Dear Ambre Salis,

Thank you for the revision of your manuscript “Acoustic cues and season affect mobbing responses in a bird community”. I have examined the revision as has an independent reviewer (Reviewer #2 from round 1), and we are both pleased with your revisions, but we both also have some concerns that need to be addressed.

***** We thank you and the other reviewer for all the helpful commentaries and advice on our manuscript. We made changes in all parts of the manuscript, with a special emphasis on the statistical analysis and the discussion. We hope you will consider this new version to be appropriate for a recommendation in PCI.

I have three primary concerns.
1. My first concern relates to transparency of your reporting. You are reporting results only from your top models, and this practice leads to bias in the literature. Switching from frequentist to Bayesian model selection does not remove the need to report the outcomes of all models. Bias can emerge from selective reporting of top models selected through Bayesian methods as it can from frequentist model selection. Please include two additional tables (possibly in a supplement):
A. A table with the model rankings and BIC etc for the full set of models you examined
B. A table that includes the parameter estimates (and SE etc.) for all models tested.

**** As suggested, we added the table with the ranking of the models tested (Table 1) and the parameters estimates of all models (Table 2 + Sup.Mat). This concerns 24 models since we changed a little the way we selected the best models, as suggested below by the reviewer.

2. My second concern involves your interpretation of your results. You make several statements about the interpretation of your results that I think many readers will feel are too confident. For example:

Abstract: “Our results therefore support the hypothesis that birds consider both the species and the number of callers when joining a mobbing chorus in winter”

Although it is plausible that your method induced a differential response due to the perception that number of callers varied, with the current design, it seems we cannot rule out an explanation of duty cycle alone. However, I think you can build a more persuasive case that your results are not due to duty cycle alone. I write more about this below.

Discussion: “These results corroborate the hypothesis that a greater number of birds mobbing a predator represents a lower risk for a potential mobber”

I also think this claim is too strong (because there was no experimental manipulation of risk).

Let's first consider the duty cycle argument. You make the case that overlapping of calls in playback reduces the chance that calls appear to be a single bird, which is plausible. However, that remains an untested assumption, and the alternative, that birds simply respond more often and more strongly when the signal is stronger is also plausible based on basic ethological theory. The duty cycle explanation is also a more parsimonious explanation. Of course, the duty cycle argument is not inconsistent with a role for selection in favoring responses to larger groups of mobbers, but assessing a role for responding based on group size would require different evidence. Ideally this evidence would be experimental and would control for duty cycle. However, because results differed between species between seasons, you could argue that birds are not simply responding to the signal with the stronger duty cycle – they do so in some cases, but in other cases, they do not. So, I think you should (a) make this case explicitly, (b) make sure you acknowledge, where relevant in the abstract and discussion, that duty cycle was a confound that limits the strength of your inference, and (c) suggest particular experimental designs (in the discussion) that would control for the duty cycle confound.

Now on to the risk argument. If we believe the argument I just put forward in the prior paragraph, that the response differences you observed between seasons/species suggest that mobbers are doing something more than just responding differently in response to different signal intensity (duty cycle), we still don't have evidence regarding a role for risk in this behavior. It would be acceptable to present differential risk as a potential explanation for the outcomes you observed, but I would do so with more caution than you currently present. I would also suggest that you consider what evidence might be informative regarding the risk hypothesis. For instance, maybe you could discuss an experiment that used mounts of different predator types with different threat levels or in locations in which they posed more or less of a threat, or provided different amounts of cover to the mobbers, or possibly some other manipulation that would allow you to assess if varying the threat level varies the mobbing behavior. Regardless, please be more cautious overall regarding your discussion of the applicability of various causal hypotheses.

**** We agree with referee, and we therefore modified two aspects of the discussion (lines 380-419 + few sentences in the main text and abstract), following the advice above. Specifically, we modified our sentences to make it clear that even if we recorded a higher response toward playbacks with three callers, birds may in fact be focusing either on the individual voices and/or the duty cycle when responding to these playbacks. We hope this new version is now clearer regarding this point. Secondly, we agree that we did not consider other explanations as the risk assessment hypothesis when we wrote our discussion. We therefore modified the paragraph talking about risk assessment to better emphasize that this is only one way of looking at it.

3. My third concern is that one of your primary inferences involves a comparison between responses in winter and spring, but you do not actually test for this difference. You make

the case that the responses in winter and spring are not comparable, but it is not clear to me that this is the case. If the overall response rates differ between seasons, then you can just fit different intercepts or slopes for different seasons in your model.
At this point it seems that you have several options. The simplest would be to explicitly acknowledge, in the abstract and in the discussion, that you did not compare winter to spring responses statistically, and that you are making only a qualitative assessment that the patterns differ. Another option could be to conduct a post-hoc test based on a single model that included both winter and summer data to quantify the effect of season (and especially interactions with season) on the mobbing response rates of interest. The most thorough option would be a model selection procedure like the one you already conducted, but including both winter and spring data and season as a predictor variable in various forms (including presence and absence of interactions with season). I would find any of these three options reasonable.
*** This point was indeed a strong discussion when writing the latest versions of the manuscript. We would like to keep the first option, i.e., not comparing with statistical tests these two seasons, and keeping the Spring tests as a replicate to nuance our results from winter. As suggested, we added in the discussion one clear paragraph to state this more clearly (lines 443-450).

I also want to call attention to several of the suggestions made by the reviewer in this round of reviews. 1. The reviewer had concerns regarding your implementation and reporting of hurdle models. Some of these concerns can be addressed by more thorough reporting of your methods. You should consider and respond to all of these comments.
2. It is my assessment that your reporting in Table 1 is correct (you have not reversed the occurrence and intensity results as suggested by the reviewer), but please check and be certain.
3. The reviewer asks for justification for why you include different terms in different components of the model (zero inflation vs. intensity) – this problem may be partly addressed by more thorough model reporting (as I mention above), but also suggests that you should devote more space to justifying the set of models you assessed.

**** Regarding point 1, we indeed took extra care in responding to reviewer's advice on the statistical analysis and reporting, we hope everything is now clearer. Regarding point 2, we indeed mixed the occurrence and intensity parts. This error is now fixed. Regarding point 3, we wrote a complete response below.

Below are my specific suggestions for edits by line number.

36-39: please acknowledge the duty cycle difference between stimuli somewhere around here in the abstract so that the reader understands that number of callers was confounded by number of calls in the stimulus recordings.
**** We added one sentence in the abstract to precise this information (lines 33-36).

41: Change “context interacts can strongly affect” to “context can affect”
**** Sentence modified (line 42).

95: I suggest you change “stability of acoustic cues” to “stability of response to acoustic cues”
**** Sentence modified (line 96).

96: I suggest you change “as much as” to “as well as”
**** Sentence modified (line 97).

108-109: This sentence
“Each spot was selected close to a tree allowing birds’ approach and concealment of experimenters, following existing trails.”
is somewhat confusing, and would be clearer as something like:
“Each spot was selected along an existing trail but close to a tree allowing birds’ approach and concealment of experimenters.”
**** Sentence modified (lines 118-120).

128-129: some clarification would be valuable here. Were the four tests at each spot carried out on the same day spaced 5 minutes apart, or was each consecutive test at a different spot, and each of the four tests at a spot within a season conducted on a different date? I think it was the second, but please clarify.
**** We added this clarification lines 140-141.

166: “NW-A45 Sony” - is this the speaker or ??
**** This is the MP4 player, we clarified the sentence accordingly (line 178).

175: Thank you for adding this clarification: “The two observers agreed on the lowest number of birds seen simultaneously by both experimenters.”
However, I think the wording needs revision. Wouldn’t this be the “highest number of birds seen simultaneously by both observers”? If the observers each saw 1 bird at the same time, and then a few seconds later, each saw 2 birds simultaneously, wouldn’t you count this a ‘2’ (highest) rather than ‘1’ (lowest)?
**** Reviewer is right, our sentence was not correct, we modified it (lines 187-188).

180: This statement: “Since the number of responding birds during the winter cannot be strictly compared to the one observed during the spring” is not informative, and is not obviously true. If you choose to keep the spring and winter analyses separate (see my comments about this above), I suggest you change the wording here to something like:
“Since social conditions for our study species differ between winter and spring and factors influencing rates of response presumably therefor differ …”
**** We agree that our sentence was not clear enough; we changed the wording as suggested (lines 192-193). We therefore chose the option to keep the analysis separated for the two seasons, as suggested above.

182: I suggest a change from “at the community level…” to something like “of any species (“community level”)…”
**** We agree with this modification (line 195).

186: insert “us” before “to take”
**** Not applicable anymore.

186: insert “an” before “excess”
**** Word added (line199).

187: change to “zeros”
**** Mistake corrected (line 200).

188: change “a first” to “an initial”
**** Sentence modified (line 205).

188: change “determine” to “determines”
190: change “determine” to “determines”
**** Not applicable anymore.

195-198: I agree with the reviewer that more information is needed here
**** We agree with this comment, we added a clearer statement on the choice regarding the different distribution chosen in our hurdle models (lines 213-228).

198: I found this confusing. To help clarify, I suggest you change “the one of the number of callers and…” to “the effect of the number of callers, and …”
[note word change and the addition of the comma]
**** We modified the sentence accordingly (line 207).

210: change “calculate how much better is the best model compared to the other ones” to “calculate how much better the best model is compared to the other ones”
**** Not applicable anymore.

212: although I agree that delta AIC or delta BIC is often used a threshold, it is rarely referred to as a ‘significance’ threshold, so you may not wish to use this word. Instead you might say something like “models with a delta >2 are commonly considered to have substantial support …”
**** Yes, it was a poor and confusing terminology (referring to frequentist approach while using multimodel inferencing according to information theory). We finally chose to rely only on the evidence ratios.

293: some additional information would be useful in these figure headings. Close inspection and consideration leads me to conclude that categories are stacked in each bar (rather than layered, so for instance, the number of crested tits in spring responding

to the 1CO treatment is about 25 rather than slightly more than 50. If I am correct, then wording could be added as a new sentence on line 295 to say something like "Responses to each of the four treatments are stacked in sequence on each bar so that the entire bar represents the sum of all responses by a given species per treatment"
**** We agree with this comment and added the sentence suggested (lines 317-319).

324: somewhere (maybe in a supplement), you should include the full set of models examined and their model ranking statistics (BIC etc.) AND ALSO their associated parameter estimates and corresponding standard errors.
**** As suggested, we added the whole set of models and their parameter estimates in our manuscript. As suggested by the reviewer below, we changed our method: we only ran four possible models (number of emitter and emitter species with or without the interaction in both parts of the Hurdle models). We therefore chose to report the estimates of the best model in the main text (Table 2), and the three other models as a supplement.

326: clarify what 'further reduced' means (your method of model selection)
**** Since we changed our model selection strategy, we modified this sentence accordingly and this commentary is no longer meaningful.

327: change "the one of the number of callers as well as" to "the effect of the number of callers, as well as"
**** Sentence modified (line 348).

353: I think the word "corroborate" here is not ideal. I would prefer something like "These results are consistent with the hypothesis…"
I prefer this wording because your results are equally consistent with other hypotheses. For instance:
(a) a stronger signal is more likely to be detected by potential additional mobbers and is therefore more likely to attract more mobbers
(b) a stronger signal is more likely to reach the threshold necessary to trigger mobbing in an individual.
**** The discussion paragraph was rewritten for a more convenient reading. We added the ideas suggested above (lines 380-399).

I will point out that both (or either) of these two hypotheses (a and b) could be true while your hypothesis is true, but both a and b could be true while your hypothesis is not true. I think it would be useful to discuss what additional evidence you would want to examine to evaluate the plausibility of your relative risk hypothesis.
**** We added one idea of improvement to better test the risk hypothesis (lines 396-399).

398: I suggest changing "opposition" to "contrast"
**** Word changed (line 500).

402: I suggest changing “aggressivity” to “aggressiveness” (here and elsewhere in the paper)
**** We modified the words in the main text.

406: This would be more clear to the reader if you again explained how “occurrence” differed from “intensity” here (like you do in the figure 2 and 3 headings). Otherwise, it is not immediately obvious to the reader how what you have written here differs from what you wrote on line 400.
One way to do this would be to write something like “Additionally, not only did fewer individuals respond in spring than in winter, but in spring, the proportion of locations with any response was lower than in winter”
**** Sentence changed (lines 458-459).

407: I think you mean “populations” (plural) here since you are taking about the populations of multiple species
**** Yes, reviewer is right, we added the plural (line 461).

441: change “despite our” to “despite the fact that our”
**** Words added (line 493).

449: change “in adequacy with” to “consistent with” or “similar to”
**** Word changed (line 501).

455: add comma after “We have” and after “however”
**** Commas added line 507.

457: change “are” to “is” (because “status” is singular)
**** Mistake corrected (line 509).

459: delete “very” (it is not necessary to make your point)
**** Word deleted (line 513).

Reviews

Reviewed by anonymous reviewer, 21 Nov 2022 04:04
21 November 2022

I thank the authors; they have done great work addressing all the concerns and queries I raised in the first round. The statistical method used in conjunction with the experimental design (factorial design) is now appropriate. The preprint organisation has also been improved to a level the reader could easily follow.

I have made a couple of suggestions to improve the presentation of results and have a

query of the models in Table 1. Otherwise, the preprint is scientifically sound and may require editing before publishing in PCI or any other journal.

Major comments:

In Table 1, mobbing responses are listed under two different response variables: response occurrence and intensity. As far as I understood, and what I see from model syntax in the R script, the zi- zero inflation part reported under the mobbing intensity.

**** Reviewer is right there was a mistake; we inverted the two columns. Table 1 is not modified and implemented with all the information needed.

In fact, the zi part of the model helps to define what variables contribute to zero inflation; it could be one variable than the other, or both variables equally contribute to zeros inflation. As the authors mentioned, there is no distinction between true and false zeros in Hurdle models. However, for example, the authors need to justify why they think that, Emitter species + number of callers contribute to zero inflation in one model and, why only the Emitter species contribute to zero inflation in another model (this is an example, it may apply to all the models presented in the table).

In addition, authors may consider defining theta in the table caption or the statistical methods section where appropriate; otherwise, people who think in a Bayesian way might be confused with parameter estimates.

**** We believe that we may not have explained clearly enough the rationale to use a Hurdle model instead of a zero-inflation model in our latest reply. We would like to take the time to explain better our choice:

- In ZI models, the occurrence of zeros results from the mixture of structural zeros and sampling ones, while in Hurdle models, the occurrence of zeros is assumed to result only from structural zero. Hurdle model is actually a two stage process and can be modelled by two separate GLMs (in case of Poisson distribution), considering that a first process generates the structural zero using a Bernoulli probability mass function, and that the second process generate the frequency of non-zero events using a left truncated probability mass function for count data. Since we can confidently assume that focal species were present at each site, we can therefore assume that the absence of response at a given site is not caused by the absence of exposed birds at that site. As noted by Feng (2021; https://doi.org/10.1186/s40488-021-00121-4), Hurdle models are usually chosen in such a situation (see also in parasitology, when we assume that a first process generate the zero, i.e susceptibility/resistance, and the second process generate the infective load in susceptible subject). We acknowledge that we made this choice only on the conceptual viewpoint but not on the statistical ability (i.e. fat of sampling zero compared to structural zero for ZI model, or zero deflation for

Hurdle model). We rewrote the method section to clarify the rational of our choice.

- Hurdle models are often done with function hurdle() in R. In this function, we can integrate the factors wanted in both parts of the models (i.e., with a binomial distribution for the presence/absence part of the data, and a truncated Poisson distribution for the count data). The hurdle() function, however, cannot take into account random effects. We therefore had to choose function glmmTMB to obtain a hurdle model with random effects.
- The writing of the glmmTMB function is less straightforward than the hurdle() function : we place the truncated count data under the name of 'zi' which may be confusing, as this is not a zero-inflation correction per se.
- We defined the parameter theta in the statistical section when referring to the choice of having negative binomial distributions for the models regarding the community (lines 223-226).

There are a couple of issues that need to be resolved or need explanation here:

01. I agree that the presence of excess zeros does not warrant using zero inflation models. Hurdle models can use alternatively, but clarification is needed on how the zero (inflation) arises in three different analyses. The first step is to evaluate the overdispersion (either with Poisson or Negative binomial distribution), which may be done using a simulation test. If the authors did overdispersion tests before selecting the Hurdle model procedure, please mention it on Page 9, lines 185-188.
**** We added a step in which we compare Hurdle models made with a truncated Poisson distribution or with a truncated Negative Binomial distribution. For the community response, the best fit was with a Negative binomial distribution. For the isolated responses of coal tits and crested tits, we found that the Poisson distribution was more appropriate (best BIC). We added this information lines 213-228.

02. In this study, as far as I understood, zeros (inflation) may arise on two processes: 1 zero may occur when the species do not present at the point or species present at the point but did not show any mobbing responses. For example, in community-level analyses, the number of responding birds may represent single species or different species (i.e., if four birds showed a mobbing response, it could be from a single species or four different species).

Were zero responses specifically generated, either absence of the mobbing species or no mobbing response towards the soundtracks, particularly for coal tits and crested tits? Clear identification of the zero-generation process is also essential when defining the glmmTMB models.

**** As outlined in the general point above, we ran our playback tests in environments in which the density of birds is extremely high, so that we can confidently assume that focal species are always present (and thus exposed to the treatment). In spring, we can confirm that we heard the presence of coal tits in all 100 spots that we selected. In winter, the birds are less vocal, hence making it difficult to detect them before any test. Yet, the density of birds is known to be very high is this territory (dataset created through citizen science). In our analyses, we therefore consider that the zeros are absence of reaction and not absence of birds. It is true that we cannot verify that birds were always present in the vicinity in all our of our tests. In this case, the large number of replicates and the fact that each spot was tested for all of the different playbacks avoid any risk for biases in our results.

03. Using the same terms in both factorial design and the models makes sense. Model selection may help choose different distribution fittings (i.e., negative binomial vs Poisson) while keeping the fixed effects in the model. This could also extend to test the hypothesis of the additive vs interaction effect of the same model instead of dropping or adding terms.
**** We follow this advice and modified a little our models. As suggested above, we first ran models either with a Poisson or a Negative Binomial distribution to choose the best fit for our models. We then ran four different models: the number of callers and the emitter species, in each part of the model (presence/absence and count), with their interaction, or without. We compared these four models with BICs and were therefore able to consider whether interactions terms should be kept (Table 1). We then evaluated the importance of each term with the estimates from function summary (Table 2).

04. Why is glmmTMB control "BFGS" used in models? I presume this is because to account lack of convergence in some models. If that is the case, please include the relevant details which would be helpful for the reader in the statistical methods section.
**** As suggested by referee we modified our sentence to explicit the use of the quasi Newton optimization method in order to circumvent convergence failure (line 210).

Zuur,A.F and Ieno,E.. Beginner's Guide to Zero-Inflated Models with R. 2016. (Chapter 6 for Hurdle models). This book extensively discusses the statistical and practical background and updated version of the methods introduced in Zurr et al. 2009, which authors cited.

Minor comments:

Page 4, line 86: rephrase or add (unclear)
**** We rephrased this sentence (line 87-89).

Page 4, line 88: If the authors can back this with a reference, that would be great.
**** We added two reference on this subject line 91.

Page 5, line 110: what is 'X'
**** X is often used as a symbol for mean; but we supressed it as it seems it is not used by all scientists.

Page 6, line 112: please give the breakdown (n=22, coal tits? great tits?).
**** We added this information (line 122).

Page 9, line 179: R version 3.6.1 was released in 2019, not 2022. Please ensure all the package versions used in the preprint are correct and include their references.
**** Reviewer is right, we used version 4.1.1 and not 3.6.1. We checked the references of the packages.

Page 9, lines 195-197: Is the overdispersion the main reason to use negative binomial distribution??
**** A (truncated) negative binomial was indeed preferred over a truncated Poisson one in order to accommodate overdispersion in our (positive) count dataset (referee likely noted that the reason was not due to an excess of zero, since this is already handled by Hurdle models). In the revised method we propose here, we compared the fit of our models with either a (truncated) negative binomial or a (truncated) Poisson distribution for all our response variables using BIC. The truncated negative binomial distribution was retained to analyse the general community response while the truncated Poisson distribution was retained for the individual response of coal and crested tits. Furthermore, we also checked the residuals of the retained model.

Page 10, line 208: if it is due to sampling size, then AIC corrected is also helpful; perhaps it may be unobserved heterogeneity. Brewer et al. 2016 Methods. Ecol.Evol. Volume 7(6) p.679-692.
**** We rephrased the methods section in this regard and cited the article suggested by the reviewer (lines 220-221).

Page 10, line 200: were random effects introduced as an intercept?
**** Yes, they were introduced on the intercept. We added this information line 209.

Page 11, line 225: it would be helpful to provide contrast measures using the final model. The authors may use the emmeans package (Ver 1.8.2), 2022 or an equivalent package to get contrast estimates. Estimates also provide strong statistical evidence for the graphical presentations in Fig1-3.
I am sure the following reference: Ratnayake et al. 2021. Behav.Ecol Vol 32(5) pages 941-951. It may be helpful to include some necessary information in the methods section, and please note that this is not an indication to cite the reference. However, the reference may be relevant as the study used mobbing calls of noisy minors to test the occurrence and the intensity of the responses of Australian magpies.

****Unfortunately, we did not find a way to provide the contrasts of the two parts of the model (emmeans provides the estimates from the joint model, with no distinction between the occurrence and intensity parts). This is where we differ from the analysis from Ratnayake et al. who did not want to explore the two-step process in the model. We contacted the author of the package on this regard, who responded that the best solution here would be to run the two models separately (i.e., running a logistic regression with a binomial distribution, then a negative binomial GLMM with the data for which the occurrence >0). We do not, at this stage, believe this would add much to our analysis, but we are willing to follow this method is the reviewer believes this is more appropriate. In any case, we here provide the effect size (odds ratios) to quantify the differences between our different playback treatments.

**** When reading Ratnayake et al., we realised we may add a paragraph about our studies species and region in the materials & methods section (102-113), and a short ethical note (239-246). We also added the distance between the recorder and the birds when we obtained our recordings for future playbacks (line 156).

I hope comments will help improve the preprint quality, particularly parts in the statistical methods section. Finally, I congratulate the authors for their good work.
*** The authors thank the reviewer for all the helpful comments on this manuscript.