubbard J, Lukas D. 2023. Implementing a rapid geographic range

expansion - the role of behavior changes. doi: https://doi.org/10.32942/X2N30J

COMMENT 19: Minor points: L. 270 it mention 30,000 iterations. I believe this should be 300,000 (especially if you have a burnin of 90,000)

RESPONSE 19: Nice catch! We checked the R code and you are correct. We fixed the text.

COMMENT 20: L. 884-885 Authors mention the trial lasts up to 15 minutes and then mention that a usual trial is 10 minutes. They should clarify what they mean here. Is 10 minutes the average length of a trial?

RESPONSE 20: We agree that this is confusing. We added a clarification (Supplementary Material 1): "(trials end at 10 min unless the individual is on the ground at the 10 min mark, in which case they are given an extra 5 min to interact)"

Download tracked changes file

Decision by Aurélie Coulon, posted 04 April 2023, validated 04 April 2023

revision needed for the preprint "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context"

Dear Dr Logan and collaborators, I have received two reviews of the revised version of your preprint called "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context" (now renamed "Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context"). One of those reviewers, Maxime Dahirel, had reviewed the previous version of your ms; the second one, anonymous, had not. The latter is very enthusiastic about your work and M. Dahirel and myself acknowledge the work you have done on your ms and think it has improved its flow. M. Dahirel is also satisfied with all the answers you made to his other previous suggestions/comments, including statistical issues.
Both reviewers however still highlighted a few difficulties in the method section, and especially the way the deviations from the preregistration are included. They made 2 different suggestions on how to handle this problem, that both seem like good ideas to me.
They also have a few other rather minor comments.
Once these are addressed, your preprint should be ready for recommendation. Best, Aurélie Coulon.
Recommender, PCI Ecology.

Reviewed by Maxime Dahirel, 31 March 2023

Download the review

Reviewed by anonymous reviewer 1, 22 March 2023

Review of "Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context."

This manuscript investigates whether flexibility can be manipulated in one task and for this flexibility to then be generalized to a different task. This is a really interesting paper, and I think would be of interest to many readers. I also applaud the authors for their commitment to the preregistration process as well as open data and code for their manuscript. However, the organization of the manuscript leaves it difficult to follow. I was lost in all the different models, especially when the authors would bring them up in the discussion, and I think a table explaining the models and what they were for would go a long way in fixing this. Additionally,

although I think it is great how they demonstrate how their investigation differed from pre-registration, it takes away from the flow of the manuscript and makes it difficult to read. I'm fine with a small explanation of these issues and putting this in the supplemental materials. This way, the authors are fully transparent and the flow/conciseness of the manuscript is maintained.

I'm also struggling with the authors assertiveness in their results. The authors did a cross validation (which should probably be described a bit better as I first assumed they were doing a LOO/pareto smoothing), and found that by taking out one individual that had the highest number of trials to reversal changed some of their results to non-significance (again, I had to look up which models were which). When looking at the graphs, this individual 'looks' like it may be driving their results. I think that this may require them to temper their discussion.

The authors also present simulations for effect sizes given specific sample sizes. I find this great. Perhaps I am wrong about this, but their effect sizes are based on previous data and the standard deviation. Given their new data has much more noise with a larger standard deviation, I would think this would potentially mean that they need to increase their sample size.

Minor points:

L. 270 it mention 30,000 iterations. I believe this should be 300,000 (especially if you have a burnin of 90,000)

L. 884-885 Authors mention the trial lasts up to 15 minutes and then mention that a usual trial is 10 minutes. They should clarify what they mean here. Is 10 minutes the average length of a trial?

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.32942/osf.io/5z8xs

Version of the preprint: v2

Authors' reply, 03 March 2023

Dear Aurélie Coulon, Maxime Dahirel, and Aliza le Roux,

We apologize for the delay in our revision - Logan (the lead on this) was in the field for several months and unable to perform additional tasks at that time. Thank you very much for your additional feedback on this manuscript! We revised the manuscript per your comments and responded to each comment (below).

Here are the links to the various versions of the article, depending on which you prefer:

-**PDF** at EcoEvoRxiv (version 3): https://doi.org/10.32942/osf.io/5z8xs

-**html** with a floating table of contents: http://corinalogan.com/Preregistrations/g_flexmanip2.html

-**rmd** file with the text and the code (version-tracked): https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip.Rmd

For reference, here is the PDF of the **preregistration**: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanipPassedPreStudyPeerReview26Mar2019.pdf

Many thanks again for your wonderful feedback throughout this process! We think the manuscript is much improved due to your efforts.

All our best,

Corina Logan (on behalf of all co-authors)

Round #2

by Aurélie Coulon, 13 Oct 2022 12:22

Manuscript: https://doi.org/10.32942/osf.io/5z8xs version v2

Revision needed for the preprint "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context"

Dear Corina Logan and co-authors,

Two reviewers have evaluated your manuscript called "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context". One of them was a reviewer on its previous version, and of its preregistration. The latter is very enthusiastic about your ms. The former highlights the improvements made since the last version, but also has a number of comments, mainly about the form of the paper. I agree with all of them, and think that overall, the ms has improved but is still a bit tough to read. This is due to the willingness to stick to the preregistration, and report on what was different from it. It makes the material and methods part a bit difficult to follow and rather technical, in the sense that the biological (behavioural) aspects of the experiments are kept away. Besides the reviewers' comments, I have a few comments and suggestions (listed below) to work on those aspects. I think the different comments made by the reviewers and myself are rather easy to take into account, and should make the manuscript close to the recommendation stage.

Best,

Aurélie Coulon

COMMENT C1: Material and methods: I think it is very confusing to present the analyses as they were planned, and then the deviations that were made to the plan. This is because there is nowhere, in the M&M section, where the reader can easily find what was actually done (they have to read the initial plan + the deviations and then make their own synthesis of all this). I think it would be much easier for the reader if the methods were presented in their final stage (instead of as they were planned), and accompanied by the section describing the deviations to the initial plan (as they are currently written). Also, the fact that the experimental protocols are only cited, and not described at all, is contributing to this feeling that we miss something about the behavioural aspects of the study. I suggest to add a brief summary of the protocols, before the section Open materials.

RESPONSE C1: Sure, we can reformat/rewrite as requested. We changed the name of the section from "Analysis Plan" to "Analyses", moved all paragraphs titled "Deviations from the plan" to the section at the end of Analyses called "Deviations from the preregistration" and added a few more details to these paragraphs when more context was needed. We added a summary of the protocols as suggested - good idea. We show here the new text we wrote in the Methods > Summary of testing protocols:

Reversal learning with color tubes: one light gray and one dark gray tube is placed such that the openings are not visible (colors are pseudorandomized for side). One color always contains a food reward. The individual has the opportunity to choose to look inside one tube per trial. Once the individual chooses 17 out of the most recent 20 trials correct, they are considered to have a color preference, and then the food is always placed in the previously non-rewarded color and the same passing criterion is used to determine their reversal learning performance. Individuals were randomly placed in the manipulated condition (serial reversals until they pass two consecutive reversals in 50 trials or less) or the control condition (receive only one reversal and then a similar number of total trials to the manipulated individuals, but with two yellow tubes, both of which always have food).

Plastic multi-access box: a puzzlebox made of plexiglas and plastic and containing one piece of food on a post in the center of the box is placed in the aviary for up to 15 minutes per trial. Each plexiglass wall has one opening (locus) with each opening requiring a different method for obtaining the food. The individual has the opportunity to attempt (touch, but not obtain the food) or solve a locus. Once a locus has been solved three times, it is considered solved and rendered non-functional in subsequent trials. The experiment ends when an individual solves all four loci or if they do not interact with or successfully solve a locus in three consecutive

trials.

Wooden multi-access box: a puzzlebox is carved into a log and has four loci where each locus contains food and requires a different motor action to solve. Three loci are covered with a plastic door on a hinge and one locus is a drawer that must be pulled out. Trials last for up to 15 minutes. The passing criterion and experiment ending criteria are the same as for the plastic multi-access box.

Reversal learning of shapes on a touch screen: this is the same experimental design as with the color tubes, except it is carried out on a touch screen computer where the individual is presented with two white symbols that differ in shape (pentagon or diamond).

COMMENT C2: In Fig.1C, the link between the numbers, arrows and the phrases "2+ learning strategies" and "1 learning strategy" is not easy to get

RESPONSE C2: That's a great point. We modified this panel and hopefully it is clearer now (Figure 1C).

COMMENT C3: In several places there are mistakes in the figure or sup.mat. numbers (e.g. l.196, it should read sup.mat.1 instead of 6; l.224: 2 instead of 5;l.592: fig.7d instead of 5d; ...)

RESPONSE C3: We apologize for this! We changed the order of the supplementary material later in the last revision and thought we had made all of the adjustments to the in text references, however, clearly, we missed several. We have now carefully gone through all of the text to correctly align the supplementary material, table, and figure numbers.

COMMENT C4: l.216-221: I don't understand why papers are cited (without a reference to a package) at the beginning of this list which is supposed to be a list of R packages

RESPONSE C4: Thank you for spotting this inconsistency! We changed it so that the package name always comes before the reference in this paragraph. We also realized that we hadn't deleted the packages that were used for the other two articles (that are now separate) and not this one, so we made sure to only include those packages that we used for this article.

COMMENT C5: Unless I got lost, there are several unregistered analyses that are not presented in the material and methods section: l.489; unregistered analysis 1, l.504-508; unregistered analysis 2, l.517-520; P2: analyses of the effect of time in first reversal. This should be fixed.

RESPONSE C5: Thank you for pointing this out! We added the unregistered analyses in Tables 1 and 3, and Figure 6 to the Methods section as suggested.

Analyses > P1: "Post-data collection, we added unregistered analyses as follows. We evaluated whether the individuals in both conditions required a similar number of trials to pass their first reversal (dependent variable: trials to reverse in first reversal, explanatory variable: condition, random variables: ID and batch; Table 1). We evaluated whether the individuals in both conditions required a similar number of trials to pass their last reversal (dependent variable: trials to reverse in last reversal, explanatory variable: condition, random variables: ID and batch; Table 3)."

Analyses > P2: "Post-data collection, we added an additional unregistered analyses comparing first versus last reversal performance for the individuals in the manipulated group (Figure 6; see r code chunk "posthoc_conditionalimprovement" for model details)."

COMMENT C6: Table SM3: I don't understand what "a" is, in some models (e.g. models 2 and 5). Also, in model 1, there seems to be a "batch" factor, while it was said in the M&M that this factor was eventually removed from the analyses...

RESPONSE C6: "a" is always the intercept. Models 2 and 5 look a little different because they just have "a" and not "a[batch]". "a[batch]" means that each batch gets its own intercept. Therefore, when it is just "a", there is only one intercept. We attempted to clarify this by adding the following to the Table SM3 caption: "a=the intercept (a[batch] is the intercept for each batch)" and "See Supplementary Material 1 for details on model specifications." Model 1 includes batch because it was part of a series (also Model 3) where we show the reasoning behind why we removed batch from the final analyses we ran, which is discussed in Analyses >

Deviations from the preregistration.
Reviews
Reviewed by Maxime Dahirel, 13 Oct 2022 12:03 I have now read the revised version of the recommended-in-principle manuscript entitled “Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context” by Logan et al.

First of all, I have to say that this is a much improved and streamlined version of the manuscript. While it certainly is not the only reason, I believe the decision to split the original version in 3 articles has been hugely beneficial. (I will also write reviews for the other 2 articles in the next few weeks).

Right now, while I still have some suggestions on how to make the manuscript better, they are (a) comparatively minor compared to the original version (b) pertain mostly to stylistic choices, rather than the substance of the manuscript (with the notable exception of COMMENT 4). Note that “comparatively minor” does not mean “unimportant”: the original manuscript was frequently unclear as it prioritised overly faithful transcription of the preregistration over making a compelling and accurate report that the reader could follow. The current version is much, much more balanced in that respect, but there are still quite a few points that might trip the average reader.

Please find below my specific comments :

[Note on line numbers and comments about formatting: I reviewed the EcoEvoRxiv version of the manuscript]

COMMENT M1: While I very much appreciate the added definitions in the Introduction, I can’t help but notice that some may seem very circular to the general reader. For instance, if we reformulate the text line 50 to use both definitions of flexibility and innovativeness in the same sentence:

”it is predicted that [the ability to adapt behavior to new circumstances] should positively related with [the ability to create new behavior or use an existing one in a new situation]”;

then it look very much like both definitions in [brackets] are reformulated versions of each other.

Can you clarify in-text how flexibility and innovativeness refer to different things (as I assume they actually do, current version of the text notwithstanding)?

RESPONSE M1: We are glad that you find the splitting of the article into three beneficial. Thank you so so much for reviewing all of these articles! We really appreciate it :-)

Thank you for pointing out the issue with the definitions. This is indeed a muddled area and Logan previously collaborated with philosophers of animal cognition to bring some clarity to the term “behavioral flexibility” (Mikhalevich et al. 2017). We see flexibility as the ability to recognize that something about the environment has changed and decide to consider other options for deploying behavior. While innovation is the specific stringing together of particular behaviors in response to the decision to change behavior in some way. We agree that we need to be more clear about how these two traits differ from each other so we clarified the definition of flexibility as follows:

Introduction: “Behavioral flexibility, the ability to adapt behavior to new circumstances through packaging information and making it available to other cognitive processes“

COMMENT M2: Line 53: For these kinds of behavioural constructs, aren’t we always measuring behavioral proxies of the “true” latent thing? (if the true thing even exists per se and isn’t only a construct we make to make sense of individual variation). I’m not sure having to use proxies for flexibility… is that much an issue as you seem to make it in this sentence. If the proxies you’re mentioning are non-behavioral (brain volume for instance), then yes, I see your point, but maybe it is best to make it explicit?

RESPONSE M2: Good point to clarify the proxies we were referring to, which allows the reader to evaluate how our direct tests differ. We made the following change to the text:

Introduction: “However, these predictions are based on species-level data and proxies for flexibility and for innovation **(e.g., brain size, number of anecdotal reports of “novel” foods consumed)** when examining such relationships”

COMMENT M3: You decided to remove the batch effect from your analysis because, among other things, it does not actually reflect a temporal "clustering" anymore. I have two questions/comments about that.

Based on the text, you don't seem to remove it from the P1 analysis, only from P2 onwards. Why is that, if your decision to include or not a Batch effect is due to, as you say, how manipulated/control birds were shuffled into batches (as opposed to e.g., data table size constraints on the estimation of random effects)?

Second, if there is a link between willingness to participate and flexibility/innovativeness (and I wouldn't be surprised if there were), wouldn't testing individuals more willing to participate first/preferentially open the experiment to some confounding? I would appreciate a discussion of this if this is a relevant issue here, especially in the context of the STRANGE framework (https://doi.org/10.1038/d41586-020-01751-5)

RESPONSE M3: We only did the evaluation of whether batch should be included or not once we got to the analyses for P2. What we did for P1 (Table 2) was run the model without batch for comparison. We found that there were no differences between the models with and without batch that would result in different conclusions. We defaulted to the model that included batch because it was in the preregistration and we wouldn't have to write a deviation for it. Per your comment, we additionally ran a comparison for Tables 1 and 3 (Results > P1) and also found that the results were very similar such that the conclusions were the same. We decided to treat them as before, by keeping the model with the batch in the text to reduce the number of deviations we needed to add to the Methods.

We evaluated the presence of variation in personality traits (exploration and boldness) that may affect willingness to participate in the behavioral flexibility experiment in a separate preregistration. In that study we directly tested the link between reversal learning performance and exploration, boldness, motor diversity, and persistence (McCune et al. 2019, http://corinalogan.com/Preregistrations/g_exploration.html). We are currently writing up the post-study manuscript and do not yet have the results, but we will report on any potential links (or lack thereof) there.

COMMENT M4: Regarding your response to original review comment D9: I was happy to see that you say that including a individual-level random slope effect of reversal number does not change the conclusion. However, based on the text (line 297) and the code (https://github.com/corinalogan/grackles/blob/5b220d237542053bfa51673ec38116497affe55c/Files/Preregistrations/g_flexmanip.Rmd#L219), I actually see no evidence that you actually did fit the model with random slopes. See https://stat.ethz.ch/pipermail/r-sig-mixed-models/2013q1/019896.html for an example of the expected syntax for the inclusion of individual-level random slopes in MCMCglmm, and please refer back to the citations I mentioned in the original comment, for the importance of correctly accounting for this (and also Harrison et al 2018, https://peerj.com/articles/4794, chapter "Choosing random effects II: random slopes" for a review). Apologies in advance if I missed the correct model somehow.

Note that I acknowledge that in some cases the dataset may be too small to estimate these random slopes accurately (see the aforementioned Harrison et al. reference), but then, the better option is to acknowledge it explicitly, rather than ignoring it.

RESPONSE M4: We apologize for this confusion! We did implement the random slope, however it was in a different section - in the results section (r code chunk "p1table"), where the table with the model output appeared (Results > P1 > Table 2). We realized it was really confusing to have an additional r chunk for P1 in the Analysis plan section, which is the code that you saw, and which we used for data checking purposes (r code chunk "serial"). Note that, because we were able to get DHARMa to do data checking on a MCMCglmm model (thanks to your feedback), we now added the MCMCglmm model to the r code chunk "serial" so the data checking was on the actual model used in the results, which included random slopes. We checked Harrison et al (2018) - thank you for sharing that - and it appears that we have enough data from 8 individuals to include the random slopes (they recommend a minimum of 5) and the number of data points between individuals does not vary that much (6-8) so it is balanced enough for a stable model.

Here is our random slope model for ease of finding it:

prior = list(R = list(R1 = list(V = 1, nu = 0)), G = list(G1 = list(V = 1, nu = 0), G2 = list(V = 1, nu = 0)))

MCMCglmm(TrialsToReverse ~ ReverseNumber, random = ~us(ReverseNumber):ID+Batch, family = ”poisson”, data = d, verbose = F, prior = prior, nitt = 300000, thin = 500, burnin = 90000)

COMMENT M5: Line 666: while I agree that differences of outcome between manipulative experiments and observations can and should be expected, it's not simply a matter of observations being more “noisy”. First, sample size may play a big role in determining uncertainty around estimates and both type M and type S errors, and it is sometimes easier to get high sample sizes in an experimental context, sometime harder. Second, any differences between studies may be much more prosaically because different studies analyse different species with different traits (you even mention that possibility, at the population level, when you discuss urban vs rural pops one paragraph down)

RESPONSE M5: Good points. We added the following to incorporate your ideas:

Discussion: “Other potential reasons for the difference in results include using different experimental designs, and/or different serial reversal passing criteria (Griffin et al., 2013 used a preset number of reversals that resulted in a maximum of four reversals), **inherent species differences, or needing a larger sample size to help reduce noise in a non-manipulative experiment.**”

COMMENT M6: Lines 32-35: That sentence is a bit unclear due to its length, and it might be helpful to split it in two (or remove the phrases in parentheses). In addition, there is a missing parenthesis.

RESPONSE M6: We can see how this long sentence was confusing. We deleted sections of the sentence to make it clearer. It now reads:

Abstract: “All grackles in the manipulated reversal learning group used one learning strategy (epsilon-decreasing) in all reversals, and none used a particular exploration or exploitation strategy earlier or later in their serial reversals.”

COMMENT M7: Line 55: again, this sentence is too long. I would suggest to cut or reformulate (the added definition of “problem solving” is a big part of what makes the sentence unwieldy _ Maybe move it to its own sentence?). It might be useful to check through the manuscript for similar long sentences that may be hard to follow at time (this is a default I definitely share, so I am sympathetic).

RESPONSE M7: Thanks for pointing this out! We agree that the addition of the problem solving definition caused problems, so we pulled it out and added a sentence afterward as follows:

Introduction: “Those investigations that examine the relationship between flexibility and innovation or problem solving in species that are expanding their range show mixed results, with these variables correlating positively (e.g., grey squirrels: Chow et al., 2016), negatively (e.g., Indian mynas: Griffin et al., 2013), or not at all (e.g., stick tool use and string pulling in great-tailed grackles: Logan, 2016). Problem solving in this context was a type of experimental assay that did not necessarily require innovativeness to solve (e.g., the ability to solve tasks using pre-trained behaviors; Griffin & Guez 2014).”

COMMENT M8: Line 72: I suggest rephrasing as “We focused our study on one population of great-tailed grackles (Quiscalus mexicanus, hereafter grackles), a bird species that is flexible (Logan, 2016). While they are originally from Central America, great-tailed grackles have rapidly expanded their geographic range across the US since 1880…”

RESPONSE M8: Thank you! We implemented your change (Introduction).

COMMENT M9: Figure 1: The natural reading order of the figure (left to right and top to bottom: A->C->B) does not match the labelling and expected hypothesis order (ABC). I am aware that the relative sizes of the subpanels complicate things; I would suggest a layout where each subpanel occupies a whole row?

RESPONSE M9: Great idea! We implemented this (Figure 1)

COMMENT M10: Hypotheses: I get that H3 is not part of this manuscript anymore, and that you want to keep hypothesis numbers consistent with the prereg and across the manuscripts, but each manuscript

should be to some degree independent and having this paragraph go H1->H2->H4 without explanation invites questions. I would suggest adding something like "H3 [one short sentence/question describing the hypothesis]. This hypothesis from the original preregistration is now treated in a separate manuscript (reference and/or link to preprint)".

RESPONSE M10: Good idea to include this so the reader isn't left wondering. We did as you suggested (Introduction > Hypotheses > H3).

COMMENT M11: Line 232: DHARMa actually definitely supports MCMC-based models, independently of package/language used to fit them, but this does necessitate some processing/extraction work on the posteriors. It only needs a matrix containing the posterior fits and a vector of the observed responses. See for example (examples are JAGS and Stan models, but this easily transfers to MCMCglmm based models)

https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMaForBayesians.html

https://frodriguezsanchez.net/post/using-dharma-to-check-bayesian-models-fitted-with-brms/

RESPONSE M11: That's great that DHARMa can work with MCMCglmm - thank you so much for these resources. There is only one model in the article where we ran DHARMa using glmer and then conducted the actual analysis in MCMCglmm: P1 when examining the correlation between trials to reverse and reversal number. Using advice from the resources you included and using the simulate function in MCMCglmm, we were able to get it to work with the DHARMa package. We deleted the sentence stating that we were not able to do this (Analysis > Data checking).

COMMENT M12: Lines 258-262: On the choice criterion: if usable screenshots from videos are available, this is definitely something where images illustrating how it looks at the moment a choice is made would be hugely beneficial (in main text or in appendices)

RESPONSE M12: Good idea - we grabbed a screenshot from a video and added a picture - the new Figure 3.

COMMENT M13: Line 335 and elsewhere: this one is very minor, but you might consider naming this distribution as "negative Binomial" rather than "gamma-Poisson", as the former name has more traction in the literature, at least in my experience.

RESPONSE M13: We prefer to stick with the terms we currently use because they follow the terms used in McElreath's book Statistical Rethinking, which is what we used to develop these models. However, on first use, we added:

Methods > Analyses > P2: "Note that a gamma-Poisson distribution is also known as negative binomial."

COMMENT M14: Line 364: whether or not you need to include treatment in your model does not actually depend, strictly speaking, on whether manipulated birds have by definition faster reversal speed, but on what is the (causal) hypothesis you want to test with this specific model.

If P2 is specifically "more flexible birds, independent of manipulation, are more innovative", then yes your model is correct, (and I insist on this point) even if manipulation did not result in changes in reversal speed. Please check carefully your predictions and hypothesis against your implied or explicit causal assumptions, and update the text accordingly

RESPONSE M14: Thank you for pointing us back to our hypothesis structure to make sure there is a clear link with the model structure. We referred back to Hernán & Robins (2006) to help us get clear about this. The assignment to a condition (manipulated or control) was the instrumental variable (Z) in a randomized experiment, where individuals experienced a treatment (X; serial reversals or only 1 reversal) with MAB performance as the outcome variable (Y). Z should not be included as a variable in statistical models testing the relationship between X and Y because it is an instrumental variable, which has the following properties: it causes X, it affects Y only via X, and Z and Y do not have a common cause (Hernán & Robins 2006).

Our hypothesis 2 states: "Manipulating behavioral flexibility (improving reversal learning speed through serial reversals using colored tubes) improves flexibility (rule learning and/or switching) and problem solving in a

new context".

This means that condition is central to the hypothesis, but should not be included in models that also include reversal learning speed because of its role as an instrumental variable. We looked at the effect of condition on both response variables (average latency and total loci solved) on both MABs in Table SM3 in the absence of other independent variables, and in separate models we also looked at the response variables in the context of reversal learning speeds (without condition).

Now that we have gone into the theoretical aspects here, thanks to your comment, we realize that we do not need the P2 models with condition. However, we will keep them in (Figure 7 and Table SM3) because in P1 we showed that the manipulation worked because it changed the reversal speed and in P2 we also show that the manipulation worked because it changed performance on the MABs. So it serves as additional evidence that the manipulation worked (supporting P1).

We made the following changes to the text:

Analyses > Deviations from the preregistration:

Previously: "We also realized that Condition (manipulated or control) does not need to be a variable in any of our models because the manipulated birds have, by definition, faster reversal speeds."

Currently: "We also realized that Condition (manipulated or control) does not need to be a variable in any of our models because **our analyses in P1 demonstrate that the manipulation causally changes reversal speeds, which is the key assumption in P2**"

COMMENT M15: I would recommend to check the formatting for consistency. In particular, there are consistently missing new line shifts between prediction subtitles and the "planned analysis" paragraphs (see eg line 367). Some of it may stem from copy pasting the original prereg text without checking formatting. In my experience, some of it may stem from slight differences in the way the rendering engine treats pdfs and html, if you only proof-read one of the formats (I reviewed the EcoEvoRxiv pdf). There are also cases of special characters that are not rendered correctly, e.g. in Table SM1: see how the R- squared was not correctly formatted.

RESPONSE M15: These are rendering differences between html (which we usually use) and pdf (which we post at the preprint server). We do check the pdfs to correct rendering issues, but it seems that some errors got through so thanks for pointing these out. We checked the final pdf and think we fixed all of these formatting issues. We also changed the superscript issue with the R-squared by spelling it out rather than trying to use a superscript (which wasn't obvious how to do).

COMMENT M16: On a related formatting comment, I suggest that if you quote the "planned analysis" paragraphs as is from the prereg (without adjusting tense for instance), it may be good to put them between quotes so that it is clear they are direct citations.

RESPONSE M16: Thank you for the suggestion. We ended up rewriting the planned analyses sections per Comment C1, so please see our Response C1 for how this section changed.

COMMENT M17: Line 382: it is unclear from the text why there is now only one point per bird vs several in the preregistration. Is it because the response variable has changed? The preregistration was wrong? Another thing? Please make it explicit

RESPONSE M17: This was actually a mistake we made in the preregistration - neither dependent variable in the preregistration for this analysis would have had repeated measures per bird. We made this clear in the text:

Analyses > Deviations from the preregistration: "(note that there would have been only one data point per bird in the preregistration as well, but we didn't realize this until after in principle acceptance)"

COMMENT M18: Line 392-411: not convinced at all this detailed quote with equations is needed, vs a simpler summary in a couple sentences and a citation to the original reference. If you keep it, and if it is indeed a direct quote, please again place it between quotation marks to make it clear it is a quote and not newly written text

RESPONSE M18: That's a good point. We ended up deleting all of the text from Logan et al. (2016) and added a description to the epsilon-decreasing sentence (the epsilon-first sentence already had a description).

Analyses > P4: "Other patterns were classified as the epsilon-decreasing strategy where individuals gradually increase their number of successes as the number of trials increases"

COMMENT M19: Line 453: given the numerous post-hoc changes to the protocol, are these simulation outcomes still valid, useful and meaningful?

RESPONSE M19: The simulation outcomes for Table SM1.1 are still valid because the only modifications to the model structure in the final analyses involved removing batch, which is something we did previously to see if it would make a difference. The values between the models with and without batch were very similar to each other (see R code chunk "modelparameters1" > mwobatch).

COMMENT M20: Figure 3: the caption states that "Individuals in the manipulated condition (who received serial reversals) did not linearly decrease their reversal passing speeds with increasing reversal number". This directly contradicts the text line 490 "there was a significant negative correlation between the number of trials to reverse (...) and the reversal number for those grackles in the flexibility manipulation condition" as well as Table 2. Please correct or explain

RESPONSE M20: Thank you for catching this mistake! It was an error in the figure caption text because the analysis in Table 2 indicates there was a linear decrease. We corrected the caption for the fig, which is now Figure 4:

"Individuals in the manipulated condition (who received serial reversals) decreased their reversal passing speeds with increasing reversal number (n=9 grackles)"

COMMENT M21: Line 518: there is at least one reference to a table (table 7) that do not exist in the current version of the manuscript. Please check very carefully throughout the manuscript for others

RESPONSE M21: Sorry about that! We went through the figure numbers in the manuscript and believe we have corrected all of the errors.

COMMENT M22: Lines 610-612 and Table 4: just as importantly, you might say that the null model remains the best supported model

RESPONSE M22: Yes, it is true that there was also no relationship between the response variable (latency to attempt a locus on the wooden MAB) and the number of trials to reverse in the last reversal, however the only reason we conducted the analysis for P2 alternative 2 was because there was no relationship here. Therefore, we will not mention that the null model is the best supported here because the only purpose of this analysis was to test the motor diversity variable.

COMMENT M23: Regarding Figure 7, and the authors' reply to my comment 14 in the previous review: I appreciate the authors' reply that the seeming outlying individual should not be excluded from the final analysis as there was no difference of treatment that would justify it. I 100% agree with that. However, this was not the intent of my comment. It was about model diagnostics (which usually include leverage checks), and suggesting a form of leave one out cross validation. If removing that point (or any other, for that matter) does change the conclusion, then confidence in the main analysis (which should include all points, indeed) must be adjusted accordingly. Indeed, you can imagine the converse: if removing a single individual can change things that much, what would have happened if you had done just one more bird, as there was scope to do in the preregistration? So to clarify, my suggestion is:

- 1/ to keep the analysis including all points as the main/actual analysis used and discussed;
- 2/ nonetheless to re-do the analysis without that point (without presenting it in full; just as we don't usually present all residual normality/heteroscedasticity checks in papers);
- 3/ and to have in mind the outcome of 2/, when discussing the results of 1/

RESPONSE M23: Ah, now we see what you meant by your comment on the previous revision - sorry for misinterpreting. Taquito (the bird with the 160 trial reversal speed) did not have a latency for the plastic MAB, but he did have data in the other three models so we ran the cross validation checks for these models. We also conducted another set of cross validation checks by removing a randomly selected bird (not Taquito) from the data to determine whether removing any bird makes a difference or only Taquito (if his removal makes a difference).

We found that removing Taquito or the random bird **did not change the results** for:
Model 12: wooden MAB latency, Fig 8b - there remained no relationship
Model 6: wooden MAB loci solved, Fig 8d - there remained no relationship

We found that removing Taquito or the random bird **did change the result** for:
Model 2: plastic MAB loci solved, Fig 8c - there was no relationship (whereas there is a relationship when the whole dataset is analyzed)

Because the changes were consistent based on removing a random data point and not specific to Taquito, we conclude that his high reversal learning value was not difficult for the model to account for. Therefore, when a model result changes due to the removal of a random bird (whether it was Taquito or not), it means that we need to be less confident in the main model. We added an asterisk to Figure 7, which indicates that the result is not as confident due to the small sample size, and we added this to the discussion accordingly as follows:

Discussion: "Because of the variation in our small sample (Taquito was by far the slowest to reverse a preference), we conducted a cross validation check to determine whether removing a bird from the data set changed the model results. We found that there was no difference in results when removing Taquito or a random bird. However, removing either from the data set changed the conclusions for one of the three models (Model 2, but not Models 6 or 12). This change in results after removing a data point indicates that we should be less confident in the conclusion that individuals who are faster to reverse a preference in their last reversal also solved more loci on the plastic multi-access box."

COMMENT M24: Figure 8 and equivalent figures in appendices: The way the figure is flattened is not flattering. I would suggest to increase the height of each subpanel and/or to reduce their width. Also do not feel you need to have all subpanels on the same column.

RESPONSE M24: Great idea - we implemented the 2 column format (now Fig 9, and in SM6) and it looks much better, thanks!

COMMENT M25: Line 803: "were", not "weAre". Please check carefully through the manuscript for typos

RESPONSE M25: We corrected the typo, thanks for catching this!

COMMENT M26: Table SM2: all or part of this table may be more helpful to the reader as one or several figures?

RESPONSE M26: This table is actually SM3 - we had misnumbered it in two instances, but corrected that now. The key model results are shown as figures in the Results section (the four panels in figure 8). We would like to keep Table SM3 as it is to ensure that readers have the raw numbers from the full models.

Reviewed by Aliza le Roux, 03 Oct 2022 11:13

COMMENT A1: The preprint "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context." by Logan et al. (https://doi.org/10.32942/osf.io/5z8xs) is a product of much problem- solving and flexibility on the part of the researchers investigating flexibility itself as a biological salient and scientific construct. It is based on a preregistered "in principle" acceptance of their proposed research (reviewed by Maxime Dahirel and Andrea Griffin; 10.24072/pci.ecology.100019). I did not read through the extensive comments on earlier drafts of this paper, as I wanted to have a completely fresh perspective on the

research.

The primary finding of this research on grackles, a bird species with a rapidly expanding geographic range, is that behavioural flexibility is not a static ability – whether expressed or not – but that animals can be "primed" for flexible responses by past experiences. Furthermore, the expression of flexibility appears to also influence other expressions of cognitive ability in the problem-solving domain. I think their approach is fascinating and as novel as they claim it to be. It is emerging that the expression of various cognitive traits is influenced by past or recent experience or environmental context (e.g., le Roux et al. 2019). The links between different cognitive abilities/constructs are currently ambiguous, to say the least; for example, exploration diversity may (Huebner et al. 2015) or may not (Lermite et al. 2017) be positively associated with problem solving ability. Therefore, this kind of research that explicitly examines the link between multiple traits is central to our understanding of how cognition works. It is clear that we cannot make assumptions about abilities or links between traits unless we examine numerous species and start to discern common patterns.

Logan et al. have here demonstrated that behavioural flexibility can be trained or improved in individual grackles, which then appears to influence innovative abilities in solving a multi-access puzzle box. These results have many potential implications, suggesting that experience in a fast-changing environment will affect problem-solving abilities in wild grackles. This aligns with research showing more innovative tendencies in urban vs rural species, based on the assumption that an urban environment is subject to more rapid changes in food availability and/or risk (e.g., Lowry et al. 2013).

In examining effect sizes, assessing innovation in more than one context, and adjusting their predictions to their experimental constraints (birds could not be tested indefinitely), they have done commendably thorough and transparent research on behavioural flexibility and innovation in wild- caught birds. They also remain cautious in their interpretations of the results. I agree with their interpretations and note that perhaps the only major constraint that was not considered is the potential influence of simple experience/ exposure to trials. I am assuming that the manipulated group of grackles had many more trials with the choice tests than the control group did, leading to a potential imbalance. I don't think absolute number of trials in which any bird participated was examined anywhere. The possible influence of exposure is something that could either be examined statistically or noted as a constraint/consideration in the interpretation of results. This is the only real caveat I have about the interpretation of their findings. My statistical skills are not significant enough to thoroughly interrogate all the modeling conducted in this dense, detailed work. As an ecologist with hands-on mixed modeling experience and an understanding of Bayesian principles, I can say that the statistical reasoning, interpretation, and multiple checks and balances appear to be thorough. I'd be happy if someone with more specific statistical expertise than mine could indeed confirm whether mistakes or statistical misinterpretations slipped in; this may have already been done with the approval of the original registered report.

Logan and her colleagues also demonstrate, perhaps unintentionally, how challenging it is to conduct research on behavioural flexibility and problem-solving in species not reared in a lab, and how important it is to constantly assess assumptions and predictions as the data unfolds. This thorough reassessment and flexibility on the part of the researchers would not have been clear without a previously evaluated registered report as guideline, and, to me, their entire project shows the value of peer-review and transparency at all stages of the research process. I recommend this preprint for acceptance.

RESPONSE A1: We love your comment about it requiring flexibility on our part to conduct this research - we say this to ourselves often! It has definitely been a test of our cognitive abilities along the way! We really appreciate your thoughts on our research and are happy that you feel like it is a good contribution that was made more rigorous by the registered report format. Thank you very much for your support! Regarding your point about the reversal manipulated individuals having more trials than the reversal control individuals, we balanced the number of trials for the control group by giving them the average number of trials it took a manipulated individual to pass the serial reversal. The trials for the control group after their first reversal consisted of 2 yellow tubes on the table in the same position as the light and dark gray reversal tubes, and both yellow tubes contained food so it did not matter which choice they made. This is clarified in the Hypothesis

section > Hypothesis 1 > Prediction 1, the Methods section > Analyses > Deviations from the preregistration > Reversal learning color tube choice criterion, and the new “Summary of testing protocols” section in the Methods.

I have some minor comments, below:

COMMENT A2: In the title, I would lean towards using the word “manipulable” rather than “manipulatable”, but this is more a personal preference. Both words are applicable, though “manipulable” is the more common term and also, apparently, used in context of psychology rather than engineering.

RESPONSE A2: That is an interesting difference - we weren’t aware of this so thank you for bringing it to our attention. We changed the word to manipulable throughout the text.

COMMENT A3: Line 61: This paragraph discusses manipulative experiments, juxtaposing it with correlations/observational evidence. Is that what the studies in the preceding paragraph employed, or did they also do manipulative experiments? The way this is set up suggests that the experiments are in contrast with what was just mentioned, but it’s not made explicit; it will help the reader to understand the novelty (or not) of the approach.

RESPONSE A3: Good point. You are correct in that the studies in the preceding paragraph did not manipulate flexibility. To clarify, we added a last sentence to preceding paragraph:

Introduction: “However, none of these experiments manipulated flexibility.”

COMMENT A4: Line 346: “analyseis” is misspelled.

RESPONSE A4: Thank you for catching this! We fixed the typo.

COMMENT A5: Table 4, p19 is unclear – what does each successive line represent?

RESPONSE A5: We now clarified in the table caption that “Each row represents one model that includes different independent variables (motor actions and/or trials last reversal)”

COMMENT A6: Line 803 – “at that time” is misspelled, as is “were” earlier in the same sentence

RESPONSE A6: Thank you for catching this! We fixed the typos.

COMMENT A7: Line 849: it is confusing to see “probability” as equivalent to “rate” per bird. Is this the likelihood of solving a particular locus, based on random probability, or based on observed rates of solving loci?

RESPONSE A7: Lambda in each second is a random probability. We removed the word “rate” from this sentence and clarified in the text:

SM1: “λi is the **random** probability of attempting a locus in each second per bird”

— Additional changes that were not brought up by the reviewers:

We realized that we should refer to the reversal learning of a “color” preference as reversal learning of a “shade” preference because light and dark gray are two shades of the same color and not two different colors. Therefore, we changed the term “color” to “shade” throughout the manuscript.

We also realized that we were using “innovativeness” and “problem solving” interchangeably even though

we distinguished them in the introduction. For clarity, we changed most of the “problem solving” terms to “innovativeness”.

We updated the Figure 7 legend to say that we are comparing abilities and that the plus and minus signs refer to the relationship between flexibility and innovativeness. This is because it is confusing to use pluses and minuses when referring to the variables we measured because the two flexibility measures both indicate that more flexibility is achieved when solving in fewer trials or in fewer seconds, however the innovativeness measure has the opposite relationship because solving more loci indicates more innovation.

We updated the shading in Figure 8 to make it so that darker shading indicates relationships we found and we added this clarification to the legend.

Download tracked changes file

Decision by Aurélie Coulon, posted 13 October 2022

Revision needed for the preprint ”Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context”

Dear Corina Logan and co-authors, Two reviewers have evaluated your manuscript called “Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context”. One of them was a reviewer on its previous version, and of its preregistration. The latter is very enthusiastic about your ms. The former highlights the improvements made since the last version, but also has a number of comments, mainly about the form of the paper. I agree with all of them, and think that overall, the ms has improved but is still a bit tough to read. This is due to the willingness to stick to the preregistration, and report on what was different from it. It makes the material and methods part a bit difficult to follow and rather technical, in the sense that the biological (behavioural) aspects of the experiments are kept away. Besides the reviewers’ comments, I have a few comments and suggestions (listed below) to work on those aspects. I think the different comments made by the reviewers and myself are rather easy to take into account, and should make the manuscript close to the recommendation stage. Best, Aurélie Coulon - Material and methods: I think it is very confusing to present the analyses as they were planned, and then the deviations that were made to the plan. This is because there is nowhere, in the M&M section, where the reader can easily find what was actually done (they have to read the initial plan + the deviations and then make their own synthesis of all this). I think it would be much easier for the reader if the methods were presented in their final stage (instead of as they were planned), and accompanied by the section describing the deviations to the initial plan (as they are currently written). Also, the fact that the experimental protocols are only cited, and not described at all, is contributing to this feeling that we miss something about the behavioural aspects of the study. I suggest to add a brief summary of the protocols, before the section Open materials.

- In Fig.1C, the link between the numbers, arrows and the phrases “2+ learning strategies” and “1 learning strategy” is not easy to get
- In several places there are mistakes in the figure or sup.mat. numbers (e.g. l.196, it should read sup.mat.1 instead of 6; l.224: 2 instead of 5;l.592: fig.7d instead of 5d; ...)
- l.216-221: I don’t understand why papers are cited (without a reference to a package) at the beginning of this list which is supposed to be a list of R packages
- Unless I got lost, there are several unregistered analyses that are not presented in the material and methods section: l.489; unregistered analysis 1, l.504-508; unregistered analysis 2, l.517-520; P2: analyses of the effect of time in first reversal. This should be fixed.
- Table SM3: I don’t understand what “a” is, in some models (e.g. models 2 and 5). Also, in model 1, there seems to be a “batch” factor, while it was said in the M&M that this factor was eventually removed from the analyses...

Reviewed by Maxime Dahirel, 13 October 2022

I have now read the revised version of the recommended-in-principle manuscript entitled "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context" by Logan et al.

First of all, I have to say that this is a much improved and streamlined version of the manuscript. While it certainly is not the only reason, I believe the decision to split the original version in 3 articles has been hugely beneficial. (I will also write reviews for the other 2 articles in the next few weeks).

Right now, while I still have some suggestions on how to make the manuscript better, they are (a) comparatively minor compared to the original version (b) pertain mostly to stylistic choices, rather than the substance of the manuscript (with the notable exception of **COMMENT 4**). Note that "*comparatively* minor" does not mean "unimportant": the original manuscript was frequently unclear as it prioritised overly faithful transcription of the preregistration over making a compelling and accurate report that the reader could follow. The current version is much, much more balanced in that respect, but there are still quite a few points that might trip the average reader.

Please find below my specific comments :

[Note on line numbers and comments about formatting: I reviewed the EcoEvoRxiv version of the manuscript]

COMMENT 1:

While I very much appreciate the added definitions in the Introduction, I can't help but notice that some may seem very circular to the general reader. For instance, if we reformulate the text line 50 to use both definitions of flexibility and innovativeness in the same sentence:

"it is predicted that [the ability to adapt behavior to new circumstances] should positively related with [the ability to create new behavior or use an existing one in a new situation]";

then it look very much like both definitions in [brackets] are reformulated versions of each other.

Can you clarify in-text how flexibility and innovativeness refer to *different* things (as I assume they actually do, current version of the text notwithstanding)?

COMMENT 2:

Line 53: For these kinds of behavioural constructs, aren't we *always* measuring behavioral proxies of the "true" latent thing? (if the true thing even exists per se and isn't only a construct we make to make sense of individual variation). I'm not sure having to use proxies for flexibility… is that much an issue as you seem to make it in this sentence. If the proxies you're mentioning are non-behavioral (brain volume for instance), then yes, I see your point, but maybe it is best to make it explicit?

COMMENT 3:

You decided to remove the batch effect from your analysis because, among other things, it does not actually reflect a temporal "clustering" anymore. I have two questions/comments about that.

Based on the text, you don't seem to remove it from the P1 analysis, only from P2 onwards. Why is that, if your decision to include or not a Batch effect is due to, as you say, how manipulated/control birds were shuffled into batches (as opposed to e.g., data table size constraints on the estimation of random effects)?

Second, if there is a link between willingness to participate and flexibility/innovativeness (and I wouldn't be surprised if there were), wouldn't testing individuals more willing to participate first/preferentially open the experiment to some confounding? I would appreciate a discussion of this if this is a relevant issue here, especially in the context of the STRANGE framework (https://doi.org/10.1038/d41586-020-01751-5)

COMMENT 4:

Regarding your response to original review comment D9: I was happy to see that you say that including a individual-level random slope effect of reversal number does not change the conclusion. However, based on the text (line 297) and the code (https://github.com/corinalogan/grackles/blob/5b220d237542053bfa51673ec38116497affe55c/Files/Preregistrations/g_flexmanip.Rmd#L219), I actually see no evidence that you actually did fit the model with random slopes. See https://stat.ethz.ch/pipermail/r-sig-mixed-models/2013q1/019896.html for an example of the expected syntax for the inclusion of individual-level random slopes in MCMCglmm, and please refer back to the citations I mentioned

in the original comment, for the importance of correctly accounting for this (and also Harrison et al 2018, https://peerj.com/articles/4794, chapter "Choosing random effects II: random slopes" for a review). Apologies in advance if I missed the correct model somehow.

Note that I acknowledge that in some cases the dataset may be too small to estimate these random slopes accurately (see the aforementioned Harrison et al. reference), but then, the better option is to acknowledge it explicitly, rather than ignoring it.

COMMENT 5:

Line 666: while I agree that differences of outcome between manipulative experiments and observations can and should be expected, it's not simply a matter of observations being more "noisy". First, sample size may play a big role in determining uncertainty around estimates and both type M and type S errors, and it is sometimes easier to get high sample sizes in an experimental context, sometime harder. Second, any differences between studies may be much more prosaically because different studies analyse different species with different traits (you even mention that possibility, at the population level, when you discuss urban vs rural pops one paragraph down)

COMMENT 6:

Lines 32-35: That sentence is a bit unclear due to its length, and it might be helpful to split it in two (or remove the phrases in parentheses). In addition, there is a missing parenthesis.

COMMENT 7:

Line 55: again, this sentence is too long. I would suggest to cut or reformulate (the added definition of "problem solving" is a big part of what makes the sentence unwieldy _ Maybe move it to its own sentence?). It might be useful to check through the manuscript for similar long sentences that may be hard to follow at time (this is a default I definitely share, so I am sympathetic).

COMMENT 8:

Line 72: I suggest rephrasing as "We focused our study on one population of great-tailed grackles (*Quiscalus mexicanus*, hereafter grackles), a bird species that is flexible (Logan, 2016). While they are originally from Central America, great-tailed grackles have rapidly expanded their geographic range across the US since 1880..."

COMMENT 9:

Figure 1: The natural reading order of the figure (left to right and top to bottom: A->C->B) does not match the labelling and expected hypothesis order (ABC). I am aware that the relative sizes of the subpanels complicate things; I would suggest a layout where each subpanel occupies a whole row?

COMMENT 10:

Hypotheses: I get that H3 is not part of this manuscript anymore, and that you want to keep hypothesis numbers consistent with the prereg and across the manuscripts, but each manuscript should be to some degree independent and having this paragraph go H1->H2->H4 without explanation invites questions. I would suggest adding something like "H3 [one short sentence/question describing the hypothesis]. This hypothesis from the original preregistration is now treated in a separate manuscript (reference and/or link to preprint)".

COMMENT 11:

Line 232: DHARMa actually definitely supports MCMC-based models, independently of package/language used to fit them, but this does necessitate some processing/extraction work on the posteriors. It only needs a matrix containing the posterior fits and a vector of the observed responses. See for example (examples are JAGS and Stan models, but this easily transfers to MCMCglmm based models)

https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMaForBayesians.html

https://frodriguezsanchez.net/post/using-dharma-to-check-bayesian-models-fitted-with-brms/

COMMENT 12:

Lines 258-262: On the choice criterion: if usable screenshots from videos are available, this is definitely something where images illustrating how it looks at the moment a choice is made would be hugely beneficial (in main text or in appendices)

COMMENT 13:

Line 335 and elsewhere: this one is very minor, but you might consider naming this distribution as “negative Binomial” rather than “gamma-Poisson”, as the former name has more traction in the literature, at least in my experience.

COMMENT 14:

Line 364: whether or not you need to include treatment in your model does not actually depend, strictly speaking, on whether manipulated birds have by definition faster reversal speed, but on what is the (causal) hypothesis you want to test with this specific model.

If P2 is specifically “more flexible birds, independent of manipulation, are more innovative”, then yes your model is correct, (and I insist on this point) even if manipulation did not result in changes in reversal speed. Please check carefully your predictions and hypothesis against your implied or explicit causal assumptions, and update the text accordingly

COMMENT 15:

I would recommend to check the formatting for consistency. In particular, there are consistently missing new line shifts between prediction subtitles and the “planned analysis” paragraphs (see eg line 367). Some of it may stem from copy pasting the original prereg text without checking formatting. In my experience, some of it may stem from slight differences in the way the rendering engine treats pdfs and html, if you only proof-read one of the formats (I reviewed the EcoEvoRxiv pdf).

There are also cases of special characters that are not rendered correctly, e.g. in Table SM1: see how the R-squared was not correctly formatted.

COMMENT 16:

On a related formatting comment, I suggest that if you quote the “planned analysis” paragraphs as is from the prereg (without adjusting tense for instance), it may be good to put them between quotes so that it is clear they are direct citations.

COMMENT 17:

Line 382: it is unclear from the text why there is now only one point per bird vs several in the preregistration. Is it because the response variable has changed? The preregistration was wrong? Another thing? Please make it explicit

COMMENT 18:

Line 392-411: not convinced at all this detailed quote with equations is needed, vs a simpler summary in a couple sentences and a citation to the original reference. If you keep it, and if it is indeed a direct quote, please again place it between quotation marks to make it clear it is a quote and not newly written text

COMMENT 19:

Line 453: given the numerous post-hoc changes to the protocol, are these simulation outcomes still valid, useful and meaningful?

COMMENT 20:

Figure 3: the caption states that “Individuals in the manipulated condition (who received serial reversals) did not linearly decrease their reversal passing speeds with increasing reversal number”. This directly contradicts the text line 490 “there was a significant negative correlation between
the number of trials to reverse (...) and the reversal number for those grackles in the flexibility manipulation condition” as well as Table 2. Please correct or explain

COMMENT 21:

Line 518: there is at least one reference to a table (table 7) that do not exist in the current version of the manuscript. Please check very carefully throughout the manuscript for others

COMMENT 22:

Lines 610-612 and Table 4: just as importantly, you might say that the null model remains the best supported model

COMMENT 23:

Regarding Figure 7, and the authors' reply to my comment 14 in the previous review: I appreciate the authors' reply that the seeming outlying individual should not be excluded from the final analysis as there was no difference of treatment that would justify it. I 100% agree with that. However, this was not the intent of my comment. It was about model diagnostics (which usually include leverage checks), and suggesting a form of leave one out cross validation. If removing that point (or any other, for that matter) does change the conclusion, then confidence in the main analysis (which should include all points, indeed) must be adjusted accordingly. Indeed, you can imagine the converse: if removing a single individual can change things that much, what would have happened if you had done just one more bird, as there was scope to do in the preregistration? So to clarify, my suggestion is:

- 1/ to keep the analysis including all points as the main/actual analysis used and discussed;
- 2/ nonetheless to re-do the analysis without that point (without presenting it in full; just as we don't usually present all residual normality/heteroscedasticity checks in papers);
- 3/ and to have in mind the outcome of 2/, when discussing the results of 1/

COMMENT 24:

Figure 8 and equivalent figures in appendices: The way the figure is flattened is not flattering. I would suggest to increase the height of each subpanel and/or to reduce their width. Also do not feel you need to have all subpanels on the same column.

COMMENT 25:

Line 803: "were", not "weAre". Please check carefully through the manuscript for typos

COMMENT 26:

Table SM2: all or part of this table may be more helpful to the reader as one or several figures?

Reviewed by Aliza le Roux, 03 October 2022

Download the review

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.32942/osf.io/5z8xs
Version of the preprint: 1

Authors' reply, 15 August 2022

— Cover letter for ALL 3 ARTICLES —

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

We are grateful for your comments on our post-study manuscript - they were so wonderfully useful! We cannot overstate this enough. We took your advice and split the article into three separate articles, which is the reason for the delay in our response.

We placed all three articles in one google doc and show all of the additions and changes in track changes. This is an editable document so feel free to leave comments and suggest changes there if it makes it easier for you.
https://docs.google.com/document/d/1pp-olcx0e2-4N-cANo9MV0bxqLKpWpZuU7Y_3ntssJ4/edit?usp=sharing

We have one author response document for all three articles where we detail what changes we made in which article as a response to which comment.
https://docs.google.com/document/d/1wp2YkjV191VROD--7CHiosyiVtQgwuSjbG5SgNyj_DM/edit?usp=sharing (and also at PCI Ecology)

For your convenience, here are links to the relevant documents:

1) The pre-study preregistration that passed peer review at PCI Ecology https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanipPassedPreStudyPeerReview26Mar2019.pdf

2) The post-study manuscripts

Article 1: (Version 2) https://doi.org/10.32942/osf.io/5z8xs (or maybe you prefer the html version: http://corinalogan.com/Preregistrations/g_flexmanip2.html)

Article 2: (Version 1) https://ecoevorxiv.org/kevqp/ (html version: http://corinalogan.com/Preregistrations/g_flexmanip2rep.html)

Article 3: (Version 1) https://doi.org/10.32942/osf.io/4ycps (html version: http://corinalogan.com/Preregistrations/g_flexmanip2post.html)

3) The version-tracked post-study manuscripts are also available in rmarkdown at GitHub:

Article 1: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip.Rmd.

Article 2: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip2rep.Rmd

Article 3: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/g_flexmanip2post.Rmd

In case you want to see the history of track changes for these documents at GitHub, click the link and then click the "History" button on the right near the top. From there, you can scroll through our comments on what was changed for each save event and, if you want to see exactly what was changed, click on the text that describes the change and it will show you the text that was replaced (in red) next to the new text (in green).

NOTE that in Article 2, as a result of one of the deviations to the preregistration we added in this revision (Deviation #6 where we realized we mis-specified the model for analyzing individual consistency in performance across contexts), the results for H3b have changed from the previously submitted version.

Photo credit goes to Corina Logan (CC-BY 4.0).

Please let us know if you have any questions or need further information. Many thanks for your attention!

All our best,

Corina, Kelsey, and Dieter (on behalf of all co-authors)

– Response from authors —

Recommender

Comment C1: Dear Corina Logan and co-authors,

The two reviewers who evaluated the first version of your pre-registration "Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context?" evaluated the preprint of the corresponding post-study, titled "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context" that you submitted for recommendation to PCI Ecology.

They both highlighted the fact that this preprint is, in its current form, very difficult to read, and I agree with them. It is a very dense paper, which combines the results of several complex experiments, and assumes the reader is familiar with the theoretical and technical background required to understand the results. As

a consequence, it is currently almost impossible for the reader to get all the outcomes of this otherwise interesting study. I hence recommend you follow the detailed suggestions and requests provided by Maxime Dahirel to make your ms easier to read (including result presentation), and also the more general comments provided by Andrea Griffin.

Response C1: We really cannot overstate how grateful we are for this feedback! It was so incredibly useful! We took this advice and divided the article into three separate articles, all resubmitted at the same time so you and the reviewers can see all at once how we revised and addressed changes.

Comment C2: Another important point raised by the reviewers is the need for more theoretical and technical background, in the introduction, and in the discussion too. And last, the need to make sample sizes more visible in the preprint, and to discuss their potential limits.

Response C2: We really appreciate the wonderful comments and we revised accordingly. Please see the specific points below for details.

Comment C3: M. Dahirel's 9th point highlights a problem in the statistical models that, unfortunately, I had not detected in the preregistration when I recommended the last version. Given the importance of this point, I am afraid I have to ask you to take it into account. I know this is not the type of request one is supposed to get on a post-study and am sorry about that. I guess this experience will be instructive for future pre-registration handlings.

Response C3: We don't have any problem updating the analysis as suggested. We did so and the results did not change (see Response D9 below for details).

Comment C4: Finally, I want to join Andrea Griffin in acknowledging the substantial amount of work the experiments described in the preprint must have required and hope you can take into account the reviewers' comments, so that this preprint can then be recommend by PCI Ecology.

Best,

Aurélie Coulon

Response C4: Thank you very much!

Reviewer 1 Maxime Dahirel

I have read the manuscript entitled "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context" by Logan et al., following up on the initial preregistration, which was recommended in principle in PCI Ecology.

I reviewed the initial preregistration with a lot of interest, so was happy to see the final paper arrive. My opinion of the submitted manuscript is mixed. I acknowledge that the authors are breaking new grounds in terms of publication procedure for our disciplines here, and that growing pains are inevitable, but there are substantial issues in the submitted manuscript that could definitely have been avoided with more polishing.

The initial preregistration pointed out that the results might need to be split in several reports. That suggestion was abandoned along the way; given the density of the present manuscript and how it is often very hard to follow, I am not convinced it was a good idea (the reference list starts line 1380!).

The very large quantity of post-hoc non-preregistered analyses adds to that difficulty. In addition, it is also hard to determine throughout the results when the authors meant to analyse their data quantitatively, qualitatively, or did switch to qualitative description (without stats) because there was not enough data to substantiate quantitative analyses. This is not a slight against qualitative or descriptive analyses, which are valid, but choosing to apply them needs to be intentional.

Please find below more specific comments below, which may help improve the manuscript, while hopefully still staying within the bounds of acceptable changes to make on an already recommended- in-principle manuscript.

MAJOR COMMENTS

COMMENT D1: The introduction (excluding the Hypotheses) is far too short, especially coming just after an Abstract that is too long (see minor comments). It does not provide nearly enough context, except maybe for the readers deeply familiar with this very specific research avenue. Important terms are also not defined:

the authors define what flexibility means in their specific context, but not innovativeness nor problem solving. More generally, a lot of what I would expect from an Introduction is missing, and I truly feel the entire 1st paragraph can be substantially expanded and split along the following lines, each with its own substantiated paragraph:

- What is behavioural flexibility and why is it important, in general and in the context of range expansions? (but see comment 2)
- how is it expected to relate to innovativeness and why/ why is it important?
- is there consistent or mixed evidence for this link?
- What can we do to improve our understanding? Are manipulative experiments useful and feasible?

Response D1: Thank you very much for your advice to break up the article into separate articles. We made a separate article specifically for the posthoc results so now they won't get confused with the results from the preregistered analyses in the other manuscripts. Breaking the article into three articles also meant that it was much easier to have shorter abstracts, which we implemented.

Article 1: We expanded the introduction according to your points (though not every point has its own paragraph - we think it works better for flow the way we revised) and we agree that it is much easier to follow the logic now. We added definitions for terms, and made it clearer where there were qualitative analyses and why in the two hypotheses that involved qualitative analyses (P1 and P4). For P1, we inserted a reference in the text to the table where the quantitative analysis is shown. We hadn't previously run an analysis for reversal 1 to quantitatively back up our statement that the control and manipulated individuals passed in similar trials, but there is no reason why not to add an unregistered analysis here to provide empirical evidence for this statement, which we have now done. Thank you for this feedback! We added the following to the RR:

Introduction:

"The behavioral flexibility (hereafter referred to as flexibility) of individuals is considered an important trait that facilitates the capacity for learning, which is then associated with problem solving ability (applying what one has learned about the world to then attempt to access a resource that is not readily accessible) [see review in @lea2020behavioral]. It is hypothesized that, through flexibility, individuals can increase the diversity of their behaviors either via asocial learning (innovativeness) or social learning, leading to the establishment of the population in a new area [@wright2010behavioral]."

"It is predicted that flexibility should positively relate with innovativeness, the ability to create a new behavior or use an existing behavior in a new situation [@griffin2014innovation]."

"or problem solving - a type of experimental assay that does not necessarily require innovativeness to solve, e.g., the ability to solve tasks using pre-trained behaviors; @griffin2014innovation"

"Manipulative experiments go beyond correlations to infer a cause and effect relationship between the manipulated variable and the variable(s) measured after the manipulation [@hernan2006instruments]. A manipulative experiment combined with the random assignment of subjects to a condition (manipulated group or control group), eliminates many confounds associated with internal and external variation (for example, season, motivation, sex, and so on). Such manipulative experiments in behavioral ecology have primarily been conducted in laboratory settings because of the increased feasibility, however such experiments are now also being conducted in wild settings [@aplin2015experimentally]."

Methods > P4: "Analysis 1 (qualitative) … We used the criterion for the epsilon-first strategy of learning the correct choice after one trial and then choosing correctly thereafter. Other patterns were classified as the epsilon-decreasing strategy. This method of qualitative inspection of learning curves is standard for this type

of learning strategy assessment [@mcinerney2010]. The variable for visual inspection was the proportion of correct choices in a non-overlapping sliding window of 4-trial bins across the total number of trials required to reach the criterion of 17/20 correct choices per individual."

Methods > P4: "Analysis 2 (quantitative): We then quantitatively determined to what degree each bird used the exploration versus exploitation strategy using methods in [@federspiel2017adjusting]..."

Results > P4: "Analysis 1 (qualitative) ..."

Results > P4: "Analysis 2 (quantitative) ..."

Results > P1: "The birds in the manipulated group required a similar number of trials during their first reversal (R1 median=75 trials) as the birds in the control group needed during their first and only reversal (R1 median=70 trials) (see unregistered analysis in Table 1)."

COMMENT D2: Continuing in the Introduction from comment 1, there is, starting line 80, a slightly jarring focus shift from expanding populations to at-risk populations. If you frame your Introduction in terms of expanding populations, saying your results may be applicable to at-risk populations does not seem appropriate, at least not without providing more context than is currently provided to substantiate the analogy/transfer. Or, the entire Introduction needs to be framed more generally in terms of exposure to new environments, rather than just range expansions (including e.g. the role of behavioural flexibility in urban environments, for which there exists recent literature, based on quick Scholar and WoS search for the terms "behavioural flexibility" and "urban")

Response D2: This is a great point - thank you so much for bringing this up. We can see now that the logic in this paragraph was not laid out clearly. We now focus the paragraph on range expansions and omit the piece about at-risk populations. The paragraph now reads:

Introduction: "If grackle flexibility is manipulatable using serial reversals, this would provide us with a useful tool for investigating the relationship between flexibility and any number of other variables implicated in geographic range expansions. It would provide researchers with a way to examine the direct links between, for example, flexibility and exploration, to determine whether they are connected and in which direction, which could provide insights into how populations establish in a new location if cross-population manipulations were conducted. If the flexibility manipulation is not successful, this could indicate either that we did not manipulate the right aspect of flexibility (e.g., perhaps training them to solve a variety of different types of tasks quickly would be more effective) or that grackle flexibility is not a trait that is trainable."

COMMENT D3: The manuscript could be more appealing if it included examples pictures of the experimental devices/animals as Figure 1. More generally, I acknowledge and appreciate that the authors made a substantial amount of ancillary data and information available online. But we cannot expect the reader to have to go through Youtube, OSF and other external sites to get a reasonable mental image of what was done (they may want to, to get more details, they shouldn't *have to*). As much as possible, details that would be written/illustrated in the main text in a "traditional" paper (read: non PCI/non pre-reg) should be present in the main text here (see comments below for the converse: that some details should be kept to appendices/ancillary materials).

Response D3: Good idea. Article 1: We added a new figure (Figure 2) that shows pictures of the reversal tube and touchscreen experiments and both of the multi-access boxes. Article 2: We added the new Figure 1 to show pictures of the apparatuses.

COMMENT D4: The current Figure 1 is very helpful to sum up the hypotheses, but it has some issues that make it hard to follow and a bit messy:

- Please keep all panels (i.e. A/B/C/D) at least the same width, ideally the same width and height
- Please keep font size consistent (always the same size for the same-level items, and always larger size for top-level titles than low-level captions. See eg how the blue text in panel A differ from all other blue texts, and is as big as the titles)
- Maybe put frames around each panel
- This one is a bit trickier because I don't have an answer to propose, but: panel D mentions convergence among individuals but only shows the time series from 1 individual. If possible, find a way to illustrate convergence by showing several individuals as in B and C
- Cf Comments 3 and 5, the illustration for the multiaccess box is hard to read in the absence of photographs, especially given the Methods text is at the end of the manuscript.

Response D4: Thanks for this feedback.

Article 1: We modified figure 1 to make all panels the same height, and A and D (now C) have the same width while B has double the width. We put frames around each panel and standardized font sizes and colors. We added a multi-individual comparison for panel D (now C). The methods text is now moved up above the results so this should help clarify how the multi-access box works.

Article 2: We made Figure 1 to show clear pictures of all of the apparatuses and we arranged the panels in a symmetrical way.

COMMENT D5: Is there any rationale for putting the Methods in the end? This cannot be to avoid changing the pre- registered text, since the hypotheses were split from the Methods
An Introduction> Hypotheses>Methods>Deviations From Pre-Reg>Results>Discussion structure would be much better (others are possible, for instance having hypotheses>methods>deviations>results for H1, before moving to the same for H2,H3,H4). In any case, it does not make any sense to put the ways in which what was actually done deviates from the pre-reg *before* the text describing what was actually done.

Response D5: Thank you for this feedback! Our rationale for putting the preregistration (with the hypotheses moved up above results) in a section called Methods at the end of the article was to try to keep clear what was the preregistration and what happened afterward. However, it does seem to be more confusing to readers than anything and it also means that we would need to do more summarizing of the methods in the main article, thus adding more text. We like the structure you propose and we have now implemented it in Articles 1 and 2 (Introduction>Hypotheses>Methods>Deviations From Pre-Reg>Results>Discussion). Article 3 is the article for the posthoc analyses so it wasn't part of the preregistration and just has an Introduction>Research questions>Methods>Results>Discussion.

COMMENT D6: The Hypotheses paragraphs need additional formatting to be clearer. As they are currently written and formatted, it is very hard to distinguish between hypotheses and predictions.
I suggest to either use bullet lists, or to put hypotheses and predictions in a table/ box where the hierarchical structure would be more obvious.
Please also update tense, here and throughout the manuscript. The hypotheses paragraphs and the Methods as presented are in the future tense, and for instance mention things that *will* be done *in 2019* which is clearly a holdover from the preregistration. While I acknowledge that accepted text from the prereg should be carried over to the final manuscript as much as possible without changes, common sense should prevail and anyway tense is not part of this: the guidelines of PCI RR, for instance, state that changes from future to past tense to reflect the fact the experiment is now done are perfectly OK.

Response D6: All articles: Thank you for these suggestions. We updated the formatting so the predictions are bulleted lists (Articles 1 and 3) or incorporated into the main text (Article 2). We also updated the tense from future to present or past throughout the manuscripts.

COMMENT D7: Substantial amounts of results are provided in-text that would be better as Appendices. I am thinking, for instance, of Tables 1,4,6,7,8,9, Fig 6, among many other elements. Many details of the Methods,

especially the unregistered, post-hoc methods, may be more suitable as appendices too. Generally speaking, the density of text and the ways results are structured do not make it easy to grasp any of the results (see comment 5 for possible suggestions). As a result, I have to confess I may have missed some possible comments to make on Results from P2 to P4

In addition, preregistered and unregistered models are lumped together in model tables without ways to distinguish them, which only adds to existing confusion. Generally speaking, I would advise to clearly separate preregistered from non-registered analyses throughout, which is not the case here

Response D7: Article 1: Thank you for your detailed suggestions! The posthoc methods and results are now their own article along with the previously numbered tables 7 and 8 (we don't see a table 9) (Article 3), and the previously numbered table 6 is now in a separate article (Article 2). We moved Tables 1 and 4 to Supplementary Material Tables SM5 and SM3, respectively. We additionally moved the unregistered Interobserver Reliability analyses, the section on Ability to detect actual effects, and details on the touchscreen reversal experiment to Supplementary Material. For Figure 6, we split it into one panel per reversal per bird and moved these to Supplementary Material 6. We kept one bird's panels in the main text for illustrative purposes. We then described in the text how we distinguished between the two strategies (see Response D18 for more details). We also clarified which analyses were unregistered by starting those paragraphs with "Unregistered analysis".

COMMENT D8: Unless I missed something, the actual sample sizes (in number of birds) is substantially lower than the preregistered one. While this is perfectly understandable due to the vagaries of field and experimental work, this is something that should definitely be mentioned in the final text. But, again unless I missed something, it is not. This is especially critical since the preregistered sample size was already low.

Response D8: Article 1: Thank you for pointing this out! We see now that we were very unclear about what happened with the minimum sample size calculation. We originally planned to test approximately 32 individuals (with half in the control group and half in the manipulated group), however we also originally planned to determine the minimum sample size using simulations from bespoke Bayesian models following Richard McElreath's Rethinking book and stats course. The Bayesian models happened after In Principle Recommendation of the preregistration because they took a long time to learn. We were able to develop these models and, according to the simulations, the minimum sample size is 15. We determined this before we stopped collecting data at the Arizona site (the site whose data is in this article). We ended up testing 20 individuals (11 control, 9 manipulated but only 8 actually passed the manipulation). We feel confident in this number because it is also the minimum sample size for four out of five analyses in a different article, which compares reversal learning, multi-access box, exploration, and persistence in grackles across populations (Logan et al. 2020 http://corinalogan.com/Preregistrations/gxpopbehaviorhabitat.html#Sample_size_rationale). We clarified the text:

Methods > Planned Sample: "Deviation from the plan: we were able to test a total of 20 individuals: 11 in the control condition and 9 in the manipulation condition. This met our minimum sample size criterion (see next section)."

Methods > *Data collection stopping rule: "We stopped testing birds after we completed two full aviary seasons because the sample size was above the minimum suggested boundary of 15 (to detect a medium effect size) based on model simulations (see Supplementary Material 1)"

COMMENT D9: Regarding modelling choices: Here we arrive to a particularly awkward point, since I am going to go explicitly *against* the preregistration as accepted. First, I remind you that I was only invited to review the initial pre-registration, not any revised version, or I would have made these comments at the time (the guidances I am going to follow and mention already existed).

- Regarding H1/P1, I was the one who pointed the need to include individual random effects, but since I did not review the revisions, could not check this was done properly. If the analyses done reflect the preregistration, this probably was not. The model in the pre-reg is TrialsToReverse ~ ReverseNumber + (1 | ID) + (1 | Batch).

This model assumes that individual only differ in their baseline/intercept number of TrialsTo Reverse, where there is every reason to expect they differ in the effect of the number of Reversals. The better model structure would be TrialsToReverse ~ ReverseNumber + (ReverseNumber | ID) + (1 | Batch) (O'Dea et al., 2022; Schielzeth & Forstmeier, 2009). As the title of the Schielzeth & Forstmeier reminds us, failure to account for this source of non-independence may lead to overconfident predictions. Given the width of the credible intervals Table 2, and your limited sample size (in number of individuals), I wouldn't be surprised if an updated analysis came back with no fixed effect of ReverseNumber. (as an aside: given the limited number of individuals, I am also not sure the Batch effect can be estimated meaningfully without informed/informative priors). Importantly, this remark about random effects apply to all models in the manuscript in which there repeated individual measurements and a "time" fixed effect (eg number of trials);

- There are hypotheses for which model dredging is used in the final prereg, for which it wasn't proposed in the original prereg I reviewed, and for which I would have recommended against it have I had the opportunity. There has been substantial discussion and much debate about the use of dredging and all-subset multimodel inference, going back at to the original Burnham and Anderson book (Burnham & Anderson, 2002; Grueber et al., 2011). Independently on one's opinion on the topic, one must at least be aware that trying to fit many models to a small size dataset is bound to yield spurious results at some points, and discuss that possibility if going ahead anyway

Response D9: Article 1: We see your point and we don't have a problem nesting ID in reversal number as a random effect because it is something you would have suggested before we conducted the study. We did this and the result did not change: there is still a significant negative correlation between the number of trials to reverse and reversal number. We replaced the results in Table 2 with the results from this nested analysis (Results > P1 > Table 2, and r chunk: p1table). There are no other analyses in Article 1 where the response variable has repeated measures.

We ran a model without batch and the results were the same (Results > P1 > r chunk: p1table), so we decided that it doesn't hurt to keep it in given that that was the preregistered analysis. We agree that batch is an issue for an additional reason that we describe for P2 in Deviations from the preregistration:

"The original model for P2 (Table SM3: Model 1) included the covariate aviary batch, however this ended up confounding the analysis because control and manipulated individuals, while randomly assigned to these conditions, ended up in particular batches as a result of their willingness to participate in tests offered during their time in the aviary (Table SM3: Model 3). Several grackles never passed habituation or training such that their first experiment could begin, therefore we replaced these grackles in the aviaries with others who were willing to participate. This means that batch did not indicate a particular temporal period. Therefore, we **removed batch from the model**."

Additionally, we had already added a posthoc analysis that assumed that individuals do not only differ in their intercept, but also in the effect of the number of reversals to compare the number of trials needed during the first and the last reversal. We moved this text from the separate posthoc section (that came at the end of the Results section) to the P1 results section to make this clearer (Results > P1 > Unregistered analysis: A pooled...).

Regarding the dredging to compare models, in the preregistration, the Akaike weight comparison was in the R code, but not explained in the text. We don't quite know why this was in the code and we don't think it adds anything to the assessment of P1 so we deleted it.

Article 2: We used multiple measures for individuals to determine how likely it was that performance on these tasks represented consistent individual differences in the degree of behavioral flexibility. Therefore, in our models, the random effect (random intercept) is included not to account for the effect of pseudoreplication of data points on the estimate of fixed effects, but rather to estimate repeatability of performance through

partitioning of among vs between individual variance. In these models we are not interpreting the fixed effects, but they are included to control for a potentially confounding factor when calculating the repeatability. Including a random slope, as you suggest, would test a different question that we were not interested in - Was there variation in behavior change across tests (or, did individuals habituate/learn at different rates across reversals and contexts; Reale et al. 2007, Nakagawa & Schielzeth 2010)?

COMMENT D10: I have no problem in principle with non-standard width credible/confidence/prediction... intervals, but the choice must be principled and consistent: here the authors switch between 89, 95 and 97% intervals between paragraphs, figures and tables without explanation. Please stick to a single interval width throughout the entire paper, and/or provide clear rationale for why a narrower/wide interval width is warranted.

In addition, please check throughout for the correct use of prediction and credible intervals. "Prediction interval" has a precise meaning, which is different from any of the meanings of "credible"/"compatibility"/"confidence" intervals. In particular, model parameters don't have prediction intervals (since they refer to *data*), but you describe their intervals as "prediction intervals". Please do not assume this is just a text problem, and check throughout code that you used the intended and correct interval(s) each time

Response D10: Article 1: Our only rationale is that we used the default intervals that went with each R function. We have now updated all intervals to 89% (because these are what we used for our bespoke Bayesian models) and we labeled them appropriately (Methods > P2 > Unregistered analysis; Results > P1 > Table 2 & Table 3; Results > P2 > Figure 7).

COMMENT D11: Lines 277-283: given the high r values, the low p values (even if p>0.05), and the very low df (especially with switching), saying that the tests are **not** correlated is a very strong conclusion and in my opinion wholly unwarranted. I would not be surprised at all that the tests are actually correlated and that there is just not enough power to detect this. Please clarify your choice to analyse separately (or if you decide to reverse your choice, to analyse them together) in the light of this

This feeds back to a broader point about the manuscript: please remember that your sample size (in terms of individuals) is low. This does not means your results are meaningless, but you must be extremely careful in your interpretations not to overreach

Response D11: Article 1: Thanks for catching this! We had stated before conducting the multiaccess box wooden experiment that we would analyze the data from both MABs separately if they were not correlated with each other. While the latency to switch measure did have a higher R2 value, we needed to use a standardized way of determining whether this was enough or not, so we chose the 0.05 p-value threshold. We can't say for sure whether the wooden and plastic multi-access boxes are the same, which is why it is better to keep them in separate analyses. If they are the same, we would run into issues of pseudoreplication when combining the two measures. If they are different, we would lose information that might exist in the variation. Please see our Response D8 showing that we met our minimum sample size. We changed the wording to:

Methods > P2:

"We found that they did not *statistically significantly* correlate with each other on either variable measured".

"Therefore, while the performance on the two multi-access boxes might not be completely independent as indicated by the high r values, the two boxes appear not to be completely interchangeable either as indicated by the lack of statistical significance and high uncertainty in the r values. We therefore analyzed the plastic and wooden multi-access boxes separately."

COMMENT D12: Line 293, you write: "the Akaike weight of the full model was 0.94, which means that including condition in the model explains the bulk of the variation in the number of trials to reverse in the last reversal" No, this is categorically and emphatically not what it means.

Having a model with a high Akaike weight simply means that this model is the best of the candidate models

(for one definition of "best"), but to put it simply, the best of a set of bad things can still be bad. A model can perfectly explain only 5% of the total variance (so not "the bulk" by any sense of the word) and get an Akaike weight >0.9, if all the other models in the chosen set perform way worse.

Please, please rewrite this statement and **all** similar ones to remove all statements, explicit and implicit, that high Akaike weight = high variance explained. Then, if you want to actually estimate the quantity of variance explained by a model, use R-squared, any of its extensions and related metrics (Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013; O'Dea et al., 2022).

Response D12: Article 1: We agree and have now removed this Akaike weight comparison because it was extra and didn't directly relate to the questions (Results > P1 > Unregistered analysis). We did keep the Akaike weight comparison in for Results > P2 > Alternative 2 because we specified this in the preregistration and we made sure to use appropriate language when interpreting the results (we only refer to the model comparison rather than to the amount of variance explained).

Results > P2 > Alternative 2: "Adding the number of motor actions (wooden: average=13, sd=4) did not improve the model fit when examining the relationship between the latency to switch loci on the wooden multi-access box (wooden: average=463, sd=481) and the number of trials to reverse in the last reversal (wooden: average=60, sd=38) because the Akaike weights were similar for both models"

COMMENT D13: Line 291: "There was additionally a difference between manipulated and control reversal speeds when comparing their last reversals". You are making a strong statement that is not supported by any model or test here, and may not even be warranted from the data: sure, the difference figure 2 looks strong, but it is still ultimately only 20 birds total. I strongly advise to strongly tone down that statement, whether or not you add a statistical analysis to quantify the difference (I would add one, a Poisson/negative binomial GLM).

Between this and comment 9, I am wondering if the authors do not overstep when claiming without nuance "The flexibility manipulation worked" in the title of one of the parts of the Discussion.

Response D13: Article 1: We removed the Akaike weight comparison analysis on which this original statement was based (per Response D12) and we conducted this analysis as you suggested (Results > P1 > Unregistered analysis). This new analysis still shows that the manipulated group solved in significantly fewer trials than the control group in the last reversals. We believe that our conclusion that the flexibility manipulation worked is warranted given this response and our Responses D8, D9, and D11.

Additionally, passing the manipulation condition was unlikely to happen by chance. Randomly reaching 50 trials in a given reversal is likely, but not in two consecutive reversals unless the individuals are able to do it on purpose, which is why we used the two consecutive reversals in 50 trials or less criterion. If they randomly pass two consecutive reversals in 50 trials, we would expect to see more variation in the number of reversals it took to reach this criterion. However, eight out of nine birds passed in six to eight reversals, indicating that there is some consistency in how birds reach this criterion.

COMMENT D14: Figure 5: First, why are the values on the x-axis different between the two *rows* (I get that the columns reflect different groups)? Shouldn't they be the same because the individuals are the same? If not, then this again reflects that the manuscript is unclear in places, and should be updated to make the explanation clearer.

Second: in most of these subpanels there is clearly an outlier individual with much higher values on the x-axis than others. Given your low N, have you tested whether your conclusions are robust to the removal of that individual?

Response D14: Article 1: For the first point, the top row is the average latency to switch between loci, which may not have any data for some birds if they solved 0-1 loci and didn't attempt another option afterward. This was the case for Taquito (160 trials to reverse, solved 0 loci on the MAB plastic). This resulted in an NA being entered for latency. For the analyses, we exclude NAs from the data set, which means that there can be different individuals in the top row versus the bottom row, which is the total number of loci solved. For Taquito,

solving 0 loci meant that he showed up in the bottom row for the MAB plastic, which is why that figure's x-axis goes up to 160 rather than the figure above for latency, which goes up to 100. We made sure to note the sample size per plot in Figure 7 (previously Figure 5) in the legend and in the text to help clarify this.

For the second point, we explicitly stated in the preregistration that we would not exclude any data (Analysis Plan) and this individual, Taquito, who was in the control group and passed reversal learning in 160 trials, is within the normal passing range for grackles. At each of the three sites we have conducted reversal learning tests on grackles, at least one individual has taken around 150 trials to pass, therefore we consider this a normal amount of trials to pass for this species (Santa Barbara: Logan 2016; Sacramento: Logan et al. in prep.). Additionally, there was nothing different about the way Taquito behaved or was treated during the experiment, which indicates that his behavior was normal for a grackle and there were no confounds due to the test procedure. We specifically want to capture individual variation in these analyses so it is important to keep Taquito in the data set.

COMMENT D15: Tables3,5,6: Please process your tables to be presentable for a manuscript (see also comment 6):

- Abbreviations must be explained
- update your md/latex code so that text stays in its cell and does not spill over neighbouring cells cell
- Table 6 spills out of the page and not all columns are fully included
- And notation conventions that are internal to R code must be removed. Here for instance, the parentheses around "intercept", the persistence of what are clearly R-specific column names "d$reversalstopass", show that model outputs have barely been formatted for publication. If you intend to produce tables directly from an Rmarkdown code without external processing (which is both laudable and doable), please be aware that it is perfectly possible to produce publication-level quality outputs, look at the options of the kable function in the knitr package , and the kableextra package, among others.

Response D15: Article 1: We have had so much trouble with tables in RMarkdown. We have tried a variety of packages to get the formatting to work, but often we are not able to get it exactly the way we want it. We have resorted to making a csv file for each table and using KableExtra to format the table from the csv rather than directly from the model output. This allowed us to remove the R conventions. Table 3 ended up being deleted and we fixed Table 5 (now labeled Table 4).

Article 2: Table 1 (previously Table 6 when articles were combined) now contains only the results from the two repeatability models, formatted in the same way as the tables in Article 1.

COMMENT D16a: Please correct me if I'm wrong (but again, bear in mind previous comments re: clarity), but the analyses on the idea that intermediate birds could perform worse are all unregistered and post-hoc? And also provided without theoretical justification (biological or statistical?).
I am quite skeptical of both their usefulness and their validity.

Response D16a: Article 3: You are correct in that the ideas around the intermediate birds were all post-hoc. The justification was that we had observed correlations between the predictor variables phi and lambda that we thought could lead to trade-offs. We have now made it clearer that these are exploratory analyses, and moved their interpretation to the discussion.

Article 3 > Research questions: "However, in the serial reversals, birds need to be able to quickly learn the new reward location and also be ready to explore the other option. Accordingly, birds might end up with one of two solutions: they might adopt a strategy of weighting recent information more heavily while also showing low exploration, or they might show high exploration while being slow at updating their attractions…We predict that birds that are more flexible, presumably those who have a high phi (faster learning rate), have shorter latencies to attempt a new locus and solve more loci on the two multi-access boxes. Given that birds might use

different strategies to be flexible (see prediction 3), we also explore whether the relationship between ϕ or λ and the performance on the multi-access boxes is non-linear."

COMMENT D16b: First: "For the manipulated birds, we found that during their last reversal there was a positive correlation between and , with individuals with higher values also showing higher values". What is the credible/confidence interval around that correlation? What is its value? Note that in a Bayesian model, it is perfectly possible, and I would recommend, to directly estimate that correlation and its interval while fitting the model.

Response D16b: Article 3: We have now added the estimate of the association between phi and lambda to the results. While doing this, we noticed that we had made a mistake in reporting this association. We mentioned this association twice, first correctly as a negative correlation and the second time (as quoted by you here) incorrectly as a positive correlation because we had misremembered that larger lambda values mean less deviation. We now corrected this.

Article 3 > Results: "The manipulation changed both ϕ and λ, such that, across all birds, there was a negative correlation between ϕ and λ (mean estimate -0.39, 89% compatibility interval: -0.72 to -0.06). … For the manipulated birds, we found that during their last reversal there was a negative correlation between ϕ and λ, with individuals with higher ϕ values showing lower λ values. This negative correlation could lead to worse performance on the multi-access boxes for birds with intermediate values. Exploration of our data shows that, for the number of loci solved on both the plastic and the wooden multi-access boxes, there is a U-shaped association…"

COMMENT D16c: Second: assuming the above holds, and the U relationship is plausible, I am sorry but I don't see any evidence for these U relationships from Figure 10, and I don't see anything that is distinguishable from noise. I suspect the fitting of an appropriate regression model, with e.g. a quadratic effect, would confirm my visual intuition

So, please provide stronger support for this U-shaped relationship, or remove all mention of it throughout the manuscript

Response D16c: Article 3: We had previously reported the results of analyses that tested for the U-shaped relationship by transforming the phi and lambda values. We now explain this in the methods section. The results of these analyses are reported in Article 3 Table 2. In addition, we now added analyses that include a quadratic effect in the correlation. These gave identical results to the previous analyses, again suggesting that birds with intermediate values of λ solved fewer loci on both multi-access boxes, and birds with intermediate values of ϕ having shorter latencies to attempt a new locus.

Article 3 > Methods: "To represent the potential U-shaped relationship which assumes that birds with intermediate phi and lambda values perform differently, we first transformed phi and lambda to calculate for each individual how far their value is from the median. Second, we ran the models squaring phi and lambda. Both approaches gave the same results, and we only report the estimates from the models with the transformed values."

COMMENT D17: Since the analyses using lambda and phi are all unregistered, I would strongly recommend to describe into more detail the theoretical rationale. Especially since quantatitively minded people will recognize those parameters as the two parameters of a negative-binomial (aka Gamma-Poisson) distribution, but making the link from this to the biological interpretation ("learning rate of attraction to either option and the rate of deviating from learned attractions") is not trivial, and new (you cite a 2021 paper), so you should not expect readers to automatically "get it" and ideally you should not expect them to go to another paper for such important details.

Response D17: Article 3: Thanks for pointing this out! As a result of making the post-hoc analyses a separate article, we now have a much longer introduction for this piece where we describe the background in much more detail. While the two parameters of lambda and phi have a similar connotation to the parameters in a negative-binomial distribution, they reflect different processes. We now provide the equations that describe how these two parameters are estimated from the data as follows:

Methods > Bayesian reinforcement learning model: "We used the version of the Bayesian model that was developed by @blaisdell2021causal and modified by @logan2020xpop [see their Analysis Plan > "Flexibility analysis" for model specifications and validation]. This model uses data from every trial of reversal learning (rather than only using the total number of trials to pass criterion) and represents behavioral flexibility using two parameters: the learning rate of attraction to either option (ϕ) and the rate of deviating from learned attractions (λ). The model repeatedly estimates the series of choices each bird made, based on two equations

Equation 1 (attraction and □): Ai,j,t+1=(1−□j)Ai,j,t+□jπi,j,t

Equation 1 tells us how attractions to different behavioral options Ai,j,t+1 (i.e., how preferable option i is to the bird j at time t+1) change over time as a function of previous attractions Ai,j,t and recently experienced payoffs πi,j,t (i.e., whether they received a reward in a given trial or not). Attraction scores thus reflect the accumulated learning history up to this point. The (bird-specific) parameter □j describes the weight of recent experience. The higher the value of □j, the faster the bird updates their attraction.

Equation 2 (choice and λ): P(i)t+1=exp(λjAi,j,t)∑m=exp(λjAm,j,t).

Equation 2 expresses the probability an individual j chooses option i in the next round, t+1, based on the latent attractions. The parameter λj represents the rate of deviating from learned attractions of an individual . It controls how sensitive choices are to differences in attraction scores. As λj gets larger, choices become more deterministic, as it gets smaller, choices become more exploratory (random choice if λj=0).

We implemented the Bayesian reinforcement learning model in the statistical language Stan [cite stan development team], calling the model and analyzing its output in R. The values for phi and lambda for each individual are estimated as the mean from 2000 samples from the posterior."

COMMENT D18: P4, Results: claims made in text are not obvious at all from figure 6, despite what the authors imply. The overlap of many lines of many colours make it extremely hard to see anything. I would suggest to gray out all lines except the ones in interest.

In addition, if there is a clear quantitative criterion to say whether a line correspond to epsilon-first vs increasing, please use it as your base to decide which lines to grey out, mention it in the text and methods, and make that clear. If not and if your evaluation is qualitative, please make that clear in the Results, the Methods, and use the appropriate degree of caution when discussing the Results.

Response D18: Article 1: Please see our Response D7 above for details about how we made the figure clearer. We moved all figures to Supplementary Material 6 except for Burrito's because we now use his figures as an example to show how we visualize the learning curves. We also clarified which analyses were quantitative and qualitative by stating this in the beginning of the appropriate paragraphs of P4 in the Methods and Results. Thank you so much for prompting us to recheck the epsilon-first threshold! We realized in doing so, that this cut off is after the first trial: if an individual chooses incorrectly in the first trial and then correctly after that, this is the epsilon-first strategy. Other variations are the epsilon-decreasing strategy. According to this criterion, no grackle used the epsilon-first strategy and we updated the Abstract, Results, and Discussion accordingly. We added to the text:

Abstract: “All grackles in the manipulated reversal learning group used one learning strategy (epsilon-decreasing: long exploration period) in all reversals and did not use the (epsilon-first strategy: quickly shift their preference), and none used a particular exploration or exploitation strategy earlier or later in their serial reversals.”

Methods > P4 “Analysis 1 (qualitative): “We used the criterion for the epsilon-first strategy of learning the correct choice after one trial and then choosing correctly thereafter. Other patterns were classified as the epsilon-decreasing strategy. This method of qualitative inspection of learning curves is standard for this type of learning strategy assessment [@mcinerney2010].”

Results > P4: “Analysis 1 (qualitative): Using the criterion for the epsilon-first strategy of learning the correct choice after one trial and then choosing correctly thereafter, no grackle in this study used this strategy in any reversal. All grackles used an epsilon-decreasing strategy in all reversals. We use Burrito’s figures to illustrate the epsilon-decreasing strategy (Figure 8): the proportion of trials he gets correct wanders up and down (epsilon-decreasing) until an asymptote at 0.8 is reached and held.”

Discussion: See Response G12 for details.

COMMENT D19: Where does the graph figure 9 come from? Is it the result of a proper attempt at causal inference, with e.g. a clear a priori DAG, and models written to evaluate the causal claims implied by the DAG? Or is it an attempt to summarise graphically all the results from the previous analyses? If the former, which I doubt given the preregistration, please make that extremely clear, as it is not from the manuscript as it is now

If the latter, I would strongly advise you to drop it, *and* to preface all your mentions of “causal” by “putative” or “hypothetical”. First because you may confuse the reader into believing you did proper causal inference. Second, because some of the relationships drawn in that graph may not be supported by data (see all previous comments). Third, because even if the correlations are supported by data, the only directions of causality that are certain are “manipulation influences the rest of the graph nodes”; to determine the direction of causation between the other nodes, you’d need to have a priori causal hypotheses, which bring us back to the above paragraph.

Response D19: Article 3: Though we had a DAG in mind a priori, we did not include it in the preregistration, as you pointed out. Therefore, we removed it from the manuscript. We now refer to potential pathways that are best supported by our data:

Results > Observed effects of the manipulation on reversal performance, ϕ, and λ: “Results from this simultaneous estimation of the potential pathways shows that our data best support that the ϕ from the initial learning and first reversal link to the number of trials to pass the first reversal, which, in turn, appear associated with how many trials they need to pass their last reversal. The ϕ for the last reversal does not appear to provide any additional information about the number of trials in the last reversal, and λ is not directly associated with the number of trials birds need to reverse (Table 7: Model 20) (Figure 9).”

MINOR COMMENTS

COMMENT D20: Abstract: this feels like an overly long abstract (about 2x longer than is typical). Ideally, lines 29 to 49 can be reduced to a maximum of 3-4 sentences summarising the key results.

Response D20: Article 1: We cut the abstract down to <270 words by taking your advice and separating the article into three separate articles, and then additionally reducing wordiness.

COMMENT D21: line 23: should be “works and predicts”? or “works to predict”? Line 25: I would add the species’ scientific name here

Response D21: Article 1: thanks for catching this! We meant the latter and updated accordingly. We also added the scientific name.

COMMENT D22: Line 56 [link to video summary]: consider archiving a copy of the video to a permanent archive, as the long term persistence of youtube links cannot be expected

Response D22: Article 1: That's a great point, thanks for bringing this up. We additionally uploaded the video to the KNB data repository with our data set.

COMMENT D23: Line 74: I would add a sentence here to explicit what "rapidly expanding its range" means: what is the native range, what is the time scale of the expansion, what is the expanded range? This could also be a map, this could also be as an Appendix (these possibilities are not mutually exclusive)

Response D23: Article 1: Good point. We changed the sentence to say "We focused our study on great-tailed grackles (*Quiscalus mexicanus*, hereafter grackles), a bird species that is flexible [@logan2016behavioral] and, while they are originally from Central America, they have rapidly expanded their geographic range across the US since 1880 [@wehtje2003range; @summers2022xpop]."

COMMENT D24: Line 260-261: Detailed instructions on which specific files to use to reanalyse data are not expected in a manuscript. The best place for such instructions is in a README/tutorial, provided with the DOI-archived code and/or the data, or in comments within the code itself. Please check throughout

Response D24: Article 1: We deleted these details from the manuscript because, you are right, our data package already lists this information.

COMMENT D25: Figure 5: please back-transform the x-axis variables for plotting so that the scales are readable by the readers

Response D25: We did as you suggested (Results > P2 > Figure 7) - note that the new figure number is 7 rather than 5.

Reviewer 2 Andrea Griffin

General comments

COMMENT G1: This study examines flexibility in a range-expanding passerine. Flexibility is operationalized using a decrease in the number of trials to criterion across serial reversals in a colour conditioning task. The fact that grackles learn to reverse faster and faster, that is, that they learn a rule, is not novel. There is a body of old comparative psychology research using the paradigm and showing the effect in a variety of species. However, the study is novel in that it tests whether rule learning generalises to other reversal paradigms (motor tasks) and also the individual repeatability of rule learning. Even though the touch screen task did not work, I commend the authors on the extensive behavioural trials they have undertaken. This is truly impressive and unprecedented to my knowledge.

Response G1: Thank you so much for this praise! It really means a lot coming from you!

COMMENT G2: In general, even as someone familiar with the pre-registration, I found the paper much to dense to read and the results too disjointed from the methods to be accessible. I suggest the manuscript would benefit from a much more synthetic write up where the final methods and ensuing results are presented (for example, the touch screen removed up front) and the information regarding changes to the pre-registration placed in appendix. For example, just explaining what the 'reversal criterion' was next the results would be helpful.

Response G2: Please see the cover letter and Responses D1 and D7 for more information about how we split the article into three separate articles and moved some pieces to Supplementary Material to address the density issue. We also moved the Methods sections to the usual position just under Introduction to help with the flow.

Article 1: Because we split the article into three and moved so many pieces to Supplementary Material, we decided to...

1) keep very few touchscreen details in the main article because it is mentioned in the hypotheses, while moving the rest of the details to Supplementary Material, and

2) keep the Deviations from the preregistration section in the main article. However, we moved it to the

Methods section.

COMMENT G3: Hence, my comments have more to do with the pitch than the detail of the study's findings. First, I think the authors should acknowledge the classic work on reversal learning and what is known about rule learning in the introduction and refer back to it in the discussion. Related to this, in my view, presenting the experimental group as a 'flexibility manipulation' is confusing and should be avoided. Experimental birds were given the opportunity to learn a rule while control birds were not; I think it would make the reading easier to refer to the groups in these terms, while also anchoring the study into a body of existing work via shared terminology.

Response G3: Article 1: Very good point - we revised the introduction to go more in depth into the previous literature on reversal learning, and how it and serial reversals relate to rule learning. We also now refer back to this in the discussion.

Previous research on serial reversal learning has used this experiment to investigate rule learning, however they usually do not use serial reversal learning as an instrument to attempt to modify behavior in a different context. We argue that the latter makes it a manipulation. Reversal learning is commonly thought of as a measure of flexibility, and this is how we consider it in our hypotheses, therefore we termed it a flexibility manipulation, which was included in the pre-study in principle recommendation. We want to keep this language because our next project builds on our findings by implementing flexibility manipulations (serial reversals) in species that are successful in human modified environments and species that are threatened or endangered to determine whether manipulating flexibility improves their success. Our revision per your comment has now resulted in our clarifying the link between reversal learning theory and the flexibility manipulation so our terminology is now better grounded in its connection with the previous literature, as well as highlighting how we are expanding the thinking of this research in new ways. We added the following:

Article 1 > Introduction: "Reversal learning is a common way of measuring flexibility that has been used for many decades across many species, therefore lending itself well to comparative analyses and generalizations [see review in @lea2020behavioral]. In this test, an individual learns to prefer the rewarded option, which differs from the non-rewarded option in color, shape, space, or another obvious feature. Once this initial preference is formed, the previously non-rewarded option becomes the rewarded option and vice versa, and the preference is reversed. Individuals who are faster to reverse their preference are considered more flexible - better able to change their behavior when the circumstances change. Serial reversal learning involves continuing to reverse the preference back and forth to determine whether individuals learn a "win-stay, lose-shift" rule that, when the reward is no longer in the expected option, they should switch to preferring the other option [@spence1936nature; @warren1965primate; @warren1965comparative]. Once this rule is learned, it can then be applied to new contexts and result in improved performance over individuals who have not learned this rule [@warren1965comparative]. We randomly assigned individuals to a manipulated or control condition and used serial reversals (for the manipulated group) to attempt to manipulate flexibility and determine whether the manipulated individuals were then more flexible and more innovative in other contexts."

Article 1 > Discussion: "This indicates that the flexibility manipulation was effective in that it manipulated reversal learning speeds, suggesting that these individuals shifted toward a "win-stay, lose-shift" rule to learn to reverse faster after more experience with reversing. [@spence1936nature; @warren1965primate; @warren1965comparative]. The manipulated individuals who increased their reversal learning speed, were then apparently able to apply this to a new context, which resulted in better performance when compared with control individuals who did not have the opportunity to learn. Previous research has also exploited the fact that most individuals can learn to learn and have used serial reversals to show that such experience usually improves performance when transferring to reversals involving different stimuli (e.g., visual vs. spatial, visual vs. visual in a new combination) [@rayburn2013reversal; @schusterman1962transfer; @warren1965primate;

@warren1966reversal]."

COMMENT G4: Second, I don't think the results are relevant to the fact that the grackle is an invasive and range expanding species. There is no comparison of individuals on the front of an invasion wave with those on the back, for example. Neither is there a comparison of grackles with a non range-expanding species. I suggest it is fine to mention the ecological characteristic of the study system but suggest removing the impression that the study is testing something particular to an invasive species.

Response G4: Article 1: Great point! We changed the language to make it clear that this study only focused on one population of grackles and did not test anything about how flexibility relates to a range expansion.

Introduction: "The first step to improving our understanding of whether and how flexibility relates to innovativeness, and the focus of the current investigation, is to start with one population and perform a manipulative experiment on one of the variables to determine whether there is an associated change in the other. Once this association is known, future research can then investigate whether flexibility and innovativeness are involved in a range expansion."

Discussion > Conclusion: we added qualifying language to indicate that we are referring to "future research", and we indicated that using the manipulation in at risk species is the topic of our next project "This is the focus of our new research program, ManyIndividuals, where we manipulate flexibility using serial reversals in the wild in species that are successful and at risk and determine whether the manipulation improves their success in human modified environments [@Logan2022mi1]"

COMMENT G5: Third, I find the conclusions around conservation applications unrealistic. Captive breeding programmes deal typically with so few individuals that I can't see how selecting them based on the type of measure used here would be possible. The type of animal training implemented here is hugely labour intensive and I cannot see that happening as part of cash-poor conservation programmes.

I think all these pitch issues can be addressed.

Response G5: Article 1: We agree that implementing a flexibility manipulation could be unfeasible for many captive breeding programs. However, we have found a strong interest from conservation managers in implementing a flexibility manipulation like ours to attempt to improve the success of individuals from endangered species in the wild. The training techniques (in captivity and in the wild) they currently invest extensive amounts of time into have not been very successful in changing behavior and they are looking for alternatives. These individuals are often part of captive breeding programs and so are already in captivity and needing enrichment, which these tests provide so it would be a form of multitasking as well. A flexibility manipulation would not need to precisely replicate our experiment, but simply have a setup that leads to a change in the environment which the individuals have to learn (e.g., feed one week at the front of the enclosure, the next week at the back). Some of the co-authors on this article (Logan, McCune, and Lukas) are part of a new project called ManyIndividuals (https://github.com/ManyIndividuals/ManyIndividuals) where two endangered species will undergo a serial reversal flexibility manipulation in the wild to determine whether it improves their success in human modified environments. This project will be able to provide even more information to conservation managers about what they might want to implement in their species and how.

COMMENT G6: A few detailed comments are as follows: L192: please specify see discussion so that the reader knows that you will address what you think it measures

Response G6: Article 1: Great point! We added: "(see Discussion for details about what we think this might measure)"

COMMENT G7: L238: please specify what dependent variable you are referring to

Response G7: Article 2: Sorry for the confusion! We added the text below to clarify:

"The distribution of values for the "number of trials to reverse" response variable in the [P3a analysis](#p3a-repeatable-within-individuals-within-a-context-reversal-learning) was not a good fit for the Poisson distribution because it was overdispersed and heteroscedastic. We log-transformed the data to approximate a normal

distribution and it passed all of the data checks. Therefore, we used a Gaussian distribution for our model, which fits the log-transformed data well. (24 Aug 2021)"

COMMENT G8: L248: I suggest the terminology flexibility manipulation is unclear and should be avoided.

Response G8: Please see Response G3.

COMMENT G9: L277: I understand that the non-interchangeable nature of these two tasks is desirable from the experimental point of view but where does the lack of performance correlation leave the conclusion that flexibility generalises?

Response G9: The original text at line 277 is:

"Because the wooden multi-access box was added after in principle recommendation, we conducted an unregistered analysis to determine whether the plastic and wooden multi-access box results correlated with each other, which would indicate that these tests are interchangeable. We found that they did not correlate with each other on either variable measured: the average latency to attempt a new locus or the total number of loci solved. Therefore, these two tests are not interchangeable and we analyzed them separately"

We are very sorry, but we are not quite sure what "flexibility generalises" refers to so we made a couple of revisions based on our interpretation of what your comment might mean. We think you might be pointing out some of our confusing terminology: we are correlating population-level performance on the two MABs (in Article 1) to determine if the two tests are interchangeable. Whereas, in Article 2, we are interested in whether individual performance is repeatable across tests to indicate that the trait behavioral flexibility is used for tasks that generally involve behavior switching contexts. We hope that now that we separated these two ideas (population-level correlation in performance and individual-level repeatability of performance across tasks) into two separate manuscripts that it is clearer.

COMMENT G10: L290: please remind the reader here what is meant by 'pass the manipulation condition'

Response G10: Article 1: Good idea! We added "which included Memela who did not pass the manipulation condition of passing two consecutive reversals in 50 trials or less"

COMMENT G11: Table 1: please specify what X means. Please remind the reader what "to pass the criterion if they were in the flexibility condition means" and "reversals to pass".

Response G11: Article 1: Thank you for catching this! We added to the legend: "Reversals to pass" indicates how many serial reversals it took a bird to pass criterion (passing two consecutive reversals in 50 trials or less) if they were in the flexibility manipulation condition. X indicates the bird attempted, but did not pass that experiment.

COMMENT G12: Discussion: I recommend the discussion should relate/compare the present findings to existing data on the topic. By referring to serial reversal learning as a 'manipulation of flexibility', and not mentioning classic work on the paradigm in other species, the uninformed reader is likely to over-estimate the novelty of this aspect of the work. Similarly, others worthy of mention have attempted to decompose reversal learning into its various phases including studies in pigeons and common mynas. This previous work could be mentioned and then the progress made by the new approach (lamda and epsilon) explained.

Response G12: Thank you so much for this really useful feedback! We implemented your changes, including discussing the myna research, however we were unable to find research examining the exploration-exploitation learning strategies for reversal learning in pigeons. We would be happy to add this if you pointed us to the article you were thinking of. We added:

Article 1 > Discussion:

"This indicates that the flexibility manipulation was effective in that it manipulated reversal learning speeds, suggesting that these individuals learned a "win-stay, lose-shift" rule to learn to reverse faster after more experience with reversing. [@spence1936nature; @warren1965primate; @warren1965comparative]. The manipulated individuals who learned this rule, were then apparently able to apply this rule to a new context,

which resulted in better performance when compared with control individuals who did not have the opportunity to learn this rule. Previous research has also exploited the fact that most individuals can learn this rule and have used serial reversals to show that such experience usually improves performance when transferring to reversals involving different stimuli (e.g., visual vs. spatial, visual vs. visual in a new combination) [@rayburn2013reversal; @schusterman1962transfer; @warren1965primate; @warren1966reversal]."

"Our results are in contrast with previous research on the correlation between flexibility performance, using serial reversals, and innovation: Indian mynas that were faster to reverse, were slower to innovate [@griffin2013tracking]. However, the @griffin2013tracking investigation was designed to evaluate the correlation between the variables and not whether manipulating flexibility using serial reversals influenced innovativeness. This difference could explain the differing results because correlational research can become noisy if there are unmeasured variables, which is something that a manipulation can help reduce. Other potential reasons for the difference in results could be due to using different experimental designs, and/or different serial reversal passing criteria [@griffin2013tracking used a preset number of reversals that resulted in a maximum of four reversals]."

"None of the flexibility manipulated individuals converged on using an epsilon-first learning strategy (learn the correct choice after one trial) as they progressed through serial reversals. All used the epsilon-decreasing strategy (explore options before forming a preference) throughout their reversals. Additionally, no grackle used a particular exploitation or exploration strategy earlier or later in their reversals. Learning theory on serial reversal experiments predicts that all individuals in the manipulated group used the "win-stay, lose-shift" rule because their reversal speeds improved [@spence1936nature; @warren1965primate; @warren1965comparative]. In contrast, learning theory on multi-armed bandit (a paradigm often used in reversal learning) decision making has a stricter criterion, predicting that the optimal strategy is to maximize the cumulative reward, which, in this case would result in individuals using the epsilon-first learning strategy immediately after the first trial [@mcinerney2010]. Both learning theories consider one trial learning the optimal solution. Perhaps these wild-caught grackles relied solely on the epsilon-decreasing strategy because these individuals are used to an environment where information about the probability of what the optimal options are varies [@mcinerney2010]. Therefore, maximizing information gain via continued exploration of the available options is likely of more use in the less predictable environment in the wild. Other investigations of the exploitation vs. exploration learning strategies involved in reversal learning have found that these strategies can vary by individual and relate to differences in reversal performance. For example, urban common mynas were slower to reverse a preference than rural mynas because they spent more time exploring their options [@federspiel2017adjusting]. Perhaps we found no such differences in the grackles because all of the individuals we tested came from an urban area. If a rural population of grackles could be found, it would be interesting to compare learning strategy use between rural and urban individuals."

Article 3 > Introduction:

"As their name implies, Bayesian reinforcement learning models (doya2007reinforcement) assume that individuals will gain from learning which of the options leads to the reward. This learning is assumed to occur through reinforcement because individuals repeatedly experience that an option is either rewarded or not. The approach is represented as Bayesian because individuals continuously update their knowledge about the reward with each choice (@deffner2020dynamic). At their core, these models contain two individual-specific parameters that we aim to estimate from reversal performance: how quickly individuals update their attraction to an option based on the reward they received during their most recent choice relative to the rewards they received when choosing this option previously (their learning rate, termed "phi"), and whether individuals already act on small differences in their attraction or whether they continue to explore the less attractive option (the deviation rate, termed "lambda"). Applied to the serial reversal learning setup, the model assumes that, at the beginning of the experiment, individuals have equally low attractions to both options. Depending

on which option they choose first, they either experience the reward or not. Experiencing the reward will potentially increase their attraction to this option: if phi is zero, their attraction remains unchanged; if phi is one, their attraction is completely dominated by the reward they just gained. In environments that are predictable for short periods of time, similar to the rewarded option during a single reversal in our experiment, individuals are likely to gain more rewards if they update their information based on their latest experience. While the performance in the reversal learning task has sometimes been decomposed between the initial association learning and the reversal learning phase (e.g. @federspiel2017adjusting), the reinforcement learning model does not make such a distinction. However, it does predict a difference between phases because individuals' internal states, in particular their attraction toward the different options, are expected to continuously change throughout the experiment. We also expect individuals to "learn to learn" over subsequent reversals (@neftci2019reinforcement), changing their learning and deviation rate over repeated reversals, which should also lead to changes in how much they explore versus exploit."

Article 3 > Discussion:

"The Bayesian reinforcement learning model we applied in these post-hoc analyses appears to be an accurate representation of the behavior of grackles in the serial reversal experiment. In the previous application of this model to reversal learning data from a different population, @blaisdell2021more had found that the choices of grackles were consistent with what this model predicts. Here, we add to this by showing that the model can identify variation in performance, and in particular reveal how individuals change their behavior through the manipulation series of multiple reversals. Previous analyses of reversal learning performance of wild-caught animals have often focused on summaries of the choices individuals make (@bond2007serial), setting criteria to define success and how much individuals sample/explore versus acquire/exploit (@federspiel2017adjusting). These approaches are more descriptive, making it difficult to predict how any variation in behavior might transfer to other tasks. While there have been attempts to identify potential rules that individuals might learn during the serial reversal learning [@spence1936nature; @warren1965primate; @warren1965comparative], it is unclear how to use these rule-based approaches for cases like the grackles, who, while apparently shifting toward a win-stay/loose-shift rule, did not fully land on this rule [@loganflexmanip2022]. More recent analyses of serial reversal learning experiments of laboratory animals have specifically focused on determining when individuals might switch to more specialized rules (@jang2015role). By contrasting both the reinforcement learning models we applied here and reversal learning models, these analyses indicate that some individuals do indeed learn more specific rules about the serial reversal, ultimately switching toward the win-stay/loose-shift strategy rather than continuously updating their attractions. However, these specialized strategies only seem to emerge in over-trained animals (several thousand trials) (@bartolo2020prefrontal), whereas individuals such as the grackles in our experiment are more likely to use these more general learning strategies that are reflected in the reinforcement learning models. Accordingly, the changes in behavior that can be observed in the serial reversal experiments are better captured by the changes in the learning rate and the deviation rate than by switches in rules.

Overall, these post-hoc analyses indicate the potential benefits of applying a more mechanistic model to the serial reversal learning paradigm. Inferring the potential underlying cognitive processes can allow us to make clearer predictions about how the experiments link to behavioral flexibility. In particular, we could expect that the previously observed differences in whether reversal learning performance links with performance in other traits [e.g. positively in grey squirrels: @chow2016practice, negatively in Indian mynas: @griffin2013tracking, depending on the other trait in great-tailed grackles: @logan2016behavioral and this article] could be linked to differences in whether the learning rate or the deviation plays a larger role in the reversal performance in a given species and in particular for the other trait. The mechanistic model can also help with setting criteria to better design the serial reversal experiments, as the changes in attraction can be used to reflect whether individuals have formed a sufficient association to reverse the rewarded option (@logan2022manyindividuals)."

Decision by Aurélie Coulon, posted 07 March 2022

Revision needed

Dear Corina Logan and co-authors, The two reviewers who evaluated the first version of your pre-registration "Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context?" evaluated the preprint of the corresponding post-study, titled "Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context" that you submitted for recommendation to PCI Ecology. They both highlighted the fact that this preprint is, in its current form, very difficult to read, and I agree with them. It is a very dense paper, which combines the results of several complex experiments, and assumes the reader is familiar with the theoretical and technical background required to understand the results. As a consequence, it is currently almost impossible for the reader to get all the outcomes of this otherwise interesting study. I hence recommend you follow the detailed suggestions and requests provided by Maxime Dahirel to make your ms easier to read (including result presentation), and also the more general comments provided by Andrea Griffin.

Another important point raised by the reviewers is the need for more theoretical and technical background, in the introduction, and in the discussion too. And last, the need to make sample sizes more visible in the preprint, and to discuss their potential limits.

M. Dahirel's 9th point highlights a problem in the statistical models that, unfortunately, I had not detected in the preregistration when I recommended the last version. Given the importance of this point, I am afraid I have to ask you to take it into account. I know this is not the type of request one is supposed to get on a post-study and am sorry about that. I guess this experience will be instructive for future pre-registration handlings.

Finally, I want to join Andrea Griffin in acknowledging the substantial amount of work the experiments described in the preprint must have required and hope you can take into account the reviewers' comments, so that this preprint can then be recommend by PCI Ecology.

Best, Aurélie Coulon.

Reviewed by Maxime Dahirel, 02 February 2022

Download the review

Reviewed by Andrea Griffin, 14 February 2022

General comments

This study examines flexibility in a range-expanding passerine. Flexibility is operationalized using a decrease in the number of trials to criterion across serial reversals in a colour conditioning task. The fact that grackles learn to reverse faster and faster, that is, that they learn a rule, is not novel. There is a body of old comparative psychology research using the paradigm and showing the effect in a variety of species. However, the study is novel in that it tests whether rule learning generalises to other reversal paradigms (motor tasks) and also the individual repeatability of rule learning. Even though the touch screen task did not work, I commend the authors on the extensive behavioural trials they have undertaken. This is truly impressive and unprecedented to my knowledge.

In general, even as someone familiar with the pre-registration, I found the paper much to dense to read and the results too disjointed from the methods to be accessible. I suggest the manuscript would benefit from a much more synthetic write up where the final methods and ensuing results are presented (for example, the touch screen removed up front) and the information regarding changes to the pre-registration placed in appendix. For example, just explaining what the 'reversal criterion' was next the results would be helpful.

Hence, my comments have more to do with the pitch than the detail of the study's findings. First, I think the authors should acknowledge the classic work on reversal learning and what is known about rule learning in

the introduction and refer back to it in the discussion. Related to this, in my view, presenting the experimental group as a 'flexibility manipulation' is confusing and should be avoided. Experimental birds were given the opportunity to learn a rule while control birds were not; I think it would make the reading easier to refer to the groups in these terms, while also anchoring the study into a body of existing work via shared terminology. Second, I don't think the results are relevant to the fact that the grackle is an invasive and range expanding species. There is no comparison of individuals on the front of an invasion wave with those on the back, for example. Neither is there a comparison of grackles with a non range-expanding species. I suggest it is fine to mention the ecological characteristic of the study system but suggest removing the impression that the study is testing something particular to an invasive species. Third, I find the conclusions around conservation applications unrealistic. Captive breeding programmes deal typically with so few individuals that I can't see how selecting them based on the type of measure used here would be possible. The type of animal training implemented here is hugely labour intensive and I cannot see that happening as part of cash-poor conservation programmes.

I think all these pitch issues can be addressed.

A few detailed comments are as follows:

L192: please specify see discussion so that the reader knows that you will address what you think it measures

L238: please specify what dependent variable you are referring to

L248: I suggest the terminology flexibility manipulation is unclear and should be avoided.

L277: I understand that the non-interchangeable nature of these two tasks is desirable from the experimental point of view but where does the lack of performance correlation leave the conclusion that flexibility generalises?

L290: please remind the reader here what is meant by 'pass the manipulation condition'

Table 1: please specify what X means. Please remind the reader what "to pass the criterion if they were in the flexibility condition means" and "reversals to pass".

Discussion: I recommend the discussion should relate/compare the present findings to existing data on the topic. By referring to serial reversal learning as a 'manipulation of flexibility', and not mentioning classic work on the paradigm in other species, the uninformed reader is likely to over-estimate the novelty of this aspect of the work. Similarly, others worthy of mention have attempted to decompose reversal learning into its various phases including studies in pigeons and common mynas. This previous work could be mentioned and then the progress made by the new approach (lamda and epsilon) explained.