Dear Authors,   
  
Thank you for your submission which has now been seen by two reviewers. In light of their comments, and my own reading of it, I regret that I am unable to accept your paper for publication in Animal Behaviour in its present form. However, I would be prepared to consider for publication a revised version that takes into account the suggestions of the reviewers. Such a revised paper would likely be reviewed and there would be no guarantee of acceptance.   
  
I join the reviewers in appreciating the question under focus; that is, the effect of early life conditions (temperature and prenatal CORT) on behavioural flexibility in two lizards differing in invasiveness. The reviewers have identified certain points that would need to be addressed. First, both reviewers have asked for clarifications and additional information in the Methods. Of these, one important clarification is about your confidence in the success of the CORT treatment. Second, I would also suggest incorporating Reviewer #1’s suggestion to introduce reversal learning in the Introduction. Third, Reviewer #1 has asked for additional information from the data, especially about how lizards learnt. Fourth, I would like to add to the reviewer comments about statistical analyses:  
  
i) It would be very helpful for a reader if the analysis section could briefly explain how the statistical model is set up to test hypotheses about the impact of stress hormones and temperature on learning rates. I assume that the interaction term between trial and hormone/temperature is key? It would be helpful to state this more explicitly (also I assume a three-way interaction was included, and if so, it would be helpful to explicitly state that).

Thanks for the suggestion. The Editor is correct, the key hypothesis test is evaluating whether the rate of learning (i.e., trial slopes) are different between early developmental treatment (i.e., CORT and temperature). We now include this in the Statistical Analyses on Methods: “We modelled correct choices [correct (1) or not (0)] as the response variable, and trial, hormone (CORT versus Control), incubation temperature (Cold versus Hot), along with the three-way interaction between trial, hormone, and temperature. If early environments impact learning then we would predict that the rate of learning (i.e., trial slope) varies by treatment as captured by the interactions.”

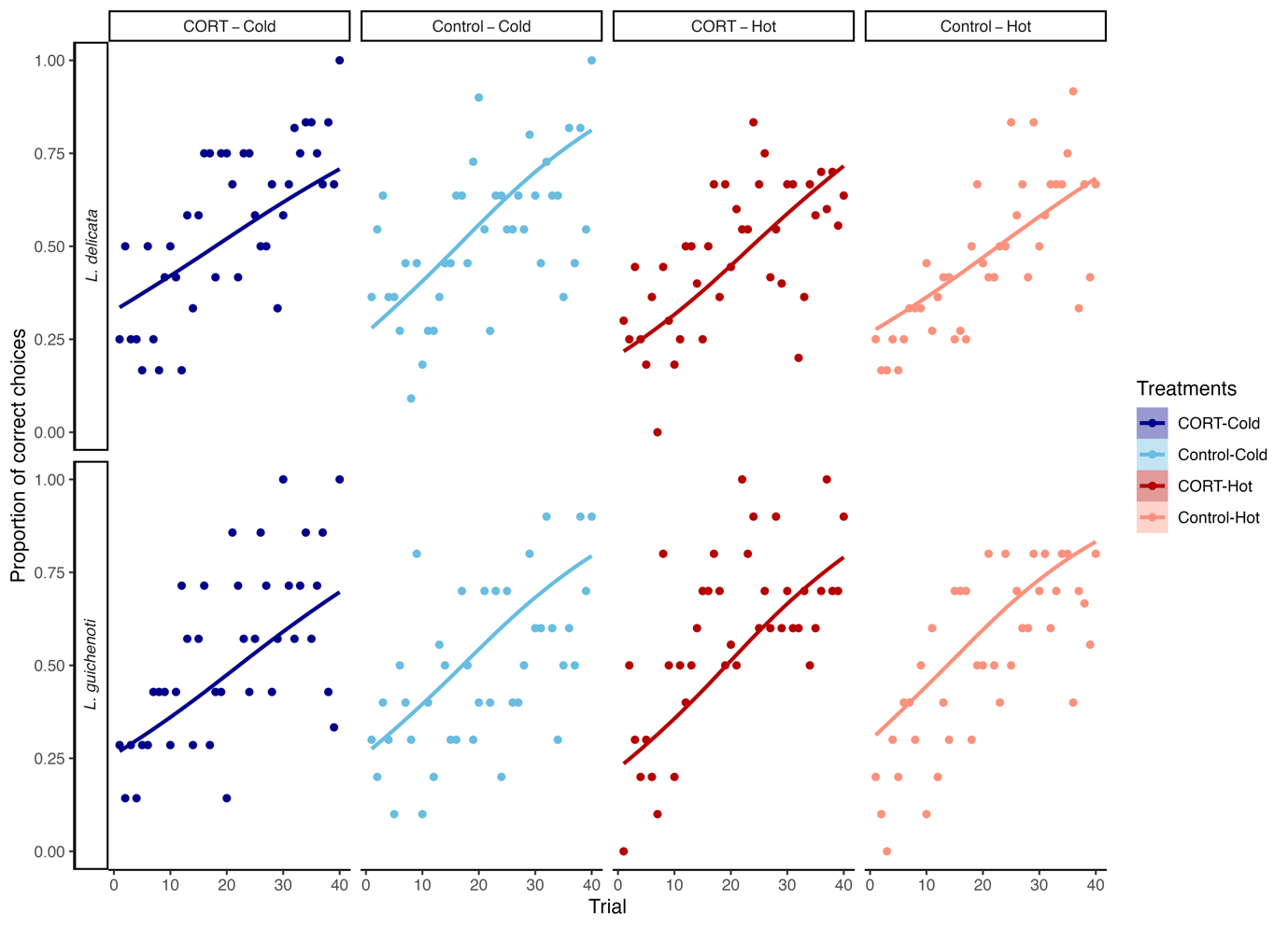
ii) Please can you mention the error structure and link function used in the statistical model? I assume that a binomial error structure was used (since the data are binary – each trial yields one of two outcomes - correct, incorrect choices)?

Absolutely. We now include this information in the Statistical Analyses in the Methods: “The error structure was modelled using a Bernoulli distribution with a logit link function (family = Bernoulli(link = 'logit')).”

ii) In the text of the results, the first time a contrast between slopes is shown within parentheses, it would be helpful to the reader to state that these are slope contrasts.

Thank you for the suggestion. We have modified the results to address this concern. We now include β before each contrast to clarify that the contrasts are between slopes.

iv) Figure 2 visualises model predictions but does not display data, and so it is difficult to assess the fit of model predictions to the data. I suggest incorporating data into figures when showing the relationships estimated in statistical models. For  binomial data, one could think of ways to aggregate the data.  
We did not include the raw data originally because incorporating it into the figure resulted in increased complexity that hindered the clarity and interpretation. However, we have made new figures with the raw data (below) that we have incorporated in the Supplementary Material. We also mentioned this information in the text in the Results section (“Figures for both species with the raw data were included in the Supplementary Material”).



v) Lines 189 – 191 – the definition of p mcmc is not clear. If, as given, it is “the probability that slopes or contrasts between slopes differed from zero (p mcmc ) using the posterior distributions,” then it is not clear how a small value indicates a statistically detectable relationship. Perhaps you mean that the p mcmc tests the null hypothesis that the slope or slope contrasts are zero?   
Yes. This is exactly what we mean. Sorry that this was unclear. Essentially, pMCMC, values are testing what proportion of the posterior distribution encompasses zero. If this value is a slope, then the entire posterior distribution of the slope (i.e., the probability density) is used to assess what proportion of slope values below zero, which tests the null hypothesis that the slope itself is different from zero. In contrast, if this is a ‘contrast’ between two slopes then the difference between the two slopes of interest becomes the posterior distribution and we are testing whether distributions overlap with zero to test the null hypothesis. This is multiplied by two for a two-tailed test which makes it analogous to a frequentist p-value.

We have now added this detail into our revised manuscript as suggested: “pmcmc test the hypothesis that slopes and slopes contrasts are different from zero.”

As you revise your manuscript, please note that the journal’s guidelines require that you address any animal welfare issues arising from your study within the Methods section, in a separate subsection headed Ethical Note. Please address all ethical implications of the experimental design and procedures, including any procedures taken to minimize adverse impacts on the welfare of subjects or to enhance their welfare (for vertebrates and invertebrates alike). For further details on what ethical information to include, please consult the “Welfare of nonhuman animal and human subjects” and “Methods” sections of the journal’s “Guide for Authors”.  
We have included in our Ethical Note another paragraph about the conditions in which the crickets were held: “Crickets were maintained in big communal enclosures (68.5 L x 49 W x 39 H cm) in the same room where the big colony is, and at the same temperature and light conditions. They were provided vegetables and water ad libitum, and they were given several egg carton shelters to refuge. All lizards were fed alive crickets except during the experiment, when we used crickets frozen at -21 ºC for 24 hours.”

With all best wishes,  
Kavita Isvaran  
Editor

Reviewer #1: The manuscript deals with relevant issues on two factors that can lead to stress during development: temperature and CORT changes in the lizard species Lampropholis delicata and L. guichenoti. L. delicata is an invasive and thus authors expected these individuals to show greater behavioural flexibility under changing temperature and CORT levels. However, the results show no difference across species and treatment groups. The study is timely and the manuscript is written well. I just have a few major and minor suggestions that will hopefully make the manuscript a better read.   
  
Line 50-53: The sentence is very confusing here so it might be better to break it down into simpler sentences.  
Thank you. We agree. We have modified the sentence to: “Animals’ responses to abrupt temperature changes are mediated by GCs (Crino et al., 2023), which can influence animals’ decision-making while nesting (Kolbe & Janzen, 2002). As a result, GCs transmission and early thermal environment are expected to interact and shape offspring traits.”

Line 52: typo in "thermal"

Corrected

Line 67: Check the spelling of 'supported'  
Corrected

Overall Introduction:   
I think you should introduce the reversal learning paradigm in the introduction and talk about how reversal learning is used to test for behavioural flexibility. You are using this as the main experiment so it might be better to give the readers some idea about how reversal learning works and some literature behind the use of this paradigm to test for cognitive/ behavioural flexibility. Here are some relevant papers:  
Hurtubise, J. L., & Howland, J. G. (2017). Effects of stress on behavioral flexibility in rodents. Neuroscience, 345, 176-192.  
Gapp, K., Soldado-Magraner, S., Alvarez-Sánchez, M., Bohacek, J., Vernaz, G., Shu, H., ... & Mansuy, I. M. (2014). Early life stress in fathers improves behavioural flexibility in their offspring. Nature communications, 5(1), 5466.  
We have included a sentence about the importance of reversal learning in measuring behavioural flexibility: “Reversal learning, in particular, is a widely employed tool to measure behavioural flexibility (Gapp et al., 2014; Hurtubise & Howland, 2017), as it assesses an individual's ability to reverse a previously learnt behaviour providing researchers with a clear indicator of their ability to adjust to new conditions (Brown & Tait, 2010).” in the Introduction

Line 97-98: How did you measure the egg identities? I can understand you might be able to track the female who laid eggs but did you know who she mated with as there were multiple males in the terrarium? Can you please explain this?

Eggs can be assigned to a given clutch when collected because females lay eggs in groups, often stuck together. We know that *L. delicata* usually lays one clutch per season, while *L. guichenoti* lays two (Chapple et al. 2011, 2015; Kar et al. 2024). Since we collected eggs during half of the reproductive season, it is very unlikely that, in those cases where we used clutches from the same terrarium, the clutches belong to the same female. As such, each clutch will reflect the mother’s identity. However, we were not able to genotype offspring due to logistic constraints to verify the identity of males, however, our previous work has shown that clutches are generally sired by a single male, but sperm storage can occur (Kar et al. 2024. Heredity). Given our partial split-clutch design, and the fact that maternal effects are expected to be stronger than paternal effects in these species, we included the clutch as a random factor to account for eggs coming from the same clutch.

We have clarified this by adding:

- “We also assigned identities to the clutch and each egg” in the Egg collection and incubation section on Methods.

- We also included the next paragraph in the Statistical Analyses section on Methods: “*L. delicata* lays one clutch per year, while *L. guichenoti* lays two (Chapple 2011, 2014). Since eggs were collected during half of the breeding season, clutches likely come from different mothers. Additionally, previous research has shown that clutches are generally sired by a single male, but sperm storage can occur (Kar et al. 2024). Given our partial split-clutch design, and the fact that maternal effects are expected to be stronger than paternal effects in these species, including the clutch as a random factor should account for the effects of parental condition”

Chapple DG, Simmonds SM, Wong BBM (2011) Know when to run, know when to hide: Can behavioral differences explain the divergent invasion success of two sympatric lizards?: Invasion Success of Two Sympatric Lizards. *Ecology and Evolution 1*:278–289.

Chapple, D. G., Miller, K. A., Chaplin, K., Barnett, L., Thompson, M. B., & Bray, R. D. (2015). Biology of the invasive delicate skink (Lampropholis delicata) on Lord Howe Island*. Australian Journal of Zoology, 62(6),* 498-506.

Kar, F., Nakagawa, S., & Noble, D. W. (2024). Heritability and developmental plasticity of growth in an oviparous lizard. *Heredity, 132(2*), 67-76.

Line 116: Check spelling of acclimatisation

Corrected

Line 124-129: Is there a way for you to be certain that the CORT treatment (topical application) actually changed the CORT concentration inside and is an effective way to manipulate CORT concentration in early life?   
Yes, our previous work has already shown that our dosing methods do change CORT concentrations inside the egg (Crino et al. 2024). While we did not specifically validate it here we applied the same dosing methods as in Crino et al. 2024. We have made this clearer in our revised manuscript:

- “This method has been validated before in *L. delicata* (Crino et al. 2024), increasing CORT concentration in eggs by approximately 2 standard deviations above the mean natural concentration.” in Manipulating Early Thermal and CORT Environments section on Methods

- We included Crino et al. 2024 reference in the Discussion, where we mentioned the effects of the same dose in other traits: “in a previous experiment, we observed that a similar dose affected a multitude of other traits including growth and baseline CORT levels in\_*L. delicata* (Crino et al. 2024)”

Line 167-171: When you were measuring whether lizards chose the correct path or not for associative and reversal learning tasks did you set a goal for learning? I mean if lizards made 5 correct choices consecutively out of 40 trials then you consider that the animal has learnt. After reading your statistical analysis section I understand that you do not have such a set goal. But can you analyse your data in a way to incorporate this? Did all treatment groups learn at the same time/trial ? Can you get the mean trial number for each group post which the individuals show 5 consecutive correct choices? I am not sure if this will give any better understanding of the learning rate but you can check this and see if this differs across groups?

We agree that this method has been traditionally employed to assess learning in different taxa, under the assumption that, those individuals that choose the correct option more often than random, have learned the task. The main problem with learning criteria is that these cutoffs (i.e., probabilities) are often constructed under the assumption that choices are independent of each other, which of course is not true. Comparing the within-individual choices and modelling the rates of learning circumvents these challenges and avoids having to make arbitrary cutoffs about whether an animal has learnt or not.

Nonetheless, we agree that such analyses can be informative. We have now included a table in the supplementary material with that information, using, as advised here, the 5 consecutive correct trials as the learning criterion. We observed that more lizards reached that criterion under cold incubation temperatures in *L. delicata*, while for *L. guichenoti* hot incubated lizards qualified under that criterion more often. However, the sample sizes of learnt versus unlearnt are too small, and we believe that our measure of learning rate is more appropriate to avoid having to make arbitrary cutoffs. As such, we have added these analyses to the supplement (see also the new figures with raw data in the Suppl. Mat. section).

Reviewer #2: Comments to authors  
  
Summary  
In this study the authors explore the impact of GC and temperature during incubation on the later expression of learning and behavioural flexibility in two species of skinks. Overall, the study is well framed and the methods well described, although I found that some information was missing. The analyses are appropriate and follow what is generally done in the field, however, I suggest an alternative look at behavioural flexibility that the authors might want to consider (but I do not insist on it). Finally, the discussion of the results is appropriate and complete. Overall, I believe that this study is of high quality and will be of interest to the broad audience of Animal Behaviour. Below I provide some suggestions to improve the MS, but these are just some small changes.  
  
Detailed comments  
L2: I think you should briefly define behavioural flexibility here.

Behavioural flexibility briefly defined in the abstract

L5: I think you mean "offspring" not "animals".

Corrected

L8: Replace the "they" with "offspring" is it is not clear who you are talking about.  
Corrected

L28-30: I think this is not a great example. In the literature, people often confuse behavioural flexibility (which is more about cognition, hence the reversal learning, problem-solving and innovation) and behavioural plasticity (which is what this example shows). I think there are plenty of examples in the literature that focus on reversal learning, innovation or problem solving and would fit better.  
We do agree with this. We have excluded that example from the text

L41: You are missing the scientific name of the sticklebacks. Also, it should be "stickleback".  
Corrected

L45-46: I would write this differently. I would write: "Glucocorticois can be transmitted directly from the parents to their offspring and influence development of the next generation but might also have transgenerational effects." When people talk about transgenerational effects they are often talking about more than just one generation.  
Corrected

L64-65: Maybe you could state previous findings on the differences between species? I think the readers would appreciate this information as not all will be familiar with previous studies.  
We have added here some information about studies examining behavioural differences between both skinks: “Previous studies exploring behavioural differences between the two species have found *L. delicata* to be more exploratory than *L. guichenoti* (Chapple et al., 2011), but no differences in learning were observed between the skinks in an associative learning task (Bezzina et al., 2014). Disparities in behavioural flexibility may be driving the differences in invasion success between both skinks, a prediction supported in other invasive species”

L71: Replace "we are uncertain" with "uncertainty still exists"  
Corrected

L91: "the warm side"

Corrected

L94: What extreme are you talking about here? The temperature?  
We meant on one side. We have clarified this in the manuscript (“on one side” rather than “in one extreme”)

L106-107: I am assuming you also provided a shelter?  
We did, we have included that in the text

L108: The scientific name of the crickets should be in italic.

Corrected

L112: What were the environmental conditions in these rooms?  
We have added to the text that “Conditions in these rooms where identical to the main room”

L113: I would use "that could be monitored by" rather than "associated to"  
Corrected

L116: "on the rack"  
Corrected

L122: "To empirically test the effect of the early environment"  
Corrected

L124: I assume you used a split clutch design?  
In most of the cases we did, but there were some individuals from the clutch that died or were used for different experiments. However, we did include the clutch identity as a random effect. We added some information in this regard in Statistical Analyses on Methods.

L129-130: You do not provide the temperatures of the treatments above. Please add here, or above.

Temperatures were included on lines 123-124: Cold – 23ºC ± 3ºC or Hot – 30ºC ±

3ºC

L159: "on the ramp", "on all three ramps"

Corrected

L170-171: 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.

L200: With "as string biosafety" do you mean stringent?  
Corrected

L201-202: I think rehoming is outside of research, so, I would remove the "for experiments"

Corrected

L213-220: You provide CI below but not here. Is there a reason?  
The reason is that the estimated mean slopes per treatment for both species are in Table 2 in the Supplementary material. We have indicated this in the second paragraph of the results: “[mean slopes (denoted as β throughout) per treatment for both species are provided in Table 2 in Supplementary Material]”

L228: "the early environment"

Corrected

L233-234: I do not see any p-values above. If you mean by "significant" that the CI are not overlapping 0, then you need to say that. Otherwise, I suggest removing the significant here.  
pmcmc of the contrast testing the difference between species is provided at the end of line 224: *L. delicata - L. guichenoti = -0.008, pmcmc = 0.636*

L239: "our results contrast with previous studies"

Corrected

L247: "elicit changes in the brain"  
Corrected

L269: I would add the incubation temps here again to remind the reader.  
R2.14

L278: "of the thermal environment"

Corrected

L288: I was a bit confused at first what you mean by sex-reverse. I suggest adding "under certain incubation regimes" to make clear that this is related to incubation temperatures.  
We have replaced the last part of the sentence with “under cold incubation temperatures” to clarify this

Table 1: there is a type. I says "Specie" instead of "Species".

Corrected  
  
I very much enjoyed reading this MS.  
Birgit Szabo