Manuscript review (Round 1)

Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context: post-hoc analyses of the components of behavioral flexbility

Dear Editor and Authors,

I reviewed the paper “Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context: post-hoc analyses of the components of behavioral flexbility” by Lukas et al, submitted to Peer Community in Ecology.

Behavioural flexibility is central to understanding the importance of cognition in animal evolution. Among the many learning pathways, individual associative learning may be the backbone of some of the most sophisticated behaviours (see publications by Johan Lind et al. - “Can associative learning be the general process for intelligent behavior in non-human animals?” and “What can associative learning do for planning?”, for a demonstration). An interesting question in this regard is to understand how animals are able to balance previously constructed associations (i.e. knowledge) with currently experienced situations (i.e. cues), and how this affects the ability of individuals in different contexts. In this paper, the authors bring an interesting perspective to this question, by not only linking behavioural flexibility to individual performance, but by trying to understand the mechanism by which such a link is made. In particular, they distinguish between the tendency to stick to previous experience and the tendency to explore new possibilities as constraints/drivers of behavioural flexibility and performance in solving cognitive tasks.

I found this article very interesting, especially in its methodological approach. It necessarily suffers from a small sample size (as is the case with many studies of animal behaviour, and which I would not comment, for authors have tackled the statistical power of their design in previous publications, and the sampling effort was already considerable!). I have also raised a few subsequent points that I would like to discuss. Specifically, the authors have tried to be as transparent as possible in their scientific approach (pre-submission paper, availability of data, etc.), but I fear that this has reached a level that detracts from the understanding of the article. I have therefore suggested ways in which the article can be a fully self-contained article, thereby improving clarity. In addition, careful editing for typos and consistency of writing is necessary, especially if the article is submitted to PCI journal that does not have editing. Overall, however, I find this to be a very nice illustration of the value of mechanistic modelling in elucidating the causalities linking animal cognition, observed behaviours and animal success (even if it is still correlation in the end). To help improve the manuscript, I have divided my comments into two main categories, each ordered according to the structure of the main text (abstract, introduction, etc.). The first (Comments on content) is devoted to improving the substance of the article, covering issues ranging from the readability of the article to methodological concerns. A second part (Comments on the form) is devoted to minor aspects, such as typos, the arrangement of equations or figures, or the flow of the text. As I am not a native English speaker myself, my proposed editing of the text is only intended to improve clarity and to be more concise. I trust the authors more than myself to make the text accurate in English if what I have suggested was not. Finally, although my work focuses on animal cognition, I base my studies on behavioural observations in nature, not on experiments. Therefore, I cannot firmly evaluate the experimental protocol, which has however already been peer-reviewed at the pre-submission stage.

Although I have highlighted the weaknesses of the paper that I have identified, I do not want this to overshadow the many qualities that the document also has,
I hope that these comments will nevertheless be appreciated by the authors.
Sincerely.

I Comments on content

I.1 Title

- I understand that these are follow-up analyses to previous work, but the authors might consider a more “stand-alone” title. Indeed, it shows new (but complementary) results to previous publications. Therefore, the authors could consider giving a title that explicitly states the results of these analyses, and only state at the end of the introduction that this can be considered as a follow-up analysis to Logan et al. (2022).

I.2 Abstract

I am afraid that, in my opinion and in its current state, the abstract needs to be rewritten, especially to make it accessible to neophytes who are unaware of the experimental set-up put in place by the research team. In particular:

- The initial sentence is, in my opinion, still too similar to what has already been published by the team (Blaisdell et al., 2021), and has lost some precision along the way (e.g. flexibility is not associated with adaptation to a new environment per se, which I currently understand. Moreover, flexibility can also be deleterious if it is ‘blind’).

- Many terms are at this stage unclear to the reader (context/multiple access box, the "components" of behavioural flexibility - not performance, I suppose - which I am not sure refer to what is stated afterwards without having read the whole document, locus/loci).
- L25-26, the authors might consider subcategorising the "flexibility" they are studying, as there is no "one behaviour" and therefore no one "behavioural flexibility". In particular, the authors are working in the context of associative learning during foraging, and it might be important to make this clear (this could be also true in the introduction).
- L30-31 "This result was supported in simulations". Are the authors referring to the model test? If so, I would suggest that the authors do not state it in this way, as the simulations were only done to evaluate the methodological approach. The authors may only write that "this result was supported by cognitive experiments on wild grackles".
- The summary lacks a broader perspective outside the world of grackles. Perhaps the author could consider adding a sentence about how the work may fit into theories around behavioural flexibility (even if this is only speculation, provided it is clearly stated).
- Overall (and like the rest of the paper), the authors might consider presenting the text as a stand-alone article. This could be done by not systematically referring to the results of the previous experiments (or by linking to the previous publications without giving a 'short' summary), although the authors could consider this study as a post-hoc analysis. I think this would reinforce the ideas that are put forward by this paper and that are really distinct from their previous work.

I.3 Introduction

- To me, the definition of ϕ as "rate of learning" is at odds with equation 1. Instead, equation 1 describes what is called "irrationality", i.e. the tendency to rely on long-term knowledge (low ϕ) rather than on recent signals. Perhaps the authors could opt for another name to avoid this confusion.
- I don't understand what it means that "individuals act on small differences in their attraction". In particular, beyond the meaning itself, I also don't understand how this refers to λ. Can the authors think of a way to make this clearer?
- L70 "less attractive", the authors might consider specifying "perceived as less attractive", as this depends on the individual's knowledge, right?
- L73 At this stage, a naive reader (as I was) may not understand why it is about the "both" options, as the experimental setting was not yet detailed. The authors may therefore consider rewording.
- L75, the authors give examples of ϕ values. However, it has not been defined that it is between 0 and 1 (which is only understood later from equation 1).

- May I ask why λ is considered a rate, since with equation 2, it seems that there is no constraint on its values, which are therefore not between 0 and 1.
- L80, perhaps the authors could also refer to Dunlap and Stephens, in their 2009 paper "Components of change in the evolution of learning and unlearned preference", who studied how the predictability of the environment can select learning. In this respect, the wording 'more stable' is vague: does it mean that the environment is predictable in the long run? If so, there may be two 'types of instability' to differentiate between: a succession of periods of short-term predictability, each involving different knowledge rules, or no predictability at all. Perhaps the authors could discuss how this would (or would not) count in their reasoning. In particular, I myself would expect λ to matter (and thus be subject to selection) only in totally unpredictable environments, so not if there is only a succession of predictable phases (as was done in the experiment presented, and thus consistent with the results found).
- When the authors refer to "serial inversions" (L96), they may specify how many inversions were performed.
- As far as the abstract is concerned, many terms are introduced but not defined at this stage (yet later in the article), which prevents a correct understanding at the beginning. The authors may therefore consider defining terms like "locus" (L100) beforehand.
- The authors might consider rewording L94 "serial reversal learning - reversing individuals", as it is difficult for me to understand. Do the authors mean that they have performed further reversals on individuals who have shown reversal ability?

I.4 Research questions

- Although the model test itself is necessary, I would not consider it in the main text, but rather place it in the supplementary material, as a proof of concept. This would reduce the already dense article. And as such, the authors might consider deleting prediction 1, as it seems very strange to 'predict' that the statistical approach adopted is reliable, as otherwise this approach would not have been considered for conducting the analysis.
- In prediction 4, the flexible individuals are those with high ϕ, but perhaps also those with low λ, aren't they?

I.5 Methods

- Overall, in the method, I am not sure that the sample sizes have been clearly stated. So the authors might consider clearly stating them here, as well as in the results, when the sample sizes change.
- L148, is there a practical reason for not setting initial attractions to 0?
- L223, I am puzzled by the choice of "89%" for the compatibility intervals. I agree that the use of these thresholds is arbitrary, but may I ask why not use a rounded value? In other words, it gives the impression that the authors have "tricked" the analysis to

fit their expectations, which is certainly not the case. Thus, the authors can further elaborate their reasoning for this choice. Given the low sample sizes, 80% could even be used.

- Unless I am mistaken, the ‘batch’, L227, has not yet been defined. Also, L171, what does the term “criterion” really refer to? Is it the criterion defined later in 2) L194?
- I am puzzled by the choice of criterion (17 out of 20 seems to me an arbitrary but unjustified choice), as it seems to be at odds with the existence of highly exploratory individuals relying little on prior knowledge (low λ, high ϕ), for whom such a criterion can only be met by chance (and thus, it is unclear to me how a bird with low λ could reverse at all and that this is “cognitively” meaningful). Could the authors expand on this criterion, and the possible consequences on their analysis that it triggers?
- The authors might consider writing a short paragraph in which they detail more carefully the experimental setting, and the different criteria for individuals to be retained or not, to be considered successful or not (both in the task itself, or to be considered “reversed”, etc.).
- L215, could the authors explain why they used two different settings (wooden box and plastic box)? Personally, I don’t understand at the moment.
- L246-248, the authors indicate that they take the distance to the median to characterize a U-shaped pattern. I have many questions about this: is it missing in the text that they actually take the absolute difference from the median (thus they use distance in its true mathematical definition)? The tables suggest that the authors did as well as the script. However, assuming they did, this does not evaluate a U-shaped pattern, but only a “triangle-shaped pattern”. Secondly, why does it have to be centred on the median of the group. Can the authors justify this? I would expect centering on the mean (as would do the polynomial regression). From the script, the authors used the function *standardize*, but I could not find from which package it is (is it the *standardize* package?), there what it did and whether the “median” was a typo or not. As a second exploration, they used the square of ϕ or λ. This approach assumes that the U-shaped model is 0-centred, unless a linear term is added (in which case the fitted estimates may imply that the polynomial is not 0-centred; from the script I don’t see any linear terms). If this is the case, then I don’t understand why model first a linear relationship and then a polynomial relationship, and discuss both in the results. Can the authors explain why they did this? From my point of view, I would encourage the authors to adopt a stepwise approach (which is generally not to be done in linear modelling when considering different variables, as specified by Mundry and Dunbar, 2009 “Stepwise model fitting and statistical inference: turning noise into signal pollution”, but not the case here). First, I would consider the polynomial term (with the linear term), and if it is not significant, I would simply transform the model into a linear model (since the second order term is, in fact, unnecessary).
- Finally, why were the two “box contexts” treated differently, instead of in a unique model, adding the box as a control variable?

I.6 Results

- L253, the authors might consider a quick reminder of what ϕ and λ refer to.
- As the authors point out, the values estimated from the reversal phase alone are not those simulated. However, we do see a linear relationship. So why can't these values be used in a practical way for further research? Also, do the authors have an explanation as to why the combination of initial and reversal (and not just the initial, which was not tested), is more conducive to deducing the parameters? I am sorry that this is not clear to me from the current explanations.
- The claimed linear relationship (Figure 2) does not seem linear to me (although the median, and not the mean on which statistics are conducted, is plotted). Yet, linear modelling could lead to significance as it seems that the relationship is always monotonically decreasing. Is this "non-linearity" due to the fact that the data have been grouped into categories (x axis)? If not, the authors may consider non-linear regression (i.e. generalized linear models).
- In L287, the authors refer to the "last two reversals": it is not clear whether the initial and the last two are used, or only the last two. In this case, I do not understand because the authors have shown that using only one reversal can lead to biased estimates. Although the "penultimate" reversal serves as a "new initial", is it not already biased as well? Would it change the results if we also considered the true initial, and the last two reversals for the other manipulated individuals?
- In L290, the authors say that the changes in λ are small compared to the changes in ϕ. As these two items are on different scales (one is linear, the other is embedded in an exponential), I'm not sure this is as straightforward as expected.
- The authors may consider removing from the results (and inserting into the discussion) the paragraphs between : L296-298, L303-305, L361-367, L373-378. In particular, I would add in the discussion (or in the introduction) the parallel with previous work, but keep it separate from the description of the results.
- L303-305, although I tend to agree with the authors' statement (summarised in a way by the idea that ϕ is a driver of response, and λ a constraint), I feel that, presented as it is, the analysis is biased by the lack of interaction between ϕ and λ in the models. Yet, looking at the tables and scripts, these interactions have been tested, haven't they? Many models, not mentioned in the main text (or by mistake) are available in the various summary tables. Is this a mistake?
- L306-308, this statement seems strong to me, as the two variables are not scaled. Could the authors consider further quantifying the changes (e.g., relative changes) to support their assessment?
- L314-316, the lack of correlation is consistent for ϕ, I think, because the changes in ϕ depended on the value of ϕ at the start. However, this explanation is not valid for λ. Unless I have forgotten, this is not explained further in the discussion. Do the authors have a possible explanation?

- L329-331, is it based on a visual assessment or on additional statistics not shown here? If the former, the authors might consider softening things by adding “tends to” to the various assessments. For example, I can only see that for both parameters, one individual is not following the group trend.
- L334, the author might consider replacing “With the Bayesian approach, we used one model to estimate…” with “We used a Bayesian structural equation modelling approach…”, as I think what the authors have done fits the structural equation modelling framework.
- I am puzzled by the results of this modelling. Indeed, from what I understood, it seems that the initial value of ϕ conditions the whole response of the bird during the whole experiment. However, as I understand it, the changes in ϕ (if I am right, these changes are evaluated by comparing the initial/first reversal and the last two reversals in the birds going through multiple reversals, which is not clearly indicated) also depend on this value, which finally implies that there is no correlation between the first and the last value of ϕ. Why then is the first value the trigger for the whole response, including subsequent performance which should, likely, be related to the ϕ at the time action is performed?
- Maybe I missed it, but for point 3), is there, at least visually, a difference between the birds tested for one or more inversions? The authors could add this information in Figure 5 (e.g. by differentiating the types of points).

I.7 Discussion

In my opinion, the discussion does not comment sufficiently on the results. Although the first paragraph does a good job of summarising the main results, they are still insufficiently commented on and compared to the literature (without giving a picture of the literature without explicit comparison to the study, as I felt in L430-450 or simply repeating the results, as I felt in L453-463 which could be elegantly linked to the idea of behavioural types). In addition, I find that there is a too strong emphasis on the importance of mechanistic modelling, which, while useful and interesting, remains a widespread approach (e.g. in the field of movement ecology). To help solve this issue, it would be beneficial to remove some paragraphs from the results (as indicated above) and add them to the discussion. In addition, the authors may comment on several of these points (and others that I might have forgotten):

- Why is there a difference in performance (and how is this affected by the two parameters of interest) between the boxes? Does this have any relevance? Furthermore, I still wonder why two boxes and why they were treated separately.
- How does the environment of the grackles differ from the experiments (which might explain why the changes in ϕ and λ differ between individuals in the experiments, both in their absolute values and in the magnitude of their change).
- Why don’t the simulations and observations match perfectly, as far as λ is concerned? For my part, I suspect that (1) there is a joint effect of ϕ and λ. Although the authors mention the idea of a trade-off (which might echo the idea of a behavioural syndrome

that might be worth discussing), the analysis, description of results and discussion suffer in my opinion from the fact that these two parameters are constantly separated. (2) A second interesting point of discussion might be to see how the magnitude of the changes in the simulation corresponds to the magnitude of the changes in the observations. Perhaps the decrease in λ observed for the observations is largely offset by an increase in ϕ which is larger than in the simulations.

- The authors might consider discussing the multiple pairwise comparisons further, as it seems to me that they are not entirely consistent (perhaps due to the sample size). For example, the initial ϕ is correlated with many variables (of performance or of the change in ϕ itself) but not with the last one (which is nevertheless U-shaped related to some performance). On the contrary, λ has no such relationship, although it is negatively correlated with ϕ. To me, this is confusing. As I suspect the sample size (as well as the experimental setting, which seems to select for a very specific type of long-term instability) to be one of the reasons for this, this could be discussed and reported with caution. Parallels with results in other species could help identify results that are likely to be erroneous (either missed or significant by chance).
- It is not entirely clear to me that ϕ should have a U-shaped relationship with the latency to resolve a new locus. In particular, when ϕ is very large, if λ is large enough, a bird should immediately move to another solution. Thus, this U-shape may not be interpreted as claimed by L469-470, but may be the simple consequence of the negative relationship between the two parameters (which may prevent λ from being large enough). It seems to me therefore important to discuss this negative relationship (i.e. the trade-off) further.
- Why is ϕ (and in particular the initial ϕ, as questioned in my point on the results) the main driver of the response? Are there related results in the published literature? Authors may wish to consider the theoretical literature on the exploitation-exploration trade-off, which is somewhat similar to what is stated here, where ϕ tends to correspond to the "memory" component, hence exploitation, and λ to the "exploration" component (e.g. Berger-Tal et al., "The Exploration-Exploitation Dilemma: A Multidisciplinary Framework", as a starting point for reading).

In addition, in order to improve the flow of the discussion, the author could consider adding sub-headings, indicating the outcome that is being discussed, and these sub-headings could in fact echo the research questions/outcomes. It seems that the order of the discussion already reflects this idea. Making it more explicit could help to identify the key points to be discussed, for the reader, but also for the authors themselves.

Finally, to broaden the scope of the article and make these results more relevant from an eco-evolutionary point of view, the authors could consider assisting the reader with an additional figure linking flexibility, behavioural patterns and environmental characteristics, which could be fully elaborated in a "speculation" paragraph (as happens in some journals, such as "Oikos"). This could be useful both for the introduction/prediction and for their discussion. For example, I am thinking of a figure similar to the one published by Riotte-Lambert and Matthiopoulos ("Environmental Predictability as a Cause and Consequence of

Animal Movement") on the effect of environmental contingency and constancy (the underlying components of predictability) on the spatiotemporal memory of animals. Here, instead of using contingency and constancy, ϕ and λ could be used to create two matrices describing respectively (1) which type/level of behaviour/flexibility and (2) environment correspond to which values of the two parameters (e.g. environment "predictable in the short term", "predictable in the long term", etc.), highlighting the area of this two-dimensional landscape in which the grackles under study are located. This could allow the authors to broaden the link between behavioural flexibility and the performance, and ultimately the evolutionary success, of the species, as well as pointing out the limit of the applicability of their findings. This is only a suggestion, however, and the authors can think of something else that would place their work more strongly in an eco-evolutionary framework.

I.8 Miscellaneous

- Although I believe from previous articles and team submissions that this project has met animal ethics guidelines and has been approved by official entities, the authors might consider writing such a paragraph (unless I missed it, my apologies).

II Comments on the form

II.1 Main text

- Authors should stick to the Greek letter (ϕ, λ) in text, figures and tables (unless there is a reason I didn't understand for the differences, but it sometimes happened in the same line, e.g. L426). For the figure, this can be done in *R* using the "expression" function of the *base* package immediately available. I would like to point out some typos "/phi" and "/lambda" L253, L309. As this is written in *Rmarkdown*, I have seen that the "/" is inverted (you should use the opposite one).
- Why label the paragraphs that follow the discussion with letters? Perhaps the authors are considering deleting it (as PCI does not edit for publication).

II.2 Abstract

- L24, the authors may add "gaining" after "allows".

II.3 Introduction

- L68-69, the authors might consider simplifying the sentence "based on the reward they perceived during their most recent choice relative to the rewards the perceived when choosing this option previously" to "based on the reward of the most recent choice relative to the previous rewards of that option".

II.4 Methods

- In the equations, some spaces seem to be missing (e.g. in equation 1 of L141), before and after the "=" and "+".
- In L142, the authors may write that i takes the value 1 or 2 as follows: $i \in \{1, 2\}$.
- In equation 2, authors may consider writing the exponential function in its classical form (e.g., e^x), to improve readability.
- References to the author C. Logan are different (e.g. L135, Logan CJ et al., L 185 C. Logan et al.). The authors might consider making the different citations more consistent, both in the main text and in the reference list. I suspect that this is due to differences in writing in the bib file associated to the *Rmarkdown* document. In addition, there are two Logan et al. 2022 references, but they are cited the same way in the text.
- In the main text, I would consider using italicised variables when they are quoted (e.g. when explaining variables in equations, t, P, j, i, etc.).
- In L157, the reference to R (which I personally would have italicised) is in square brackets, instead of parentheses. Again, if the authors use *Rmarkdown*, this is due to the ";" used to separate the version from the reference. A simple "," would solve the problem.
- L184, I find the reference to the PDF unusual. Authors can only consider adding the book reference, which would be listed with all the other citations.
- For the equations in subparts 1) and 3), the authors may consider: (i) writing everything in a fully mathematical way (e.g., $\mathcal{N}(\mu, \sigma)$, $\mu = a + b\phi + c\lambda$, α_{batch} etc.). Furthermore, it seems to me to be unnatural (though somewhat understandable) to label the underlying distribution used for the likelihood test as "likelihood", as it is not the likelihood itself. I would consider specifying only the formula for the linear model (and if not, at least labelling it as "model" or "the model", but only one of them, see the difference between 1) and 3)), and then specifying where the parameter used as the output variable comes from. For example, for 3) it would be: "We modelled the probability to solve a locus p as a function of ϕ and λ such as $ln\frac{p}{1-p} = \alpha_{batch} + \beta\phi + \gamma\lambda$ where p is the probability in a binomial distribution such that $N_{loci\,solved} = B(4, p)$.
- In addition, I would call (Equation 3, 4, etc.) the equations just mentioned, for consistency with equations 1 and 2.

II.5 Results

- In L314, I think there is a space missing before "Table 1".
- References to models sometimes capitalise the "m" (e.g. L369), and sometimes not (L385). The authors might consider copying this (and other typos, e.g. L367, no space between the comma and the 3).

II.6 Discussion

- The authors could consider deleting “go-no go”, L480, which does not provide any information, and only mentioning an “inhibitory task”, or further detail the layout of this task.

II.7 Figures

- In Figure 1, the authors might consider capitalising the first word of each axis, as is done in other figures.
- In Figure 2, the authors might consider adding the mean, as statistics are working on those, and note the median depicted by the boxplot. Furthermore, they may consider removing the legend and colouring the Greek letters ϕ and λ as in the boxplot, to make the figure simpler. As the figures appear to have been made with *ggplot*, this can easily be done by using the *ggtext* package, and labelling the y axis with html, for example (the colours are not those used, however).

```
ylab("<span style='font-size:20pt;font-weight:bold'>Standardised
<b style='color:#458b74'>&Phi</b>
<b style='color:#ffc125'>&Lambda</b>
of simulated individuals </span>")
```

- In Figure 3, the authors may consider keeping the background white (for consistency with the other figures), as well as labelling the y axis rather than the graph itself. In addition, to be consistent with Figure 2, the title of the legend could be capitalised.
- In Figure 4, the authors can distinguish between significant and non-significant pairwise comparisons (e.g., dashed or plain lines) to make it more readable (perhaps the coefficient of estimates and its compatibility interval could also be added to the middle of the arrow).

II.8 Tables

- The authors may rename the models, and specify the output variable (while being consistent, the model is sometimes capitalized, e.g. model 6, or not, other models), so that the reader can quickly identify which model it corresponds to, as there are many presented, instead of just naming (“model 7 plastic” etc.). Also, I would like to point out that the model numbers in the text and in the tables seem to me to be different (e.g. L346, I believe the authors are referring to model 6 and not 20). The model numbers in the text do indeed start at 17, if I am not mistaken.
- The authors may want to double-check the different names of the variables: intercept is sometimes called intercept per bird or simply intercept (whereas it is referred to as a or α in the equations), the other variables include estimates (b * lambda), * is not a

mathematical term (which is ×; yet I believe estimates should not be included in the label).

- The authors could consider highlighting the bold lines where significant. This would make the tables easier to read. An alternative could be also to provide forest plots, instead of such tables.