Journal of Animal Ecology

Hutfluss Alexander (Orcid ID: 0000-0001-8514-6452)
Briffa Mark (Orcid ID: 0000-0003-2520-0538)
Dingemanse Niels Jeroen (Orcid ID: 0000-0003-3320-0861)

Male song stability shows cross-year repeatability but does not affect reproductive success in a wild passerine bird

Alexander Hutfluss[1], Eira Bermúdez-Cuamatzin[2], Alexia Mouchet[1,3], Mark Briffa[4], Hans Slabbekoorn[2], Niels J. Dingemanse[1]

[1] Behavioural Ecology, Department of Biology, Ludwig Maximilian University of Munich (LMU), 82152 Planegg-Martinsried, Germany

[2] Behavioural Biology, Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE Leiden, The Netherlands

[3] Laboratoire Evolution Génomes Comportement et Ecologie (EGCE), UMR Université Paris-Saclay-CNRS-IRD, 91190 Gif-sur-Yvette, France

[4] School of Biological and Marine Sciences, Animal Behaviour Research Group, University of Plymouth, Plymouth, Devon, PL4 8AA, UK

Correspondence

Alexander Hutfluss, Behavioural Ecology, Department of Biology, Ludwig Maximilian University of Munich, Großhaderner Straße 2, 82152 Planegg-Martinsried, Germany

Email: alexander.hutfluss@gmx.de

Tel: 49 89 2180 74 214

Fax: 49 (0) 89 2180 74 204

Short Title: Individual differences in bird song stability

Abstract

1. Predictable behaviour (or "behavioural stability") might be favoured in certain ecological contexts, e.g. when representing a quality signal. Costs associated with producing stable phenotypes imply selection should favour plasticity in stability when beneficial. Repeatable among-individual differences in degree of stability are simultaneously expected if individuals differ in ability to pay these costs, or in how they resolve cost-benefit trade-offs.

2. Bird song represents a prime example, where stability may be costly yet beneficial when stable singing is a quality signal favoured by sexual selection. Assuming energetic costs, ecological variation (e.g. in food availability) should result in both within- and among-individual variation in stability. If song stability represents a quality signal, we expect directional selection favouring stable singers.

3. For a three-year period, we monitored 12 nest box plots of great tits *Parus major* during breeding. We recorded male songs during simulated territory intrusions, twice during their mate's laying stage, and twice during incubation. Each preceding winter, we manipulated food availability. Assuming that stability is costly, we expected food-supplemented males to sing more stable songs. We also expected males to sing more stable songs early in the breeding season (when paternity is not decided), and stable singers to have increased reproductive success.

4. We found strong support for plasticity in stability for two key song characteristics: minimum frequency and phrase length. Males were plastic because they became more

stable over the season, contrary to expectations. Food-supplementation did not affect body condition but increased stability in minimum frequency. This treatment effect occurred only in one year, implying that food supplementation affected stability only in interaction with (unknown) year-specific ecological factors.

5. We found no support for directional, correlational, or fluctuating selection on the stability in minimum frequency (i.e., the song trait whose stability exhibited cross-year repeatability): stable singers did not have higher reproductive success. Our findings imply that stability in minimum frequency is not a fitness quality indicator unless males enjoy fitness benefits via pathways not studied here. Future studies should thus address the mechanisms shaping and maintaining individual repeatability of song stability in the wild.

Introduction

Within-individual variation in behaviour occurs because individuals vary their behaviour in response to environmental changes they experienced within their life time (Dingemanse et al., 2010; Nussey et al., 2007). Individuals can also vary their behaviour within a given (set of) environmental condition(s). The resulting "intra-individual variation" (Stamps et al., 2012; Westneat et al., 2015) affects how predictable individuals are to others. Behaving (un)predictably can be advantageous but costs and benefits are trait- and context-specific. Consequently, behavioural stability and the ability to flexibly adjust it ("plasticity in stability") is likely under selection. For example, behaving unpredictably can increase survival during predator attacks

(Brembs, 2011; Briffa, 2013; Domenici et al., 2008; Jones et al., 2011), while behaving predictably can also be favoured, e.g., when representing a quality indicator under sexual selection (Sakata et al., 2008; Schuett et al., 2010). Despite its importance, few studies have estimated the fitness consequences of behavioural stability (Byers, 2007).

Behavioural stability may evolve if repeatable and heritable (Martin et al., 2017; Prentice et al., 2020; Stamps et al., 2012; Westneat et al., 2013). Variation in stability might exist for various reasons (Hertel et al., 2021; Jolles et al., 2019; Mitchell et al., 2021). First, trade-offs between mean and variance (within the same or across traits) might be resolved differently among individuals. Those may occur between social/energetic costs vs. signal values in communication (Grava et al., 2013; Rivera-Gutierrez et al., 2010), or between strength vs. precision/skill in agonistic interactions (Lane & Briffa, 2020; Reichert & Gerhardt, 2012). Second, behavioural stability might be beneficial yet costly, and individuals might vary in ability to pay these costs. Finally, if stability is a fitness indicator trait, selection would act to erode variation, and variation would then persist because of mutation-selection balance (Auld et al., 2010; Sakata et al., 2008; Schuett et al., 2010).

Bird song is highly suitable for testing hypotheses on behavioural stability (Patricelli & Hebets, 2016; Sakata & Vehrencamp, 2012; Westneat et al., 2015). The stereotypic, ongoing production of acoustic signals requires repeated activity of syringeal and respiratory muscles and is energetically costly (Grava et al., 2012; Sakata & Vehrencamp, 2012; Suthers & Zollinger, 2004). Individuals should thus upregulate song stability of certain traits only when benefits outweigh costs. For example, if a male's song stability reduces how much its breeding partner engages in extra-pair matings, males should produce stable songs before (but not necessarily after) clutch completion. Similarly, if beneficial yet energetically costly, access to supplementary food

should increase stability. Ultimately, we expect performance limits to song stability if traded-off with other traits (Podos, 1997). We then expect repeatable variation when individual differences in quality result from variation in early-life conditions causing silver spoon effects ("developmental stress hypothesis"; Holveck et al., 2008; MacDougall-Shackleton & Spencer, 2012; Van De Pol et al., 2006) . Some studies reported that song stability affects predictors of fitness, such as dominance, or fitness components, such as extra-pair offspring (Botero et al., 2009; Byers, 2007; De Kort et al., 2009; Taff et al., 2012). Nevertheless, relatively few studies have directly tested for a pattern of positive selection which is expected when song stability is indeed a quality indicator. Unfortunately, past studies often relied on methodological and statistical practices that produce biased estimates (Cleasby et al., 2015), such on one-year studies or "statistics-on-statistics" (reviewed in Text S1 and Table S1). We thus require field studies specifically designed to acquire unbiased estimates of individual variation in stability, and selection acting on this variation. This represents the aim of the current study.

In our study species, the great tit (*Parus major*), song stability research has focused on minimum frequency and phrase length. Minimum frequency is repeatable and may be an important signal trait under sexual selection (Cardoso, 2012; Halfwerk et al., 2011; Patel et al., 2010). In some species, birds producing stable frequencies are larger (Bartsch et al., 2016) and produce more extra-pair offspring (Byers, 2007). Furthermore, minimum frequency plastically varies among songs (Cardoso & Atwell, 2011), phrases within songs (Cardoso & Atwell, 2011), song types (Logue et al., 2007; Slabbekoorn & den Boer-Visser, 2006), and elements within song types and phrases (Marler & Isaac, 1960). Great tits sing low songs mostly prior to egg fertilization (Halfwerk et al. 2011) and those with lower minimum frequencies produced more stable songs (Lambrechts, 1997). This suggests that stability of the minimum frequency of low-frequency songs

could either be a direct or indirect target of sexual selection by female preference (Cramer, 2013b). Along the same lines, selection is thought to favour stability in phrase length (Lambrechts & Dhondt, 1986; Weary, 1989). Great and blue tits (*Cyanistes caeruleus*) producing stable phrase lengths are found to be dominant and have higher reproductive success (Lambrechts & Dhondt, 1986; Poesel & Kempenaers, 2000). The stable production of phrase lengths may further be costly because fatigue negatively affects it (Lambrechts & Dhondt, 1986).

We estimated within- and among-year repeatability of stability in two song traits, and quantified environmental factors moderating their stability. We repeatedly recorded songs of the same birds, elicited during simulated territory intrusions. We expected males to up-regulate song stability during their mate's laying stage, thus for it to change seasonally within individuals. This assumes that singing stable songs is costly, which we tested using supplementary feeding experiments that we conducted outside the breeding season: winter food should improve body condition, in turn enabling males to produce more stable songs. Finally, we tested whether song stability affected key components of reproductive success.

Methods

Data were collected over three years (2017-2019) in 12 nest box plots between Starnberg and Herrsching am Ammersee (Bavaria, Germany) (Nicolaus et al., 2015). Each plot contained 50 boxes in a grid (50 m apart). Boxes were monitored April-July to quantify lay date, clutch size, and fledgling number (Nicolaus et al. 2015). We caught breeders in their box when offspring were 10-12 days old, measured morphology and behaviour, and ringed those not previously ringed. Fourteen days after hatching, we measured the nestlings. The number of nestlings on that day was

then compared to the deceased nestlings found 19-21 days after hatching to quantify fledgling number.

We manipulated food availability from July through March. We placed feeders baited with sunflower seeds in half of the plots ("treated"); supplementary food was not provided in the remaining plots ("control"). We randomly assigned plot treatment; half of the plots switched treatment across years in a stratified random design to ensure each plot received each treatment at least once over the three years. In treated plots, we placed a feeder at four locations to avoid resource monopolisation. We estimated natural winter food availability by counting beech seeds (the species' main winter food source; van Balen, 1980), in four 30×30-cm sections arranged in a transect under ten European beeches (*Fagus sylvatica*) in the study area. Counts showed a distinct binary pattern, and natural winter food availability was thus categorized as "no seeds" (2018) vs. "seeds" ("mast years"; 2017 and 2019). This pattern was expected because mast years are normally followed by years without seeds (Van Balen, 1980).

Simulated territory intrusions

We acquired repeated song recordings by subjecting each male in each year to four simulated territorial intrusions during its first breeding attempt: two times during egg-laying (1 and 3 days after the first egg was observed) and two times during incubation (1 and 3 days after clutch incubation was confirmed). We placed a taxidermic model of a conspecific male as visual stimulus (on a 1.2 m pole one meter in front of the box), while playing back a conspecific song as acoustic stimulus (using a Foxpro Shockwave speaker placed under the model on the ground) (Strauß et al. 2020).

We randomly assigned models (n=23) and playbacks (n=174). In 2017, we used 13 songs as playback stimuli from 13 different locally recorded birds. In 2018 and 2019, we added 161

playback songs from recordings made during the 2017-experiments. To create song variation that mimicked natural variation, we used—from the same recording—complete strophes followed by their respective silent intervals. Playbacks had an average strophe length of 1.85 s (min-max: 0.84 - 3.58 s) and an average minimum frequency of 3335 Hz (2528 - 4153 Hz). During playback creation and broadcasting, amplitudes were adjusted via normalization and speaker settings to similar levels. Playbacks were repeated until the test ended. After the focal male entered a 15-m radius around the box, we recorded its behaviour and song for three minutes. The observer (15 meters from the box) took recordings with a windscreen-covered directional microphone (Sennheiser ME66/K6) attached to a recorder (TASCAM DR-05, recording 44.1 kHz, 16-bit WAV files). Subjects not arriving within 15 min were scored "non-responsive" (Araya-Ajoy and Dingemanse 2014; 2017).

Song analysis

Acoustic parameters were extracted using Avisoft SAS Lab Pro (Avisoft Bioacoustics, Specht 2002). We applied band-pass filters to remove background noise outside the species-specific frequency song range (2.0-8.0 kHz). We subsequently increased amplitudes using the normalization function. Concurrently, we produced a spectrogram per strophe (FFT length: 1024, Overlap: 87.5%, Window: Hamming) to detect and exclude low quality (e.g. noisy) recordings. We extracted spectral and temporal features per element using the automatic measurements function (three thresholds: -24 dB, start: -16 dB, end: -12 dB, hold time: 10 ms, minimum length: 30 ms). This function recorded, for each element, start time, end time, and peak frequency for both time points; we used the lower point to measure minimum frequency (Fig. 1). We manually excluded elements overlapping spectrally and temporally with the (readily identifiable) playback stimulus song. We manually deleted noises and playback songs without temporal overlap when

affecting automated element detection. Observers performing checks were trained by scoring >2800 measurements of sixteen recordings; unknown to the observer, those consisted of two duplicates of eight files in random order. Observers started scoring data when their between- and within- observer repeatability was ≥0.9 (Table S2). Each phrase was assigned to one of 53 song types defined for this population by one observer (A.H.). Song types were classified beforehand using structural measurements, i.e. the number of elements per repeated phrase (one, two, three, or more), element structure within phrase (tonal, frequency modulated or resembling harmonics), and the combination of elements within a phrase (e.g. tonal plus "harmonics", or two frequency modulated plus a tonal). Differences in frequency or length (of single elements or the complete phrase) were not used to differentiate song types.
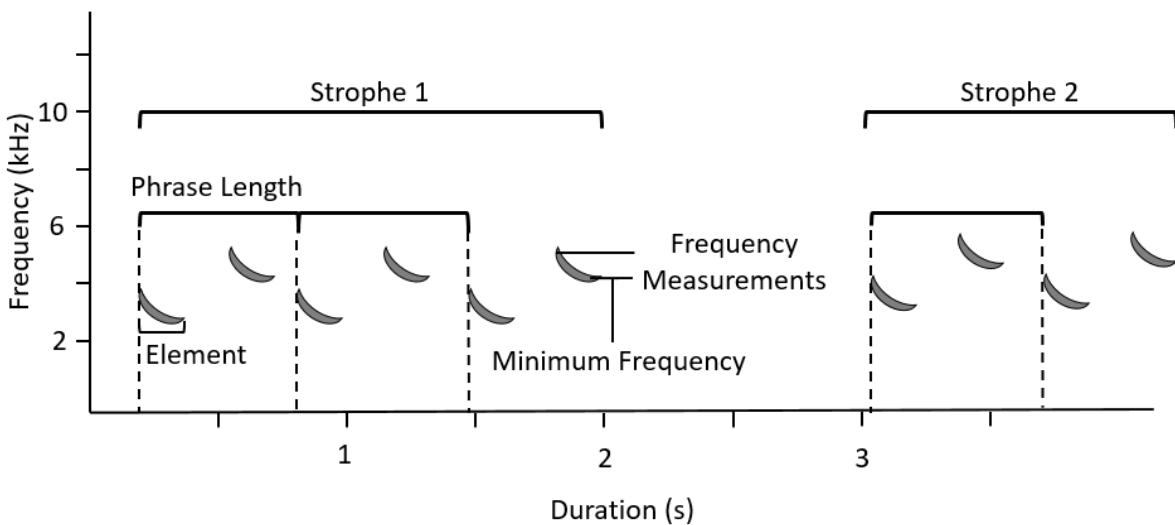

Figure 1: *Illustration of song components. We show two schematic songs (strophes) of one song type within one recording. The first strophe consists of three, the second of two, phrases; each phrase contains two elements (grey). Phrase length was defined as time elapsed between the start of a focal and the subsequent phrase within a strophe (dotted lines).*

Statistical analysis

We quantified, first, patterns of within- and among-individual variation in song stability, and second, links between individual stability and reproductive parameters among-individuals. To address the first question, we used univariate double hierarchical generalized linear mixed models (DHGLMs) to simultaneously estimate random and fixed effects associated with the "mean" and "variance" parts of the model (Lee & Nelder, 2006; Rönnegård et al., 2010). We analysed variation in (1) minimum frequency (the lower of the two peak frequencies per element) and (2) phrase length (time elapsed between the onsets of subsequent phrases; by nature missing for the strophe's last phrase). Previous studies analysed lowest elements per strophe/phrase (e.g. Bermúdez-Cuamatzin et al., 2011). We did not have specific hypotheses/predictions about variance within or among strophes/phrases; we thus analysed variation among all elements (within and among phrases) as an integrative measure of spectral variability.

We ran three models per response ("base", "environmental", and "song type"). All fitted random intercepts for individual (n=273 levels) and recording identity (n=585 levels) for both parts. Following Araya-Ajoy et al. (2015), we included random intercepts for each combination of individual and year ("individual-year"; n=334 levels). This separated transitory (repeatable within-year) from long-term (repeatable among-year) individual effects. Variance among individual-years is attributable to environmental factors that are individual-year specific; this causes repeatability due to environmental confounds in single-year repeated measures datasets (i.e., pseudo-repeatability). Variance among individuals instead reflects whether an individual's year-specific mean is repeatable across years (i.e., long-term repeatability). Finally, by adding recording identity, we used the repeated measures data (for elements, or phrases) within recordings to estimate whether average levels of behaviour differed between recordings within individuals

within years. All models included the unique combination of plot and year (plot-year; n=34 levels) as a random effect for both parts; fitting this dummy variable avoided pseudo-replicated estimates for food-supplementation treatment, which varied among plot-years. We performed explorative analyses to investigate observer (n=10 levels) and playback stimulus effects (n=158 levels, 16 playbacks were not used in recordings selected for the analyses), fitted as additional random effects for both parts. Neither observer nor playback identity were important (Table S3) and therefore excluded from further analyses. All models included year, fitted as a fixed factor because it contained few levels (n=3) (Zuur et al., 2009). Models estimated the covariance between "mean" and "variance" among individuals, individual-years, and recordings. Among individuals, this covariance estimates whether an individual's average behaviour correlates with its stability.

We expanded this **Base model** in two steps. First, we included nest stage (egg-laying vs. incubation), test number (first vs. second test within nest stage), and supplementary feeding treatment (vs. control) as fixed effects. Supplementary feeding is often not effective in mast-years (Perdeck et al., 2000); we considered this by fitting a year-treatment interaction. All fixed effects were fitted to both parts. We expanded this **Environmental model** by including random intercepts for song type (varying within and among individuals; n=53 levels) for both parts, while additionally fitting the covariance between "mean" and "variance" parts at this level (**Song type model**). This tested whether variation in mean and/or residual variance was attributable to song type. Do individuals differ in stability because they use different song types, and do they change stability by switching song types? If so, point estimates of other variance estimates should shrink when including song type in the model.

We collected 124,921 data points (number of elements). To manage computational time, we analysed only the first two phrases per strophe. This resulted in 21,578 and 20,618 data points

for the minimum frequency and phrase length, respectively. For minimum frequency, we ran the base model also on the full database, but this did not yield different results compared to the reduced dataset (Table S4). We thus used the reduced dataset throughout.

To address links between fitness and song stability, we used multivariate DHGLMs; those combined our base models with a linear-mixed effect model to avoid "statistics-on-statistics". However, these models did not converge. As an alternative, we extracted the Best Linear Unbiased Predictors (BLUPs) for each individual's mean and stability from the base model and used those as covariates in a set of linear mixed-effect models. We focused exclusively on the trait exhibiting among-individual variance in stability, (minimum frequency); analyses were conducted for four breeding parameters: 1) lay date 2) clutch size, 3) probability to produce at least one fledgling ("nest success"), and 4) number of fledglings for successful nests. In our population complete nest failures are caused by extreme events such as cold spells or predation (Marques-Santos & Dingemanse, 2020); we thus limited model 4 to nests with at least one fledgling. For each response, we fitted random intercepts for individual, plot, and plot-year identity, and included year as fixed factor.

Model set-up and interpretation of estimates

Analyses were performed in R v.3.6.1 (R Core Team 2019). DHGLMs were implemented using the "brms" package (Bürkner, 2017). Each DHGLM fitted default priors for all parameters. See Text S2 and Table S5 for model fitting procedures, performance, and diagnostics statistics. Linear mixed-effect models for our reproductive success analyses were implemented using the "lme4" package. Posterior distributions of model estimates were derived from 2000 simulations conducted using the "sim" function of the package "arm". We present mean (β for fixed effects in mean part, γ for fixed effects in variance part, and δ for random effects) and 95% credible intervals

(CIs) for all parameters. All estimates are based on the posterior distributions of the models. For residual variance, estimates are presented on the log-scale. Fixed effects are regarded "strongly supported" when their 95% CIs did not overlap zero; the probability of the effect size being unequal zero is then <0.05 (Hadfield 2010). We regarded 95% CIs overlapping zero but showing substantial skew in one direction as "moderate support" for an effect. We viewed variation attributable to random effects supported when their lower 95% CI did not include zero.

Animal Ethics

The study was approved by the Regierung Oberbayern (ROB-55.2-2532.Vet_02-17-215) and followed ethical guidelines for the treatment of animals in research ("Guidelines for the Treatment of Animals in Behavioural Research and Teaching," 2016).

Results

Descriptive statistics

We conducted 2,740 simulated territory intrusions; 1,582 resulted in a recording, 586 (of 273 males) were suitable for analysis. These included 6,544 songs, and 21,578 analysable elements (106,663 for the full database, Table S4). Birds sang (mean±SE) 23.9±0.6 songs, with 8.1±0.2 phrases. Males contributed on average, 2.45 recordings (range: 1-9) to the analyses. Fifty-two individuals were recorded in two years, five in all three years.

Great tits sang with an average minimum frequency of 4090 Hz (variance 700 Hz) and an average phrase length of 0.46 s (variance 0.09 s). During egg laying, minimum frequency was 4015 Hz (variance 658 Hz) compared to 4124 Hz (variance 711 Hz) during incubation. Average phrase lengths were similar for the two contexts (0.46 s), but variance differed (egg laying: 0.10 s; incubation: 0.08 s).

Sources of variation in behavioural mean

For the mean part of our models (Table 1a), patterns of variation differed between the two song traits. For minimum frequency, all random effects (except for plot-year) were supported (with CIs not including zero; Table 1a). Frequencies were more similar within vs. among recordings of the same individual, implying day-specific environmental effects within individuals (recording identity effects). Average values over all elements within recordings were also more similar for recordings taken in the same year (individual-year identity effects), implying year-specific environmental effects within individuals (i.e., "pseudo-repeatability"). Finally, an individual's grand mean annual value was repeatable among years (individual identity effects). This implied genuine (long-term) repeatable differences. For phrase length, by contrast, effects of neither individual-year nor individual were supported (owing to CIs including zero) (Table1a). Thus, phrase length showed neither pseudo- nor genuine repeatability. As above, recording identity effects were supported, implying day-specific environmental effects within individuals. Neither trait showed spatiotemporal variation (plot-year identity).

Environmental models (Table S6a) moderately supported seasonal changes in minimum frequency: for nest stage, CIs were skewed away from, but still slightly overlapped, zero (β = -0.08, 95% CI: -0.17, 0.01). Birds produced lower frequencies when their mate was incubating. For phrase length, by contrast, shortening with repeated exposure was strongly supported (effect of test number within nest stage: β = -0.01, 95% CI: -0.03, -0.00). Minimum frequency did not vary with test number, and phrase length did not vary with nest stage. Food supplementation affected neither trait nor were treatment effects year-specific (Table S6a).

Minimum frequency and phrase length both varied among song types (**song type models**; Table S7a): certain song types were shorter and had, on average, lower minimum frequencies than others. There was more variance among song types than among levels of other random effects. Importantly, inclusion of song type did not reduce point estimates for any random or fixed effect (Table S7a). It did increase the CIs of the random effect individual-year identity. This may signal increased uncertainty due to increased model complexity or suggest that year-specific individual differences were partly mediated by song type change.

Sources of variation in behavioural stability

For the variance part of our models (Table 1b, Fig. 2), patterns of variation also differed between the song traits. Stability in minimum frequency exhibited both short- and long-term repeatability. Though stability in phrase length showed short-term repeatability (i.e., a individual-year effect), as above, 95% CIs for individual identity effects touched zero, suggesting lack of support for cross-year repeatability.

Against expectations, our models strongly supported, and for both traits, a pattern where birds became more—not less—stable during incubation (**environmental models**; Table S6b). Food supplementation made males sing with higher stability in terms of minimum frequency but not phrase length. This treatment effect was year-specific: a negative treatment effect on residual variance was only supported in 2017 (main effect of food treatment effect; Table S6b). In the other years, the effect of food supplementation was significantly more positive in comparison (2018 × food treatment, 2019 × food treatment; Table S6b). However, adding the interaction effects to the

estimates for 2017 revealed no support for positive effects in either of the two years. Thus, food supplementation did not affect stability in those years.

For both traits, stability varied among song types (**song type models**; Table S7b, Fig. 2): certain song types had higher spectral (minimum frequency) and temporal (phrase length) stability. Most variance was among recordings or song types. Importantly, the inclusion of song type only explained residual variance in stability (Table S7b; Fig. 2), within- and among-individual variation in stability were not affected: neither among- nor within-variation in stability arose from variation in song type.
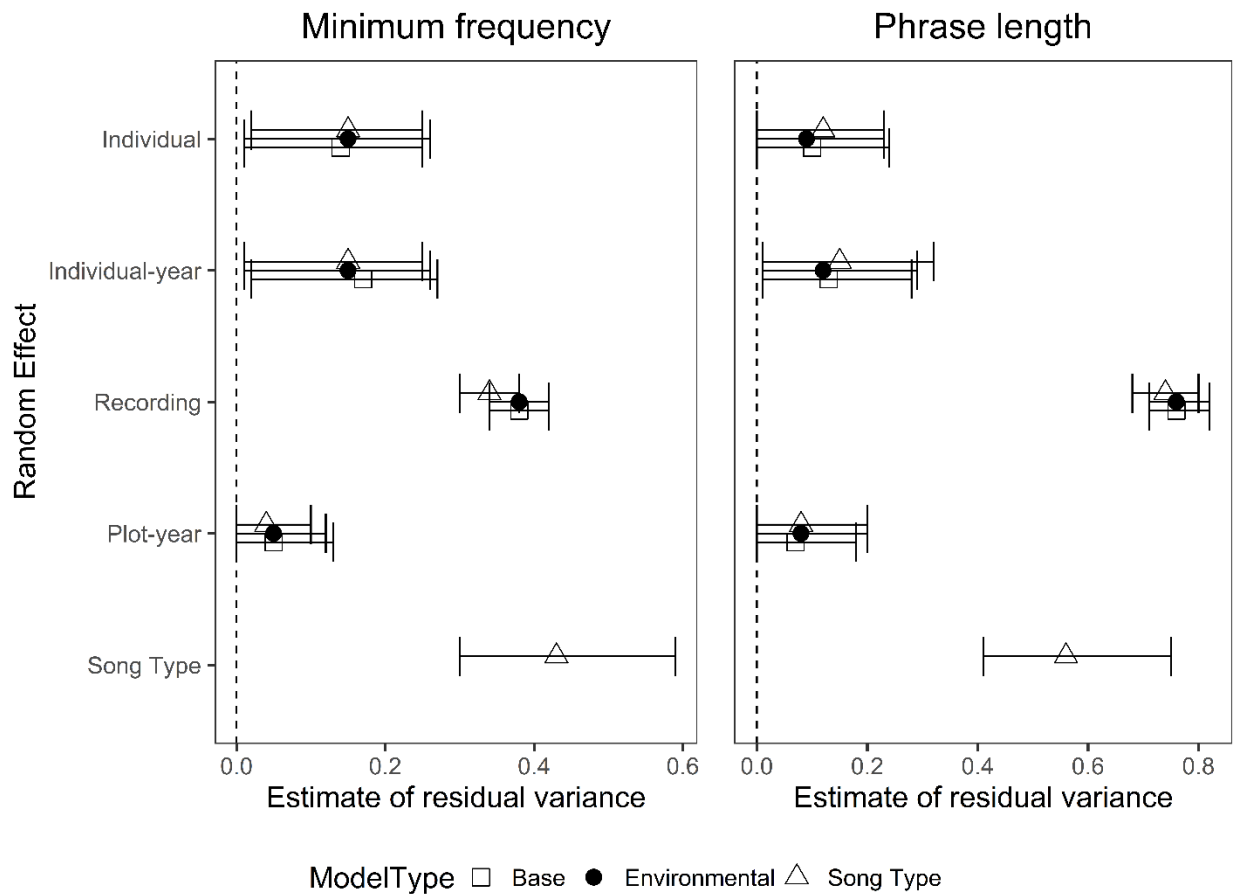

Figure 2: Estimates of variance in residual variance among five random effects for minimum

frequency (A) and phrase length (B). Estimates of base (empty square), environmental (filled triangle) and song type models (crossed out circle) are presented. Error bars indicate upper and lower credible intervals, vertical dotted lines represent zero to illustrate estimates not touching zero and thus supporting variation.

Correlations between mean and stability

For both song traits, mean and variance positively correlated among recordings: when individuals increased minimum frequency (or phrase length) across recordings, they also became less stable (Recording identity correlation) (Table 1c). Mean-variance correlations were very uncertain at the individual and individual-year identity level, thus not interpretable, though mean-variance relationships were similar across most levels (Table 1c): individuals producing higher minimum frequencies likely also produced less stable songs. Mean-variance correlations among song types were also supported: song types of higher frequency (or with longer phrase lengths) were less stable (Table S7c).

Table 1: Effect sizes and 95% credible intervals (CI) for predictors of mean (a) and residual variance (b) in minimum frequency and phrase length, and (c) mean-variance correlations (Base models).

(a) Means	Minimum frequency N = 21578	Phrase length N = 20618
Fixed Effects	β (95% CI)	β (95% CI)
Intercept	-0.01 (-0.10, 0.09)	0.45 (0.43, 0.46)
Year 2018	-0.04 (-0.17, 0.09)	0.01 (-0.01, 0.03)
Year 2019	-0.04 (-0.19, 0.10)	0.02 (-0.00, 0.04)
Random Effects	σ^2 (95% CI)	σ^2 (95% CI)
Individual	0.23 (0.07, 0.32)	0.03 (0.01, 0.04)
Individual-year	0.11 (0.01, 0.27)	0.01 (0.00, 0.03)
Recording	0.47 (0.42, 0.51)	0.07 (0.07, 0.08)
Plot-year	0.05 (0.00, 0.13)	0.01 (0.00, 0.02)
(b) Residual variances		
Fixed Effects	φ (95% CI)	φ (95% CI)
Intercept	-0.43 (-0.52, -0.35)	-3.86 (-3.99, -3.73)
Year 2018	0.03 (-0.08, 0.14)	0.06 (-0.11, 0.24)
Year 2019	0.05 (-0.07, 0.17)	-0.07 (-0.26, 0.12)
Random Effects	σ^2 (95% CI)	σ^2 (95% CI)
Individual	0.14 (0.01, 0.25)	0.10 (0.00, 0.24)
Individual-year	0.17 (0.02, 0.27)	0.13 (0.01, 0.28)
Recording	0.38 (0.34, 0.42)	0.76 (0.71, 0.82)
Plot-year	0.05 (0.00, 0.13)	0.07 (0.00, 0.18)
(c) Mean-variance Correlations (a × b)	r (95% CI)	r (95% CI)
Individual	0.37 (-0.77, 0.95)	0.28 (-0.85, 0.97)
Individual-year	0.25 (-0.90, 0.97)	-0.11 (-0.92, 0.96)
Recording	0.44 (0.33, 0.55)	0.22 (0.12, 0.31)

Food manipulation and body condition

We found no support for the prediction that winter food supplementation bettered the males' condition. Following beech mast years, body mass during breeding increased by 0.34g (0.02 – 0.63) but neither body mass during winter, nor body mass during breeding were affected by food supplementation (Table S8).

Reproductive success

We found either no—or at best weak—support for fitness effects of bird song. Effects were neither found with respect to an individual's average level of singing nor with respect to its stability. Briefly, neither lay date, clutch size, probability to produce fledglings, nor the number of fledglings varied with an individual's mean or stability in behavior (Table 2, Fig. 3). Analyses of fitness are detailed in the supplementary, where we also describe further models exploring correlational and heterogeneous selection, which also did not fit the data (Text S3, Tables S9 & S10).

Table 2: Effect sizes and 95% credible intervals (CI) for long-term individual-specific estimates of mean and variance (=instability) in minimum frequency on reproductive parameters. All models control for effects of year.

(a) Means	Lay date N = 362	Clutch size N = 362	Nest success N = 362	Number fledglings N = 321
Fixed effects	β (95% CI)	β (95% CI)	β (95% CI)	β (95% CI)
Intercept[a]	7.80 (6.43, 9.22)	8.39 (8.05, 8.72)	2.19 (1.59, 2.80)	5.00 (4.60, 5.39)
Year 2018	13.01 (11.60, 14.46)	0.05 (-0.31, 0.41)	0.92 (-0.10, 1.94)	2.23 (1.76, 2.71)
Year 2019	4.99 (3.46, 6.45)	-0.68 (-1.11, -0.28)	-0.96 (-1.72, -0.20)	-0.76 (-1.32, -0.21)
Minimum frequency (mean)*	-0.37 (-5.13, 4.37)	-0.19 (-1.73, 1.28)	-1.20 (-4.08, 1.67)	-0.08 (-1.75, 1.51)
Minimum frequency (variance)[c]*	9.28 (-0.37, 19.14)	-0.24 (-3.24, 2.79)	2.52 (-3.20, 8.03)	0.25 (-2.91, 3.43)
Random effects	σ^2 (95% CI)	σ^2 (95% CI)	σ^2 (95% CI)	σ^2 (95% CI)
Individual	3.27 (2.64, 4.05)	0.44 (0.36, 0.53)	0 (0, 0)	0.22 (0.17, 0.27)
Plot-year	0.43 (0.24, 0.66)	0.003 (0.001, 0.004)	0 (0, 0)	0.11 (0.06, 0.18)
Plot	4.33 (1.72, 8.19)	0.14 (0.05, 0.26)	0 (0, 0)	0.11 (0.04, 0.21)
Residual	23.16 (20.12, 26.63)	1.94 (1.67, 2.24)	3.29[b]	2.32 (1.97, 2.70)

[a] Reproductive parameters for birds breeding in 2017 with population level mean and RWV of minimum frequency

[b] Residual variance fixed to $\pi^2/3$ for binary models

[c] Higher values imply decreased behavioral stability

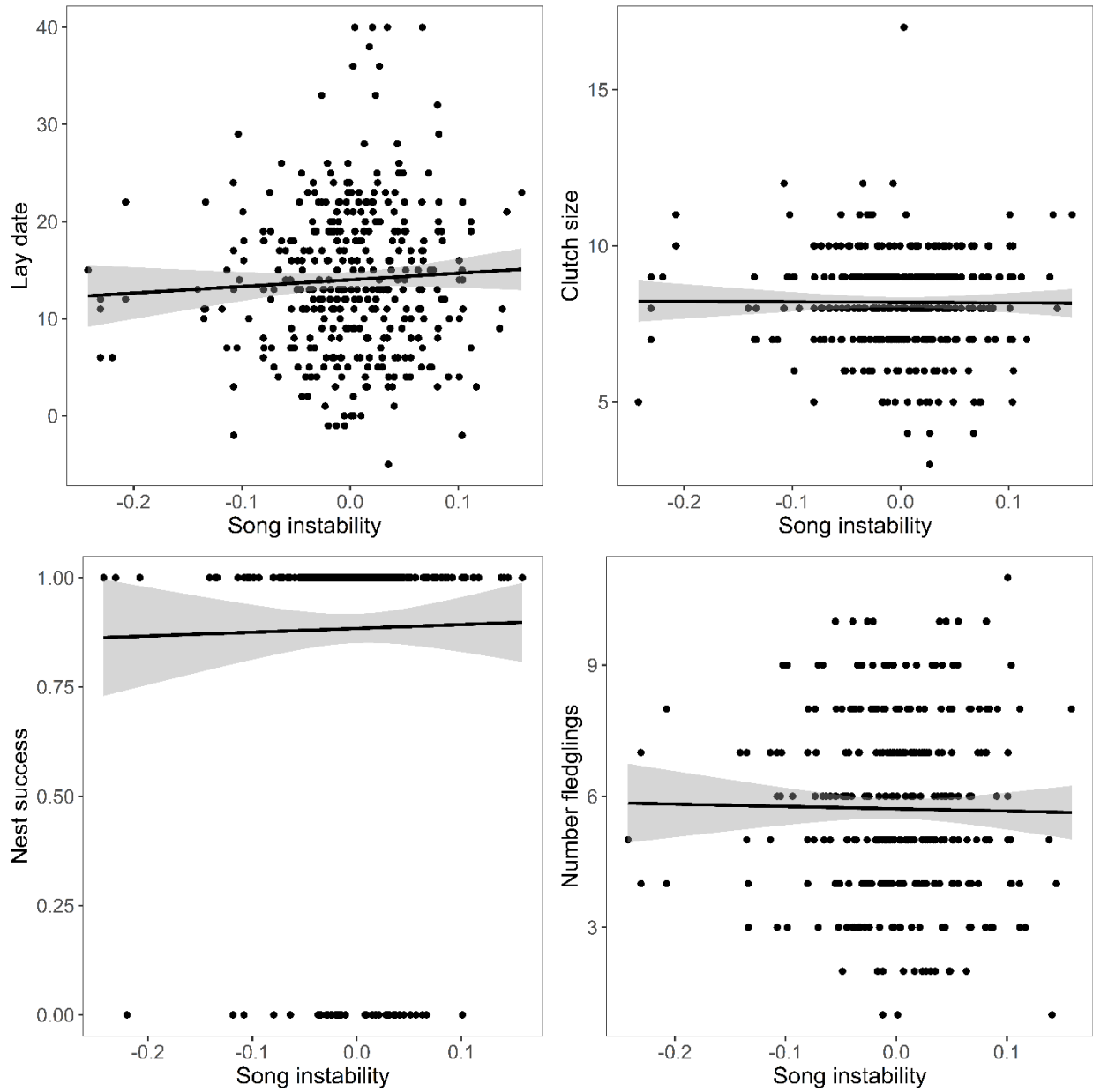* An individual's merit for each trait was estimated as its Best Linear Unbiased Predictor

Figure 3: Four reproductive parameters as a function of the Best Linear Unbiased Predictor (BLUP) of instability in minimum frequency for each individual. Black lines represent linear slopes, grey areas confidence intervals.

Discussion

We showed that great tits were plastic and repeatable in song stability both within and across years. Within years, phenotypic plasticity occurred because males changed song stability over the season. Across years, individuals varied their stability with individual-specific environmental factors that were stable within years but changed across years. Individual differences in stability were thus partly repeatable over short time spans (causing within-year repeatability). Such effects would lead to individual repeatability in stability in single-year studies because testing conditions are repeatable (thus pseudo-repeatability). Simultaneously, we found strong evidence for individual differences in stability that were repeatable over long time spans (causing cross-year repeatability).

Individual differences in singing did not correlate with any of our measures of reproductive success. This either means that song stability does not represent a quality indicator, or that it affects fitness via other pathways, such as survival, or extra-pair paternity. This raises the question of which mechanisms cause and maintain individual variation in bird song stability. Finally, these patterns of variation characterized only one of the studied traits (minimum frequency), implying that biological processes shaping song stability are, in addition, trait-specific. Overall, our study thereby exemplifies the utility of multi-trait and multi-year studies in the specific context of song stability but also in bird song studies in general.

Plasticity in behavioural stability

We demonstrated plasticity in bird song stability: males changed stability over days and years. Some of this temporal variation resulted from seasonal plasticity as individuals increased

stability from laying to incubation. We expected benefits of stability to peak prior to egg laying, when great tits invest more in low-frequency song types during dawn-chorus singing at their mate's nest box (Halfwerk et al., 2011). Opposite to expectations, stability increased over the season. During egg laying, our great tits respond to intrusions by alarm calling and using physical aggression; they shift to singing during incubation (Araya-Ajoy & Dingemanse, 2014; Strauß et al., 2020). Stability may therefore at best only represent a quality signal when stakes are not peaking but when birds still strongly rely on singing. We also detected unexplained within-individual variation in stability. Birds may face day-to-day changes in food availability or micro-climatic factors affecting energy reserves, thus their production of high-quality songs (Barnett & Briskie, 2007; Strauß et al., 2020; Thomas, 1999). Among years, song stability may vary with spring temperature, territory quality, breeding density, or age (Botero et al., 2009; De Kort et al., 2009; Grava et al., 2012; Strauß et al., 2020). Importantly, males did not plastically shift their stability by changing song types, though song types differed in length, minimum frequency and stability (see also Logue et al., 2007; Slabbekoorn & den Boer-Visser, 2006).

Our experimental food manipulation, furthermore, revealed year-specific effects of food supplementation on stability. Stability increased in one of two beech mast years. This finding implies that factors other than natural winter food availability (e.g. snow cover, temperature) mediated food supplementation effects. We considered food supplementation because we assumed it would improve the condition of supplemented birds at the onset of the breeding season, such that fewer additional resources would have to be gathered to develop high quality (stable) songs. Indeed, birds were heavier during breeding after natural beech crop years. We thus tested whether body mass (during winter or breeding) was affected by food supplementation, which was not the case. We should therefore also consider processes other than phenotypic plasticity. For example,

if stable singers are more competitive (Botero et al., 2009), they might have prioritized access to high-density environments created by our manipulation. Regardless of whether experimental effects were caused by plasticity vs. non-random settlement, our results do not easily match the idea that stability acts as a general signal of quality (Zsebők et al., 2017): supportive evidence was absent in two out of three years.

Several other studies also manipulated environmental conditions to test for adaptive shifts in stability. For example, variance-sensitive foraging theory posits that selection can favour individuals that accept risks during foraging when faced with demanding or unfavourable conditions; several studies show experimental increases in brood demand increasing within-individual variance in parental provisioning (Mathot et al., 2017; Westneat et al., 2013, 2017). Individuals also adaptively downregulate behavioural stability to reduce predation risk (Briffa et al., 2013). Finally, exploring many behavioural options early in life, can be adaptive when high behavioural variance facilitates learning optimal behaviour (Brembs, 2011), which may then be followed by increasing stability in adulthood when the optimal behaviours are adopted (Biro & Adriaenssens, 2013; De Kort et al., 2009). Interestingly, previous experiments also revealed year-specific effects on behavioural stability when studied in a foraging (rather than communication) context (Mathot et al., 2017). This and our study combined thus demonstrate a key role for year-specific ecological factors as moderators of behavioural stability, a finding that should inform future experimental study designs.

Individual differences in stability

For minimum frequency, we found cross-year repeatability in stability. In great tits, as little as half of the breeders survive between years (Bauchau & Van Noordwijk, 1995), implying that individuals were repeatable over biologically relevant timespans. Previous studies did not measure song stability over multiple years, and neither accounted for within-individual variation among years; nonetheless, several assumed that individuals were repeatable (Table S1). Of course, individual repeatability is required for song stability to serve as a signal of 'individual quality', whether shaped by early-life conditions and/or by genetic make-up (Bartsch et al., 2016; Holveck et al., 2008). As repeatability sets an upper limit to heritability (Falconer & Mackay, 1996), our analogous estimates for individuality in stability imply it might be able to evolve. Importantly, because signals of individuality in stability were present for minimum frequency but not for phrase length, different aspects of song stability likely vary in evolvability (Houle, 1992).

Repeatability of behavioural stability is receiving increasing attention (Biro & Adriaenssens, 2013; He et al., 2017; Hertel et al., 2021; Highcock & Carter, 2014; Jolles et al., 2019; Mitchell et al., 2016; Prentice et al., 2020; Stamps et al., 2012). Surprisingly little theory exists that formally predicts the emergence of repeatability of stability. For signalling behaviour, individual variation in stability might persist if traded-off with other costly traits, such as song length or duration (Sakata et al., 2008; Schuett et al., 2010), leading to a shallow fitness landscape where selection cannot easily erode standing variation (e.g. Mangel & Stamps, 2001). Mean-variance correlations are potentially important candidate mechanisms here (Briffa et al., 2013; Highcock & Carter, 2014; Prentice et al., 2020; Rönnegård et al., 2010). An example is the trade-off between accuracy (mean) and precision (variance) (Duckworth et al., 2018; Sih & Del Giudice, 2012).

A possible functional explanation for our findings is, however, that "individual quality" attributable to genetic make-up or early life conditions (silver spoon and developmental stress effects) explains why some birds can sing both low-frequency and stable songs. If both traits are favoured, their correlation assists evolution (Schluter, 1996), and directional sexual selection would then act to erode standing variation (Lande & Arnold, 1983). The persistence of repeatable individual variation may then result from mutation-selection balance (Desai & Fisher, 2007) or involve trade-offs detailed above. However, as we did not find evidence for selection, the former explanation (mutation-selection balance) hinges on the notion that directional selection would act via other pathways. For example, stable singers might lose less paternity within their nests or acquire more extra-pair offspring (Araya-Ajoy et al., 2016). Regarding the latter explanation (trade-offs), we know that some behavioural types are less plastic, thus behaving more predictably (He et al., 2017; Mitchell et al., 2016; Stamps et al., 2012; but see Hertel et al., 2021). Ultimately, future studies should focus on identifying trait correlations, and study direct vs. indirect pathways of selection (Martin et al., 2017; Prentice et al., 2020). Unpredictable behaviour may also persist because of fluctuating selection (e.g. costs varying with predation risk). Understanding the maintenance of individual variation in stability will thus require integrating studies of quantitative genetics and selection (Martin et al., 2017; Prentice et al., 2020).

Selection on behavioural stability

Song stability can affect within-pair reproductive success via two non-independent pathways, none of which were supported. First, pairs might mate assortatively with respect to quality, either by mutual mate choice, preference, or competition (Holveck & Riebel, 2010; Jiang et al., 2013). If

stable singing signals male quality, and assortative mating for quality occurs, we would expect stable singers to have mated with females that breed early and (thus) produce larger clutches (Araya-Ajoy et al., 2016; Hatchwell, 1991; Abbey-Lee & Dingemanse 2019). Second, if song stability predicts quality in terms of competitiveness, we expect them to occupy high-quality territories, or provide the best parental care: stable singers should then be more successful in raising offspring, independent of female quality. None of those predictions were supported, casting doubt on the validity of the idea that song stability represents a fitness indicator by signalling quality.

Some studies imply that stable singers mate with early-breeding females, or produce more offspring, but that they are not more able to attract a mate (Byers et al., 2015; Taff et al., 2012). It is possible that great tits are an exception where unstable singers are less likely to become breeders. Testing this explanation would require comparisons of bird song between single and settled males. We view such explanations as unlikely for two reasons. First, the empirical evidence for song stability representing a proxy for male quality is scarce. Second, datasets biased towards stable singers apparently still harbour sufficient individual variation in stability to demonstrate nonzero repeatability; the same should be true for fitness effects. Third, selection acting through female preference is highly species- and trait-specific (Macdougall-Shackleton, 1997; Soma & Garamszegi, 2011). In great tits, repertoire size, rather than other song traits may affect female choice, survival and reproduction (Baker et al., 1986; McGregor et al., 1981; Rivera-Gutierrez et al., 2010). Finally, meta-analyses imply that the available estimates demonstrating effects on within-pair reproductive success might be biased upward due to publication biases and experimental confounds (Soma & Garamszegi, 2011). Our finding that stability and reproductive success are not linked might thus genuinely match the true patterns in nature.

Single vs. multi-year studies and types of repeatability

Individual differences in aspects of bird song stability were both repeatable within and across years. These two aspects of individuality have different proximate underpinnings. The former is fully attributable to reversible plasticity, the latter to permanent environmental effects plus genetic variance (Araya-Ajoy et al., 2015; Araya-Ajoy & Dingemanse, 2017). Study designs enabling the estimation of these two components thereby greatly further our understanding of proximate underpinning of bird song. We wholeheartedly recommend that bird song research starts broadly implementing such multi-year sampling schemes, and associated statistical approaches, introduced in this paper. For example, to what extent is repertoire size repeatable within vs. among years? And do other metrics of stability in other species show the same patterns we demonstrated?

Importantly, evidence for within-year repeatability can arguably be interesting in its own right. Even if the trait is not repeatable over years, it implies that male phenotypes are stable within breeding seasons and thus may be accurately estimated by receivers (e.g., mates, competitors). Indeed, any (sexual) selection imposed by actions of conspecifics will likely be in response to the male's behaviour(al stability) expressed in a given year, in part because receivers cannot estimate a male's overall long-term mean. Single-year studies can thus be useful, particularly in identifying mechanisms of selection. Those should, however, be interpreted with extreme caution: recent studies show that estimates of selection on repeatable traits that are plastic (or measured with error) can be highly biased when based on this type of data (Dingemanse et al., 2021; Ponzi et al., 2018). The solution would therefore be to ensure that individual-specific environment effects are estimated regardless, which strictly requires multiyear studies.

Conclusion

We show that song stability is repeatable across years in wild great tits. Song stability was also plastic, changing over the season and with supplementary food manipulations. We sampled males for their songs both within and among years, which revealed substantial pseudo- and genuine individual repeatability. Great caution is thus required in studying individuality in bird song, for which single-year studies are ill advised. Repeatability and plasticity in stability differed among traits, implying that stability may not always be functional or serve trait-specific functions. Finally, we found no evidence for selection on song stability, which questions the validity of the common assumption that song stability signals male quality. Further work is required to test whether this conclusion holds for other ways by which birds differ in song stability, and thereby address whether reported patterns are species- or context-specific versus generally applicable.

Data availability

Data available from the Dryad Digital Repository https://doi.org/10.5061/dryad.tmpg4f519 (Hutfluss et al. 2022).

All authors contributed to the writing and approved submission. Data will be deposited on Dryad following manuscript acceptance.

Conflict of Interest

No conflicts of interest apply

Author Contributions

Alexander Hutfluss, Hans Slabbekoorn and Niels Dingemanse conceived and designed the study. Eira Bermudez-Cuamatzin, Mark Briffa, Hans Slabbekoorn and Niels Dingemanse gave conceptual and methodological advice and help on the collection, extraction and analysis of the data. Alexander Hutfluss and Alexia Mouchet collected the data, Alexander Hutfluss analysed the data. Alexander Hutfluss and Niels Dingemanse led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Abbey-Lee, R. N., & Dingemanse, N. J. (2019). Adaptive individual variation in phenological responses to perceived predation levels. *Nature communications*, *10*(1), 1-8.

Araya-Ajoy, Y. G., & Dingemanse, N. J. (2014). Characterizing behavioural 'characters': an evolutionary framework. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*, 20132645. https://doi.org/10.1098/rspb.2013.2645

Araya-Ajoy, Y. G., & Dingemanse, N. J. (2017). Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *Journal of Animal Ecology*, *86*(2), 227–238. https://doi.org/10.1111/1365-2656.12621

Araya-Ajoy, Y. G., Kuhn, S., Mathot, K. J., Mouchet, A., Mutzel, A., Nicolaus, M., Wijmenga, J. J., Kempenaers, B., & Dingemanse, N. J. (2016). Sources of (co)variation in alternative siring routes available to male great tits (Parus major). *Evolution*, *70*(10), 2308–2321. https://doi.org/10.1111/evo.13024

Araya-Ajoy, Y. G., Mathot, K. J., & Dingemanse, N. J. (2015). An approach to estimate short-term, long-term and reaction norm repeatability. *Methods in Ecology and Evolution*, *6*(12), 1462–1473. https://doi.org/10.1111/2041-210X.12430

Auld, J. R., Agrawal, A. A., & Relyea, R. A. (2010). Re-evaluating the costs and limits of adaptive phenotypic plasticity. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1681), 503–511. https://doi.org/10.1098/rspb.2009.1355

Baker, M. C., Bjerke, T. K., Lampe, H., & Espmark, Y. (1986). Sexual Response of Female Great Tits to Variation in Size of Males' Song Repertoires. *The American Naturalist*, *128*(4), 491–498. https://doi.org/10.1086/284582

Barnett, C. A., & Briskie, J. V. (2007). Energetic state and the performance of dawn chorus in

silvereyes (Zosterops lateralis). *Behavioral Ecology and Sociobiology*, *61*, 579–587. https://doi.org/10.1007/s00265-006-0286-x

Bartsch, C., Hultsch, H., Scharff, C., & Kipper, S. (2016). What is the whistle all about? A study on whistle songs, related male characteristics, and female song preferences in common nightingales. *Journal of Ornithology*, *157*(1), 49–60. https://doi.org/10.1007/s10336-015-1245-y

Bauchau, V., & Van Noordwijk, A. J. (1995). Comparison of survival estimates obtained from three different methods of recapture in the same population of the great tit. *Journal of Applied Statistics*, *22*(5–6), 1031–1038. https://doi.org/10.1080/02664769524775

Bermúdez-Cuamatzin, E., Ríos-Chelén, A. A., Gil, D., & Garcia, C. M. (2011). Experimental evidence for real-time song frequency shift in response to urban noise in a passerine bird. *Biology Letters*, *7*(1), 36–38. https://doi.org/10.1098/rsbl.2010.0437

Biro, P. A., & Adriaenssens, B. (2013). Predictability as a personality trait: consistent differences in intraindividual behavioral variation. *The American Naturalist*, *182*(5), 621–629. https://doi.org/10.1086/673213

Botero, C. A., Rossman, R. J., Caro, L. M., Stenzler, L. M., Lovette, I. J., de Kort, S. R., & Vehrencamp, S. L. (2009). Syllable type consistency is related to age, social status and reproductive success in the tropical mockingbird. *Animal Behaviour*, *77*(3), 701–706. https://doi.org/10.1016/j.anbehav.2008.11.020

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1707), 930–939. https://doi.org/10.1098/rspb.2010.2325

Briffa, M. (2013). Plastic proteans: reduced predictability in the face of predation risk in hermit

crabs. *Biology Letters*, *9*(5), 20130592. https://doi.org/10.1098/rsbl.2013.0592

Briffa, M., Bridger, D., & Biro, P. A. (2013). How does temperature affect behaviour? Multilevel analysis of plasticity, personality and predictability in hermit crabs. *Animal Behaviour*, *86*(1), 47–54. https://doi.org/10.1016/J.ANBEHAV.2013.04.009

Bürkner, P.-C. (2017). Advanced Bayesian Multilevel Modeling with the R Package brms. *R Journal*, *10*(1), 395–411. http://arxiv.org/abs/1705.11123

Byers, B. E. (2007). Extrapair paternity in chestnut-sided warblers is correlated with consistent vocal performance. *Behavioral Ecology*, *18*(1), 130–136. https://doi.org/10.1093/beheco/arl058

Byers, B. E., Akresh, M. E., & King, D. I. (2015). A proxy of social mate choice in prairie warblers is correlated with consistent, rapid, low-pitched singing. *Behavioral Ecology and Sociobiology*, *69*(8), 1275–1286. https://doi.org/10.1007/s00265-015-1940-y

Cardoso, G. C. (2012). Paradoxical calls: the opposite signaling role of sound frequency across bird species. *Behavioral Ecology*, *23*(2), 237–241. https://doi.org/10.1093/beheco/arr200

Cardoso, G. C., & Atwell, J. W. (2011). On the relation between loudness and the increased song frequency of urban birds. *Animal Behaviour*, *82*(4), 831–836. https://doi.org/10.1016/j.anbehav.2011.07.018

Cleasby, I. R., Nakagawa, S., & Schielzeth, H. (2015). Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, *6*(1), 27–37. https://doi.org/10.1111/2041-210X.12281

Cramer, E. R. A. (2013a). Physically Challenging Song Traits, Male Quality, and Reproductive Success in House Wrens. *PLoS ONE*, *8*(3), e59208.

https://doi.org/10.1371/journal.pone.0059208

Cramer, E. R. A. (2013b). Measuring consistency: spectrogram cross-correlation versus targeted acoustic parameters. *Bioacoustics*, *22*(3), 247–257. https://doi.org/10.1080/09524622.2013.793616

De Kort, S. R., Eldermire, E. R. B., Valderrama, S., Botero, C. A., & Vehrencamp, S. L. (2009). Trill consistency is an age-related assessment signal in banded wrens. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2315–2321. https://doi.org/10.1098/rspb.2009.0127

Desai, M. M., & Fisher, D. S. (2007). Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics*, *176*(3), 1759–1798. https://doi.org/10.1534/genetics.106.067678

Dingemanse, N. J., Kazem, A. J. N., Réale, D., & Wright, J. (2010). Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution*, *25*(2), 81–89. https://doi.org/10.1016/J.TREE.2009.07.013

Dingemanse, N. J., Araya-Ajoy, Y. G., & Westneat, D. F. (2021). Most published selection gradients are underestimated: Why this is and how to fix it. *Evolution*, *75*(4), 806–818. https://doi.org/10.1111/evo.14198

Domenici, P., Booth, D., Blagburn, J. M., & Bacon, J. P. (2008). Cockroaches Keep Predators Guessing by Using Preferred Escape Trajectories. *Current Biology*, *18*(22), 1792–1796. https://doi.org/10.1016/j.cub.2008.09.062

Duckworth, R. A., Potticary, A. L., & Badyaev, A. V. (2018). On the Origins of Adaptive Behavioral Complexity: Developmental Channeling of Structural Trade-offs. In *Advances in the Study of Behavior* (Vol. 50, pp. 1–36). Academic Press Inc.

https://doi.org/10.1016/bs.asb.2017.10.001

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. Pearson Education India.

Grava, T., Grava, A., & Otter, K. A. (2012). Vocal performance varies with habitat quality in black-capped chickadees (Poecile atricapillus). *Behaviour*, *149*(1), 35–50. https://doi.org/10.1163/156853912X625854

Grava, T., Grava, A., & Otter, K. A. (2013). Habitat-induced changes in song consistency affect perception of social status in male chickadees. *Behavioral Ecology and Sociobiology*, *67*(10), 1699–1707. https://doi.org/10.1007/s00265-013-1580-z

Guidelines for the treatment of animals in behavioural research and teaching. (2016). In *Animal Behaviour* (Vol. 111, pp. I–IX). Academic Press. https://doi.org/10.1016/S0003-3472(15)00461-3

Halfwerk, W., Bot, S., Buikx, J., Van Der Velde, M., Komdeur, J., Ten Cate, C., & Slabbekoorn, H. (2011). Low-frequency songs lose their potency in noisy urban conditions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(35), 14549–14554. https://doi.org/10.1073/pnas.1109091108

Hatchwell, B. J. (1991). An Experimental Study of the Effects of Timing of Breeding on the Reproductive Success of Common Guillemots (Uria aalge). *The Journal of Animal Ecology*, *60*(3), 721. https://doi.org/10.2307/5410

He, R., Pagani-Núñez, E., Chevallier, C., & Barnett, C. R. A. (2017). To be so bold: boldness is repeatable and related to within individual behavioural variability in North Island robins. *Behavioural Processes*, *140*, 144–149. https://doi.org/10.1016/j.beproc.2017.04.014

Hertel, A. G., Royauté, R., Zedrosser, A., & Mueller, T. (2021). Biologging reveals individual

variation in behavioural predictability in the wild. *Journal of Animal Ecology*, *90*(3), 723–737. https://doi.org/10.1111/1365-2656.13406

Highcock, L., & Carter, A. J. (2014). Intraindividual Variability of Boldness Is Repeatable across Contexts in a Wild Lizard. *PLoS ONE*, *9*(4), e95179. https://doi.org/10.1371/journal.pone.0095179

Holveck, M.-J., Vieira de Castro, A. C., Lacklan, R. F., ten Cate, C., & Riebel, K. (2008). Accuracy of song syntax learning and singing consistency signal early condition in zebra finches. *Behavioral Ecology*, *19*(6), 1267–1281.

Holveck, M. J., & Riebel, K. (2010). Low-quality females prefer low-quality males when choosing a mate. *Proceedings. Biological Sciences / The Royal Society*, *277*(1678), 153–160. https://doi.org/10.1098/rspb.2009.1222

Houle, D. (1992). Comparing evolvability and variability of quantitative traits. *Genetics*, *130*(1).

Hutfluss, A., Bermúdez-Cuamatzin, E., Mouchet, A., Briffa, M., Slabbekoorn, H., Dingemanse, N.J.. (2022). Male song stability shows cross-year repeatability but does not affect reproductive success in a wild passerine bird, Dryad, Dataset, https://doi.org/10.5061/dryad.tmpg4f519

Jiang, Y., Bolnick, D. I., & Kirkpatrick, M. (2013). Assortative mating in animals. *American Naturalist*, *181*(6). https://doi.org/10.1086/670160

Jolles, J. W., Briggs, H. D., Araya-Ajoy, Y. G., & Boogert, N. J. (2019). Personality, plasticity and predictability in sticklebacks: bold fish are less plastic and more predictable than shy fish. *Animal Behaviour*, *154*, 193–202. https://doi.org/10.1016/j.anbehav.2019.06.022

Jones, K. A., Jackson, A. L., & Ruxton, G. D. (2011). Prey jitters; protean behaviour in grouped prey. *Behavioral Ecology*, *22*(4), 831–836. https://doi.org/10.1093/beheco/arr062

Lambrechts, M. M. (1997). Song frequency plasticity and composition of phrase versions in great tits Parus major. *Ardea*, *85*(1), 99–109.

Lambrechts, M. M., & Dhondt, A. A. (1986). Male quality, reproduction, and survival in the great tit (Parus major). *Behavioral Ecology and Sociobiology*, *19*(1), 57–63. https://doi.org/10.1007/BF00303843

Lande, R., & Arnold, S. J. (1983). The Measurement of Selection on Correlated Characters. *Evolution*, *37*(6), 1210. https://doi.org/10.2307/2408842

Lane, S. M., & Briffa, M. (2020). Perceived and actual fighting ability: determinants of success by decision, knockout or submission in human combat sports. *Biology Letters*, *16*(10), 20200443. https://doi.org/10.1098/rsbl.2020.0443

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *55*(2), 139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Logue, D. M., Droessler, E. E., Roscoe, D. W., Vokey, J. R., Rendall, D., & Kunimoto, R. M. (2007). Sexually antithetical song structure in a duet singing wren. *Behaviour*, *144*(3), 331–350. https://doi.org/10.1163/156853907780425749

Macdougall-Shackleton, S. A. (1997). Sexual Selection and the Evolution of Song Repertoires. In *Current Ornithology* (pp. 81–124). Springer US. https://doi.org/10.1007/978-1-4757-9915-6_3

MacDougall-Shackleton, S. A., & Spencer, K. A. (2012). Developmental stress and birdsong: Current evidence and future directions. In *Journal of Ornithology* (Vol. 153, Issue SUPPL. 1, pp. 105–117). Springer. https://doi.org/10.1007/s10336-011-0807-x

Mangel, M., & Stamps, J. (2001). Trade-offs between growth and mortality and the maintenance

of individual variation in growth. In *Evolutionary Ecology Research* (Vol. 3, Issue 5). Evolutionary Ecology, Ltd.

Marler, P., & Isaac, D. (1960). Physical Analysis of a Simple Bird Song as Exemplified by the Chipping Sparrow. *The Condor*, *62*(2), 124–135. https://doi.org/10.2307/1365677

Marques-Santos, F., & Dingemanse, N. J. (2020). Weather effects on nestling survival of great tits vary according to the developmental stage. *Journal of Avian Biology*, *51*(10), jav.02421. https://doi.org/10.1111/jav.02421

Martin, J. G. A., Pirotta, E., Petelle, M. B., & Blumstein, D. T. (2017). Genetic basis of between-individual and within-individual variance of docility. *Journal of Evolutionary Biology*, *30*(4), 796–805. https://doi.org/10.1111/jeb.13048

Mathot, K. J., Olsen, A.-L., Mutzel, A., Araya-Ajoy, Y. G., Nicolaus, M., Westneat, D. F., Wright, J., Kempenaers, B., & Dingemanse, N. J. (2017). Provisioning tactics of great tits (Parus major) in response to long-term brood size manipulations differ across years. *Behavioral Ecology*, *28*(6), 1402–1413. https://doi.org/10.1093/beheco/arx083

McGregor, P. K., Krebs, J. R., & Perrins, C. M. (1981). Song Repertoires and Lifetime Reproductive Success in the Great Tit (Parus major). *The American Naturalist*, *118*(2), 149–159. https://doi.org/10.1086/283811

Mitchell, D. J., Beckmann, C., & Biro, P. A. (2021). Understanding the unexplained: The magnitude and correlates of individual differences in residual variance. *Ecology and Evolution*, ece3.7603. https://doi.org/10.1002/ece3.7603

Mitchell, D. J., Fanson, B. G., Beckmann, C., & Biro, P. A. (2016). Towards powerful experimental and statistical approaches to study intraindividual variability in labile traits. *Royal Society Open Science*, *3*(10). https://doi.org/10.1098/rsos.160352

Nicolaus, M., Mathot, K. J., Araya-Ajoy, Y. G., Mutzel, A., Wijmenga, J. J., Kempenaers, B., & Dingemanse, N. J. (2015). Does coping style predict optimization? An experimental test in a wild passerine bird. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1799), 20142405–20142405. https://doi.org/10.1098/rspb.2014.2405

Nussey, D. H., Wilson, A. J., & Brommer, J. E. (2007). The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, *20*(3), 831–844. https://doi.org/10.1111/j.1420-9101.2007.01300.x

Patel, R., Mulder, R. A., & Cardoso, G. C. (2010). What Makes Vocalisation Frequency an Unreliable Signal of Body Size in Birds? A Study on Black Swans. *Ethology*, *116*(6), 554–563. https://doi.org/10.1111/j.1439-0310.2010.01769.x

Patricelli, G. L., & Hebets, E. A. (2016). New dimensions in animal communication: the case for complexity. In *Current Opinion in Behavioral Sciences* (Vol. 12, pp. 80–89). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2016.09.011

Perdeck, A. C., Visser, M. E., & Van Balen, J. H. (2000). Great tit Parus major survival and the beech-crop. *Ardea*, *88*, 99–106.

Podos, J. (1997). A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution*, *51*(2), 537–551. https://doi.org/10.1111/j.1558-5646.1997.tb02441.x

Poesel, A., & Kempenaers, B. (2000). When a bird is tired from singing: a study of drift during thedawn chorus. *Etología*, *8*, 1–7.

Ponzi, E., Keller, L. F., Bonnet, T., & Muff, S. (2018). Heritability, selection, and the response to selection in the presence of phenotypic measurement error: Effects, cures, and the role of repeated measurements. *Evolution*, *72*(10), 1992–2004. https://doi.org/10.1111/evo.13573

Prentice, P. M., Houslay, T. M., Martin, J. G. A., & Wilson, A. J. (2020). Genetic variance for behavioural 'predictability' of stress response. *Journal of Evolutionary Biology*, *33*(5), 642–652. https://doi.org/10.1111/jeb.13601

Reichert, M. S., & Gerhardt, C. H. (2012). Trade-offs and upper limits to signal performance during close-range vocal competition in gray tree frogs Hyla versicolor. *American Naturalist*, *180*(4), 425–437. https://doi.org/10.1086/667575

Rivera-Gutierrez, H. F., Pinxten, R., & Eens, M. (2010). Multiple signals for multiple messages: Great tit, Parus major, song signals age and survival. *Animal Behaviour*, *80*(3), 451–459. https://doi.org/10.1016/j.anbehav.2010.06.002

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H. A., & Strandberg, E. (2010). Genetic heterogeneity of residual variance - Estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution*, *42*(1), 1–10. https://doi.org/10.1186/1297-9686-42-8

Sakata, J. T., Hampton, C. M., & Brainard, M. S. (2008). Social Modulation of Sequence and Syllable Variability in Adult Birdsong. *Journal of Neurophysiology*, *99*(4), 1700–1711. https://doi.org/10.1152/jn.01296.2007

Sakata, J. T., & Vehrencamp, S. L. (2012). Integrating perspectives on vocal performance and consistency. In *Journal of Experimental Biology* (Vol. 215, Issue 2, pp. 201–209). The Company of Biologists Ltd. https://doi.org/10.1242/jeb.056911

Schluter, D. (1996). Adaptive radiation along genetic lines of least resistance. *Evolution*, *50*(5), 1766–1774. https://doi.org/10.1111/j.1558-5646.1996.tb03563.x

Schuett, W., Tregenza, T., & Dall, S. R. X. (2010). Sexual selection and animal personality. *Biological Reviews*, *85*(2), 217–246. https://doi.org/10.1111/j.1469-185X.2009.00101.x

Sih, A., & Del Giudice, M. (2012). Linking behavioural syndromes and cognition: a behavioural ecology perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2762–2772. https://doi.org/10.1098/rstb.2012.0216

Slabbekoorn, H., & den Boer-Visser, A. (2006). Cities Change the Songs of Birds. *Current Biology*, *16*(23), 2326–2331. https://doi.org/10.1016/j.cub.2006.10.008

Soma, M., & Garamszegi, L. Z. (2011). Rethinking birdsong evolution: meta-analysis of the relationship between song complexity and reproductive success. *Behavioral Ecology*, *22*(2), 363–371. https://doi.org/10.1093/beheco/arq219

Stamps, J. A., Briffa, M., & Biro, P. A. (2012). Unpredictable animals: individual differences in intraindividual variability (IIV). *Animal Behaviour*, *83*(6), 1325–1334. https://doi.org/10.1016/J.ANBEHAV.2012.02.017

Strauß, A. F. T., Hutfluss, A., & Dingemanse, N. J. (2020). Great tits responding to territorial intrusions sing less but alarm more on colder days. *Ethology*, *126*(4), 445–454. https://doi.org/10.1111/eth.12989

Suthers, R. A., & Zollinger, S. A. (2004). Producing Song: The Vocal Apparatus. *Annals of the New York Academy of Sciences*, *1016*(1), 109–129. https://doi.org/10.1196/annals.1298.041

Taff, C. C., Steinberger, D., Clark, C., Belinsky, K., Sacks, H., Freeman-Gallant, C. R., Dunn, P. O., & Whittingham, L. A. (2012). Multimodal sexual selection in a warbler: Plumage and song are related to different fitness components. *Animal Behaviour*, *84*(4), 813–821. https://doi.org/10.1016/j.anbehav.2012.07.002

Thomas, R. J. (1999). Two tests of a stochastic dynamic programming model of daily singing routines in birds. *Animal Behaviour*, *57*(2), 277–284. https://doi.org/10.1006/ANBE.1998.0956

Van Balen, J. H. (1980). Population fluctuations of the great tit and feeding conditions in winter. *Ardea*, *55*(1–2), 143–164.

Van De Pol, M., Bruinzeel, L. W., Heg, D., Van Der Jeugd, H. P., & Verhulst, S. (2006). A silver spoon for a golden future: Long-term effects of natal origin on fitness prospects of oystercatchers (Haematopus ostralegus). *Journal of Animal Ecology*, *75*(2), 616–626. https://doi.org/10.1111/j.1365-2656.2006.01079.x

Weary, D. M. (1989). Categorical perception of bird song: how do great tits (Parus major) perceive temporal variation in their song? *Journal of Comparative Psychology*, *103*(4), 320–325. https://doi.org/10.1037/0735-7036.103.4.320

Westneat, D. F., Mutzel, A., Bonner, S., & Wright, J. (2017). Experimental manipulation of brood size affects several levels of phenotypic variance in offspring and parent pied flycatchers. *Behavioral Ecology and Sociobiology*, *71*(6), 1–13. https://doi.org/10.1007/s00265-017-2315-3

Westneat, D. F., Schofield, M., & Wright, J. (2013). Parental behavior exhibits among-individual variance, plasticity, and heterogeneous residual variance. *Behavioral Ecology*, *24*(3), 598–604. https://doi.org/10.1093/beheco/ars207

Westneat, D. F., Wright, J., & Dingemanse, N. J. (2015). The biology hidden inside residual within-individual phenotypic variation. *Biological Reviews*, *90*(3), 729–743. https://doi.org/10.1111/brv.12131

Zsebők, S., Herczeg, G., Blázi, G., Laczi, M., Nagy, G., Szász, E., Markó, G., Török, J., & Garamszegi, L. Z. (2017). Short- and long-term repeatability and pseudo-repeatability of bird song: sensitivity of signals to varying environments. *Behavioral Ecology and Sociobiology*, *71*(10), 154. https://doi.org/10.1007/s00265-017-2379-0

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer New York. https://doi.org/10.1007/978-0-387-87458-6