Measurement artefacts lead to false positives in the study of birdsong in noise

Henrik Brumm*[,1], Sue Anne Zollinger[1], Petri T. Niemelä[2] and Philipp Sprau[2]

[1]*Communication and Social Behaviour Group, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany; and*
[2]*Behavioural Ecology, Department of Biology, Ludwig-Maximilians-University Munich, 82152 Planegg-Martinsried, Germany*

Summary

1. Numerous studies over the past decade have reported correlations between elevated levels of anthropogenic noise and a rise in the minimum frequency of acoustic signals of animals living in noisy habitats. This pattern appears to be occurring globally, and higher pitched signals have been hypothesized to be adaptive changes that reduce masking by low-frequency traffic noise. However, the sound analysis methods most often used in these studies are prone to measurement errors that can result in false positives. In addition, the commonly used method of measuring frequencies visually from spectrograms might also lead to observer-expectancy biases that could exacerbate measurement errors.

2. We conducted an experiment to (i) quantify the size and type of errors that result from 'eye-balling' frequency measurements with cursors placed manually on spectrograms of signals recorded in noise and no-noise conditions, and (ii) to test whether observer expectations lead to significant errors in frequency measurements. We asked 54 volunteers, blind to the true intention of our study, to visually measure the minimum frequency of a variety of natural and synthesized bird sounds, recorded either in noise, or no-noise conditions. Test subjects were either informed or uninformed about the hypothesized results of the measurements.

3. Our results demonstrate that inappropriate methodology in acoustic analysis can yield false positives with effect sizes as large, or even larger, than those reported in published studies. In addition to these measurement artefacts, psychological observer biases also led to false positives – when observers expected signals to have higher minimum frequencies in noise, they measured significantly higher minimum frequencies than uninformed observers, who had not been primed with any expectation.

4. The use of improper analysis methods in bioacoustics can lead to the publication of spurious results. We discuss alternative methods that yield unbiased frequency measures and we caution that it is imperative for researchers to familiarize themselves both with the functions and limitations of their sound analysis programmes. In addition, observer-expectancy biases are a potential source of error not only in the field of bioacoustics, but in any situation where measurements can be influenced by human subjectivity.

Key-words: animal communication, anthropogenic noise, ecological novelty, observer bias, repeatability, song frequency, sound analysis, spectrogram, urban ecology

Introduction

The study of the effects of anthropogenic noise on animal communication is currently receiving increasing interest in the fields of behavioural ecology and evolution. This is because a better understanding of noise pollution has far reaching implications for the mechanisms of signal production and perception, the behavioural ecology of signalling, the evolution of communication systems, and conservation issues (Endler 1992; Brumm 2013; Wiley 2015).

Numerous recent studies have reported elevated minimum frequencies of bird vocalizations in areas with intense anthropogenic noise (reviewed in Gil & Brumm 2014). This phenomenon appears to be widespread, as it has been observed in many bird species all over the world from Europe to Asia and the Americas (reviewed in Brumm & Zollinger 2013). Higher minimum frequencies have been suggested to be adaptive because they may reduce acoustic masking by low-frequency anthropogenic noise (e.g. Slabbekoorn & Peet 2003; Hu & Cardoso 2010; Montague, Danek-Gontard & Kunc 2013). The majority of studies on birdsong in noisy environments measured song frequencies visually from spectrograms using, for example, on-screen cursors that are featured in most sound analysis programmes. For instance, of the 40 published field studies reviewed by Brumm & Zollinger (2013) and Roca *et al.* (2016), 19 used visual measurements and a further nine did not mention the method used (cf. Brumm & Bee 2016). However, the practice of extracting frequency measurements visually from spectrograms is potentially prone to bias (Greenewalt 1968; Beecher 1988) as there are numerous problems with it (Zollinger *et al.* 2012). A particularly relevant problem for the study of birdsong in noisy environments is that two sounds,

*Correspondence author. E-mail: brumm@orn.mpg.de

which are identical in both frequency and amplitude, can appear markedly different in a spectrogram if there is some other, higher amplitude sound in the background of one of the recordings that is not present in the other. Because the amplitude scaling of uncalibrated spectrograms is adjusted to the highest amplitude, high levels of noise will result in the lower-amplitude signal being displayed with a smaller frequency bandwidth compared with a spectrogram of the signal without the noise (Zollinger *et al*. 2012). Another problem arises from the fact that masking noise can make it difficult to detect the actual lower frequency end of the signal from the spectrogram tracing. In this case, the lowest frequencies detectable for measurement would increase with increasing noise level. Both problems may result in a measuring artefact giving the false impression of a positive relationship between minimum signal frequency and noise (Zollinger *et al*. 2012; Grace & Anderson 2015; Rios-Chelén, Lee & Patricelli 2016).

In addition to these inherent problems with the practice of 'eye-balling' cursor placement, visual measurements from spectrograms are likely to also be prone to observer bias if the person taking the measurements has certain expectations in mind. Such psychological observer biases, although often unconscious, are classic examples of sources of measurement errors in animal behaviour studies (Martin & Bateson 2007). None of the published studies of birdsong in noise that visually extracted minimum song frequencies mention whether or not uninformed observers measured the songs. In cases in which informed observers eye-balled the spectrograms, it cannot be ruled out that an (unconscious) observer bias affected the results, exacerbating any other measurement errors.

Although a potential observer-expectancy bias has not been quantified so far, recent studies have shown that the systematic measurement error of the eye-balling practice can be as high as 0·3–1·8 kHz (Zollinger *et al*. 2012; Grace & Anderson 2015; Rios-Chelén, Lee & Patricelli 2016; Rios-Chelén *et al*. 2017), which is substantial in relation to the reported noise-related frequency shifts in birdsongs. We know of more than two dozen published studies on birdsong in noise-polluted areas that visually extracted song frequencies from spectrograms and several additional studies that do not describe how frequencies were measured at all. While we do not want to point fingers at individual studies, the mean noise-related increase in minimum frequencies that they reported were between 0·03 and 0·9 kHz, which means that the reported effects tend to be smaller than the potential systematic error. This problem raises the question of whether the phenomenon of increased minimum song frequencies in urban birds might be less widespread than is commonly assumed (cf. Brumm & Zollinger 2013; Rios-Chelén *et al*. 2013).

Understanding the methodological concerns involved in research is crucial for assessing the validity of data and the conclusions drawn from them. However, many studies on birdsong in noise do not seem to be aware of the biases introduced by inappropriate methods and despite the many problems with the eye-balling practice, studies using this method continue to be published. The continuing publication of potential measuring artefacts may, at least partly, be due to the fact that researchers are still being encouraged to eye-ball acoustic parameters from spectrograms (e.g. Cardoso & Atwell 2012; Job, Kohler & Gill 2016; Narango & Rodewald 2016).

Here, we provide evidence for the magnitude of the errors resulting from the practice of visually extracting minimum song frequencies from spectrograms, using examples from one of the most extensively studied bird species in the field of urban bioacoustics, the great tit (*Parus major*) as well as synthesized signals. We asked human test subjects to visually measure the minimum frequency of a set of identical signals, recorded either in noise or no-noise conditions. In addition to the mean measurement errors that occur from the masking of the signal, we also quantified for the first time (i) a potential observer-expectancy bias, which might be introduced by observers who are informed about the presumed effects of noise on minimum song frequencies, and (ii) the repeatability of eye-balling frequencies. Repeatability in this context provides information on how much variation in the data is explained by differences between observers, i.e. to what extent individuals differ from each other in measuring minimum frequencies.

Materials and methods

CONSTRUCTION OF TEST SOUNDS

As source material for the song measurement tests, we used high-quality recordings from different great tit populations in Germany, the Netherlands and Switzerland. These were made with a digital recorder (Edirol R-09 (Roland Corporation, Hamamatsu, Japan) or Marantz PMD 660 (Kawasaki, Japan), sampling frequency: 44·1 kHz) in combination with a Telinga Stereo DAT microphone (Tobo, Sweden) mounted in a parabolic dish or a Sennheiser ME66 directional microphone (Wedemark, Germany). For details of the recording procedures see Ritschard *et al*. (2012). From these recordings, we selected 10 songs, each from a different male, in which the element with the lowest frequency was frequency modulated. These songs were edited using the software Avisoft SASLab Pro v. 5.2.08 (Avisoft Bioacoustics, Berlin, Germany). First, the peak amplitudes of each of the 10 songs were normalized to obtain equal maximum amplitudes for all songs while maintaining the relative amplitude differences between the elements within each song. We then copied the lowest-frequency element of each song into one single WAVE file retaining the original sample rate of 44·1 kHz. In addition to these natural song elements, we also synthesized eight elements (Table 1) using Avisoft SASLab Pro with a sampling rate of 44·1 kHz, with 16-bit depth. These

Table 1. Acoustic properties of the eight synthesized elements. The element shape is the characteristic of the frequency modulation according to a sine curve for angles changing in the denoted range

Stimulus ID	Minimum frequency (Hz)	Shape	Amplitude (dB)
1	1500	$\sin(\frac{3}{2}\pi - 2\pi)$	0
2	1500	$\sin(0 - \frac{1}{2}\pi)$	−6
3	1500	$\sin(\frac{1}{2}\pi - \pi)$	−3
4	1300	$\sin(\frac{1}{2}\pi - \pi)$	−9
5	1550	$\sin(\pi - \frac{3}{2}\pi)$	−12
6	2200	$\sin(\pi - \frac{3}{2}\pi)$	−9
7	1800	$\sin(\frac{1}{2}\pi - \pi)$	−6
8	1100	$\sin(\frac{1}{2}\pi - \pi)$	−3

synthesized song elements allowed us to assess the robustness of the measuring method itself because they provide ground-truth data of known spectral content that can be compared with the values measured by the test subjects, i.e. observers (details see below).

We also created synthesized background noise that was used for the preparation of the test stimuli (see below). This noise was based on recordings from a total of 50 min of traffic noise recorded at five different locations between 08.00 and 19.00 h in bird habitats in the city of Munich, Germany. The traffic noise recordings were made with a Sennheiser ME62 microphone and a Marantz PMD660 solid-state recorder (44·1 kHz, 16 bit). From these 50 min of traffic noise recordings we calculated an averaged power spectrum (using the function 'Power spectrum (averaged)' in Avisoft), which was then used as a filter for synthesized white noise (using the Frequency Domain Transformation function in Avisoft). Thus, the resulting filtered noise had the same spectral shape as the average natural traffic noise. To produce the test stimuli, all song elements (natural and synthetic) were re-recorded twice in an anechoic room (located at the Max Planck Institute for Ornithology in Seewiesen, Germany), with and without the synthesized noise (Fig. 1). The anechoic room was a custom-built floating room (3·5 m × 3·1 m and 2·4 m high, completely lined with 30-cm-deep pyramidal sound absorbers). A mesh grille over the sound absorbers on the floor allowed access to the chamber and setting up the equipment. By using an anechoic chamber we were able to re-record the test stimuli under acoustic free-field conditions with extremely low background noise levels (<20 dB(A), re. 20 μPa). All sounds were played from a computer and fed through an amplifier (Technics SU-V300M2, Panasonic Corporation, Kadoma, Japan) to two loudspeakers (JBL Control 1 Pro, Los Angeles, CA, USA) placed next to each other at a height of 1·2 m. The broadcast sounds were recorded with a digital recorder (Marantz PMD 660, 44·1 kHz sampling frequency) connected to a microphone (Sennheiser ME62) that was placed at a distance of 1·4 m from the loudspeakers facing the loudspeaker membranes. The loudspeakers and the microphone inside the anechoic room were connected by cables to the amplifier and recorder, which were both placed outside of the anechoic room. Peak amplitudes of the song playback were set at L_{FA} = 70 dB SPL at the position of the recording microphone. Given the natural song amplitudes of great tits (Blumenrath & Dabelsteen 2004), the playback amplitude simulated birds approx. 8 m away, which is about the typical recording distance in studies on urban great tit song (Mockford & Marshall 2009; Huffeldt & Dabelsteen 2013). The noise amplitude was set at 60 dB(A) SPL, which is within the range of traffic noise levels encountered by great tits and other birds in urban habitats (Slabbekoorn & Peet 2003; Brumm 2004; Dorado-Correa, Rodríguez-Rocha & Brumm 2016).

Finally, the test stimuli (i.e. test files for the observers) were created by combining the 18 song elements (10 natural and 8 synthetic) that were re-recorded in two different treatments (with and without noise) in one single file containing 36 song elements. Each rendition of each element (with and without noise) was included twice in the file to obtain repeated measurements (see below). Therefore, one test stimulus contained altogether 72 song elements. The order of the elements was systematically varied between three different test stimuli (A, B and C).

FREQUENCY MEASUREMENTS

The minimum frequency of each song element was measured by 54 students and researchers from the Ludwig Maximilians University Munich and the Max Planck Institute for Ornithology. To this end, the test persons were asked to visually extract the minimum frequencies using the on-screen cursor in Avisoft SASLab Pro, which was placed on a spectrogram of the recording (Fig. 1). To increase the frequency resolution of the measurement we down sampled all recordings to 16 kHz and spectrograms were calculated using a FFT size of 512, frame size 100%, Hamming window. These settings resulted in a temporal resolution of 32 ms and a frequency resolution of 31 Hz. We choose this frequency resolution because it is higher than the resolution used in published studies that visually extracted minimum frequencies of birdsongs in anthropogenic noise. In the eye-balling studies that we know of, the frequency resolution of the spectrograms used ranged from 43 to 344 Hz (in some cases, however, the frequency resolution could not be calculated because the studies did not mention the spectrogram settings and the sampling frequency of the recordings). Because of the higher frequency resolution used in our study, any difference in minimum frequency would be easier to detect visually, which means that our test most likely underestimates the potential error of previous studies.

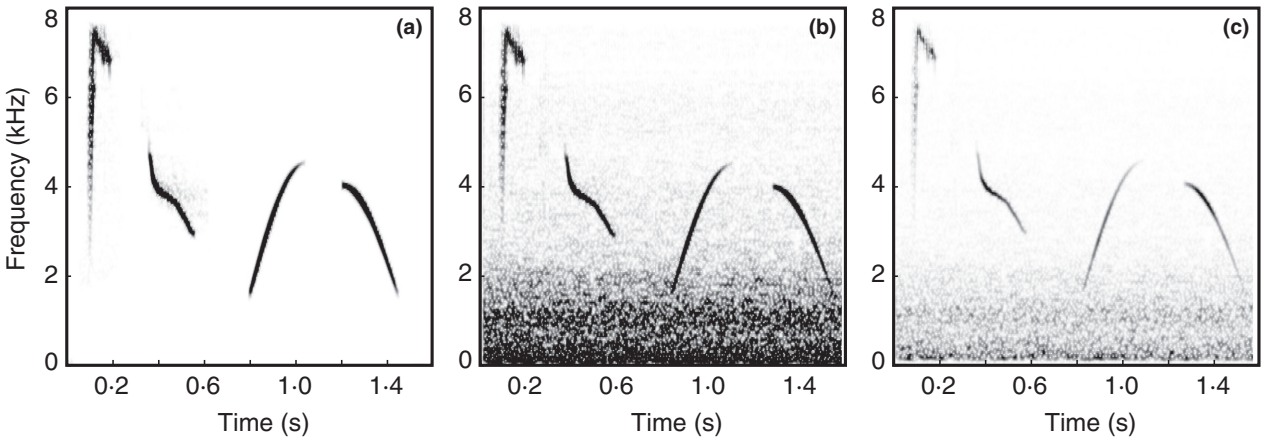

Fig. 1. Exemplary test sounds measured by the test subjects. (a) two great tit song element and two synthesized elements with no background noise (in total 10 different great tit song elements from different individuals and eight different synthesized elements were used) are shown in spectrograms of the same resolution used in the experiment (16 kHz sampling rate, 512 point FFT, Hamming window). Test subjects could adjust the spectrographic display contrast and darkness to their liking. Panel (b) shows the same four test elements in masking noise, drawn with the same spectral and display settings as the spectrogram in panel (a). (c) A second example of the same four elements in noise, demonstrating how adjusting the spectrogram contrast and display settings can lessen the intensity of noise displayed, but that this also results in reduced visualization of the elements, particularly at the lowest amplitude portions of each element.

Observers (i.e. test subjects) were randomly assigned to two treatment groups: informed observers ($N = 27$) and uninformed observers ($N = 27$). The informed observers were told that their measurements are part of a study testing the hypothesis that birds sing at higher frequencies in noisy environments. The subjects of both groups were not informed (i.e. naïve) about the true goal of the study prior to their measurements nor were they told that they measured each element several times. All observers were instructed by the experimenter to measure the minimum song frequencies by placing the cursor on the lowest song frequency in the spectrogram. Each observer could change the intensity threshold of the spectrogram to yield the best visibility of the song tracing for each measurement. The participation of observers in the study was voluntary and observers gave their informed consent to have their anonymous data used in this study.

STATISTICAL METHODS

Estimating mean minimum frequencies

To assess average differences in measured minimum frequencies between the two treatments (informed vs. uninformed), noise levels (noise vs. no-noise) and their interaction, we first constructed a univariate mixed-effects model with minimum frequency as response variable (Table 2). Stimulus order (three levels of element arrangement: A, B and C; factor), signal type (natural vs. artificial; factor) and sequence (sequence in which the 72 elements were scored for minimum frequency within individuals: squared and unsquared to control for linear and nonlinear learning or fatigue during the trial; mean centred covariate) were also included as fixed effects to control for potential bias caused by the experimental design. Observer and element identities were fitted as random effects to control for among-observer and among-element variation.

Table 2. Results of univariate linear mixed-effects model for pooled minimum frequency data to study the mean differences between different treatment groups and the interaction between groups. We present fixed (β) parameters and random (σ^2) parameters with their standard errors, and F-statistics for the fixed parameters with their *P*-values

Fixed effects*	β (SE)	$F_{(\text{NUMdf, DENdf})}$	*P*
Intercept	2753 (74·45)	$1394{\cdot}91_{1,17{\cdot}2}$	<0·001
Sequence	2·50 (0·90)	$7{\cdot}78_{1,3812{\cdot}4}$	0·006
Sequence squared	−0·04 (0·01)	$9{\cdot}80_{1,3812{\cdot}1}$	0·002
Signal type	−1225 (105·8)	$134{\cdot}22_{1,16{\cdot}0}$	<0·001
Stimulus order		$1{\cdot}74_{2,50{\cdot}5}$	0·185
B	−39·73 (27·43)		
C	−47·87 (27·43)		
Noise	266·5 (14·31)	$346{\cdot}83_{1,3809{\cdot}0}$	<0·001
Observer-expectancy bias	−20·80 (24·81)	$0{\cdot}70_{1,66{\cdot}9}$	0·405
Observer-expectancy bias × Noise	137·40 (18·29)	$56{\cdot}48_{1,3809{\cdot}0}$	<·001
Random effects	**σ^2 (SE)**		
Individual ID	5696·5 (1354·1)	–	–
Stimulus ID	49 360 (17 578)	–	–
Residual	77 162 (1768·1)	–	–

*In trait signal type, reference group is natural song; In noise, reference group is no-noise; in Stimulus order, reference group is A; in Observer-expectancy bias, reference group is uninformed; covariates are mean centred.

In some cases, mixed-effects models assuming equal slopes across the levels within a random effect (e.g. assuming that the magnitude of change between two measurements is equal across individuals) might yield optimistic standard errors around the parameter estimates of fixed effects. This leads to upwards bias in *P*-values, facilitating type I errors (Schielzeth & Forstmeier 2009). To test if such bias was present in our models, we also fitted a random slopes model in which this potential bias is controlled for by allowing slopes to vary among the levels within a random effect by including the interaction term between the focal random effect and the focal fixed effect in the model (Schielzeth & Forstmeier 2009). In our case, these interactions terms were as follows: observer and noise level, observer and signal type, observer and sequence, element and noise level, element and stimulus, element and treatment and element and sequence. As neither the point estimates nor the *P*-values of the fixed effects changed when fitting a random slopes model, we present here only the simpler model with fewer estimated parameters (as described above).

Estimating variance components

To assess individual-level variance components and their repeatabilities for each treatment group separately, we further constructed four univariate mixed-effects models, i.e. one for each treatment group (informed/noise, informed/no-noise, uninformed/noise and uninformed/no-noise) with minimum frequency as the response variable (Table 3a). Measuring repeatability, i.e. the amount of total phenotypic variance explained by the individual, is a standardized way to express individual variance and thus allows comparisons of the relative bias caused by the individual between groups or populations. In each case, we fitted random intercepts for individual and song element identity; this enabled us to decompose the phenotypic variance into variance attributable to individual identity, element identity (fitted to control for between-element variation) and within-individual within-element residual. In other words, we were able to estimate to which degree individuals differed from each other in their average minimum frequency measurements. In all models, we also included signal type (natural vs. synthetic element; factor), stimulus order (three levels of element arrangement: A, B and C; factor) and sequence as fixed effects. This enabled us to control for potential bias caused by the experimental design. Repeatability was calculated by dividing a focal variance component by the total phenotypic variation not attributable to fixed effects (Nakagawa & Schielzeth 2010; Dingemanse & Dochtermann 2013). Statistical significance of a focal random effect was assessed by applying a likelihood ratio test (LRT) assuming an equal mixture of $\chi^2(0)$ and $\chi^2(1)$ degrees of freedom, as suggested by Self & Liang (1987) and Visscher (2006). This χ^2-distributed test statistic was calculated as twice the difference in Log Likelihood between the initial model (detailed above) and a model excluding the focal random effect (Meyer 1992; Wilson *et al.* 2010). Statistical significance of fixed effects was based on the Wald F-statistics and numerator and denominator degrees of freedom (Gilmour, Gogel & Cullis 2009).

Comparing individual variance across treatments

One multivariate mixed-effect model was constructed to test whether variance attributable to individual identity differed between treatment groups. We were interested in differences in between-observer variance between treatments because such variation causes different magnitudes of potential observer bias. Variation between observers simply means that some observers overestimate minimum frequencies, some underestimate them and some provide unbiased estimates. Thus, if

Table 3. Results of (a) four univariate linear mixed-effect models: one for each unique treatment group combinations to estimate individual-level variance components (V_i) and repeatabilities (R) for each group and (b) comparison of individual-level variance (V_i; lower diagonal) and repeatability (R_i; upper diagonal) estimates across all four treatment groups (derived from one multivariate model: full model output not shown here, see Methods). We present fixed (β) parameters and random (σ^2) parameters and repeatabilities (R) for individual level with their standard errors, and F-statistics (for the fixed parameters) and χ^2-values (for the random parameters) with their P-values

(a)	Informed/noise			Informed/no-noise		
Fixed effects*	β (SE)	$F_{(NUMdf, DENdf)}$	P	β (SE)	$F_{(NUMdf, DENdf)}$	P
Intercept	3177·09 (97·38)	$1349{\cdot}52_{1,24{\cdot}6}$	<0·001	2711·72 (90·20)	$895{\cdot}33_{1,16{\cdot}2}$	<0·001
Sequence	−0·26 (1·58)	$0{\cdot}03_{1,1139{\cdot}6}$	0·865	2·35 (1·45)	$2{\cdot}61_{1,1140{\cdot}3}$	0·109
Sequence squared	<0·01 (0·02)	$<0{\cdot}01_{1,1138{\cdot}9}$	0·981	−0·03 (0·02)	$3{\cdot}01_{1,1140{\cdot}5}$	0·085
Signal type	−1262·73 (118·98)	$112{\cdot}63_{1,16{\cdot}0}$	<0·001	−1152·43 (133·58)	$74{\cdot}43_{1,16{\cdot}0}$	<0·001
Stimulus order		$0{\cdot}67_{2,30{\cdot}0}$	0·518		$5{\cdot}87_{2,30{\cdot}3}$	0·007
B	−83·01 (77·09)			−73·21 (21·50)		
C	−73·87 (78·67)			−47·12 (21·99)		
Random effects	σ^2 (SE)	$\chi^2_{0/1}$	P	σ^2 (SE)	$\chi^2_{0/1}$	P
Individual ID	30 766 (8366·6)	370·14	<0·001	912·14 (655·44)	3·4	0·026
Stimulus ID	62 105 (22 266)	–	–	78 405 (28 035)	–	–
Residual	58 825 (2470·4)	–	–	57 920 (2431·4)	–	–
Repeatability	R (SE)	$\chi^2_{0/1}$	P	R (SE)	$\chi^2_{0/1}$	P
Individual ID	0·203 (0·053)	370·14	<0·001	0·007 (0·005)	3·4	0·026
	Uninformed/noise			Uninformed/no-noise		
Fixed effects	β (SE)	$F_{(NUMdf, DENdf)}$	P	β (SE)	$F_{(NUMdf, DENdf)}$	P
Intercept	3033 (76·74)	$1722{\cdot}12_{1,19{\cdot}2}$	<0·001	2721 (99·16)	$798{\cdot}63_{1,18{\cdot}0}$	<0·001
Sequence	−1·92 (1·51)	$1{\cdot}63_{1,719{\cdot}3}$	0·205	2·99 (1·69)	$3{\cdot}15_{1,720{\cdot}1}$	0·078
Sequence squared	0·01 (0·02)	$0{\cdot}23_{1,718{\cdot}5}$	0·629	−0·04 (0·02)	$2{\cdot}89_{1,720{\cdot}4}$	0·092
Signal type	−1320·41 (106·82)	$152{\cdot}90_{1,16{\cdot}0}$	<0·001	−1184·59 (141·36)	$70{\cdot}42_{1,16{\cdot}0}$	<0·001
Stimulus order		$1{\cdot}03_{2,18{\cdot}0}$	0·387		$0{\cdot}72_{2,18{\cdot}0}$	0·503
B	41·33 (44·85)			6·11 (48·65)		
C	−24·49 (43·00)			−46·97 (46·64)		
Random effects	σ^2 (SE)	$\chi^2_{0/1}$	P	σ^2 (SE)	$\chi^2_{0/1}$	P
Individual ID	5971·3 (2297·8)	72·1	<0·001	6737·1 (2704·4)	51·42	<0·001
Stimulus ID	49 908 (17 928)	–	–	87 596 (31 394)	–	–
Residual	33 116 (1751·4)	–	–	49 442 (2613·1)	–	–
Repeatability	R (SE)	$\chi^2_{0/1}$	P	R (SE)	$\chi^2_{0/1}$	P
Individual ID	0·067 (0·028)	72·14	<0·001	0·047 (0·021)	51·52	<0·001

(b) V_i/R_i

	Informed/noise	Informed/no-noise	Uninformed/noise	Uninformed/no-noise
Informed/noise	–	$\chi^2_{1d.f.} = 37{\cdot}71$, $P < 0{\cdot}001$	$\chi^2_{1d.f.} = 5{\cdot}14$, $P = 0{\cdot}023$	$\chi^2_{1d.f.} = 8{\cdot}37$, $P = 0{\cdot}004$
Informed/no-noise	$\chi^2_{1d.f.} = 40{\cdot}02$, $P < 0{\cdot}001$	–	$\chi^2_{1d.f.} = 12{\cdot}59$, $P < 0{\cdot}001$	$\chi^2_{1d.f.} = 8{\cdot}25$, $P = 0{\cdot}004$
Uninformed/noise	$\chi^2_{1d.f.} = 10{\cdot}68$, $P = 0{\cdot}001$	$\chi^2_{1d.f.} = 7{\cdot}92$, $P = 0{\cdot}005$	–	$\chi^2_{1d.f.} = 0{\cdot}42$, $P = 0{\cdot}517$
Uninformed/no-noise	$\chi^2_{1d.f.} = 8{\cdot}96$, $P = 0{\cdot}003$	$\chi^2_{1d.f.} = 8{\cdot}68$, $P = 0{\cdot}003$	$\chi^2_{1d.f.} = 0{\cdot}06$, $P = 0{\cdot}806$	–

*In trait signal type, reference group is natural song; in Stimulus order, reference group is A; covariates are mean centred.

estimates are collected by a single observer, average estimates would potentially be more biased in treatments that express more individual variation in minimum frequencies. Minimum frequencies measured in informed/noise, informed/no-noise, uninformed/noise and uninformed/no-noise treatment were fitted as the first, second, third and fourth response variables. We used the same fixed and random effects structure as detailed for the univariate models above. Note, however, that we only show whether the individual-level variance components and their repeatabilities differ between the treatments instead of the full model outputs (see 'Results' and Table 3b). Following

Dingemanse & Dochtermann (2013), covariances at the residual level were constrained to zero because they were non-estimable. Moreover, covariances at the individual level were only estimated between the groups informed/noise and informed/no-noise as well as uninformed/noise and uninformed/no-noise because individuals were crossed across noise groups. Moreover, at the element identity level, all covariances were estimated because the same elements were crossed across all four groups (i.e. informed/noise, informed/no-noise, uninformed/noise and uninformed/no-noise). We took a two-step approach: First, we tested whether treatments generally differed from each other in individual variance by comparing the unconstrained model (which estimated a separate variance for individual identity for each treatment) with a model where individual variances were constrained to be the same across all treatments (Dingemanse & Dochtermann 2013). After finding a significant general difference, we tested in more detail whether the individual variance components differed across each of the two focal treatment group combinations. This was done by comparing the unconstrained model with a model that was constrained to be identical for the two focal treatment groups (Dingemanse & Dochtermann 2013), while the rest of the treatment groups were still free to vary. We further applied the same approach to variance-standardized data (Dingemanse & Dochtermann 2013) to test whether the repeatability of a focal variance component differed across our treatments. The significance of treatment specificity of a variance component (and R) was determined by applying a LRT (see above), with which we compared the fit of the unconstrained model (see above) with one where the focal variance component was constrained to be identical among the treatments. The χ^2-distributed test statistic was calculated as twice the difference in Log Likelihood between the two models over three (first step) and one (second step) degrees of freedom. All models were fitted with Gaussian error distributions; visual inspection confirmed that residuals did not deviate from the normal error distribution. All statistical models were fitted in ASReml 3.0.5 (Gilmour, Gogel & Cullis 2009).

Results

SOURCES OF VARIATION IN MEASURED MINIMUM FREQUENCIES

Informed and uninformed individuals did not differ from each other in their measures of average minimum frequencies in no-noise recordings (Table 2 and Fig. 2). However, in recordings with noise, minimum frequencies were scored on average 267 Hz higher compared to noise-free conditions. Moreover, there was a significant interaction between the level of information and noise level: informed individuals in noisy conditions measured minimum frequencies on average 137 Hz higher compared to uninformed individuals in noisy conditions (Table 2 and Fig. 2). It is worth noting that in the no-noise condition both informed and uninformed observers measured minimum frequencies of synthetic signals on average relatively close to the true value (informed: 1516 Hz, uninformed: 1522 Hz, true value: 1556 Hz). In noise, however, informed observers measured synthetic signals on average to be 304 Hz higher than they actually were and uninformed observers measured them on average to be 161 Hz higher.

The overall arrangement of the song elements in the test file (i.e. 'stimulus order' in Table 2) did not affect average minimum frequency measures. The sequence in which the randomly assigned song files were presented to the observers had an effect on measured minimum frequency (Table 2). Elements presented towards the end of the test files were measured to have higher minimum frequencies compared to elements in the beginning (linear term), although this relationship was slightly concave (quadratic term) (Table 2).

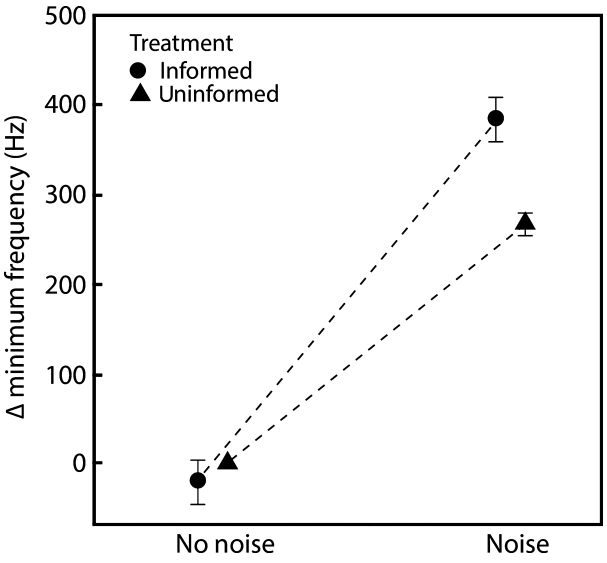

Fig. 2. Differences in estimated minimum frequencies (± SE) across all treatment groups extracted from the model presented in Table 2. The reference treatment is uninformed/no-noise.

VARIANCE COMPONENTS AND THEIR COMPARISON BETWEEN TREATMENTS

Individual identity and element identity explained significant amounts of variation in all treatment groups (informed/noise, informed/no-noise, uninformed/noise and uninformed/no-noise; Table 3a). In other words, individuals differed consistently from each other in their average minimum frequencies. Element identities also differed significantly from each other in their average measured minimum frequencies, which is to be expected by definition as elements were chosen (in case of natural elements) or built (in case of artificial elements) to differ from each other.

A multivariate model testing for general differences in individual-level variance components across all treatment groups revealed that individual-level variances differed across treatment groups ($\chi^2_3 = 46{\cdot}64$, $P < 0{\cdot}001$). More detailed *post hoc* comparisons showed that all treatment groups except uninformed/no-noise and uninformed/noise differed from each other in V_I and R estimates (Table 3b) and that individual differences in scored average minimum frequencies were highest in the informed treatment with noise and lowest in the informed treatment without noise, compared to the other treatment groups (Table 3).

Discussion

Our study reveals that the inappropriate use of acoustic analysis programmes can yield false positives in the study of birdsong in noise. When extracting minimum song frequencies visually from spectrograms, at least two causes contributed to the deviation from true values: a systematic measurement artefact in masked signals and a psychological observer bias. Although each observer measured identical recordings, the signals that were mixed with synthesized low-frequency traffic noise were erroneously assessed as having higher minimum frequencies. This artefact can be accounted for by the fact that the lower frequency end of the measured signals was difficult to see in the spectrograms because of the masking traffic noise (Zollinger *et al.* 2012; Grace & Anderson 2015).

Our findings demonstrate that observers measured higher minimum frequencies in noise compared to no-noise conditions and that this effect was significantly stronger among informed observers who expected an increase in minimum frequency. In no-noise conditions, informed and uninformed observers did not differ in their minimum frequency measures. Thus, noisy recordings combined with *a priori* expectations of the data at hand (which is the norm in the field) can cause an upwards bias in measured average frequencies, leading to the false impression of raised minimum frequencies in noise. Our results are a vivid example of observer-expectancy bias (Martin & Bateson 2007), a type of cognitive bias where individuals who have certain expectations about the outcome of the experiment seem to find those expectations even though they are not necessarily real. Moreover, the observers in our study differed consistently from each other in how high or low they measured minimum frequencies irrespective of treatment and noise group combination. Individual differences were strongest among informed individuals scoring noisy samples, where individuals explained as much as ~20% of the variation in measured minimum frequencies. This among-individual variation introduces a potentially severe bias: different observers would measure different mean minimum frequencies, especially in situations with high background noise levels and an observer with previous knowledge about the study question. If the frequency measures are taken only by a single observer – which is usually done in studies extracting frequency measures by hand – the data are potentially biased due to individual variation, as the average minimum frequencies depend on who is extracting them from the recordings. However, average estimates that are pooled from the estimates made by multiple individuals, whether collected by experts or not, are generally thought to be unbiased (Conradt & Roper 2005; Dyer *et al.* 2008; Sumpter & Pratt 2009). This so-called 'wisdom of crowds' effect is based on the phenomenon that groups of individuals make collective decisions that are less prone to error compared to those taken by a single individual, even if the single individual is an expert (Conradt & Roper 2005; Dyer *et al.* 2008; Sumpter and Pratt 2009). Indeed, a data collection protocol using average minimum frequencies that are pooled from the data extracted by many individuals would reduce the bias caused by individuals consistently differing from each other in their frequency estimates. However, as the among-individual variation in noisy conditions in our study is focused around the biased averages, the simple inclusion of multiple observers in the data collection would not significantly increase the quality of the estimates from noisy recordings compared to the estimates from non-noisy recordings. Instead of using the wisdom of a biased crowd to better faulty measurements, the way forward is to make objective measurements in the first place. In the next section, we will advise how this can be done.

Both measurement artefacts, the individual error introduced by visual scoring and the observer-expectancy bias, can be avoided by using power or amplitude spectra (or zero-crossing counts from waveforms for sounds with constant frequencies) rather than visually extracting minimum frequencies from spectrograms. From power spectra, minimum and maximum frequency can be measured reliably at a set amplitude threshold below the peak amplitude (Zollinger *et al.* 2012). This method is well-established and is often used among bioacousticians studying animal sounds (e.g. Podos 1997; Fischer, Hammerschmidt & Todt 1998; Templeton, Greene & Davis 2005; Siemers & Kerth 2006; DuBois, Nowicki & Searcy 2009; Hanna *et al.* 2011). In recordings with high levels of low-frequency noise, such as those from heavily noise-polluted areas, the measuring threshold needs to be set at a value at which the signal-to-noise ratio is positive; otherwise the measurement of the minimum frequency would be biased by the noise. This means that signal components that are lower in amplitude than the noise at the same frequency cannot be included in the measurement. This is certainly a limitation, but the only way to remedy this drawback is to make better quality recordings in the first place. The skilful use of acoustic recording equipment can reduce the amount of noise in the recordings and yield high signal-to-noise ratios that allow inspecting a wide range of signal frequencies in power spectra even in the presence of high levels of low-frequency noise. Particularly high signal-to-noise ratios can be achieved with radio microphones placed near the song posts of birds (Nemeth *et al.* 2012) or acoustic recording devices fixed on the animals themselves (Zollinger, Goller & Brumm 2011; Anisimov *et al.* 2014; Gill *et al.* 2015).

Some researchers may be reluctant to use threshold measurements because they can see signal frequencies on the spectrogram that cannot be captured by the threshold. However, it is important to bear in mind what these low-frequency components actually represent. A threshold of only 10 dB below the peak already comprises about 90% of the signal energy and a 20 dB threshold, which is often used for high-quality recordings, captures 99% of the signal energy. Frequency components visible at lower frequencies might look persuasive on the spectrogram but are negligible in terms of signal transmission.

Our findings do not necessarily suggest that the results of studies using the eye-balling method are completely invalid. To assess the magnitude of the measurement error in these studies, however, the audio recordings need to be re-analysed with appropriate methods. As mentioned above, measuring minimum frequencies at a set threshold below the peak amplitude

is one way forward. There are a number of textbooks and manuals available that will be helpful to those wishing to make valid measurements (e.g. Beecher 1988; Rossing 1989; Bradbury & Vehrencamp 1998; Hopp, Owren & Evans 1998; Tohyama & Koike 1998).

On a broader note, we are concerned by the continuing publication of studies that rely on eye-balling frequencies from spectrograms by informed observers. The problem of observer biases in behavioural studies is long known (Kazdin 1977; Caro *et al*. 1979; Balph & Balph 1983) and the need to avoid such bias has been repeatedly emphasized (reviewed in Traniello & Bakker 2015; Forstmeier, Wagenmakers & Parker 2016). In bioaocoustic research, observer biases can be an issue too, for example, when spectrograms are visually assessed (Jones, ten Cate & Bijleveld 2001). Therefore, it is relatively customary to use several uninformed observers when applying visual scoring (e.g. Houx & ten Cate 1999; Janik 2000; Beecher *et al*. 2007; Geberzahn & Gahr 2013). However, avoiding observer-expectancy biases cannot entirely solve the problems with the visual extraction of measures from spectrograms because the practice is inherently flawed, as shown by this and other studies (Zollinger *et al*. 2012; Grace & Anderson 2015; Rios-Chelén, Lee & Patricelli 2016; Rios-Chelén *et al*. 2017). Although this pitfall has been recognized since the early days of spectrographic analyses of birdsong (Greenewalt 1968), the need for appropriate methods is not always heeded.

To date, acoustic measurements are usually done with the help of analysis software, a practice that has considerably advanced the field of bioacoustics. On the other hand, acoustic analysis programmes can easily be misused, especially as measurements appear to be only one click away. Without an understanding of acoustic principles, the use of analysis software can lead to unfavourable results, and in the worst case to the publication of spurious findings. We therefore encourage researchers to familiarize themselves with the physics of sound and the methodological principles of bioacoustics before using acoustic analysis programmes to make measurements.

Authors' contributions

H.B. conceived the study. All authors contributed to the design of study. P.S. synthesized the noise masker and S.Z. the test signals. H.B. and P.S. produced the test files and collected the data. P.T.N. and P.S. designed and performed the data analysis. H.B. led the writing of the manuscript, with critical input from all authors.

Acknowledgements

We thank all test persons for taking part in this study. Mathias Ritschard provided the great tit songs and Adriana Dorado-Correa provided the traffic noise recordings. We also thank Holger Schielzeth and three anonymous reviewers for very constructive comments. Financial support was provided by the German Research Foundation (awards BR2309/7-1, BR2309/8-1, BR2309/8-2 and SP1450/3-1). The authors declare they have no conflict of interest.

Data accessibility

Data available from https://doi.org/10.5061/dryad.590jh (Brumm *et al*. 2017).

References

Anisimov, V.N., Herbst, J.A., Abramchuk, A.N., Latanov, A.V., Hahnloser, R.H.R. & Vyssotski, A.L. (2014) Reconstruction of vocal interactions in a group of small songbirds. *Nature Methods*, **11**, 1135–1137.

Balph, D.F. & Balph, M.H. (1983) On the psychology of watching birds: the problem of observer-expectancy bias. *Auk*, **100**, 755–757.

Beecher, M.D. (1988) Spectrographic analysis of animal vocalisations: implications of the uncertainty principle. *Bioacoustics*, **1**, 187–208.

Beecher, M.D., Burt, J.M., O'Loghlen, A.L., Templeton, C.N. & Campbell, S.E. (2007) Bird song learning in an eavesdropping context. *Animal Behaviour*, **73**, 929–935.

Blumenrath, S.H. & Dabelsteen, T. (2004) Degradation of great tit (*Parus major*) song before and after foliation: implications for vocal communication in a deciduous forest. *Behaviour*, **141**, 935–958.

Bradbury, J.W. & Vehrencamp, S.L. (1998) *Principles of Animal Communication*. Sinauer, Sunderland, MA, USA.

Brumm, H. (2004) The impact of environmental noise on song amplitude in a territorial bird. *Journal of Animal Ecology*, **73**, 434–440.

Brumm, H. (2013) *Animal Communication and Noise*. Springer-Verlag, Heidelberg, Berlin, Germany.

Brumm, H. & Bee, M. (2016) A meta-analytic castle built on sand? A comment on Roca *et al*. *Behavioral Ecology*, **27**, 1277–1278.

Brumm, H. & Zollinger, S.A. (2013) Avian vocal production in noise. *Animal Communication and Noise* (ed H. Brumm), pp. 187–227. Springer-Verlag, Heidelberg, Berlin, Germany.

Brumm, H., Zollinger, S.A., Niemelä, P.T. & Sprau, P. (2017) Data from: Measurement artefacts lead to false positives in the study of bird song in noise. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.590jh

Cardoso, G.C. & Atwell, J.W. (2012) On amplitude and frequency in birdsong: a reply to Zollinger et al.. *Animal Behaviour*, **84**, e10–e15.

Caro, T.M., Roper, R., Young, M. & Dank, G.R. (1979) Inter-observer reliability. *Behaviour*, **69**, 303–315.

Conradt, L. & Roper, T.J. (2005) Consensus decision making in animals. *Trends in Ecology and Evolution*, **20**, 449–456.

Dingemanse, N.J. & Dochtermann, N.A. (2013) Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology*, **82**, 39–54.

Dorado-Correa, A.M., Rodríguez-Rocha, M. & Brumm, H. (2016) Anthropogenic noise, but not artificial light levels predicts song behaviour in an equatorial bird. *Royal Society Open Science*, **3**, 160231.

DuBois, A.L., Nowicki, S. & Searcy, W.A. (2009) Swamp sparrows modulate vocal performance in an aggressive context. *Biology Letters*, **5**, 163–165.

Dyer, J.R.G., Ioannou, C.C., Morrell, L.J., Croft, D.P., Couzin, I.D., Waters, D.A. & Krause, J. (2008) Consensus decision making in human crowds. *Animal Behaviour*, **75**, 461–470.

Endler, J.A. (1992) Signals, signal conditions, and the direction of evolution. *American Naturalist*, **139**, S125–S153.

Fischer, J., Hammerschmidt, K. & Todt, D. (1998) Local variation in Barbary macaque shrill barks. *Animal Behaviour*, **56**, 623–629.

Forstmeier, W., Wagenmakers, E.-J. & Parker, T.H. (2016) Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, published online ahead of print.

Geberzahn, N. & Gahr, M. (2013) Song learning in male and female *Uraeginthus cyanocephalus*, a tropical songbird species. *Journal of Comparative Psychology*, **127**, 352–364.

Gil, D. & Brumm, H. (2014) Acoustic communication in the urban environment: patterns, mechanisms, and potential consequences of avian song adjustments. *Avian Urban Ecology* (eds D. Gil & H. Brumm), pp. 69–83. Oxford University Press, Oxford, UK.

Gill, L.F., Goymann, W., Ter Maat, A. & Gahr, M. (2015) Patterns of call communication between group-housed zebra finches change during the breeding cycle. *eLife*, **4**, e07770.

Gilmour, A.R., Gogel, B.J. & Cullis, B.R. (2009) *ASReml User Guide Release 3.0*. VSN Int Ltd, Hemel Hempstead, UK.

Grace, M.K. & Anderson, R.C. (2015) No frequency shift in the 'D' notes of Carolina chickadee calls in response to traffic noise. *Behavioral Ecology and Sociobiology*, **59**, 253–263.

Greenewalt, C.H. (1968) *Bird Song: Acoustics and Physiology*. Smithsonian Institute Press, Washington, DC, USA.

Hanna, D., Blouin-Demers, G., Wilson, D. & Mennill, D. (2011) Anthropogenic noise affects song structure in red-winged blackbirds (*Agelaius phoeniceus*). *Journal of Experimental Biology*, **214**, 3549–3556.

Hopp, S., Owren, M. & Evans, C. (1998) *Animal Acoustic Communication. Sound Analysis and Research Methods*. Springer, Heidelberg, Germany.

Houx, B.B. & ten Cate, C. (1999) Song learning from playback in zebra finches: is there an effect of operant contingency? *Animal Behavior*, **57**, 837–845.
Hu, Y. & Cardoso, G. (2010) Which birds adjust the frequency of vocalizations in urban noise? *Animal Behaviour*, **79**, 863–867.
Huffeldt, N. & Dabelsteen, T. (2013) Impact of a noise-polluted urban environment on the song frequencies of a cosmopolitan songbird, the Great Tit (*Parus major*). *Ornis Fennica*, **90**, 94–102.
Janik, V.M. (2000) Whistle-matching in wild bottlenose dolphins (*Tursiops truncatus*). *Science*, **289**, 1355–1357.
Job, J.R., Kohler, S.L. & Gill, S.A. (2016) Song adjustments by an open habitat bird to anthropogenic noise, urban structure, and vegetation. *Behavioral Ecology*, **27**, 1734–1744.
Jones, A.E., ten Cate, C. & Bijleveld, C.T.J.H. (2001) The interobserver reliability of scoring sonagrams by eye: a study on methods, illustrated on zebra finch songs. *Animal Behavior*, **62**, 791–801.
Kazdin, A.E. (1977) Artifact, bias, and complexity of assessment: the ABCs of reliability. *Journal of Applied Behavior Analysis*, **10**, 141–150.
Martin, P. & Bateson, P. (2007) *Measuring Behaviour. An Introductory Guide*. Cambridge University Press, Cambridge, UK.
Meyer, K. (1992) Variance-components due to direct and maternal effects for growth traits of Australian beef-cattle. *Livestock Production Science*, **31**, 179–204.
Mockford, E.J. & Marshall, R.C. (2009) Effects of urban noise on song and response behaviour in great tits. *Proceedings of the Royal Society B-Biological Sciences*, **276**, 2979–2985.
Montague, M., Danek-Gontard, M. & Kunc, H. (2013) Phenotypic plasticity affects the response of a sexually selected trait to anthropogenic noise. *Behavioral Ecology*, **24**, 342–348.
Nakagawa, S. & Schielzeth, H. (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, **85**, 935–956.
Narango, D.L. & Rodewald, A.D. (2016) Urban-associated drivers of song variation along a rural–urban gradient. *Behavioral Ecology*, **27**, 608–616.
Nemeth, E., Kempenaers, B., Matessi, G. & Brumm, H. (2012) Rock sparrow song reflects male age and reproductive success. *PLoS ONE*, **7**, e43259.
Podos, J. (1997) A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution*, **5**, 537–551.
Rios-Chelén, A.A., Lee, G.C. & Patricelli, G.L. (2016) A comparison between two ways to measure minimum frequency and an experimental test of vocal plasticity in red-winged blackbirds in response to noise. *Behaviour*, **153**, 1445–1472.
Rios-Chelén, A.A., Quirós-Guerrero, E., Gil, D. & Macías Garcia, C. (2013) Dealing with urban noise: vermilion flycatchers sing longer songs in noisier territories. *Behavioral Ecology and Sociobiology*, **67**, 145–152.
Rios-Chelén, A.A., McDonald, A.N., Berger, A., Perry, A.C., Krakauer, A.H. & Patricelli, G.L. (2017) Do birds vocalize at higher pitch in noise, or is it a matter of measurement? *Behavioral Ecology and Sociobiology*, **71**, 29.
Ritschard, M., van Oers, K., Naguib, M. & Brumm, H. (2012) Song amplitude of rival males modulates the territorial behaviour of great tits during the fertile period of their mates. *Ethology*, **118**, 197–202.
Roca, I.T., Desrochers, L., Giacomazzo, M. *et al*. (2016) Shifting song frequencies in response to anthropogenic noise: a meta-analysis on birds and anurans. *Behavioral Ecology*, **27**, 1269–1274.
Rossing, T.D. (1989) *The Science of Sound*. Addison-Wesley Publishing Company, Reading, MA, USA.
Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, **20**, 416–420.
Self, S.G. & Liang, K.Y. (1987) Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.
Siemers, B.M. & Kerth, G. (2006) Do echolocation calls of wild colony-living Bechstein's bats (*Myotis bechsteinii*) provide individual-specific signatures? *Behavioral Ecology and Sociobiology*, **59**, 443–454.
Slabbekoorn, H. & Peet, M. (2003) Birds sing at a higher pitch in urban noise. *Nature*, **424**, 267.
Sumpter, D.J.T. & Pratt, S.C. (2009) Quorum responses and consensus decision making. *Philosophical Transactions of the Royal Society of London B-Biological Sciences*, **364**, 743–753.
Templeton, C.N., Greene, E. & Davis, K. (2005) Allometry of alarm calls: black-capped chickadees encode information about predator size. *Science*, **308**, 1934–1937.
Tohyama, M. & Koike, T. (1998) *Fundamentals of Acoustic Signal Processing*. Academic Press, San Diego, CA, USA.
Traniello, J.F.A. & Bakker, T.C.M. (2015) Minimizing observer bias in behavioral research: blinded methods reporting requirements for Behavioral Ecology and Sociobiology. *Behavioral Ecology and Sociobiology*, **69**, 1573–1574.
Visscher, P.M. (2006) A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Research and Human Genetics*, **9**, 490–495.
Wiley, H. (2015) *Noise Matters – The Evolution of Communication*. Harvard University Press, Cambridge, UK.
Wilson, A.J., Reale, D., Clements, M.N., Morrissey, M.M., Postma, E., Walling, C.A., Kruuk, L.E.B. & Nussey, D.H. (2010) An ecologist's guide to the animal model. *Journal of Animal Ecology*, **79**, 13–26.
Zollinger, S.A., Goller, F. & Brumm, H. (2011) Metabolic and respiratory costs of increasing song amplitude in zebra finches. *PLoS ONE*, **6**, e23198.
Zollinger, S.A., Podos, J., Nemeth, E., Goller, F. & Brumm, H. (2012) On the relationship between, and measurement of, amplitude and frequency in birdsong. *Animal Behaviour*, **84**, E1–E9.

Received 9 December 2016; accepted 22 February 2017
Handling Editor: Holger Schielzeth