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 1

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?

COMMENT 2

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”)

COMMENT 3

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).

COMMENT 4

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.

COMMENT 5

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.

COMMENT 6

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.

COMMENT 7

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

COMMENT 8

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.

COMMENT 9

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

COMMENT 10:

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

COMMENT 11

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

COMMENT 12

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).

COMMENT 13

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.

COMMENT 13

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?

COMMENT 15

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.

COMMENT 16

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.

First: “For the manipulated birds, we found that during their last reversal there was a positive correlation between ϕ and λ, with individuals with higher ϕ values also showing higher λ 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.

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

COMMENT 17

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.

COMMENT 18

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.

COMMENT 19

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.

MINOR COMMENTS

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.

line 23: should be "works and predicts"? or "works to predict"?

Line 25: I would add the species' scientific name here

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

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)

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

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

REFERENCES

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd edition). Springer.

Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: Challenges and solutions. *Journal of Evolutionary Biology*, *24*(4), 699–711. https://doi.org/10.1111/j.1420-9101.2010.02210.x

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

O'Dea, R. E., Noble, D. W. A., & Nakagawa, S. (2022). Unifying individual differences in personality, predictability and plasticity: A practical guide. *Methods in Ecology and Evolution*, *n/a*(n/a). https://doi.org/10.1111/2041-210X.13755

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145