Comments to the authors

Thanks for the opportunity to review the preprint for PCI. The preprint addressed an essential issue in heterospecific mobbing behaviour: how the number of mobbers and the probability of mobbing changed with the number of individuals and emitter species identity between two seasons within a community (outside the breeding season: May -July). To answer this question authors used two resident species: Crested tits and Coal tits species. Then, using a playback experiment with one and three callers of Crested and Coal tit exemplars at 100 locations(points), the authors recorded the community's mobbing responses, including Crested and Coal tit species. In addition, the authors did the same experiment in two seasons using the exact points and the "crossover design". Finally, they assigned the location instead of the subjects (Crested tits and Coal tits) for treatment playbacks.

Authors found that three caller playback treatments had more mobbing responses (mobbers) than the single caller playback treatments within the community. Further, Coal tit playbacks had more mobbing responses than the Crested tit playbacks. Overall, the preprint is written well, and the organisation is clear, but there is room to improve.

I highly value the author's effort to conduct the experiment with minimal bias and allow open access to the data and the statistical analyses conducted for the preprint. It was an interesting read. However, using the subset of data for specific analyses may raise some considerations. For example, all the mobbing intensity analyses focused on the subset of analyses removing zero occurrences. Zeroes may represent selected community species absent or did not respond to the playbacks. The zero percentage is nearly 50% of 800 trials. Following an alternative, zero-inflated GLMM procedure would strengthen the analyses and conclusions.

Major comments

Crossover design and underlined statistical data analyses require justification and clarification.

1. In general, crossover designs require the same subject with treatment playbacks within a season and exposed the same in the next season or in-between two washout periods (in this experiment, two treatments in one season and the other two in the next season). Instead, the authors used the location to crossover the treatments. One of the caveats in this approach is that the community composition at each point may vary between seasons, and the individuals exposed to the acoustic cues in the previous season may not be present at the site (see below point 2). As a result, even though the sample number is equal between two seasons, using filtered data sets (either absence of the species or no response at each point) for the current analyses require statistical justification. I am not entirely convinced why the authors conducted partial analyses for both seasons separately when interactions were evident (Table 2). However, I believe the authors may have a good reason,

and it may be helpful for the reader if it is spelt out in the methods or the statistical analyses sections (please see comment # 3).

If the same number of treatments (four) were repeated in the next season, I find it challenging to understand the design as a crossover design. Ideally, a cohort (in this case, a community) of individuals exposed to two treatments in the winter (i.e., ICR, 3CO) and the same cohort getting the other two treatments (i.e., 3CR, 1CO) in the spring may result in a crossover design.

2. It is also unclear how the correlation was done using the subset of the data to confirm the presence of both species at the exact location. For example, if I understood the table in the R script correctly, in spring, out of 400 trials, only six trials had both species present/mobbed and 313 trials with both species' absence. The same applies to the winter data; out of 400 trials, only 68 trials had both species and 250 trials with both species absent. This data suggests zero inflation (~70% of zeros). Is the correlation reported between 30% of the occurrence of both species and or using the complete data set? Figure 3 based on this result (proportion of points), are SE and confidence intervals present in figure 3 based on the model estimates?

In Figures 2 c and 2 d, I presume that dots represent the median as the data based on counts; and are those values based on predicted or raw data?

Authors could have done alternative analyses: zero-inflated Poisson or Negative Binomial GLMM considering the zero inflation while adhering to the experimental design ("crossover design") or without losing the design structure. Otherwise, it becomes an exploratory data analysis as it currently stands in the preprint.

The suggested alternative analysis procedure is only possible for a crossover design if the authors successfully identified the responded individuals with colour bands or another individual-identifying method at each location. Otherwise, it would be wise to disregard the crossover design; a zero-inflated GLMM still account for the zero inflation in mobbing intensity analyses. It also combines the binomial and count parts currently present in the preprint. One of the best references I have come across is Zurr and Leno, Beginner's Guide to Zero-Inflated Models with R,2016.

Minor comments

3. The results section can organise into two sections: 1) to show the community mobbing occurrence (presence vs absence), mobbing intensity and the difference between the two seasons (Table 2 community; community). Then, 2) specific Coal and Crested tit mobbing occurrence, the mobbing intensity, and seasonal differences in the separate section (Table 2, Coal tits and Crested tits). However, it is somewhat difficult to follow the results section, at least for me.

4. LRT analysis showed that Coal tits alone analysis with 800 trials showed a marginal difference between complex interaction and the additive model (p=0.045, with Singularity = TRUE and model convergence errors). So, it would be helpful for the reader to present these results for both species on page 10, lines 215 -216. Table 2 presents the Analysis of Deviance results (Type 2 Wald chi-square tests). It is helpful to mention the exact tests in the statistical analyses section and the table headings where appropriate.

5. Generally, the discussion is slightly longer and can be reduced by removing the parts irrelevant to the experiment and the data presented in the preprint. Below, I try to draw a few sections that need consideration; however, once the authors carry out the analyses considering the zero inflation, the current interpretation may or may not hold. So, I am reluctant to comment at this point.

Page 19, lines 328 -333. This justification of the duty cycle idea may be helpful in the methods section where introduce and define duty cycles in this study? I think the author's discussion slightly goes beyond the evidence presented in the preprint. Please note that Landsborough et al. 2020 did not disentangle the effect of calling rate and duty cycle, so I am not sure the last line is correct here.

Page 22, lines 399-401. I think individually marked Crested tits may be helpful to strengthen the argument that not the lower presence of Crested tits shows the lack of responses. Did the authors have territoriality data of at least both the selected species?

Specific comments

Page 2, line 34; I think in the abstract, the authors need to tell the reader what the acoustic cues used in this study are: number of callers and emitter species. Then the following line 43 may also require a slight adjustment. Not only in the abstract but throughout the preprint, the "acoustic cue" and "cues" needs to define and should be consistently used as the title imply (i.e., page 4, line 85; page 5, line 103 etc.).

Page 2, line 38; it is worth mentioning that three caller playbacks attracted more mobbers than one caller.

Page 2, line 43; the study demonstrated that outside the breeding season, community response to mobbing interacted with the number of callers and emitter species and the season.

Page 3, line 65; this may not always be true; a larger group of deceptive callers increase the high risk of deceptive mobbing calls. Page 9, lines 185-186; please mention how you discarded the terms; either using the stepwise method (forward, backward, or mixed) or any other model selection method.

Page 9, line 190; is it? The *emmeans* predicted values are generally not on the response scale, or I may miss the point here.