

## 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.”

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).

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.

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)?

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)?

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?

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.

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.

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.

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.