Tariel-Adam et al. TGP
OSF Preprints 10.31219/osf.io/mr8hu (version 2)
Submitted to PCI Evolutionary Biology

Major Comments

1. Clarification about predictions for TGP.

Author Response: We rephrase the predictions at the end of the introduction and hope it is clearer.
DMS Comment: The revised predictions are very clear.

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.

Author Response: To test if specific windows were more sensitive than others, we added custom pairwise contrasts using only exposure windows that were significantly different from the control. These contrasts were added to the two tables. For instance, at the parental generation for the shell thickness corrected by snail size, Early, Middle, Late and Lifelong treatments were significantly different from the Control treatment, and we thus performed pairwise comparisons between these treatments only. We can therefore know if some exposure windows are different from the others, i.e. if one/several exposure windows were more sensitive compared to the others.
DMS Comment: With the addition of the pairwise contrasts, I can see how the experiment would allow for comparisons of specificity between different developmental windows.

    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).

Author Response: We agree that P-values do not say anything about the strength of the Treatment effect but we have not added the effect size for Treatment as it would not bring any information regarding the sensitive windows.
DMS Comment: As the sensitive windows are within the broader effect of treatment, it is important to know the overall treatment effect size to situate the pairwise contrasts for specific developmental windows. For example, a large effect for single contrast will not matter much if the overall treatment has a weak or negligible effect.

Additionally, effect sizes for the contrasts can be quantified as Cohen's *d*, which would standardize the difference between the control and developmental window of interest (Cohen, 1988; Huberty, 2002; Nakagawa & Cuthill, 2007). Cohen's d takes into account the magnitude of the difference between means and the pooled standard deviation. Effect sizes for contrasts can be calculated directly in emmeans using `eff_size()`, as noted previously in point 2.4 (below).

    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.

Author Response: The parameter estimates come indeed from the output of `summary(model)`. The estimated differences/contrasts between the control and any treatment are the same whether we use the summary(model) or the emmeans$contrasts because we either only have one factor in the model (the Treatment) or have a covariate which is centred around 0 (e.g. snail size) or factorial covariate (e.g. Test environment).
DMS Comment: In the earlier version, it was not clear to me how the parameter estimates were quantified. With this additional information and the updated caption for Table 1, I think the information is clear. Moreover, the estimates would be the same regardless of summary(model) or emmeans$contrasts for the reasons described by the authors.

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).

Author Response: We don't really see why effect sizes on contrasts would allow us to do this. A big contrast already means a big difference, and it is thus already a measure of the strength of the difference/contrast. We do not believe that the effect sizes of the contrasts provide more information or tests than the contrasts themselves on which window was more/less different from the Control treatment. But we agree that we needed to test for that and that is why we added the custom pairwise contrasts.
DMS Comment: I have touched on this point above (response to 2.3), but I think it is important to put the treatment effects in context. The authors have clarified how they can say if some development windows are more sensitive than others; however, unstandardized effect sizes - such as contrasts - do not represent how much variation is explained by the treatment (treatment effect size) or between the control and a development window (contrast effect size).

Without standardizing the treatment and contrast effect sizes (i.e., accounting for the explained variance), a large contrast could just be an artifact of the sample size. Cohen's *d* on the contrasts would show if windows were more or less sensitive compared to the control and the strength of that effect (Cohen, 1988; Huberty, 2002; Nakagawa & Cuthill, 2007). At present, the contrasts are just the result of a t-test, with the interpretation constrained by null hypothesis significance testing.

3. Calculate potential tradeoffs between behavioural and morphological responses.

Author Response: We added tests for trade-offs in the supplementary information. They did not reveal any links/trade-offs between morphological and behavioural defences. We did not perform Pearson correlation tests as our 'refuge use' variable is a binomial variable and, as you said, it would have required a lot of correlation tests (correlation between all morphological and behavioural defences for all treatments). We thus realised two linear models for each generation with the behavioural defences as the Y variable (refuge use or time to reach the refuge) and the first and second axis from a PCA done on all morphological variables. We also included interactions between this pc1 et pc2 with the Treatment fixed effects
DMS Comment: I really appreciate this additional set of analyses, and I know they would not have been easy to do. I think confirming there were no links or tradeoffs between morphological and behavioural defences makes the study more robust.

Reviewer Summary: I only have reservations about points **2.3** and **2.5**, as detailed in my comments.

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)

Cohen J (1988) *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, N.J.

Huberty CJ (2002) A history of effect size indices. *Educational and Psychological Measurement*, **62**, 227–240. https://doi.org/10.1177/0013164402062002002

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x