Response to the Recommender and Reviewers

Dear Benoit Pujol, David Murray-Stoker, Willem Frankenhuis and Timothée Bonnet,

We would like to thank you very much for your highly valuable reviews and suggestions that have greatly improved the article. It was really nice to receive such positive, clear, helpful and constructive comments. We have taken into account all suggestions in our manuscript and responded point by point below in blue font colour.

We have notably made substantial changes in the discussion to better discuss the existence of sensitive windows of WGP and TGP and the validation or not of our predictions. We have especially taken care to compare our findings to the theory. We also added additional statistical analyses to test for sensitive windows (custom contrast pairwise tests) that are included in the two tables and in a new summary figure.
Finally, we added statistical tests for trade-offs between morphological and behavioural defences in Supplementary Information. And a justification for the duration of exposure in the Material & Methods.

We hope you will find our modifications correct and valuable.

Thank again,

Juliette Tariel-Adam, Sandrine Plénet and Emilien Luquet

Recommender – Benoit Pujol

Dear Dr. Tariel-Adam and collaborators,

Thank you for submitting your preprint to PCI Evolutionary Biology for recommendation. I have read your preprint entitled “Sensitive windows for within- and trans-generational plasticity of anti-predator defences” and I have read the comments of the three reviewers who have complementary expertise on your research topic.

The three reviewers and I agree about the quality of your research and the value of this preprint. The paper is clearly written, straight to the point, and Figures participate actively to make experimental approaches easy to understand. The reviewers and I found that your research targets a valuable question on phenotypic plasticity and its adaptive significance when considering that it may be restricted to specific developmental windows that are themselves, as you state in the abstract: “highly sensitive and responsive to environmental changes". You will see in their reviews that although the three reviewers appreciated your work, they raised some concerns that I would like you to address. I have an additional request, which is to adda a paragraph at thei end of discussion on the contribution of your findings to the theory. In the introduction, you present the state of art with strong bibliographic support and outline the hypotheses that are tested in this paper. As a result, we understand what is at stake in terms of contribution of your work. In the discussion, you discuss the proximal conclusion that can be drawn from your results but do not discuss the implication of your findings to the actual theory. Such discussion would add substantially to the scope of this work.

Thank you very much for organising the peer-reviewing process and finding good reviewers. We have better discussed the validation (or not) of our predictions and the contribution of our findings to the theory. We have chosen to disseminate this information during the discussion rather than in a dedicated paragraph.

Looking forward to reading the next version of this promising paper.

Regards,

Benoit Pujol

Reviewer 1 – David Murray-Stoker

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.

We rephrase the predictions at the end of the introduction and hope it is clearer.

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.

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.

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/

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.

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.

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).

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).

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.

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.

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.

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.

Thank you very much!

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.

Done.

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.

This precision has been added in the section Statistical analysis in the M&M.

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.

Reviewer 2 – Willem Frankenhuis

2023 PCI Evolutionary Biology: Willem Frankenhuis review of "Sensitive windows for within- and trans-generational plasticity of anti-predator defences".

My expertise is well-matched to theoretical aspects but not to empirical aspects of this paper. Hence, before accepting to review, the editor and I have discussed limiting the scope of my review to theoretical aspects of the paper.

I enjoyed reading this well-written and well-structured paper describing an experimental study, with a rich variety of measures, that addresses an interesting question: how does the developmental timing of exposures to predator cues influence the development of anti-predator defences in the freshwater snail *Physa acuta*?

In the study, the F1 parents (parental generation) were randomly assigned to one of 6 treatments: 5 treatments with exposures to predator cues at different developmental windows: in the embryonic stage, in the early/mid/late post-embryonic stages, and lifelong (embryonic and all post-embryonic windows). Treatment 6 is a control condition with no exposure to predator cues in all of these windows. Following these treatments, the researchers quantified “Within-Generational Plasticity” by measuring 6 defences: 2 behavioral (refuge use, time to reach the refuge) and 4 morphological (snail mass, shell length and width, shell thickness, shell crush resistance). The F2 offspring (offspring generation) were reared in water without any predator cues during all windows; so, all of the

offspring received the same treatments. The researchers then quantified “TransGenerational Plasticity” by measuring the offspring’s defences.

Despite being very positive, I have some questions about conceptual framing, rationale for experimental design, and data interpretation. I have provided a few suggestions that the authors may consider to strengthen their excellent paper. I think all of my questions can be addressed. My comments about the conceptual framing are largely about the semantic fit with other theoretical and empirical reports in the field and may be easily addressed. My question about rationale for an aspect of the experimental design may also be easy to address. My suggestion to consider conducting an additional statistical analysis, if the authors choose to incorporate it, would require more substantive work, yet is feasible. I think it would be a worthwhile addition. Per the reviewer guidelines, I should add that I have not identified any flaws in the study.

1. Definition of sensitive windows

The term “sensitive period” is often, perhaps typically, used to refer to a developmental window in which the impact of a given experience on a particular phenotype is greater than the impact of that same experience on that phenotype in other developmental windows. The authors use a similar definition: “the same environmental change may generate different phenotypic effects depending on whether it was experienced early or late in development. Certain developmental windows are particularly sensitive to environmental changes, i.e. environmental changes during these sensitive periods generate particularly strong effects on phenotype” (lines 34-37). Whereas the former definition applies generally to any cue (i.e. any observation that reduces uncertainty about the environment or the organism), the latter definition focuses specifically on cues reflecting environmental changes. However, cues may also indicate that environmental conditions have remained the same (e.g. predator density is high from one time period to the next). I wonder whether this focus on a subset of sensitive periods is deliberate; if not, it can be easily addressed by in the authors’ definition replacing the word “changes” with “cues”, which is more general. This difference is subtle and perhaps it is unlikely to confuse readers, but I would like to mention it just for the authors to consider.

We agree that there is a difference between cue and change and it was not our intention to focus only on cues reflecting environmental change. We are now using the terms “cue”, “environment” or “environmental state”.

2. Uniform plasticity

More important is the question how to describe cases where the evidence does not reveal differences in plasticity across developmental windows. Based on the widely used definition I provided earlier, I would describe such cases as providing a lack of evidence for a sensitive window (consistent with uniform plasticity). By contrast, the authors write that the sensitive window is very wide: “Overall, none of the exposure windows was particularly sensitive compared to the others. The sensitive window for within- and trans-generational defence induction was thus very wide, which could be the result of the strong selective pressure imposed by predation at all developmental stages” (lines 11-12). This wording is fine, in principle, yet I worry that it may confuse some readers, if they would not refer to such cases (of uniform plasticity) as ‘a sensitive window’. The authors may consider slightly rewording,

for instance, as follows: “We find no evidence for a sensitive window; rather, snails responded to a similar extent across all developmental periods.” That said, the authors’ and my own wording do not capture one important way in which the data “do” support the existence of sensitive windows, discussed next.

We agree with your comment. When we wrote the manuscript, we discussed for a long time how to describe the results: as an absence of sensitive windows or the presence of a very wide one. We followed your suggestion (a lack of sensitive windows rather than a very wide sensitive window) and changed the discussion accordingly.

3. Global and specific patterns

Irrespective of whether one prefers to summarize the overall patterns as providing evidence for a “wide” sensitive window or as providing “no support” for the existence of a sensitive window (see comment 2), as the authors note, the impact of predator cues depends on their developmental timing. The Discussion section summarizes well: “the defences expressed were different depending on the window of exposure to predator cues for both WGP and TGP. At the parental generation (WGP), all exposure windows induced the expression of morphological defences; but two morphological strategies can be differentiated between exposures at embryonic or post-embryonic stages and only post-embryonic exposures altered behavioural defence. At the offspring generation (TGP), all exposure windows induced offspring to use refuge, but only certain windows altered morphological defences. These results confirm that the developmental windows at which environmental cues were perceived is an important factor driving both WGP and TGP” (lines 306-313).

In my view, if researchers employ the widely used definition I described earlier, these findings “do” support the existence of sensitive windows: in F1, “only post-embryonic exposures altered behavioural defence” (line 310); in F2, “only certain windows altered morphological defences” (line 311). The impact of a given experience (predator cue) on a particular phenotype (behavioral and morphological defences) is greater in some developmental windows than others. I think this finding will be of much interest to the community and deserves to be made more explicit, particularly in the abstract, which currently states, in my view less clearly: “Although all exposure windows impacted on the expression of offspring defences, they did so differently and the response patterns were complex for morphological offspring defences. This complexity in the temporal dynamics of transgenerational induction could result from the transmission to offspring of both cues about predator presence and parental somatic condition” (lines 14-18).

When I first read the abstract, I mainly took away that all predator exposure windows induced within- and trans-generational responses, without a clear view of the interesting fine-grained patterns—that the impact of a given cue on a particular phenotype depends on the developmental window in which exposure to the cue occurs (as summarized well in the above quote, lines 306-313, from the Discussion section). My hope is that this will be more clearly reflected in the abstract. I should add that I very much appreciated the authors’ thoughtful reflections on potential adaptive reasons for these fine-grained patterns (described in the paragraph beginning on line 331).

We also discussed that point during the writing of the manuscript. The definition of sensitive windows does not specify at which scale one looks at the phenotype: the "whole" phenotype of an individual? The defensive phenotype (all defences)? A single phenotypic trait? We followed your suggestion and focused on a fine-grained scale. We just speak briefly about the global scale in the conclusion at the end of the discussion.

4. Rationale for predictions

The authors note: “For WGP, we predict that embryonic and early post-embryonic windows should be particularly sensitive. For TGP, we predict that (1) one sensitive window of the TGP should be the late parental development; and (2) other sensitive windows of TGP should be those of WGP and so should also be early-life windows” (lines 104-107). I recommend making the grounds for each of these predictions more explicit. As it stands, the introduction discusses a relevant selection of theoretical and empirical research, but does not seem to make the link between this research and these two predictions explicit. Are both predictions based on both the theoretical and the empirical research? Or is prediction 1 based primarily on theory and prediction 2 on the empirical record?

We have reworded the predictions at the end of the introduction and added more explicitly the empirical and/or theoretical rationale behind the predictions throughout the introduction.

5. Statistical power

I am not a statistical expert, but to my understanding, the authors have conducted null hypothesis tests that afford the following inference about the overall pattern: the data provide no evidence for differences in the extent to which the phenotype responds to cues across developmental windows. In addition, it would be interesting to quantify the extent to which the data support the hypothesis that the phenotype responds equally to cues across developmental windows. A Bayesian analysis can address this question and may have merit for two reasons. First, theoretically, it is valuable to know whether the data are more likely under the hypothesis that the phenotype responds equally to cues across developmental windows (H0) versus the hypothesis that the phenotype responds differently to cues across developmental windows (H1). If such an analysis would show, for instance, that the data are 100 times more likely under H0, this would constitute evidence for the absence of a sensitive window (or, for a wide sensitive window). Second, an analysis may not lead us to reject the null hypothesis—i.e., the data provide no evidence for differences in the extent to which the phenotype responds to cues across developmental windows—for different reasons: the null hypothesis may be a good model, or the study may not be adequately powered to detect anticipated effect sizes. The current study included 240 F1 and 240 F2 adult snails (40 snails x 6 treatments). I do not have the expertise to evaluate whether this sample size implies adequate power for the current design. Regardless, I think it would be interesting to know whether, and if so to what extent, the data support H0.

If we understood well, you are speaking when we have a significant effect of “Treatment” and we would like to know whether the phenotype responded equally to cues whatever the exposure window was. We added custom post-hoc contrasts to test this in the two tables. These custom post-hoc pairwise contrasts use only exposure windows that were significantly different from the control. 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.

However, these tests only provide evidence that there are one/some exposure windows that are different from the others. They do not provide evidence that these exposure windows are all the same when all pairwise comparisons are non-significant. We cannot use power analysis as we have recently read that they are not informative in that case "All authors of this book are in agreement: post hoc power analysis is useless. Many statisticians have made this abundantly clear (Gelman 2019; Althouse 2021). […] 'On the basis of my observed effect size, did I have enough power to reject my null hypothesis?'. To avoid needless suspense, the answer is always: no." (see section 1.6 Post hoc power analysis at https://aaroncaldwell.us/SuperpowerBook/introduction-to-power-analysis.html). We could carry out a Bayesian analysis as you suggested, but this is a huge job that requires a thorough revision of the manuscript.

6. Duration of predator cues

Whereas the F1 parents were exposed to predator cues for 5 days (out of 7 days) in the embryonic stage, they were exposed to predator cues for 14 days (out of 42 days) in the early, middle, and late post-embryonic stages. There may be good (biological) reasons for this variation in the duration of exposure to the predator cue, but I did not see these stated in the paper (I'm sorry if I missed them). I think it would be good to explain why this design is preferable over, for instance, one where the F1 parents were exposed to predator cues for 5 days (out of 7 days) in the embryonic stage, 5 days (out of 14 days) in the early post-embryonic stage, 5 days (out of 14 days) in the middle post-embryonic stage, and 5 days (out of 14 days) in the late post-embryonic stage. If the global pattern supports uniform plasticity (see comment 4) "despite shorter exposures" to the predator cue in the embryonic stage compared with the post-embryonic stages, one could even argue that the embryonic stage is more sensitive to the cue than the other stages.

We have added the rationale behind the choice of exposure duration in the 'Experimental Treatment" section in the M&M, it was indeed missing information in the manuscript.
The durations of the exposure windows are relative to the development duration of *Physa acuta* at 25°C. We chose the durations to maximise the exposure duration for each developmental stage.

- The minimum duration of embryonic development is 7 days. We wanted to expose snails in the Embryo treatment during their whole embryonic development. We decided to expose snails only for 5 days to make sure no hatchlings were exposed to predator cues in the Embryo treatment.

- The minimum duration between hatching and sexual maturity is 33 days (from our personal observations). We waited 4 days after the hatching of the first snail until all snails had hatched. We then divided the remaining 29 days into two: 14 days for the early post-embryonic exposure and 14 days for the middle post-embryonic exposure before sexual maturity. Then, 14 days for the late exposure when snails started to be sexually mature to be consistent with the early and middle exposures.

7. Rate of environmental change

Mathematical theory on the evolution of sensitive windows suggests that the rate of environmental change, relative to generation time, may influence the width of sensitive windows (see Walasek et al. 2022, below). From this perspective, it would be interesting to have estimates of the rate of environmental change in predator density (or a related measure, like abundance) in the wild population from which the snails in the study were sampled. If such information is currently not available, this may be a valuable direction for future research. This information would also be relevant to the authors' theoretical claim in the Introduction section that "Late-perceived cues are indeed the most reliable about offspring environment because of the short time lag between cue exposure by parents and phenotypic selection in offspring, implying a low likelihood of environmental change if the environment is auto-correlated over time" (lines 53-55).

We added a paragraph to discuss the link between the relative rate of environmental change and the width of TGP sensitive windows in the discussion.

However, we don't quote Walasek et al (2022) in this paragraph as we don't think the model you used can explain the sensitive windows of TGP. In your model, the reliability of cues available to individuals changes across development. Parental cues are (most of the time) transmitted at the beginning of offspring development (age 0). Their reliability should change equally across offspring development whether they were highly reliable parental cues at offspring age 0 or poorly reliable. In other words, the change in parental cues' reliability across offspring development should not be linked to when these cues were perceived during parental development. Parental cues' reliability is likely to decrease over offspring development regardless of when parents perceived them. In addition, if we understood well how models work, parental cues just act as a prior and are not available anymore to offspring across their development.

8. Figure 1 is great

Figure 1 is offers a very clear and informative overview of the experimental design! My only suggestion would add to write the number of days of exposure to predator cues directly below the thick red line (e.g., in the embryonic stage, you could add "5 days"). Though this information can be read from the x-axis, it will be helpful to have it directly in view. I don't think adding this information will clutter the figure (e.g. if the font size is similar to that of the text in the figure stating "No exposure to predator-cues"), but the authors can make this call.

Figure 1 has been changed to add the number of days of exposure for each treatment.

9. Additional literature

I would like to end my review by suggesting some additional literature that the authors may consider incorporating. These papers may have the potential to enhance the paper but are not essential to include. Thus the authors should feel entirely free to include, or not include, these papers as they see fit:

Groothuis, T. G., & Taborsky, B. (2015). Introducing biological realism into the study of developmental plasticity in behaviour. Frontiers in Zoology, 12, 1-14. https://doi.org/10.1186/1742-9994-12-S1-S6

Smallegange, I. M. (2011). Complex environmental effects on the expression of alternative reproductive phenotypes in the bulb mite. Evolutionary Ecology, 25, 857-873. https://doi.org/10.1007/s10682-010-9446-6

Stamps, J. A., & Luttbeg, B. (2022). Sensitive period diversity: Insights from evolutionary models. The Quarterly Review of Biology, 97, 243-295. https://doi.org/10.1086/722637

Uller, T., Nakagawa, S., & English, S. (2013). Weak evidence for anticipatory parental effects in plants and animals. Journal of Evolutionary Biology, 26, 2161-2170. https://doi.org/10.1111/jeb.12212

Walasek, N., Frankenhuis, W. E., & Panchanathan, K. (2022). Sensitive periods, but not critical periods, evolve in a fluctuating environment: A model of incremental development. Proceedings of the Royal Society B, 289, 20212623. https://doi.org/10.1098/rspb.2021.2623

This literature was really helpful, thank you. We added citations to Groothuis & Taborsky (2025), Stamps & Luttbeg (2022) and Walasek et al. (2022) in the introduction and discussion.

Signature

I hope my review will help the authors to strengthen their manuscript.

For accountability and transparency, I would like to sign my review.

Willem Frankenhuis

Reviewer 3 – Timothée Bonnet

This manuscript presents an experiment trying to identify sensitive windows for plastic responses to predator cues, within and across generations, in a freshwater snail. Overall the experiment and data analyses appear neatly done, while the interpretation seemed fully appropriate. I had only two somewhat substantial comments that require clarifications. The writing was mostly clear, to the point and enjoyable, although I pointed to a few questionable choices of vocabulary or syntax.

Main comments:

* How where the different treatments organised in space and time (and with respect to any other potential experimental factor)? I did not find any information about this, but I think it is crucial to properly interpret the results. Some non-random experimental factor correlated with treatment could explain some of the apparently idiosyncratic results (e.g., fluctuation of direction of the effect along treatments on figure 4 F). If there is information about experimental factors, those could also be included in models in order to account for blocking and perhaps decrease noise and reveal more of the treatment signals.

We did not include any blocking effect as we did not have any blocks. The snails were randomly assigned to new aquariums every two weeks to make sure all aquariums had the same snail density (for example, if an aquarium suffers low mortality, the snail density is lower. Snail shuffling helps to maintain a constant density between aquariums). There was thus no aquarium block effect.

The snails were kept in aquariums, which were placed on plastic trays. All aquariums within a tray were of the same treatment, and there were multiple trays for each treatment. Trays were rotated at each water change to avoid any temperature gradient, which occurred twice a week.

All treatments were synchronised in time.

* I think individual random effects are necessary to analyse the "refuge use" variable which has four measurments per individual. The authors should modify their models or explain why they think it is not necessary to account for repeated measurements.

In our model, the Y variable for refuge use was the number of times the snail was out of the water, out of the total number of tests n. In R, our model looks like this:

model1 <- glm(c(Y, Y - n) ~ treatment, data = data, family = “binomial”).

Because the response variable Y represents all of the ‘refuge use’ data for a single snail, there is no need to account for repeated measurements in this model.

We noticed that the confusion stemmed from Figure 2, which presented an incorrect model equation. Therefore, we have updated the figure to reflect the correct equation.

We could have indeed done another model where the Y variable is a measure of refuge use (Y = 1 if the snail is out of the water during the test, 0 otherwise). In R, the model would have looked like this:

model2 <- glmer(Y ~ treatment + day.of.measure + (1 | snail.ID), data = data, family = “binomial”).

When I first tried the model 2, the random part was not “converging” with the following error message “boundary (singular) fit: see help('isSingular')”. I tried the package you advised and the model converged. Thank you for advising on this great package. We got the same results as model 1: No significant effect of treatment at the parental F1 generation and a significant effect in the offspring F2 generation. We got a significant effect of the day of measurement at the parental F1 generation but not at the offspring F2 generation.

Specific comments:

L.111 A short summary of the species reproductive system could be useful.

A short summary of the biology of *Physa acuta* has been added to the Supplementary Information.

L.116 "bowled" - "boiled"?

Done.

L.147 "We wanted" - "We waited"?

Done.

L.167-170 and L.201-202. From what I understand, each individual was measured four times for "refuge use". It seems necessary to account for individual repeated measurement (most likely using a random effect). Using the package lme4 that GLMM be done with glmer, but I would recommend glmmTMB (which tends to perform better with binary GLMMs; it is also compatible with emmeans).

See the response above.

L.231 "at both generations and removed from the model" - "for either generations so we removed it from the models"?

Changed to “The interaction was not significant at either generation and was therefore removed from the model.”

Figure 3 caption. "confidence interval" at which level? 95% as is the default in emmeans?

Added “95% confidence interval” in Figure 3 and Figure 4 caption.

L.305 "Contrary to what expected" - "Contrary to what we expected"?

Done.

L.323 "snails during embryonic" - "snails exposed during embryonic"?

Done.

L.333-336 Alternatively, the decreased escape behaviour could be a maladaptive change, either as a correlated side effect of another response, or an idiosyncratic plastic response (plasticity is often not adaptive!).

The decreased escape response could indeed be a correlated side effect of the response on morphological defences. We looked at the tradeoffs between morphological and behavioural responses (see Supplementary Information) but found no evidence. We also added a sentence that this decreased response might be a maladaptive change L408.

L.336 "may have orient snails to a developmental pathway" - "may have directed / lead snails toward a developmental pathway"?

Changed to “may have directed snails toward a developmental pathway”

L.338 "orient snails in" - "directed / lead snails toward"?

Changed to “while post-embryonic exposures may have been too late to direct snails towards the highly accelerated growth pathway”

L.366 "infirm" - "contradict / disprove / refute" (infirm = cripple)

Changed to “This result contradicts our first prediction that”

Timothée Bonnet (I sign all my reviews)