EDITOR’S COMMENTS

(i) Line 211. The details of the partial split clutch design are not clear in the methods. It would be useful to briefly mention how eggs from a clutch were assigned to different treatments.

Thank you for the suggestion. We now include this information at the end of the “Manipulating Early Thermal and CORT Environments” section: “We counterbalanced the number of eggs assigned to each hormone and temperature treatment using a partial split-clutch design. We assigned at least one egg per clutch to each treatment, and the remaining eggs were randomly assigned to one of the treatments. When the eggs per clutch was less than four eggs, we assigned each egg randomly to one of the treatments. The final total number of clutches per species is included in Results.”

(ii) As mentioned in the manuscript the final sample sizes were 7 to 12 per treatment per species. I appreciate the effort that goes into carrying out such an experiment. However, with the given sample sizes, one question that comes to the reader’s mind is whether the experimental design had the required precision to detect biologically meaningful differences between treatments. It would be very helpful to briefly discuss the precision around slopes (eg the width of credible intervals) and importantly, whether this indicates that the lack of differences between slopes is a robust result or a result that needs verification with an increase in sample size.

Thanks for this comment. We agree that the 95% confidence intervals are critically important which is why we present these in Table 1 in Supplementary Material for all the slopes. Based on this analysis, it does not appear we have a power issue given that all slopes are significantly different from 0.

We think the Editor is referring to the contrast between slopes when asking whether we had sufficient power. If our interpretation is correct, then we agree we have not fully reported all 95% CIs. We have now done so when we discuss contrasts in the results. However, we also acknowledge that the sample size for some contrasts is far larger than 7-12 because different questions are being asked. For example, comparing differences between species we have n = 47 and n = 37 for each L. delicata and L. guichenoti respectively, or within species across temperature (n = 23 VS 24 in L. delicata or n = 17 VS 20 in L. guichenoti) or CORT (n = 23 VS 24 in L. delicata or n = 17 VS 20 in L. guichenoti) we have nearly double the sample size.

Nonetheless, we agree that this could be acknowledged better in the discussion. As suggested, we have also modified the discussion to highlight the wide confidence intervals around comparisons between treatment slopes as follows:

*“We acknowledge that sample size may be limiting our ability to detect the small effect sizes in our experiment given the uncertainty around slope contrasts between treatments (see 95% CI of contrasts in results).* Animals came from a total of 36 unique clutches in *L. delicata* and 33 in *L. guichenoti, yet it is challenging to do more animals than this given the logistical constraints associated with running such detailed cognitive trials*. *However, it is noteworthy that our effect sizes are very small and likely of little biological significance. For example, slopes of CORT-treated animals is 100.094% the slope of Controls in L. delicata and 96.865% in L. guichenoti, while the slope in hot-incubated lizards is 111.582% the slope of those incubated at cold temperatures in L. delicata and 127.384% in L. guichenoti.”*

(iii) Regarding sample sizes, could you indicate how many different clutches the eggs come from? This will give the reader an idea of how representative the sample is of the lizard populations (and it would be nice if you could include this in your discussion of sample sizes, precision of estimates and robustness of the main findings).

Absolutely. We have included a sentence at the beginning of the Results indicating the number of clutches per species of the final sample size. “These animals came from a total of 36 clutches in *L. delicata* and 33 in *L. guichenoti*”. We also mentioned the number of clutches when discussing the potential limitations of our sample sizes (see comment above).

EDITORIAL FORMATTING ISSUES:  
  
1. Shorten the Highlights. Each point should be no longer than 85 characters including spaces.  
  
2. Take the figures out of the text and upload them separately. Figure legends should be collected together and placed after the references in the manuscript. They must not be on the same page as the figure or uploaded as figures.  
  
3. Figure labels should be lowercase letters in full parentheses (e.g. (a)) and placed inside the axes of the graph. Use lowercase letters in the text as well as the figure.  
  
4. Add a space between numbers and units, including in figures.  
  
5. The references Baker 1979, Crino et al. 2023, 2024 are incomplete. Add full details and check all the references for completeness.  
  
6. Tables should have a short one-sentence title above the table and other information should be placed below the table.  
  
7. Remove the internal horizontal lines from the tables.  
  
8. Add the title page to the manuscript as it will not be reviewed again.

All the edits have been addressed in the final version of the manuscript

REVIEWERS' COMMENTS:  
  
Reviewer #1: I appreciate the revisions made by the authors and I have no further comments.

Thank you very much  
  
  
Reviewer #2: Comments to authors  
  
I would like to thank the authors for their careful consideration of the reviewer comments. All of my previous comments were incorporated appropriately and I think the manuscript has much improved. Overall, I believe that the study is well introduced and now all methods and analysis are very clearly described. Furthermore, the discussion does an excellent job of putting the results into a broader context. Therefore, I have no major comments left, only a few very minor comments and typos which I list below. Line numbers refer to the Non-highlighted revised manuscript.

Thank you very much  
  
Keywords: I suggest not to use words in the keywords that are already in the title because it reduces findability of your study after publication

That is a good point, we have modified the chosen keywords. Thank you for the advice.  
  
L101: "identities" could you be more specific here? I assume this refers to female or tub (i.e. group of females)

What we did was assigning a clutch number and an egg number to give them an identity, while the tub where each egg belonged to was recorded separately. We have modified the text: “We also assigned a unique number to the clutch and each egg.”  
  
L134: "eggs were"

Corrected  
  
L194: By "parental condition" do you mean identity?  
Corrected

L211: "fed" not "feed"  
Corrected

L242: "no" not "not"  
Corrected

L333: "are" not "is"  
Corrected

L338: "at least from temperature and GCs"  
Corrected

Finally, as a response to one answer to a previous comment:  
  
I have been thinking about this for a while now and this is just a suggestion for you to think about. We really only measure behavioural flexibility in reversal tasks as performance relates to acquisition. We can think about the first phase as the control and then the reversal as the test. Depending on how you collect the data, one can then subtract the control from the test either using trials to criterion or number correct. This way, we take individual differences into account. Analysing trial by trial data to investigate learning can be done in addition but I think that we need to change how we look at reversal learning as a measure of behavioural flexibility. In this way, flexible individuals would be those that are faster in the reversal compared to the acquisition. I am not saying you need to change your analysis as I have not seen a single paper looking at behavioural flexibility this way but I think it is worth considering.  
  
Thanks for this comment. This is an interesting idea. We are not quite sure exactly what the reviewer is thinking, however, we have thought about this as well, and even contemplated how such a metric could be derived. However, we think that the simple metric suggested may not necessarily capture behavioural flexibility in a way that is comparable across individuals. For example, assume we have an individual whose acquisition slope is 0.8, whereas its reversal slope is 0.8. If we were to subtract this (or create a ratio) the value designated to this individual would be 0 (or a ratio of 1). In contrast, we have a second individual who is clearly not learning, has an acquisition slope of 0.01 and reversal of 0.01, then this individual would have the same designated value (or ratio), making it challenging to compare across individuals in a consistent manner.  
  
I see your point, however, I still think that taking individual differences in general learning ability into account would be important. As regarding your example: both individuals would still be considered inflexible from a behavioural flexibility perspective. Form a learning perspective one is of course more proficient in the task than the other. These are two different perspectives which both are important and need to be taken into account in my opinion. Nonetheless, this is not relevant for the current publication.

We agree with the reviewer’s opinion, and this is definitely something to take into account in future studies

I very much enjoyed reading the manuscript and I am looking forward to seeing it published.

We appreciate your kind words

Birgit Szabo