Dear Ambre Salis,

I apologize for the delay in providing these comments. By the time I received the second review, I had left on holiday for much of July.

Two independent reviewers and I have read your manuscript (Acoustic cues and season affect mobbing responses in a bird community). The reviewers and I all see value in this study. I appreciate the use of a thorough design to simultaneously evaluation multiple potential influences on the response to mobbing calls, the large sample of trials, and the evaluation of an important assumption of your experiment. However, we also all identified some important areas where improvements are merited. I have provided some detailed comments below and both reviewers provided detailed suggestions as well. Please carefully consider all these suggestions and either implement the suggestions or explain why you have not done so if you chose to resubmit a revised manuscript.

**** Dear Recommender,

We apologize for the delay to revise our manuscript; we received the review during the annual closing of the university. We thank the recommender and the reviewers for their helpful comments on the manuscript and changed it accordingly. The main change was on the statistical analysis. More specifically, we finally chose to run hurdle models as suggested by referee 2 and we also used multi-models inferencing approach for model selection. We believe the statistical framework is now clarified for the reader and also more appropriate to our data. We also provided some additional methodological details when necessary and implemented our introduction and discussion sections following the recommendations of the reviewers. We will be glad to hear about reviewers' insights on this revised version.

Before providing my detailed comments and those of the reviewers, I want to call attention to several points that are particularly important.

Both reviewers noted that group size may not be expected to correlate with the reliability of mobbing calls. I encourage you to explore the literature on this subject further, and to update your discussion of this topic

**** We agree that reliability of the calls is not necessarily increased with a greater number of birds mobbing; we therefore discarded this argument in our manuscript and only kept the general group effects decreasing individual risk.

Reviewer 1 and I both felt that your explanation in the main text of the supplementary experiment (done to assess the likelihood of overlapping responses between playback locations) is insufficient. I encourage you to either bring this experiment to the main document, or at least to provide more details in the main document.

**** We completed the information in the main text; we however feel that the additional experiment should stay as a supplementary material as they are not crucial for the reading of the article.

Reviewer 1 and I also wanted to see more information about the 3-bird playback stimuli. Besides addressing the questions of Reviewer 1, I encourage to consider adding the stimuli and the sound spectrograms of the stimuli to your supplement.

**** As suggested, we added one example of each stimulus in our supplementary material.

On a related topic, both reviewers and I share the concern about your interpretation of the 3-bird stimulus due differences in duty cycle between the treatments.

**** We completed the discussion on this part; and hope this new version is now clearer for the reader.

Reviewer 2 identified several important issues related to your statistical analyses, including a model convergence error (which I also replicated when I ran your code). I do not believe there is only a single correct way to analyse a dataset, but I would like you to seriously consider Reviewer 2's recommendations and concerns.

**** We completely modified our statistical analyses based on the critics arising from Reviewer 2; In particular we used a mixture distribution model as recommended by Reviewer, although we preferred to use Hurdle models instead of zero inflation one since the former seems more straightforward than the latter to interpret in our context (i.e. there is a first process that determines the occurrence of mobbing, using a Binomial distribution, and if there is occurrence then, a second process determines the number of mobbers, the intensity of mobbing, using a truncated count distribution).

Sincerely,

Tim Parker

What follows are some specific concerns that I noted as I read the manuscript (organized by line number):

41: "We therefore confirm the hypothesis" – should be something like "We therefore find support for the hypothesis"

**** We modified this sentence accordingly (Line 38).

161-162: when a bird was detected in the area, what was your protocol? Did you wait for it to leave or move to another location?

**** We waited for it to leave; we added this information line 168.

113: How did you determine the order in which you visited survey locations?

**** The chose the 100 points following existing trails (information line 109). Then, on each day, we chose to begin the tests at a different part of the trails distanced at least from 500m from the first spot of the day before. We also changed the direction in which we walked into the trails halfway into the experiment (line 129-130).

159: How did you resolve differences in observations between observers?

**** We noted the minimal number of birds seen simultaneously by the two observers to minimize the experimenter bias (we added this information line 175-176).

174: please explicitly state that you excluded cases of zero detections from the intensity analyses

**** We added this information line 190-191.

177: which versions of the lme4 R package? Also, please cite the package (e.g., 'Bates et al….' for lme4).

**** We cited the packages and the versions we used in main text (glmmTMB and Dharma, lines 197 & 203).

179: if you decide to remain with analyses using Poisson error (instead of taking the suggestions of reviewer 2), you should report the details of your test for overdispersion.

**** We finally followed a negative binomial distribution in our Hurdle models given their greater support according to their AIC. We added this information line 195.

183-185: This sort of step-wise procedure can lead to biased model estimates (see Forstmeier and Schielzeth. 2011. Behav Ecol Sociobiol 65:47–55). Given that you present full models, it is not clear to me why you were using a step-wise procedure.

**** We abandoned the frequentist approach (based on a step-wise procedure selection) in favour of the multi-model inferencing one. Thus all nested model within the full model were constructed and the BIC criterion was to select the potential hurdle models (we preferred the BIC criterion rather than the AIC since the former is more sensitive to the sample size, although both criterions give a similar model ranking). We report the three best models with their BIC, weighted BIC, and delta (Table 1).

189: again, more information needed about packages

**** We cited the packages used in the new analysis proposed.

Table 1 vs. Table 2. The first table in the text is labeled 'Table 2', but presumably it should be labeled 'Table 1' – it seems that when you cite Table 1 in the text, you are referring to the table currently labeled 'Table 2'

**** We supressed the first Table since we now propose to clearly differentiate the two experiments. We checked the citations of Tables and Figures in the new manuscript.

Table 2 (as currently labeled): Why do you not present number of mobbers results for coal and crested tits? You seem to have that information in another form in Figure 2.

**** We did not include this information in the previous version because the analysis on the mobbing intensity of coal and crested tits is limited to the number of mobbing events we recorded. In some cases, especially in spring, the sample size of such analyses was therefore strongly reduced. We chose to present these results in the revised version for the sake of clarity about our data (Figure 2d, 2f, 3d and 3f).

Table 1 (as currently labeled): The heading says "propensity and intensity", but the content of the table looks like only one or the other, but definitely not both.

**** Recommender is right, a mistake was present in this Table. We completely replaced it because of the new analyses.

Table 1 and Table 2/ Figure 2: You currently only present the estimates in graphical form. It would be useful to present the actual numbers (either in an expanded form of Tables 1 and 2, or maybe in a table in the supplement).

**** We added more details in the main text of the result section (e.g., lines 227, 230, 261).

Table 2: 'NB' should be defined in the heading

**** Not applicable anymore.

338: You should more explicitly acknowledge the limitations of your experiment here (duty cycle not standardized).

**** We rephrased the paragraph in the material and method to clearly explain differences in duty cycles and we also reorganised the paragraph in the discussion to clearly state that our experiment cannot conclude on this subject (lines 150-151 and 361-363).

REVIEWER 1:

General comments
I think that the intent of this paper was very interesting and a much needed approach to mobbing (i.e. looking at multiple factors). There were some very odd word choices throughout that could be fixed to help with clarity. Though this paper is clear and interesting, I have some major concerns with the methods and therefore conclusions from the results (see below).

Specific comments

51: this sentence is a bit oddly worded. I would suggest something more like "mobbing encourages the predator to give up hunting and/or move to another location, in both the short and long term."
**** We rephrased the sentence accordingly (line 50).

63-65: how does a larger group increase the reliability of the information? In some systems larger mobbing groups are not necessarily better informed. For example, a large group of babblers is often less reliable than one drongo in some flocks. Also, in many groups some of the most important species that recruit larger groups are also the ones who take advantage of the group by using false alarm calls to get better access to food resources (like great tits).
**** We agree with reviewer that this argument is not supported by any strong data. We therefore suppressed this argument and focus only on effect of lower risk due to increased group size.

110-117: I am a bit concerned about the 100m distance. In every circumstance I have heard about, individuals range much farther than that in winter and so the chance of not only having a neighbor overhear the playback (if they move even 10 m closer) and get primed to respond to any following playback, but also end up re-testing individuals multiple times is quite large. While you explain a complementary experiment that shows that the birds do not follow you, there is no information on sample size or duration of tracking the individual, or if following the bird with the binoculars might impact their movements. While this

information may be in the supplementary info, as this is such a large concern for this paper, I would like to see it much more detailed in the main body of the text. I am not convinced it is far enough with what I have read and what I have seen, heard from others, and read in other papers.

**** We are aware that most published papers use larger distances between tests to avoid a risk of pseudo replication, but we never read this risk to be assessed in these publications. For these reasons, we did a complementary experiment to test whether tests done at 100m apart from each other could possibly affect our tests. This experiment was also done on great tits in Salis et al. 2022 (Behav. Ecol. Sociobiol. Vol 76(4), 46) and we obtained the same conclusions. We thus agree that birds can range larger distances for daily foraging according to the local condition (e.g. food supply), but in our study area, we are rather confident that this 100m distance allowed us to test different birds. We also agree with reviewer that more explication is needed, we therefore included the more explanation on this experiment in the main text (lines 111-119).

137-139: why these two different stimuli? Where there differences in the responses? Were these responses balanced across the treatments (e.g., playbacks made from responses to both predator calls and mobbing calls)?

**** We are afraid that we do not understand what the reviewer is asking here; could you reformulate? To clarify our protocol: (1) we elicit a mobbing response from free ranging birds by way of a mobbing chorus sound used by Dutour et al 2016. Then (2) we recorded both coal and crested tits responding to such playback mimicking a mobbing situation. At last, we used our recordings to build the different soundtrack used in the present study.

146-149: was there any control or taking into account that if all calls were coming from one place, unless they were completely overlapped, the response could have been received as a higher call rate from one individual rather than a lower call rate from multiples? Could the 3 birds have been received as more dangerous/urgent than the 1 bird playback (rather than a larger group and therefore safer to approach)?

**** The playbacks strongly overlap, and this is the argument that make us believe that birds received the playbacks as the sum of several individuals rather than only one individual calling more than others. We discuss this line 371-373 and add the information line 152-153.

175: why the number of individuals rather than proportion of birds present that mobbed? If there were simply less birds around to receive the signal, then there may be a lower response (not due to an actual lower response but due to the fact that fewer birds were around to receive the signal). If 100% of the available birds responded, that is a strong response (but could be counted as a low response if few birds were around to respond). This is also true across seasons as I would expect fewer conspecifics to show up as they are risking a territorial dispute with the resident breeding pair. How was this discrepancy dealt with?

**** There was usually no bird in the vicinity (within 10m) of the loudspeaker before the playbacks. Since we did not work on flocks, counting the available birds (i.e. those hearing the playback) cannot be easily done, if feasible, since birds may not be visible and may have a reduced call activity especially in winter. We agree that a proportion of birds responding among the available birds at each location would be preferable, especially if we are interested to examine the difference across locations or within a location according to the season, since the number of available birds may vary spatially or temporally (over a sufficiently long period). In our study, the comparison of treatments within a same season was done by comparing the response of the available birds among the successive tests realized at each locality on a relatively short period so that one may can expect that

the number of birds available did not substantially vary between tests done in the same locality (within the same season). Moreover, if the number of available birds substantially vary between tests (within the same season), given that the test order sequences were random, it should inflate the background noise in our statistical analyses, thus reducing the effect size of the treatments, but not the inverse (artificially increase the effect size).

238-239: since the total number of species (even rare, eavesdropper species) is included, I am a bit more concerned that it is not possible to determine the difference between attendance due to a higher call rate (more dangerous predator) vs. a higher number of individuals. While other coal and crested tits may easily be able to determine the number of individuals calling, this gets less likely when talking about non-flock mates who primarily eavesdrop when they happen to be in the same area. They could, then, simply be responding to the overall amount of calling, not specifically more individuals.
**** Please see the comment below for lines 331-332 as it refers to the same matter.

322: why would a larger conspecific group be more reliable? I believe that in the Magrath paper cited here, the reliability comes from listening to multiple other species, not more of the same one.
**** We agree with this comment and supressed this argument that is not sufficiently established to be stated in our manuscript.

331-332: specifically due to relying on duty cycle and caller identity, for those species other than the one calling, it is likely that they may only pay attention to duty cycle (as caller ID can be difficult for many heterospecifics to determine) which means that the test was between a lower and higher duty cycle call (different threat information) not the number of callers. In Dutour et al.'s study the responding individuals were listening to conspecific calls, where individuals were more likely to be able to determine caller identity. When looking to heterospecific response, especially those not frequent flock members, caller ID likely does not matter as much (if heterospecifics can tell individuals apart at all) and they are only receiving duty cycle. Therefore, they are receiving a high duty cycle in the 3 caller and a lower duty cycle in the 1 caller signaling different threat and garnering a different response.
**** We agree with reviewers that there is a chance that heterospecifics focused on duty cycle rather than really detecting the number of individuals calling. We modified our paragraph to make this statement clearer than in the first version (lines 360-371).

401-402: while these three hypotheses are feasible, and reliability could be less due to song – thought we know little about how song and calls differ in these species, there are other reasons heterospecific may not listen. For example, crested tits may be highly aggressive in the spring and will chase away any other bird they find in their territory making responding to mobbing a bad idea for heterospecifics. Crested tit mobbing calls may not be relevant if they have different nest predators or different threats (Magrath). All of these hypotheses should be discussed.
**** We thank the reviewer for these two ideas to discuss, we added this in the paragraph of the discussion (lines 450-457).

REVIEWER 2:

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.
**** The term ‘crossover’ design was misused in our first version of the manuscript and we therefore suppress this term. In regards to the two reviewer‘s comments we modify our statistical analysis and formulate our experiment differently. We rather present the experiment in winter as the main experiment, since mobbing behaviour and heterospecific communication regarding mobbing appear crucial in this season. We then propose a replication of the same experiment in spring, to test whether our results are generalisable in other contexts. With this organisation, we want to show that

the same experiment done in different seasons does not lead to the same results; but we do not specifically test what are the differences between seasons. Indeed, since we have no data to interpret this difference, we believe that a statistical analysis to quantify the difference between spring and winter is not appropriate here. For this reason, we separately analyse winter and spring data and we supressed the global model.

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?

**** The correlation was initially there to test whether the coal tits and crested tits responded at the same points or not. However, this test was probably not appropriate, and we finally chose to suppress this test from our analyses. In the Figures, we present the mean and confidence intervals calculated with the raw data.

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?

**** The figures presented are based on the raw data, with mean and 95% confidence intervals.

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.

*** We agree with reviewer, and we therefore used mixture mixed models as recommended by reviewer. We preferred to use Hurdle models rather than Zero Inflated ones, since its structure is more appropriate given the objective and the context of our study. Rather than using a mixture of distribution for the zero mass (i.e. a degenerate distribution at zero) like ZIP or ZINB, the Hurdle models use a Binomial distribution that determines the occurrence of the event and a truncated count distribution (in our case a negative binomial one) that determines the number of events. This two-stage process is therefore more convenient to interpret our data (i.e. a first process that determine the probability of at least one response, and a second that determine the intensity). We also revised our methodology for models selection. Indeed, we abandoned the frequentist approach based on a step wise selection process, and used rather the multi model inference one. Thus all nested models were computed and their relative support given the data were compared using the BIC (we preferred the BIC rather than the AIC since the former is more sensitive to the sample size, although both gave similar results). Moreover, we also used gof test to verify the adequacy of our full model. All these modifications are presented in the revised statistical section (lines 180-214).

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.
**** Since we now propose a new organisation based on 1/ main experiment then 2/ replication in spring, we also reorganised this section accordingly. We changed the figures to clearly fit this new organisation and added graphs about the intensity of response of coal and crested tits. We hope this new version is easier to follow.

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.
***** Since we changed our analyses, we are not sure this comment still stands. However we followed this advice by showing the models that also fitted well our data (close BIC) in Table 1.

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.
**** Given the commentaries of recommender and reviewers, we modified slightly our discussion; we are open to new comments for this section if the new organisation of the manuscript is validated.

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.
**** We modified this paragraph following the remarks from the first reviewer who also suggested some changes regarding the presentation and discussion of the duty cycle.

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?
****** In the studied forests there are several thousands of birds from both species. Hence using colour marked birds in our study would have required to catch and band birds during several months or years which could not be done. Nevertheless, although crested tits were not marked, we can confirm that the lower response in spring from coal tits was not due to their disappearance from the territory since coal tits' songs were hearable at each point. Crested tits may simply be less vocal regarding song production in spring.

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.).
**** We used only the term acoustic cue to refer to the number of callers and the species identity. When discussing which criteria (identity of callers or duty cycle) is used by birds, we used only the word criteria. We hope this new version is now clearer for the reader.

Page 2, line 38; it is worth mentioning that three caller playbacks attracted more mobbers than one caller.
**** Given the slight changes in our statistical analyses and results, we do not add this information in the abstract.

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.
**** We modified our abstract for other reasons, and believe this sentence is now inappropriate.

Page 3, line 65; this may not always be true; a larger group of deceptive callers increase the high risk of deceptive mobbing calls.
**** We agree with reviewer and suppress this argument.

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.
**** Not applicable anymore (model selection with BIC).

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