Summary

Tariel-Adam et al. (OSF Preprints 10.31219/osf.io/mr8hu, submitted to PCI Evolutionary Biology) tested for sensitive windows of within- and trans-generational plasticity in anti-predator defences (WGP and TGP, respectively). They found evidence for both WGP and TGP, with more frequent effects of predator cues on WGP. Effects of predator cue exposure varied by behavioural and morphological defences, which further varied between WGP and TGP. I think this is a compelling experiment with clear future directions and additional experiments, and below I provide several constructive criticisms and comments that I hope will be helpful to the authors.

Major Comments

1. Clarification about predictions for TGP.
    1.1. The predictions for sensitive windows for TGP (lines 104-107) were unclear even after several re-readings. While it is just one small part of the study, I do think it is important to have clear predictions for the experiment.
    1.2. My reading of prediction 2 is that TGP should have the same sensitive windows as WGP (i.e., embryonic and early post-embryonic), but prediction 1 states that late post-embryonic as a sensitive window.
    1.3. I think the prediction could be clarified into one statement, for example: "For TGP, we predict that embryonic, early post-embryonic, and late post-embryonic windows will be most sensitive." Again, I know this is minor, but having a clear prediction could help the reader link the objective of the study to the results.
2. Quantify effect sizes for treatments and contrasts.
    2.1. A missing component from the data analysis is a quantification of effect sizes. While P-values will say if there is a significant effect, they do not say anything about the biological relevance or strength of the effect.
    2.2. On line 365, the authors state that none of the exposure windows was more sensitive than the others, but there weren't any statistical tests to support this claim. By quantifying effect sizes, the authors could actually answer this question and have a more robust set of evidence.
    2.3. Effect sizes for treatment (eta-squared and partial eta-squared for F statistics, Cohen's *w* for chi-squared statistics, and intra-class correlation coefficient for random effects) will say how much variance is explained by those factors (i.e., strength of the effect).
        2.3.1. Treatment effect sizes can be calculated using functions in the `effectsize` package (Ben-Shachar et al., 2020).
        2.3.2. Package website: https://easystats.github.io/effectsize/
    2.4. After calculating contrasts on estimated marginal means, effect sizes for each contrast can be calculated as Cohen's *d* using the `eff_size()` function within the `emmeans` package (Lenth et al., 2022).

2.4.1. The authors do report tests of parameter estimates in the tables, but my reading of the table is that these parameter estimates come from the output of `summary(model)` and are not, for example, the contrast between exposure at that development stage and the control based on emmeans.

2.5. I think the paper would benefit by reporting the treatment and contrast effect sizes. Not only would this show the biological relevance of any effect (something a P-value cannot do), and the authors would then be able to say if specific windows were more sensitive than others (i.e., compare the contrast between the control and each window to see which had the greatest difference).

3. Calculate potential tradeoffs between behavioural and morphological responses.
   3.1. This may just be my interpretation, but lines 342-344 seems to suggest that there may be tradeoffs between defensive responses. As I was reading the manuscript, I was wondering if a quantification of tradeoffs or 'conflicts' between different behavioural and/or morphological defences.
   3.2. I suggest the authors estimate tradeoffs between responses for each of the WGP and TGP experiments by quantifying the pairwise correlations between responses for each of the exposure treatments.
   3.3. To help with interpretation across traits and treatments, the data should be standardized (e.g., center and scale to a mean = 0 and sd = 1), and the pairwise Pearson correlation coefficients calculated. Negative correlations between behavioural and/or morphological responses could suggest antagonism or a tradeoff in the responses.
   3.4. There would be a lot of pairwise correlations, so do be careful of identifying a potential false positive, but I think this could be something useful to add to the story.
   3.5. I also want to add that calculating pairwise correlations is not the only method and simply one method I am suggesting to calculate potential tradeoffs.
4. Fantastic figure design for understanding the experimental design, statistics, and results.
   4.1. I just wanted the authors to know that I really appreciated all of the figures. Figure 1 helped me to better understand the experimental design, Figure 2 helped me understand the statistical analyses, and Figures 3 and 4 clearly presented the results.

Minor Comments

1. Tables: Number of digits for P-values could be limited to 0.001 or 0.0001, with any P-value below that threshold noted as < 0.001 or < 0.0001.
2. Lines 235-238: Looking at the code provided, the ANOVA tests were conducted using Type II sums-of-squares, and I think this should be noted in the main text.

Reviewed by:
David Murray-Stoker
Ph.D. Candidate
University of Toronto
dstoker92@gmail.com

Please do not hesitate to contact me directly via electronic mail if any of my comments were not clear or require further clarification during the review and revision process.

References (Peer Community Journal format from Zotero Plug-In)

Ben-Shachar MS, Lüdecke D, Makowski D (2020) effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, **5**, 2815. https://doi.org/10.21105/joss.02815

Lenth RV, Buerkner P, Herve M, Jung M, Love J, Miguez F, Riebl H, Singmann H (2022) emmeans: Estimated Marginal Means, aka Least-Squares Means.