How subtle protocol choices can affect biological conclusions? An example with how great tits respond to allopatric mobbing calls.

Ambre SALIS[1*], Jean-Paul LENA[1] & Thierry LENGAGNE[1]

[1] Univ Lyon, Université Claude Bernard Lyon 1, CNRS, ENTPE, UMR 5023 LEHNA, F-69622, Villeurbanne, France

[*] Author for correspondence: ambre.salis@univ-lyon1.fr

Acknowledgments

We thank Mylène Dutour, Toshitaka Suzuki and David Wheatcroft for their transparency, allowing us this comparison study. Collaborations with M. Dutour on other projects are currently being reviewed; but the current study was done independently from her work. We thank the Fondation Vérots for access on their property. Finally, we thank Charlotte Bourbon, Jean Capelle and Julie Ruffion for useful help in the field.

The authors comply with the ASAB guidelines for the use of animals in research. The fieldwork did not require any special permit but followed the laws of the Rhône county and the rules of the ethics committee of the University Lyon 1.

Abstract - In the last ten years, numerous replicated studies showed divergent results from the original papers, leading to the recognition that science may be facing a replication crisis. Apart from fraud, different results may emerge from unconscious bias from the experimenter. Indeed, while the original authors may be prone to p-hacking (to collect data, select data or use statistical analyses until nonsignificant results become significant), the replication-authors are on the contrary probably unwillingly prone to show opposite results (i.e., null-hacking). Two different researchers unknowingly working on the same question and on the same population could overcome this effect. We here present such a comparison: two researchers investigated the response of great tits (*Parus major*) to mobbing calls of an allopatric species, in their natural and reversed order. We show that although the effect sizes of the differences of interest are similar, biological conclusions solely based on the p-value would be opposite. We here illustrate how subtle protocol choices, especially regarding the soundtrack preparation and playback methodology, can explain variation in the results, but that another pitfall in field studies is a general problem of sample size and blinded rely on the p-value.

Keywords - Animal communication, Heterospecific communication, Mobbing, Replication crisis, Syntax

During the last decade, failure to reproduce published results in various fields or research (e.g., in psychology Bohannon, 2015, or in epidemiology Lash et al., 2018) alerted the scientific community about a potential low reliability of published results. This replication crisis, complex and heavily debated, can be explained in different ways (Fanelli, 2018; Maxwell et al., 2015). Indeed, while models have shown that the global "publish or perish" problem probably increases misconduct (Grimes et al., 2018; Higginson & Munafò, 2016), direct frauds (i.e., fabrication or falsification of data) remain scarce according to empirical evidence (Fanelli, 2018). More probable is the combined effect of several inconspicuous and ordinary factors (Ioannidis, 2005) leading to the publication of false results and/or interpretations, with one major example being conscious or unconscious p-hacking (i.e., to collect data, select data or use statistical analyses until nonsignificant results become significant, Head et al., 2015). In the field of animal behaviour and especially in animal communication, three specific factors are specifically of importance. Firstly, behaviour is by nature an external proxy of internal state of animals, needing researchers to interpret each variable under study. Such behaviours are prone to flexibility between researchers regarding the type of recording, the definition of behaviour, and their relative relevance to the question asked. Indeed, each scientist is the sum of their past experience, knowledge, and personal background, which can affect their biological conclusions at the creation of the protocol (Tang-Martínez, 2020), when analysing results (Silberzahn et al., 2018) or when interpreting those results (Tang-Martínez, 2020). Greater flexibility increases the opportunity to transform negative results into positive ones (Ioannidis, 2005), and in a general manner creates disparities between papers investigating the same question. Secondly, the difficulty to obtain large numbers of wild subjects as much as the ethics considerations often lead behaviour studies to obtain restricted sample sizes (Schwagmeyer & Mock, 1997). Small sample sizes are less likely to detect small differences between treatments (type I error,

Button et al., 2013) but are also prone to stochastic variation so that the probability of a positive result is inflated despite no biological difference (type II error, Button et al., 2013). Thirdly, fields of research such as language evolution in animal communication are quite new, with several teams working simultaneously on similar questions. This increases the risk of more spectacular positive results being published in priority, as each team aims at showing their most influential work (Ioannidis, 2005). Replicating behavioural studies should consequently be of great interest, with however one caveat: while the original author may have been prone to p-hacking, the replicating author may in opposition (probably unwillingly) possess a "null hacking bias" (i.e., the motivated pursuit of null results by replicating investigators, Bryan et al., 2019). As a result, replication studies are often as questionable as the study they wish to replicate (Schmidt & Oh, 2016). To circumvent this problem, one would need two researchers having the exact same question, at the same time, testing the same population, without being influenced by each other. Such a situation occurred in our laboratory: two independent researchers, one leaving and one arriving in the laboratory, had by chance the same idea, and communicated too late about it. This led to two datasets answering the same question, obtained with quite similar yet not exactly equal protocols. Since the disparities between protocols are relatively low and all justified by authors, this article is therefore one great opportunity to compare how slight changes of protocol between two researchers can affect, or not, the resulting biological conclusions.

The biological question at stake concerns the currently hotly debated question of compositional syntax in birds. Compositional syntax is defined as when the meaning of a sequence is related to its different parts and in the way they are combined (Suzuki et al., 2019; Suzuki et al., 2020). Recent studies have proposed that some species, when mobbing a predator (i.e., actively harass it instead of flying away, Carlson et al., 2018), use a combinatorial call in a fixed order: the first part (hereafter called FME: Frequency Modulated

Elements, Dutour et al., 2017) elicits vigilance, while the second part (called the D notes, following Hailman et al., 1985) elicits approach from the receiver (Dutour et al., 2019; Suzuki et al., 2016, Figure 1). Combined, the resulting sequence engender behaviours such as scanning, approaching and calling in receivers, typical behaviours linked to mobbing (Carlson et al., 2017; Salis et al., 2020; Suzuki et al., 2016). Furthermore, the reversed order (i.e., D notes then FME) results in lower responses from the birds (Dutour et al., 2019; Suzuki et al., 2016). Debates on whether such coding strategy can be designated as compositional syntax in the human linguistics sense have been profuse (Bolhuis et al., 2018a, 2018b; Griesser et al., 2018). The reasonable response to such critics is that this young subject deserves more studies on the same species to conclude on such potentially high cognitive abilities in birds. One way to dig into that question can be to test whether a species known to use a FME-D combinatoriality also respond to mobbing calls of an allopatric species exhibiting a similar ordering sequence in mobbing call but made up of acoustically different notes. Two adequate species for such an experiment are the great tit (*Parus major*), living in Europe and for which the use of compositional syntax have already been investigated (Dutour et al., 2019), and the North American black-capped chickadee *(Parus atricapillus*), for which the mobbing calls are also made up of a FME-D notes combination with fixed syntactic rules (although the idea that FME notes are related to vigilance behaviour and D notes to approach has not been tested yet in the black-capped chickadee, Baker & Becker, 2002; Otter, 2007, Figure 1). Great tits respond to the mobbing calls of the black-capped chickadee in the same way they do for conspecific calls: they mob to the complete sequence, are vigilant when hearing FME notes, and approach to D notes (Randler 2012, Salis et al. 2020). One could therefore expect great tits to respond with mobbing behaviour when presented to an unknown call sequence when it has the same composition, while failing to do so when the ordering of the call sequence is reversed, as they did for conspecific calls

(Dutour et al. 2019). Such questions were addressed by two studies in a few months of interval (Dutour et al., 2020 and the present study). These two experiments are one great opportunity to test whether replication without prior knowledge of the previous studies reaches the same conclusion.

In this article, we will therefore not discuss the importance of the resulting biological conclusions regarding compositional syntax in birds, since Dutour and her colleagues (2020) already did so. We will focus on whether slight differences in protocol choices engendered contrasting results, discuss which parameters may be of importance in such potential disparities and conclude on how this affect our field of research.

Methods

General organisation

Our experiment is aimed at answering two specific questions: (i) do great tits respond to allopatric mobbing sequences never eared before in the same way they do for conspecific calls (question 1, hereafter designated as the "species comparison"), and (ii) whether they would do so for allopatric calls for which order is reversed (i.e., D-FME, question 2, hereafter designated as "order comparison").

To do so, a in a field study we presented great tits with mobbing recordings of great tits, natural calls of black-capped chickadees, reversed calls of black-capped chickadees, or background noise (control). We measured their vigilance with the number of scans they produced, and whether they approached the loudspeaker. If they respond to allopatric mobbing sequences, they should scan and approach as much as when hearing conspecific calls. Secondly, if order is important in the decoding process, they should not respond anymore when the allopatric mobbing sequence is reversed (less scanning and less approaching).

We here describe our protocol and for each point, describe the similarities or difference with Dutour et al. (2020). Our protocols are similar on most points but, while Dutour et al. (2020) created two separate experiments with somewhat different protocols (see Table 1), all our different treatments were tested in the same global experiment. We will consequently separate the two questions (species comparisons and order comparison) only in the statistical analysis and results section. For clarity sake, Table 1 summarize the common ground and differences with Dutour et al. (2020).

Preparation of soundtracks

In both our experiment and in Dutour et al. (2020), 4 different types of soundtracks were built: first, soundtracks with the complete (FME-D) mobbing call sequence of the great tit (GT) or black-capped chickadee (BC), to check whether great tits responded in a similar way to allopatric calls and conspecific ones. Secondly, artificially reversed black-capped chickadees sequences (D-FME) to test the importance of order. At last, we both used a control, background noise (BN).

Our soundtracks of great tits and black-capped chickadees were built using recordings obtained from the Xeno-canto online database (www.xeno-canto.org) and the Macaulay Library (www.macaulaylibrary.org). Dutour et al. (2020) used the same websites (recordings previously used in other studies) in addition to three recordings of their own of great tits. For both species, we only conserved good quality recording files (A or B grades) under the denomination “Alarm call”/“Call”. Then, to ensure that the selected recordings truly represented mobbing call sequences, several features were controlled: first, in both studied species a mobbing call correspond to an association of FME and D notes (Fig.1). Hence, selected recordings were all made of the same FME and D notes reported in Baker and Becker (2002) and Templeton et al. (2005) for the black-capped chickadee and in Randler (2012) and Kalb et al. (2019b) for the great tit. In addition, for both species, the D notes

length is known to vary with the context (Kalb et al., 2019b; Templeton et al., 2005); we therefore checked that our soundtracks had the same length as the D notes used in mobbing calls (X = 0.05 s ± 0.01 for the great tit, X = 0,18 s ± 0,02 for the black-capped chickadee, mean ± standard deviation). Finally, for the great tit, we verified that no FME used in food or flight related contexts were the most preponderant in any of our recordings (i.e., G, H, I and M notes associated with food for the great tit, Kalb et al., 2019a).

From these recording files, we built 40 soundtracks of 1 min mobbing sequences of great tits and black-capped chickadees (20 for each species, each provided from a different emitter) using Avisoft-SASLab software (files were converted into a Wav format). To allow comparison between black-capped chickadees and great tits responses, we constructed every mobbing sequence with a similar duty cycle (Landsborough et al., 2019) and mobbing calls repetition (30 calls/minute, natural range of repetition rate, Suzuki et al., 2016). Consequently, each mobbing call emitted by both species had a total similar FME duration (0.31 ± 0.06 sec/call, mean ± SD) and D duration (0.50 ± 0.07 sec/call), but not the same number of notes, since BC D notes are longer (Figure 1). Dutour et al. (2020) created the same number of playbacks but chose to control for the number of D notes per call (8 D notes/call) and global duty cycle instead of the number of calls/min. Therefore, the number of calls per playback was lower in the black-capped chickadees playbacks (14 calls/min) compared to the playbacks of great tits (26 calls/min). In both our experiment and Dutour et al. (2020), reversed playbacks were constructed by putting the FME notes after the D notes. We made sure that the space between the FME and D notes was the same before and after the manipulation. The reversed playbacks therefore possessed the exact same duty cycle and rhythm that the natural order playbacks. We also constructed 20 background noise soundtracks extracted from the original recordings (control soundtrack hereafter referred as

BN). Each of these 80 soundtracks were cleared of any other bird calls, background noise was reduced, and amplitude homogenized.

Field tests

In our experiment, data were collected in the east of France at during the breeding season (March/April 2019), in a radius of 25 km around Lyon. Data of Dutour et al. (2020) were collected in the same territory, also during the reproductive season, but the year before (February/March for question 1, May for question 2). For each type of soundtrack, 20 fully independent tests were performed (each bird tested was tested only once, and each bird received a different playback).

In both Dutour et al. (2020) and our experiment, each test was performed by two field assistants. One of them was assigned to the soundtrack operation, while the other was kept unaware of the selected soundtrack (using headphones with music) and assigned to the behavioural recording of the focal bird. For each test, after detecting an individual using binoculars, the focal bird was observed for at least 1 min, and the pre-test behaviour (singing or foraging) was noted. If the animal was displaying an alarm behaviour, no test was performed. A loudspeaker was placed 16 m from the bird (16.79 ± 6.27 m), and at less than 3 m to a potential roost (bushes/trees) to allow the approach of the focal bird. The two field assistants were then placed in retreat (minimum of 15 m to both the bird and the loudspeaker) before launching the soundtrack with a remote control. All soundtracks were broadcast using a Shopinnov 20 W loudspeaker with an intensity of 79.8 ± 1.9 dB(C) (measured at 1 m from the loudspeaker using Lutron SL-4001, C weighting, slow settings, re: 20 µPa). The field procedure for question 1 of Dutour et al. (2020) and in our experiment were both based on a complete randomized design and very similar, excepted for three details: 1/ the main observer in Dutour et al. (2020) was aware of the playback launched, as she did not wear any sound protection, 2/ the loudspeaker was placed at ~ 30 m from the bird in Dutour et al. (2020), and

3/ the amplitude of the sound was of 83 dB in Dutour et al. (2020). In experiment 2 of Dutour et al. (2020), the tests were carried out at the nest, the loudspeaker was placed at 20 m, and birds were tested several times using a crossover design.

Tests were carried out between 06:00 and 13:00 h during calm and dry weather days. Each of the 4 soundtracks were tested each day in a different order to avoid any temporal effect. To avoid pseudoreplication, each selected focal bird was separated from each other by at least 100 m (Dutour et al., 2019). Although birds were not individually ringed, great tits are known to be strongly territorial during the breeding period (Krebs, 1971; Wilkin et al., 2006) so that spacing between neighbouring individuals is often used to ensure sampling of different specimens in field tests. As several other studies (e.g. Dutour et al. 2019), we used a distance that roughly correspond to the highest average distance expected according to territorial sizes reported in this species (c.a. 1.5 ha, Wilkin et al. 2006). Moreover, in the present study, two or three singing birds were often concurrently detected within 100 m, suggesting territorial size to be substantially inferior to 1.5 ha in the study area. As all sampled birds were at least distant by 100 m from each other, we are thus well confident that the risk of testing the same individual twice remained quite low.

Behavioural observations

In both our experiment and Dutour et al. (2020), during 1 min of playback, two types of behavioural states were assessed, respectively (1) Vigilance effort as indicated by the number of horizontal scans displayed (the number of movements that birds made with their heads from left to right or right to left, approximately a 180° turn (Dutour et al., 2019; Suzuki et al., 2016)*;* (2) Approach inferred using a dichotomic variable (approaching at least halfway from their starting point) measured with a Leica DISTO D210 telemeter. In the field, we reported for each test the distance of the bird from the loudspeaker at the beginning of the test, and the closest distance of the bird during the test. We then divided the closest distance by the

distance at the beginning and the bird was considered as approaching when the ratio was < 0.5. This way of defining approach allowed us to take into account the initial distance of the bird (even if we tried to be at 15 m from the bird, we sometimes were at 13 m or 17 m). All observations were done using binoculars and recorded on a voice recorder (Sony ICD-PX370) by the same pair of observers in Dutour et al. (2020) while two trained pairs of observers ensured the field observations in the present study. We limited birds' disturbance with two decisions: tests were of short duration, and birds were tested only once. Moreover, after our tests, we checked that all birds returned to their pre-test behaviour in less than 5min.

Statistical Analysis

We followed the same methodology as Dutour et al. (2020) to analyse our results. We therefore split our tests into two questions: first, we compared the response of great tits to natural conspecific and natural allopatric calls (species comparison). Then, we compared responses to the control (Background noise), the natural allopatric call, and the reversed allopatric call (order comparison). We used GLMM (*glmer*, package *lme4*) for both the scanning and approach behaviour, with the original soundtrack as a random effect. Posthoc comparisons were achieved with functions *emmeans* and *multcomp::cld* (packages *emmeans* and *multcomp*) with a Tukey adjustment for multiple comparisons. The number of scan produced was analysed with a Poisson distribution and log link function, since no overdispersion was detected (checked with *glmm.overdisp*, package *RVAideMemoire*). Note that Dutour et al. 2020 used a quasi-Poisson distribution because of overdispersion of their data; and that the analysis of experiment 2 took into account the identity of the bird tested, as they were tested multiple times. We also corrected the analyses for the actual observation time using the time the bird was actually seen as an offset. For the approach behaviour, we set a logistic regression (binomial distribution and logit link function). All fixed effects introduced in the models were tested using Wald tests (*Anova*, package *car*).

Since the raw data available in the supplementary material of Dutour et al. (2020) is incomplete regarding the cross over design used for the second experiment, it was not possible to embed both our dataset and the one of Dutour et al. (2020) in a same analysis in order to compare both studies. Nevertheless, the available information published in Dutour et al. (2020) was sufficient to calculate the effect sizes of each relevant comparison, and we therefore used these metrics to compare our results from those of Dutour et al. (*2020*). We computed odds ratio (hereafter OR, *odds.ratio*, package *questionr*) for the approach behaviour, and Cliff's d for the scanning behaviour as this variable does not follow a normal distribution (*cliff.delta*, package *effsize*). One should nevertheless note that the computed effect size does not take into account the non-independence of the observations done in the second experiment of Dutour et al. (2020, i.e., the cross-over design where different acoustic tests were performed on the same subjects).

Results

Question 1: Response to natural mobbing calls from a conspecific or an allopatric species

In our experiment, great tits scanned an average of 7.30 ± 3.16 scans (mean ± standard deviation) when presented with conspecific calls, and 6.80 ± 3.12 scans when presented with black-capped chickadees calls (Figure 2). No statistical difference was detected in our model ($X^2 = 0.93$, df = 1, p = 0.33), and the calculated effect size of the difference was 0.18 (Cliff's d, 95% CI [0.42; 5.24]). In Dutour et al. (2020), great tits produced 10 ± 5.33 scans in response to conspecific calls and 9.05 ± 5.62 scans in response to black-capped chickadees calls. The resulting effect size is 0.12 (Cliff's d, 95% CI [-0.22; 0.44]), hence very similar to the one we detected (Figure 3).

60% of the great tits tested (n = 20 for each treatment) approached the loudspeaker when hearing conspecific calls, but only 30% when hearing black-capped chickadees calls

(Figure 2), and the difference between both treatments approached statistical significance (n = 40, $X^2 = 3.51$, df = 1, p = 0.06), with an odds ratio of 3.5 (95% CI [0.94; 12.97]). This difference was stronger than in Dutour et al. (2020), who found an odds ratio of 1.5 (95% CI [0.42; 5.24]) between the two treatments. Nonetheless, the confidence intervals of the effect sizes being large (Cumming 2007), the difference between our two studies cannot be considered as statistically significant (Figure 3).

Question 2: Response to reversed allopatric calls

Great tits scanned differently background noise, natural allopatric calls, and reversed allopatric calls (n = 60, $X^2 = 56.04$, df = 2, $p < 0.001$, Figure 2). Indeed, they scanned less to the background noise than to either of the two allopatric soundtracks (BN vs BC Natural: z = 7.41, $p < 0.001$; BN vs BC Reversed: z = 6.59, $p < 0.001$, Figure 2). They produced on average 6.8 ± 3.12 scans toward the natural calls, and 5.9 ± 3.54 scans toward the reversed, leading to an effect size of 0.16 (95% CI [-0.20; 0.48]), which is a non-statistically significant difference as indicated by post-hoc tests (z = 1.15, p = 0.48). Birds only produced 1.55 ± 2.06 scans when hearing control tests. In contrast, in Dutour et al. (2020), great tits scanned on average 14.3 ± 6.80 scans to the natural calls, 11 ± 6.55 scans to the reversed calls, and 8.85 ± 6.33 the control tests, leading to a substantial difference between natural and reversed calls (0.27, 95% CI [0.07; 1.52]) and no significant difference between reversed calls and background noise (0.21, 95% CI [-0.15; 0.52]). Nonetheless, the effect sizes associated to these differences remain comparable between the two studies (Figure 3).

Only 5% of great tits approached the loudspeaker when hearing background noise, but 30% when hearing naturally ordered allopatric calls and 10% when hearing reversed allopatric calls (Figure 2). Even though the odds ratio of the difference between our treatments was superior to 1 (Figure 3), it was not statistically significant ($X^2 = 4.29$, df = 2, p = 0.12). Our effect sizes parallel the ones from Dutour et al. (2020) who also did not detect statistically

significant difference between natural and reversed calls (Figure 3). Nonetheless, the percentage of approach in Dutour et al. (2020) were overall higher, with 55% of birds approaching in response to natural allopatric calls, 35% for the reversed calls, and 15% for BN.

Discussion

Two researchers with the same idea, very similar protocol and statistical analysis have obtained similar effect sizes for the differences of interest, nonetheless differed about the great tit's ability to respond to allopatric calls when considering results based on the p-value. Indeed, we detected a lower response to black-capped chickadees calls compared to conspecific ones, while the responses to both calls were similar in Dutour et al. (2020). We detected no difference between responses to natural and reversed allopatric calls while Dutour et al. (2020) detected one for scanning. While the difference between great tit and black-capped chickadee natural calls can easily be explained by a subtle protocol choice; the difference regarding the second question (i.e., effect of reversion on great tits' response) could be explained both by a protocol choice and/or by the p-value fluctuating especially with low sample sizes (n = 20 for each treatment in both studies). These two disparities are therefore of different kind and will be discussed below.

Allopatric versus conspecific mobbing calls

In our experiment, great tits approached less to allopatric calls than to conspecific ones, a result different from Dutour et al. (2020) who did not detect any difference. A lower response from GT to BC has previously been detected in Randler (2012), while a similar level of response was found in Dutour et al. (2017). One could hypothesize that such difference is explained by the distance of the loudspeaker from the focal bird (30 m for Dutour et al. 2020 versus 16 m for us). Indeed, amplitude of the sound is probably a proxy for

urgency in birds (Hingee & Magrath, 2009) and calls uttered at larger distance could consequently engender lower approach. In addition, increased distance implies both the attenuation of the sound (lower sound to noise ratio) and the degradation of some sound characteristics (e.g., high frequencies are degraded more easily, Kroodsma et al., 1982). Sound attenuation and degradation have been repeatedly shown to modify birds' response, especially in studies investigating anthropogenic noise (Jung et al., 2020; Shannon et al., 2016). In our situation, the differences between the mobbing calls of the allopatric black-capped chickadee and the sympatric marsh tit (who possess a similar mobbing call) could therefore be less salient at longer distances. However, two points should be raised: firstly, sound attenuation and degradation of a sound at 30 m (Dutour et al. 2020) versus 15 m (our experiment) in a semi-open environment is probably extremely similar. Secondly, the discriminative skills of Parids are known to be particularly precise. For example, black-capped chickadees and mountain chickadees (*Poecile gambeli*) can distinguish each other calls' based on features of their D notes (Bloomfield, Farrell, & Sturdy, 2008).

We rather suggest that such difference may lie in the soundtrack preparation, and particularly in the number of D notes. Indeed, D notes possess a general recruitment function in some species of Parids (Dutour et al., 2019; Suzuki et al., 2016) and the number of D notes per call is thought to code for urgency in Parids (Kalb et al., 2019b; Templeton et al., 2005). One recent study demonstrated that the increase of D notes may not be as important as the resulting increase of duty cycle (i.e., the amount of time a signal is present over a specified time, Landsborough et al., 2019). In our case, since black-capped chickadees' notes are longer than great tits' notes, each researcher chose to either control for the duty cycle of each type of note per call or for the number of D notes per call. Dutour et al. (2020) chose to control the number of D notes, with 8 notes per call, while we chose to control the duty cycle resulting in only 2 or 3 D notes per call (Figure 1). Future experiment disentangling the effect

of D note number and duty cycle with a crossed design may be of interest. Importantly however, even if the response to BC calls was lower, the effect sizes of the differences between natural and reversed order in our second question were similar in Dutour et al. (2020) and our own experiment: different choices in protocol did not hamper subsequent differences of interest.

Difference in scan number

Our second disparity lies in the difference in scanning behaviour for the second question. The absolute number of scans was extremely different, with rarely more than 10 scans counted in our study, while most observations from Dutour et al. counted more than 10 scans. In addition, the difference between control and reversed BC playbacks was strong in our study, but not significant in Dutour et al. (2020). The scanning variable could be criticized: counting 180° head turn in real time may be difficult and is probably impacted by the observer's personal definition of scanning. However, the two observers in our study only varied in their scan number for 1 scan on average. Such a result is in accordance with Dutour et al. (2019) who tested the differences in scan count between two experienced ornithologists and detected a high concordance between observers. The difference in absolute scores between our two studies may consequently rather be explained by the context in which the birds were tested. Indeed, while we tested free ranging birds while foraging, Dutour et al. tested birds when arriving at their nest box. In particular, birds are probably more vigilant (hence increasing the number of scan) in the vicinity of their nest, and the perceived risk associated to conspecific and allopatric calls could also differ according to the distance of the caller from the nest. This subtle variation of context between both studies could thus well explain both the stronger difference between BN and Reversed playback in our study compared to Dutour et al. (2020), and the overall disparities of the absolutes scores between the two studies. A question that remains is whether, in addition to difference in absolute

scores of scanning, such difference in context may also affect the differences between treatments. Since the effect sizes of the differences between treatments were similar between our experiment and the one from Dutour et al. (2020), we think that the context overall increased the scanning behaviour but did not affect the differences between treatments.

Similar effect sizes, but dissimilar p-values

Obtaining similar effect sizes of the difference between natural and reversed calls indicates two important things. Firstly, this indicates that even if Dutour et al. (2020) were not fully blinded when doing their playback tests, they were not affected by an expectancy effect (i.e., unknowingly distorting the observations to make them fit with your hypothesis, Holman et al., 2015; Rosenthal & Fode, 1963). Secondly, obtaining similar effect sizes but dissimilar p-values between the two studies indicates a discrepancy between effect sizes and analyses based on p-values. The use of p-value is increasingly criticized (Anderson et al., 2000). Indeed, p-values are known to flicker even with great sample sizes (Halsey et al., 2015). In our case, natural variability combined with the difference of experimental design between both studies (i.e., completely random versus partly cross over design) could well have contributed to this phenomenon. Indeed, the slightly lower p-value reported by Dutour et al. (2020) may have arisen from a higher statistical power of the cross over design permitted by the subtraction of the predicted individual variability from the residual variance (i.e., through the inclusion of a random individual effect). Unfortunately, the estimate of the subject effect was not reported in Dutour et al. (2020) precluding the possibility to examine this point more formally. This emphasizes the need to rely more on effect sizes and on the biological relevance of them (Nakagawa & Cuthill, 2007), especially since they seem more stable with low sample sizes (Halsey et al. 2015). Clustering several tiny, repeated experiments may be another solution to control natural and protocol variability (von Kortzfleisch et al., 2020). More generally, the various sources of variability between the two

experiments (protocol choices and natural between-year variability) show how much replicated studies and meta-analysis approaches are needed.

Conclusion

In conclusion, we found that the context in which the birds are tested (here, different distances from the nest) as much as the playback preparation can modify the behavioural cues assessed in language related studies. These different protocol choices seem to have mainly affected the absolute scores rather than the differences between treatments, as we found similar effect sizes between the two experiments. However, relying only on the p-value would here have led to different biological conclusions regarding complex syntax use in great tits. We believe this work provides a clear demonstration that the confrontation between two similar experiments is not a matter of who did wrong and who did right, but rather that both experiments grabbed one aspect of reality, at one precise moment. In our field of research, the flexibility present in behavioural measures and the limited sample sizes are probably the major explanations for disparities between similar experiments. Repeated experiments are therefore one great opportunity to understand variability in natural experiments and to approach, at best, biological reality.

TABLES

Table 1. Protocol comparison between the experiment of M. Dutour & et al. and A. Salis et al. Experiments consisted in recording behaviour of birds when hearing specific soundtracks. We listed the factors that could potentially influence different results in the two studies. Bold text emphasizes the differences between protocols. CRD = Completely randomized design, GLM = generalized linear model.

Protocol Choices	**Salis et al. (Question 1 & 2)**	**Dutour et al. (2020); Question 1**	**Dutour et al. (2020); Question 2**
Receiver species	Great tit	Great tit	Great tit
Emitter species	Black-capped chickadee	Black-capped chickadee	Black-capped chickadee
Season	Breeding Season	Beginning of Breeding Season	Breeding Season
Location of tests	North of Lyon, France	North of Lyon, France	North of Lyon, France
Bird tested	Free ranging birds	Free ranging birds	**Birds at nest boxes**
Number of playbacks	4	2	3
N per treatment	20	20	20 (repeated measures)
Experimental design	CRD	CRD	**Crossover design**
Distance sampling (m)	100	100	50 (nest boxes)
Control(s)	Background noise	Ø	Background noise
Soundtracks origin	Xeno-canto + Macaulay Library	Xeno-Canto + own recordings	Xeno-Canto + Macaulay Library
Control for Number of Notes or Duty Cycle?	**Duty cycle per call**	Number of D notes per call	Number of D notes per call
Distance with the loudspeaker	16 m (Approach = 8 m)	**30 m (Approach = 15 m)**	20 m (Approach = 10 m)
Double blind observation	**Yes (headphones)**	Partial (unaware but can hear the playback)	Partial (unaware but can hear the playback)
Variables of interest	Scan + Approach	Scan + Approach	Scan + Approach
Statistical analysis	GLM; Poisson & Binomial	GLM; Quasi Poisson & Binomial	GLM; Quasi Poisson & Binomial

FIGURES

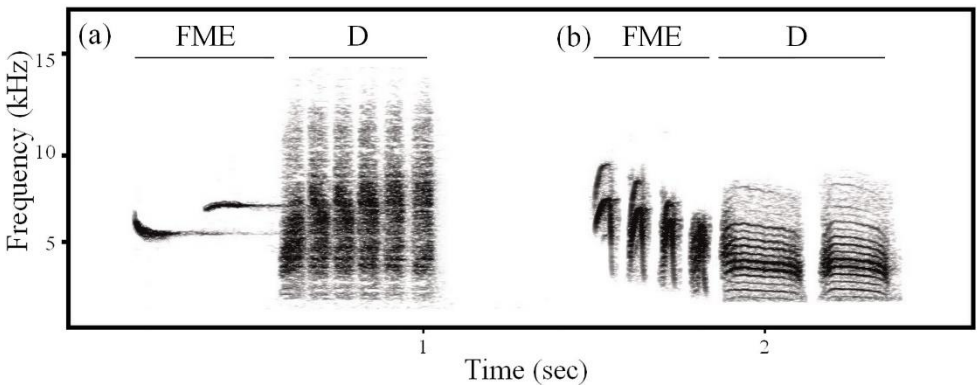

Figure 1. Spectrograms of a typical mobbing call of (a) great tits (*Parus major*), and (b) black-capped chickadees (*Poecile atricapillus*). X-axis is time (sec) and Y-axis is frequency (kHz). For the great tit, the combined FME and D notes generates a mobbing call sequence. The same principle is present for the black-capped chickadee. Made with Avisoft SASLab: Fs: 44 kHz, FFTLength 512; Bandwidth 324 Hz; Resolution 96 Hz.

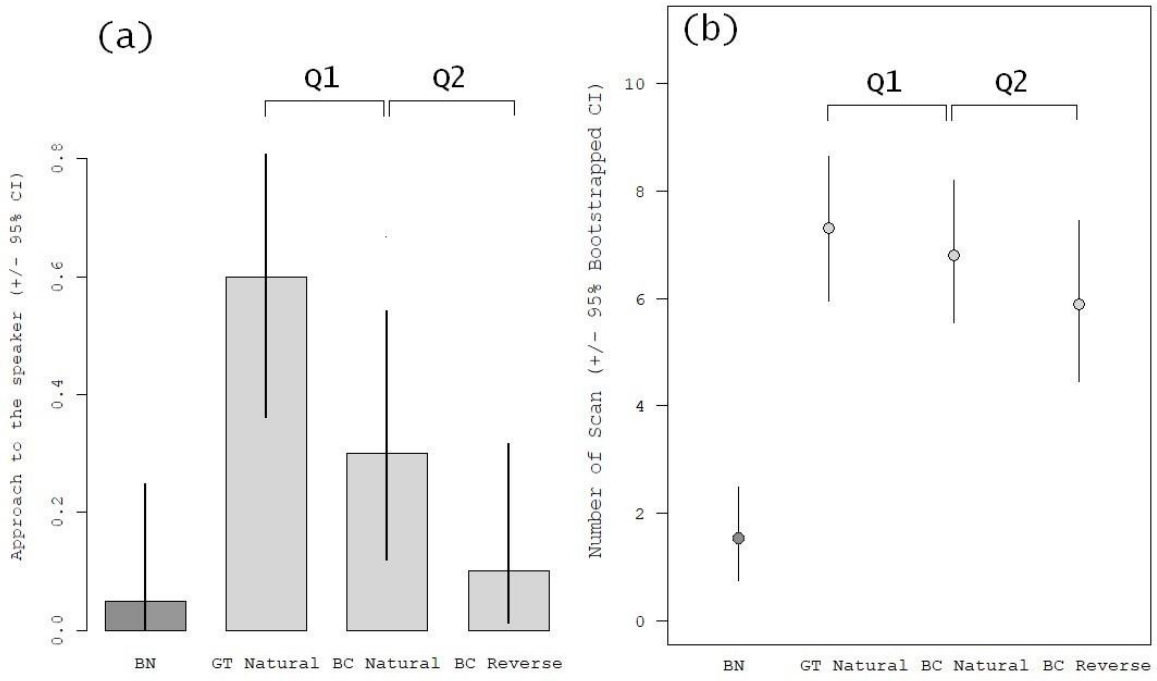

Figure 2. Results of our experiment, with (a) the proportion of individuals that approached the loudspeaker when hearing the different treatments, and (b) the number of scans (i.e., 180° turn of the head to explore the environment) made by great tits during the one minute test. For both figures, the 95% confidence intervals are given. Q1 represents the comparison of interest for the first question (emitter species comparison), comparable to the experiment 1 of Dutour et al. (2020). Q2 represents the comparison of interest for the second question (order comparison), comparable to the experiment 2 of Dutour et al. (2020). Statistical inference can be made using the overlap of such CI: if they overlap at less than halfway, the difference can be considered as statistically significant for an alpha = 5% (Cumming et al., 2007). BN = Background noise, GT = great tit, BC = black-capped chickadee.

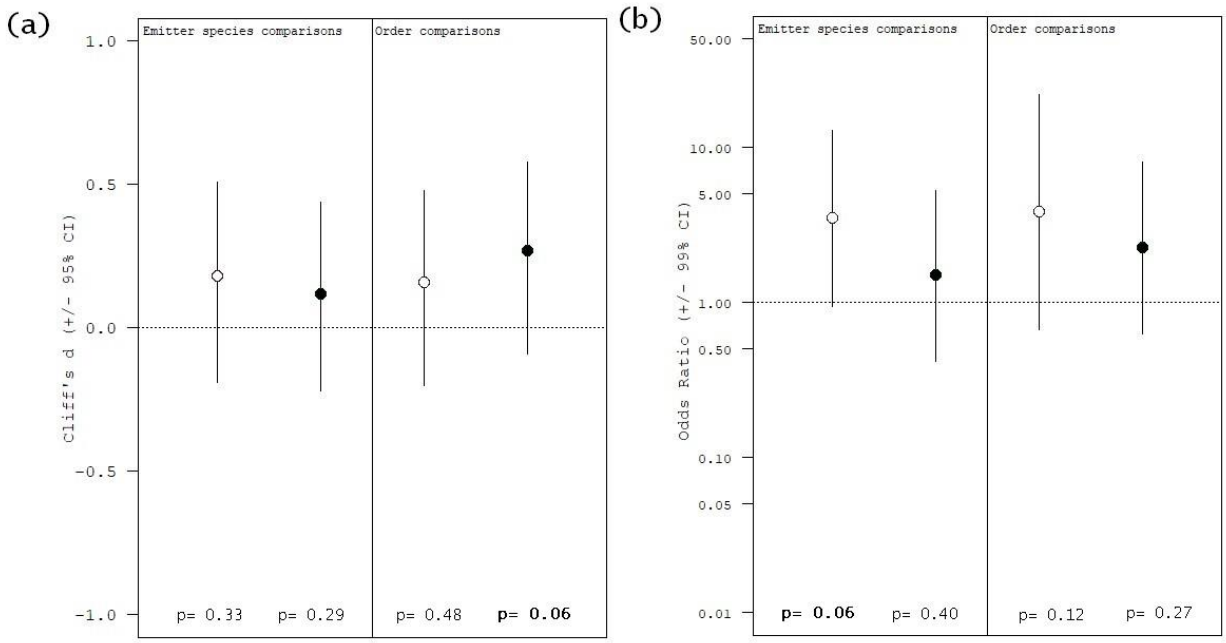

Figure 3. Comparison of effect sizes between our own experiment (white dots) and results from Dutour et al. (2020, black dots) who tested the same population with the same treatments. (a) Represents the comparisons for the scanning variable, using Cliff's D, and (b) the comparisons for the approach variable, with odds ratio. For each effect size given, the associated 95% confidence intervals are given. For both (a) and (b), left part concerns question 1 (species comparison: GT-Natural vs BC-Natural) and right part concerns question 2 (order comparison: BC-Natural vs BC-Reversed). Statistical inference can be made using the overlap of such CI: if they overlap at less than halfway, the difference can be considered as statistically significant for an alpha = 5% (Cumming et al., 2007). Associated p-value found in respective models are indicated below each comparison. A table summarizing the results and effect sizes of the two studies can be found in Sup.Mat.

SUPPLEMENTARY MATERIAL

Sup.Mat 1. Comparisons of p-value and effect sizes of both studies (Salis et al. or Dutour et al.) regarding the differences between treatments (Experiment 1: between species comparison, Experiment 2: order comparison). Effect sizes from the scanning behaviour are Cliff's d given with their 95% confidence intervals (CI). If such CI encompass 0, the difference can be considered as non-statistically significant. Effect sizes from the approach behaviour are Odds Ratio given with their 95% CI. If such CI encompass 1, the difference can be considered as non-statistically significant.

		Salis et al.		Dutour et al.	
Comparison	*Behaviour*	*Conclusion based on p-value*	*Effect size*	*Conclusion based on p-value*	*Effect size*
BC-Natural vs GT-Natural	Approach	Marginal effect (p = 0,06)	3.5 [0,94; 12,97]	No difference (p = 0,40)	1,5 [0,42; 5,24]
	Scan	No Difference (p = 0,33)	0.18 [-0.19; 0.51]	No difference (p = 0,29)	0,12 [-0,22; 0,44]
BC-Natural vs BC-Reversed	Approach	No Difference (p = 0,12)	3.85 [0,67; 22,11]	No difference (p = 0,27)	2,26 [0,63; 8,10]
	Scan	No Difference (p = 0,48)	0.16 [-0,20; 0,48]	Marginal effect (p = 0,06)	0,27 [-0,09; 0,58]
BC-Reversed vs Control	Approach	No difference (p = 0,80)	0,47 [0,04; 5,69]	No difference (p = 0,32)	0,33 [0,07; 1,52]
	Scan	Difference (p < 0,001)	0,73 [0,43; 0,89]	No difference (p = 0,35)	0,21 [-0,15; 0,52]

BIBLIOGRAPHY

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing : Problems, Prevalence, and an Alternative. *The Journal of Wildlife Management*, *64*, 912.

Baker, M. C., & Becker, A. M. (2002). Mobbing calls of black-capped chickadees: effects of urgency on call production. *The Wilson Bulletin*, *114*, 510-516.

Bloomfield, L. L., Farrell, T. M., & Sturdy, C. B. (2008). All “chick-a-dee” calls are not created equally Part II. Mechanisms for discrimination by sympatric and allopatric chickadees. *Behavioural Processes*, 77, 87-99.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science*, *349*, 910-911.

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018a). Meaningful syntactic structure in songbird vocalizations? *PLOS Biology*, *16*, e2005157.

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018b). The slings and arrows of comparative linguistics. *PLOS Biology*, *16*, e3000019.

Bryan, C. J., Yeager, D. S., & O’Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*, 25535-25545.

555 Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson,
556 E. S. J., & Munafò, M. R. (2013). Power failure : Why small sample size
557 undermines the reliability of neuroscience. *Nature Reviews Neuroscience*,
558 *14*, 365-376.
559 Carlson, N. V., Healy, S. D., & Templeton, C. N. (2018). Mobbing. *Current*
560 *Biology*, *28*, R1081-R1082.
561 Carlson, N. V., Pargeter, H. M., & Templeton, C. N. (2017). Sparrowhawk
562 movement, calling, and presence of dead conspecifics differentially
563 impact blue tit (Cyanistes caeruleus) vocal and behavioral mobbing
564 responses. *Behavioral Ecology and Sociobiology*, *71*, 133.
565 Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental
566 biology. *Journal of Cell Biology*, *177*, 7-11.
567 Dutour, M., Léna, J.-P., & Lengagne, T. (2017). Mobbing calls : A signal
568 transcending species boundaries. *Animal Behaviour*, *131*, 3-11.
569 Dutour, M., Lengagne, T., & Léna, J. (2019). Syntax manipulation changes
570 perception of mobbing call sequences across passerine species. *Ethology*,
571 *125*, 635-644.
572 Dutour, M., Suzuki, T. N., & Wheatcroft, D. (2020). Great tit responses to the
573 calls of an unfamiliar species suggest conserved perception of call
574 ordering. *Behavioral Ecology and Sociobiology*, *74*, 37.

Fanelli, D. (2018). Opinion : Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*, 2628-2631.

Griesser, M., Wheatcroft, D., & Suzuki, T. N. (2018). From bird calls to human language : Exploring the evolutionary drivers of compositional syntax. *Current Opinion in Behavioral Sciences*, *21*, 6-12.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*, 171511.

Hailman, J. P., Ficken, M. S., & Ficken, R. W. (1985). The 'chick-a-dee' calls of Parus atricapillus : A recombinant system of animal communication compared with written English. *Semiotica, 56,* 191-224.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*, 179-185.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*, e1002106.

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, *14*, e2000995.

Hingee, M., & Magrath, R. D. (2009). Flights of fear : A mechanical wing whistle sounds the alarm in a flocking bird. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 4173-4179.

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of Experimental Bias in the Life Sciences : Why We Need Blind Data Recording. *PLOS Biology*, *13*, e1002190.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*, e124.

Jung, H., Sherrod, A., LeBreux, S., Price, J. M., & Freeberg, T. M. (2020). Traffic noise and responses to a simulated approaching avian predator in mixed-species flocks of chickadees, titmice, and nuthatches. *Ethology*, *126*, 620-629.

Kalb, N., Anger, F., & Randler, C. (2019a). Great tits encode contextual information in their food and mobbing calls. *Royal Society Open Science*, *6*, 191210.

Kalb, N., Anger, F., & Randler, C. (2019b). Subtle variations in mobbing calls are predator-specific in great tits (Parus major). *Scientific Reports*, *9*, 6572.

Krebs, J. R. (1971). Territory and Breeding Density in the Great Tit, Parus major L. *Ecology*, *52*, 2-22.

Kroodsma, D. E., Miller, E. H., & Ouellet, H. (Éds.). (1982). *Acoustic communication in birds*. New York: Academic Press.

Landsborough, B., Wilson, D. R., & Mennill, D. J. (2019). Variation in chick-a-dee call sequences, not in the fine structure of chick-a-dee calls, influences mobbing behaviour in mixed-species flocks. *Behavioral Ecology*, *31*, 54-62.

Lash, T. L., Collin, L. J., & Van Dyke, M. E. (2018). The Replication Crisis in Epidemiology : Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports*, *5*, 175-183.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*, 487-498.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance : A practical guide for biologists. *Biological Reviews*, *82*, 591-605.

Otter, K. A. (Éd.). (2007). *The ecology and behavior of chickadees and titmice : An integrated approach*. Oxford ; New York: Oxford University Press.

Randler, C. (2012). A possible phylogenetically conserved urgency response of great tits (Parus major) towards allopatric mobbing calls. *Behavioral Ecology and Sociobiology*, *66*, 675-681.

Rosenthal, R., & Fode, K. L. (2007). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*, 183-189.

638 Salis, A., Léna, J., & Lengagne, T. (2020). Great tits (*Parus major*) adequately
639 respond to both allopatric combinatorial mobbing calls and their isolated
640 parts. *Ethology*, eth.13111.
641 Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings
642 in psychology : Is lack of replication the real problem? Or is it something
643 else? *Archives of Scientific Psychology*, *4*, 32-37.
644 Schwagmeyer, P. L., & Mock, D. W. (1997). How to minimize sample sizes
645 while preserving statistical power. *Animal Behaviour*, *54*, 470-474.
646 Shannon, G., McKenna, M. F., Angeloni, L. M., Crooks, K. R., Fristrup, K. M.,
647 Brown, E., … Wittemyer, G. (2016). A synthesis of two decades of
648 research documenting the effects of noise on wildlife : Effects of
649 anthropogenic noise on wildlife. *Biological Reviews*, *91*, 982-1005.
650 Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E.,
651 … Nosek, B. A. (2018). Many Analysts, One Data Set : Making
652 Transparent How Variations in Analytic Choices Affect Results.
653 *Advances in Methods and Practices in Psychological Science*, *1*, 337-356.
654 Suzuki, T. N., Griesser, M., & Wheatcroft, D. (2019). Syntactic rules in avian
655 vocal sequences as a window into the evolution of compositionality.
656 *Animal Behaviour*, *151*, 267-274.
657 Suzuki, T. N., Wheatcroft, D., & Griesser, M. (2016). Experimental evidence
658 for compositional syntax in bird calls. *Nature Communications*, *7*, 10986.

659 Suzuki, T. N., Wheatcroft, D., & Griesser, M. (2020). The syntax–semantics
660 interface in animal vocal communication. *Philosophical Transactions of*
661 *the Royal Society B: Biological Sciences*, *375*, 20180405.
662 Tang-Martínez, Z. (2020). The history and impact of women in animal
663 behaviour and the ABS : A North American perspective. *Animal*
664 *Behaviour*, *164*, 251-260.
665 Templeton, C. N, Greene, E., Davis, K. (2005). Allometry of Alarm Calls :
666 Black-Capped Chickadees Encode Information About Predator Size.
667 *Science*, *308*, 1934-1937.
668 von Kortzfleisch, V. T., Karp, N. A., Palme, R., Kaiser, S., Sachser, N., &
669 Richter, S. H. (2020). Improving reproducibility in animal research by
670 splitting the study population into several 'mini-experiments'. *Scientific*
671 *Reports*, *10*, 16579.
672 Wilkin, T. A., Garant, D., Gosler, A. G., & Sheldon, B. C. (2006). Density
673 effects on life-history traits in a wild population of the great tit Parus
674 major : Analyses of long-term data with GIS techniques: Great tit
675 breeding density. *Journal of Animal Ecology*, *75*, 604-615.
676