

SEXUAL SELECTION IN HOUSE WRENS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Emily R. Cramer

August 2012

© 2012 Emily R Cramer

## SEXUAL SELECTION IN HOUSE WRENS

Emily R Cramer, Ph. D.

Cornell University 2012

Chapters 1 and 2 focus on two aspects of trilled song that are likely physically challenge to produce: low vocal deviation and high consistency. Chapter 1 (in review at Behavioral Ecology) describes a playback study to territorial male house wrens showing that males do not respond differently to songs with different trill performance and consistency characteristics. Chapter 2 examines whether song traits correlate with male quality measures, with social pairing success, with extra-pair mating success, or with annual reproductive success. Again, I found no evidence that these song parameters are important. I conclude that house wrens do not use these song components as signals of male quality. In Chapters 3 and 4, I examine post-copulatory sexual selection in house wrens. Chapter 3 (accepted at the Journal of Ornithology) examines correlations between male quality measures and sperm morphology; we also show that sperm morphology is consistent between years, despite testicular regression and regrowth. In Chapter 4, we ask whether variation in sperm morphology is related to extra-pair paternity success, and find that it is not. Chapter 5 (accepted at Ethology) takes a more mechanistic perspective, and asks whether high circulating testosterone is necessary for aggressive behaviors in house wrens. I found no evidence to suggest that this is the case.

## BIOGRAPHICAL SKETCH

I was born on May 29, 1984 and grew up on the family sheep farm in Jefferson, Maryland before earning a bachelor's degree in biology at St. Mary's College of Maryland. After a year off doing fieldwork in the Adirondacks, learning molecular techniques at the University of Maryland Baltimore County, and doing more fieldwork in Mexico, I began graduate school at Cornell in June 2006 with a field trip to Costa Rica. At Cornell, I struggled for a year in an attempt to study northern cardinals before switching study species to the much-more-tractable house wren. I plan to pursue a postdoc in sperm morphology with Arild Johnsen in Oslo, Norway, after completing my dissertation.

## ACKNOWLEDGMENTS

Huge thanks to Paulo Llambías for establishing the house wren population and then graduating at the perfect time; to Taza Schaming for being willing to share birds with me; and to Katie LaBarbera for getting the house wren microsatellites ready to roll.

Thanks to the behavior graduate students and postdocs for being such a fabulous support group. You make an awesome, critical, helpful audience, and I'll miss having you to pass papers around to for feedback. Thanks for building a community with me. Thanks to the Behavior Faculty for putting up with my demands for professional development advice at lunch bunch, and for providing feedback on research there, too!

For funding, thanks to The Cornell RedHeads birding team, the Kramer family (not related; Lab of O donors), the American Ornithologists' Union, the Animal Behavior Society, The Cornell Sigma Xi Chapter, an NSF predoctoral fellowship with a Nordic Research Opportunity supplement, the Department of Neurobiology and Behavior; and for travel funding, thanks to the International Society for Behavioral Ecology and the Cornell University Graduate School.

For keeping me sane and happy, and making Ithaca a great place to live: my Ithaca family (Dan Fergus, Holly Menninger, Chris Wiley, Emma Greig, Gwen Glazer, Nancy Chen, and Andy Poshadel), my knitters (Kim Falbo, Alisa Royem, et al.), my bird dudes (Tim Lenz, Matt Medler, Jay McGowan, Brad Walker, Shawn Billerman, et al.), and the NBB coffee crowd (Terri Natoli, Stacey Coil, Sandra Anderson, and Howie Howland). I'd also like to thank my genetic family for all their support.

To Jan Lifjeld and his lab, for letting me come and work in Norway for a couple of months and being a great advisor and host. To Jordan Price for suggesting to me that I try grad

school, and for continuing to encourage me to aim high. To the Lovette and Webster labs, for adopting me and being a great source of intellectual support (and particularly to Rachel Vallendar and the microsatellite and big kids' reading groups, for being my first community at Cornell); also to my official labmates and labmates-in-law.

To my advisor, Sandy Vehrencamp, my advisor-on-paper, Mike Webster, and the rest of my committee: Andy Bass, Irby Lovette, and Janis Dickinson, for all your feedback and support.

Thanks for practical things to: Laura Stenzler, Amanda Talaba and Chris Makarewich for help getting started with lab work; Peter Schweitzer and the Cornell Core Facility for sequencing and SNP analysis; the Shaw Lab for bench space for running testosterone assays; Katie McGee, Katie Baird, Eileen McIver, Noelle Chaine, Natalie Koscal, and Carly Hodes for assistance in the field; Katie Baird, Caroline Rusk, Rebecca Fellman, Nikki Thompson, Noelle Chaine, and Natalie Koscal for sound analysis assistance; Bob Johnson and Noah Hamm for access to the Cornell Research Ponds; Karl Fitzke, Greg Budney, and the Macaulay Folks for last-minute assistance throughout my field seasons getting recording and playback gear together; Charles Dardia and Kim Bostwick for bird skinning and import/export help with sperm; Wendy Williams and the Cornell Animal Health Diagnostics Center for help with IACUC and measuring house wren health; Paula Cohen's lab and Bob Doran for use of and assistance with microscopy; Cynthia Marquis, Sue Taggart, and Ken Putnam for making it work out to have field assistants; Terri Natoli, Stacey Coil, Sandra Anderson, Lori Miller and Dawn Potter for making NBB an easy department to work in; Dean Hawthorne for my supped-up version of Raven; to Mari Kimura, Yula Kapetanacos, Sarah States, and Petra Deane for co-leading Expanding Your Horizons (and for passing along the workshop!); to Wes Hochachka, Françoise Vermeulen, Jay Barry, Haim Bar, Jing Yang, and Ben Risk for statistical consulting at the drop of a hat; and to

Loretta Bennett for driving a mean shuttle that builds a strong community of Lab of O grad students.

## TABLE OF CONTENTS

Biographical sketch.....	iii
Acknowledgements .....	iv
List of Figures .....	viii
List of Tables .....	ix
Preface .....	x
Chapter 1 .....	1
Vocal deviation and trill consistency do not affect male response to playback in house wrens	
References .....	28
Appendix 1.1 Supplementary Material to Chapter 1 .....	34
Chapter 2 .....	57
Vocal deviation and trill consistency do not indicate male quality in house wrens	
References .....	91
Chapter 3 .....	97
Sperm length variation in house wrens <i>Troglodytes aedon</i>	
References .....	118
Chapter 4 .....	123
Sperm morphology does not relate to extra-pair paternity success in house wrens	
References .....	140
Chapter 5 .....	144
Are androgens related to aggression in house wrens?	
References .....	163
Conclusions .....	167

## LIST OF FIGURES

Figure 1.1 Trade-off between frequency bandwidth and trill rate .....	9
Figure 1.2 Concatenated spectrogram demonstrating pitch shifting .....	17
Figure 1.3 Responses to playback .....	18
Figure 1.4 Survival analysis of latency to approach playback speaker .....	22
Figure S.1.1 Spectrogram of house wren song .....	35
Figure S.1.2 Effect of trill position on trill pitch .....	36
Figure 2.1 Correlations in song measures between trill types .....	74
Figure 2.2 Comparison of song measures of successful and unsuccessful males.....	82
Figure 3.1 Seasonal patterns in sperm morphology .....	108
Figure 3.2 Year-to-year repeatability in sperm morphology .....	110
Figure 5.1 Seasonal patterns in testosterone and egg-laying .....	155
Figure 5.2 Androgens and aggression .....	158

## LIST OF TABLES

Table 1.1 Statistical results for playback responses.....	19
Table S.1.1 Comparison of pitch measures (high vs. low frequency) .....	38
Table S.1.2 Different approaches to analyzing playback results .....	43
Table S.1.3 Evaluation of playback stimuli relative to natural songs .....	47
Table 2.1 Correlations between male quality and song measures .....	75
Table 2.2 Relationships between male mating success and male and song quality .....	79
Table 2.3 Paired comparisons of songs and male quality for pair and extra-pair males .....	83
Table 2.4 Association of reproductive success with male and song quality, mating success ...	85
Table 3.1 Comparisons of sperm across categories of males .....	111
Table 3.2 Correlations between male quality and sperm .....	113
Table 4.1 Comparison of sperm traits for successful and unsuccessful males .....	131
Table 4.2 Paired comparisons of sperm traits for pair and extra-pair males.....	133
Table 4.3 Associations between sperm traits and reproductive success.....	134
Table 5.1 Relation of testosterone to male and capture characteristics.....	154
Table 5.2 Relation of testosterone to aggressiveness.....	157

## PREFACE

One of the lessons I remember most from introductory biology class is that form matches function. We mostly applied this concept to things like cell structure and body plans, but as I continued in biology, I learned that it can be applied to other topics as well. I like to think that my dissertation applies this general rule to two questions in sexual selection. How does acoustic form relate to song function? How does sperm form relate to sperm function? Along the way, I've also investigated the physiological underpinnings of territorial aggression, an important, sexually-selected suite of behaviors typically thought to be controlled by testosterone (incidentally, a compound that functions because of how its form matches the form of its receptors).

One potential mechanism of maintaining signal honesty is to have the form of the signal match the function. If the form of the signal is excessively difficult or costly, a low quality individual will be unable to produce a high quality signal. I focused in on two aspects of acoustic form that seemed likely to follow this hypothesis, and I investigated how these signals function in male-male communication and how they relate to male mating success, which likely reflects female choice as well as male-male communication. After extensive song measurements, I found no evidence that either of these aspects of song form functions as a signal in my study species.

Sperm form is the subject of fairly intense study, and there are a variety of studies on how sperm morphology affect swimming speed and longevity, which in turn likely affect fertilization success. I was unable to directly quantify how sperm form and function relate, but I did describe sperm morphology and seasonal change in it in my study species; I also investigated whether sperm morphology relates to extra-pair mating success. Across the breeding season, sperm length changes in a way that could be adaptive, though I was unable to determine whether this change

occurred within individuals. A single individual's sperm remained constant across years, despite testicular regression and regrowth between measuring events. Sperm function did not differ between males that were more and less successful in siring offspring.

Testosterone was an interesting potential bridge between these two areas of inquiry; high levels of testosterone are thought to be necessary for both sperm and song production. I investigated how testosterone relates to territorial aggression. Being able to defend a territory (using song, in part) is a crucial first step for a male to attract a female (and get to use his sperm). Territorial defense, to my surprise, does not appear to be regulated by testosterone as it is in sparrows. Rather, my study species fits in with a growing number of species where aggression is independent of circulating testosterone concentrations at breeding-season levels.

My study species for this work was the house wren (*Troglodytes aedon*), a species I chose primarily out of convenience. Paulo Llambías, a grad student in Ecology and Evolutionary Biology, had a population of box-nesting house wrens located in Ithaca that he was finishing with, and Taza Schaming (a grad student in the Department of Natural Resources who'd already decided to work on Paulo's population), was open to sharing birds with me. Song is an obvious candidate as a sexually selected trait in this species because they are otherwise so drab, and preliminary analysis of recordings from Macaulay Library suggested that the songs did fulfill the initial requirements of the performance-based parameters I wanted to measure. Moreover, no one had worked extensively on house wren song in the past, though there is a large body of work on other aspects of house wren biology, which provides solid background information. Though sperm was not an initial interest of mine, house wrens also turn out to be an easy system to work with in sperm; they give samples fairly readily, and the sperm are short enough that they are fairly easy to measure.

Though I did not find strong relationships between form and function in my dissertation research, I still suspect that form and function are related in this system, only I by chance did not chose the aspects of form that match function. To use enzymes as an analogy, I'm not looking at the ligand-binding domain, but perhaps at a trans-membrane portion of the protein, where the fine details of the structure are less crucial than they are in other parts. I find it very interesting that these aspects of song and sperm do not match the conventional predictions, because I suspect that many behavioral ecologists would, like me, have predicted that they are important aspects of function. Finding that they are not acting as predicted opens new questions that will, I hope, further our understanding of sexual selection.

## CHAPTER 1

# VOCAL DEVIATION AND TRILL CONSISTENCY DO NOT AFFECT MALE RESPONSE TO PLAYBACK IN HOUSE WRENS

EMILY CRAMER

### ABSTRACT

Signals that require a high degree of skill to produce are expected to honestly indicate signaler quality. In trilled bird song, two parameters that likely reflect performance difficulty are vocal deviation (how rapidly sound frequency is modulated) and trill consistency (how precisely syllables are repeated). These parameters function as intra- and inter-sexual signals in most bird species tested to date, but they may not adequately capture song performance difficulty in all species. I used two playback protocols to test whether males respond differently to songs that differ in vocal deviation and trill consistency in house wrens (*Troglodytes aedon*). Despite large sample sizes ( $n = 50$  and  $24$  males), male responses did not depend on playback treatment. Males sang each trill type at a range of pitches, and the vocal deviation of the trill depended strongly on the pitch at which it was sung, consistent with models of song production mechanics. I propose that the addition of the pitch covariate may make it computationally intensive to evaluate vocal deviation, limiting the usefulness of this potential signal for this species. Moreover, producing each trill type at a range of pitches may itself serve some communication function, which could override the potential signal value of trill consistency. While vocal deviation and trill consistency are male quality indicators in several species, these results suggest that species-specific

differences in what constitutes a “challenging” song may prevent them from being universally applicable.

## INTRODUCTION

Index and handicap signals can send reliable information about individual quality because of the constraints on or costs of signal production, respectively (Vehrencamp 2000; Bradbury and Vehrencamp 2011). Reliable signaling is important because, if senders routinely produce unreliable signals that manipulate receiver behavior against the receiver’s interests, the receiver should be selected to ignore the signal (Bradbury and Vehrencamp 2000; Searcy and Nowicki 2005). Signals must therefore be reliable on average for a signaling system to persist (Bradbury and Vehrencamp 2000; Searcy and Nowicki 2005).

Signal reliability due to production costs and constraints may be most likely to arise in motor displays, such as vocalizations in birds (Byers et al. 2010). Because producing motor displays involves precise movements and may require a high degree of skill, only healthy individuals with well-constructed neuromuscular systems should be able to achieve a “good” display (Byers et al. 2010). Two aspects of birdsong where skill may be particularly important, and which are therefore candidate index or handicap signals of individual quality, are “vocal deviation” (Podos 1997, 2001, 2009) and the consistency of song or note repetition (Byers 2007; Botero et al. 2009; Sakata and Vehrencamp 2012).

Vocal deviation is a measure of the speed of frequency modulation in a trill, or a series of repeated syllables. In a trill, there is a trade off between the range of sound frequencies covered and the trill rate, or the rate at which syllables are repeated: while it is possible to produce a song with a low bandwidth and a low trill rate, it is not possible to produce a song with a very broad bandwidth and a fast trill rate (Podos 1997, 2002; see Cardoso et al. 2007 for an additional

interpretation). Mechanistically, this relationship is thought to exist because modulating the sound frequency requires the bird to modify tension on the syringeal membranes to change the fundamental frequency (Goller and Suthers 1996) and to alter the volume of the upper vocal tract (including beak gape and the oropharyngeal cavity: Hoese et al. 2000; Reide et al. 2006) so that the sound resonating chamber's dimensions match the fundamental frequency. Because a broader frequency sweep corresponds to a larger magnitude change in the vocal tract, and because the speed of motion is limited, birds cannot simultaneously cover a broad frequency bandwidth and repeat notes at a high rate (Podos 1997, 2009). The maximum (i.e., fastest physically achievable) combination of a fast trill rate and a broad frequency bandwidth can be estimated (Podos 1997), and deviation from this performance maximum is called "vocal deviation" (Podos 2001). Studies in a variety of passerine species support the hypothesis that low vocal deviation indicates higher male quality. Females prefer low deviation songs in laboratory copulation solicitation display assays (Vallet et al. 1998; Drăgănoui et al. 2002; Ballentine et al. 2004; Caro et al. 2010), males that sing lower deviation songs pair earlier in the field (Christensen et al. 2006), and extra-pair sires sing lower deviation songs than the within-pair males they cuckold (Cramer et al. 2011). Moreover, males' vocal deviation capabilities correlate with phenotypic measures of quality in some passerine species (Ballentine 2009; Sockman 2009; though not others, Cardoso et al. 2012). Males either use lower-deviation trill types or decrease the vocal deviation of a given trill type in social contexts where signaling at a high level may be more important (Beebee 2004; Trillo et al. 2005; Kunc et al. 2006; Cardoso et al. 2009; DuBois et al. 2009). Several playback studies to territorial males suggest that low deviation trills simulate high-quality rivals (Illes et al. 2006; Cramer and Price 2007; de Kort et al. 2009; Sewall et al. 2010; DuBois et al. 2011).

The consistency of song repetition has a similar logical mechanical basis, and likely requires a similarly high level of skill to attain. Producing a song requires precise coordination of respiratory, syringeal, and vocal tract muscles, as well as integration across several regions of the brain and across the two sides of the syrinx (Suthers 2001; Jarvis 2004; Sakata and Vehrencamp 2012). To coordinate all these actions precisely enough that a song or trill syllable is always produced in exactly the same way therefore appears difficult (Byers 2007). Though less studied than vocal deviation, several lines of evidence also suggest that trill consistency is an honest indicator of male quality (reviewed in Sakata and Vehrencamp 2012). For instance, males that are extra-pair sires sing with higher trill or whole-song consistency than the within-pair sires they cuckold (Byers 2007; Cramer et al. 2011). Playbacks with different levels of song consistency elicit different male responses. de Kort et al. (2009b) found a less aggressive response to more consistent songs, perhaps indicating that the focal male is more intimidated by a high-quality rival. Conversely, Rivera-Gutierrez et al. (2011) found a more aggressive response to more consistent songs, interpreted as focal males being more motivated to defend against a higher-quality rival. Song or trill consistency correlates with age or male quality in several species (Botero et al. 2009; Wegrzyn et al. 2010).

In addition to vocal deviation and trill consistency, other factors can also affect how challenging a song is to produce (e.g., Forstmeier et al. 2002; Podos et al. 2009; Cardoso and Hu 2011), suggesting that one or a few measures of song performance may not be sufficient for all species. Examining multiple performance measures within a single species could be revealing. I conducted a playback experiment on territorial male house wrens (*Troglodytes aedon*) to simultaneously assess the effects of vocal deviation and trill consistency on male responses to simulated intruders. If vocal deviation and trill consistency are reliable indicators of male quality

in house wrens, as in other species, I expected to find a differential response to playbacks that differed in vocal deviation or trill consistency. Based on previous work (de Kort et al. 2009a) and theoretical expectations (Collins 2004), I predicted that males would respond most strongly to intruders whose songs were of similar vocal deviation and trill consistency to their own songs (though other predictions would be possible: see Searcy and Nowicki 2000, Vehrencamp 2000). Intruders with high quality songs relative to the focal male may present too great a threat for the focal male to attempt an escalated interaction (e.g., Langemann et al. 2000), while intruders with relatively low-quality songs should be easily evicted from the territory and therefore would not require a strong response (as has been found with plumage signals, e.g., Greene et al. 2000). Alternatively, if either vocal deviation or trill consistency is not a male-male signal in house wrens, I expected to find no effect of that parameter on playback responses.

## METHODS

### *Study system and general field methods*

House wrens are a 10-g insectivorous, cavity-nesting species. I conducted this study in two partially wooded sites with artificial nest boxes near Ithaca, New York (lat. 42°31'N, long. 76°28'W). At this site, house wrens are migratory and polygynous, and extra-pair paternity is common (approximately 12-25% of offspring; LaBarbera et al. 2010 and unpublished data). I captured individuals, banded them with a combination of colored leg bands and a US Fish and Wildlife Service band for later identification, and took standard morphological measurements. I then monitored nesting success from April-August 2009 and 2010 (for details on the study sites, see Llambías 2009).

### *Song measurements*

I defined a note as a continuous trace on a spectrogram. A syllable is one or more notes that typically occur together, and a trill is a series of consecutive repetitions of the same syllable (Catchpole and Slater 2008). Each house wren song typically consists of a series of low-amplitude “introductory” syllables followed by one or more trills (e.g., Supplementary Figure S1.1). To control for variability due to the specific structure of different syllables (e.g., some syllables are simple one-note downsweeps while others consist of highly modulated frequencies, which may represent different performance challenges; e.g., see Ballentine et al. 2004; Cardoso et al. 2009; Podos et al. 2009), I visually separated trill syllables into categories (“types”) according to the number of notes and the general shape of those notes, which are stereotyped within and between males. While subjective assessment of trill types is less ideal than an objective, computer-based method, I used trill type only as a covariate to reduce noise in the data set, so mis-categorizations would reduce the power of my analysis but would not generate spurious effects. Of the 4817 trills I measured, 96.4% (4645 trills) were composed of syllables belonging to one of eight types. Analyses of male singing ability were restricted to these eight common trill types.

I noticed and quantified (see below, and supplementary material) a pattern in song structure that appeared to represent an additional performance dimension. Trills of the same type can occur early or late in the song. When they occur earlier in the song, they are typically higher-pitched and cover a broader frequency bandwidth (Sibley 2000; Supplementary Material Figure S.1.2, Table S.1.1). I investigated this relationship further as a potential confound of vocal deviation measurements (see below). This phenomenon also prevented me from using computer-based algorithms to classify trill types, because currently-available programs do not assess similarity of syllables at different absolute pitches and with different frequency bandwidths.

Songs and playback trials (see below) were recorded with a Marantz PMD690 and a Sennheiser ME67 or MKH816 at a 48 kHz sampling rate and a 16-bit depth. Eight of the males whose songs were used to construct playback stimuli were recorded with an M-Audio Microtrack-24/96 recorder at 44.1 kHz and 16-bit depth, but recorder type did not affect song measurements. I isolated individual songs from longer recordings in Syrinx PC (John Burt, [www.syrinxpc.com](http://www.syrinxpc.com)), choosing the first 5 renditions of each novel arrangement of syllable types that was of sufficient quality for measuring. I then measured the frequency bandwidth encompassing 99% of the sound energy in each syllable using a custom plug-in in Raven Pro 1.3 (Bioacoustics Research Program, Cornell Lab of Ornithology; Ithaca, NY; window with 80.1% overlap in the time domain giving 111 sample hop size, 4096 DFT size and 11.7 Hz grid). This plug-in measures from the 0.5<sup>th</sup> percentile to the 99.5<sup>th</sup> percentile frequency limits of sound energy. For each trill, I calculated the mean frequency bandwidth, mean high frequency, mean low frequency, duration of the trill (time from start of the first syllable to the start of the last syllable), and trill rate (one less than the total number of syllables in the trill, divided by trill duration).

Using Raven's batch tools, I stored individual trill syllables as separate files for cross correlation analysis in SoundXT (Cortopassi and Bradbury 2000). All syllables from a single trill were cross-correlated with each other using the following parameters: FFT length 1024; data length 50%; Hann window; 80% overlap; masking method broadband; 50% masking; masking adjustment bias; spec pairwise; correlator type matrix standard method. To minimize interference from background noise that occurred outside the frequency range of the trill, I bounded the cross-correlation at 200 Hz below the lowest frequency and 200 Hz above the highest frequency in the trill. I calculated trill consistency as the average cross-correlation score within each trill.

### *Vocal Deviation Calculation*

I calculated an upper-bound regression following Podos (1997), using a bin size of one syllable/sec and songs recorded in 2009 (using songs from both years to calculate the line does not produce qualitatively different results for any analyses). Because house wren trills clustered into three broad trill rate categories, many bins did not contain any trills with high frequency bandwidths (Figure 1.1). Because the goal of this upper-bound regression is to estimate the maximum combination of frequency bandwidth and trill rate, I did not include measures from bins that lacked high-performance combinations, resulting in the inclusion of only 10 (of 17) bins (Figure 1.1). I calculated vocal deviation as the orthogonal deviation from this line (Podos 1997, 2001). I assigned all trills that fell above the performance maximum a negative vocal deviation score following Ballentine et al. (2004).

### *Playback experiments*

I conducted on-territory playbacks using two different protocols (protocol 1, n = 50 males in 2009; protocol 2, n = 24 males in 2010; see differences below). For both protocols, I exposed each male to playback of three different vocal deviation treatments in random order, balanced across males. Though each male heard all three vocal deviation treatments, he heard only a single trill consistency treatment; half of the males received low consistency stimulus songs and half received high-consistency songs. I audio-recorded each male's vocalizations and spoken observations of his behaviors for a pre-playback period, during playback, and for a post-playback observation period. I noted whether the male was within 5 m of the speaker, using a ring of flagging tape as a reference. I also scored the number of times he flew across the speaker, defined as flights where the bird passed within 2 horizontal m of the speaker and landed on the opposite side of the speaker, regardless of his height above it. From these audio-recordings, I

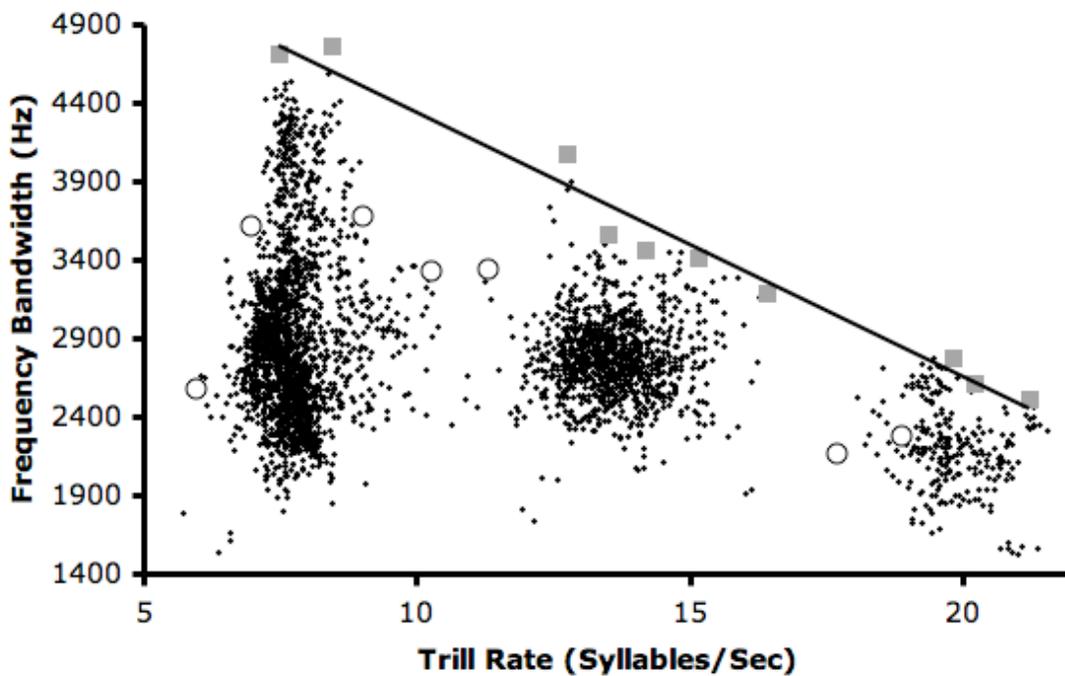


Figure 1.1. Frequency bandwidth as a function of trill rate in house wrens ( $n = 3312$  trills measured in 51 males in 2009; 2010 songs not shown). Large grey squares indicate the points used to calculate the upper-bound regression line estimating the performance limit on trill rate and frequency bandwidth (line); white circles are the data points from bins that lacked high performance trill exemplars, which I excluded from analysis (see main text for details). Black points are songs not used to estimate the performance limit.

extracted the following response measures: song rate (songs/min) during and after playback rounds, the proportion of time spent within 5 m of the speaker during and after playback, the rate of flights across the speaker during and after playback, and the latency for the male to approach within 5 m of the speaker.

To conduct playback trials, I began with a 5-min pre-playback observation period then played a “lure” call, which was a non-song chatter call sometimes given in agonistic situations (e.g., de Kort et al. 2009b). When the male responded to the lure, either by approaching the speaker or abruptly changing his vocal behavior, I began the playback trial. Some males did not respond quickly to the lure; I played it for up to 5 minutes, then allowed a 5-minute pause followed by another 5 minutes of lure. If the male still did not respond, I either did not perform a trial to that male or I changed speaker locations for a second attempt. Lure vocalizations ensure that the male can hear playback from his location and may also reduce variation in male motivation due to the social context immediately preceding playback.

To reduce the possible effects of breeding stage on responses to playback, I conducted most playback trials on males at nest boxes where they were advertising or where the female had not yet begun to incubate. One playback was conducted on a male with a laying mate, and two were done on males with incubating mates. Several others were conducted at a secondary nest box where a male was advertising, while his primary female incubated.

*Protocol 1.*—I played songs from a FoxPro FX5 speaker (with TX-200 transmitter for remote control) between 0515 and 1200 EST, using uncompressed .wav sound files. I set up the speaker approximately 5 m from a nest box where the focal male had been singing that day. For this protocol, one playback trial consisted of three playback treatments, each of 1.5 min of songs (at a rate of 10 songs/minute) and 3.5 min of silence, followed immediately by the next

treatment. An entire trial therefore lasted 15 min. In one trial, the interval between treatments was lengthened due to technical difficulties, but only 3.5 min of post-playback behavior was analyzed.

*Protocol 2.*-- Initial analysis of protocol 1 showed no difference in males' responses to the different treatments. To try to ensure that this negative result was not a byproduct of experimental design, I modified the playback procedure in several ways for protocol 2. Males may be highly aggressive to playback from the center of their territory regardless of the playback stimulus (i.e., a "ceiling effect" on aggression; Stoddard et al. 1991), so for protocol 2 I placed the speaker just outside the area encompassed by his song posts (mean  $\pm$  SE 25 m  $\pm$  1.3 m from the nest box, range: 15-40 m). Playback treatments presented in rapid succession, as in protocol 1, are considered more likely to result in order effects (see Brenowitz 1981). For protocol 2, then, I allowed at least 1 hr to elapse between treatments (range: 60-125 minutes mean: 67  $\pm$  1.87 min). I did not present treatments on different days, to avoid having the male's breeding stage advance or the neighborhood composition change (due to newly arriving males), which could have affected males' responses.

Protocol 1 used only short playback and post-playback periods. A longer playback could increase males' differentiation of playback songs, and a longer post-playback period could better reveal differences in male response. I therefore increased the playback duration to 2.5 min of song (10 songs/min) and the post-playback observation period to 12.5 minutes for each treatment in protocol 2.

I used a Nagra DH speaker and amplifier, which reproduces frequencies more faithfully than the FoxPro speaker, connected to an iPod via approximately 15 m of cable (Canare Cable 705, L-4E6S). All stimuli were stored as uncompressed .wav files. The initial playback round

began between 0530 and 1200; all playback treatments were finished before 1440. Fourteen of the 24 subjects in protocol 2 were also subjects in protocol 1. However, I treated these males as independent data points because the two protocols were each used in a different year and analyses were conducted on each year separately.

### *Playback stimuli*

I manipulated single “source” songs in six ways to create stimulus sets consisting of all three vocal deviation levels at both consistency levels. I maintained the natural sequence of note types and the approximate number of notes of each type. All stimuli were high-pass filtered at 1000 Hz to eliminate background noise; manipulations were performed in Syrinx PC.

To change the vocal deviation of the songs, I found notes of the same trill type sung by the same male but with different frequency bandwidths. I then pasted relatively broadband notes together with short inter-note intervals to create low deviation songs; I pasted low bandwidth notes together with longer inter-note intervals to create high deviation songs; and I pasted mean-bandwidth notes together with mean-duration intervals for medium deviation stimuli. Values for the bandwidth and trill rate of high, medium, and low vocal deviation songs were chosen based on the distribution of vocal deviation observed in the population: low and high deviation songs were at the extremes of the natural range (see Supplementary Materials, Table S4). For protocol 1, I chose songs with low-deviation (i.e., physically challenging) syllable types and manipulated only these types, assuming that they would be the most likely trills to reveal a vocal deviation effect. Although previous studies have used a similar approach (e.g., de Kort et al. 2009b), for protocol 2, I manipulated all of the trill types in a song.

To create the different consistency treatments, I either pasted together many different renditions of the same trill syllable (producing a low consistency song), or pasted together the

same trill syllable multiple times (high consistency). I matched the mean bandwidth of syllables for the inconsistent stimulus as closely as possible with the bandwidth of the note used for the consistent stimulus, so that songs would have the same vocal deviation.

From each set of six modified songs, I constructed playback stimuli for two focal males: a high consistency stimulus subset for one male, and a low consistency subset for the other, each with all three vocal deviation levels in the same order. Vocal deviation is therefore a within-male comparison and trill consistency a between-male comparison, but I treated trill consistency trials as paired in analyses because they were derived from the same source song. I used paired high- and low-consistency stimuli on different days and on males that were out of presumed hearing distance from each other. To offset possible time-of-day effects, I began playback of paired stimuli within 90 minutes of each other. Each playback stimulus set was used for only one focal male, and each males' songs were used to create only one stimulus set per experimental protocol.

### *Statistical analyses*

To assess the possible covariation between pitch and vocal deviation, I constructed a general linear model with trill type, pitch (measured as mean high frequency), and their interaction as predictors, and vocal deviation as the response variable. Because the interaction term was highly statistically significant, I constructed individual models for each of the eight common trill types, using only pitch as a predictor. For subsequent analyses, I used the residual of these separate vocal deviation-pitch relationships as the measure of vocal deviation. My rationale for using this corrected measure is that vocal deviation was strongly negatively correlated with pitch, which depended mostly on whether the trill was sung early or late in the song (Supplementary Material Figure S.1.2). Results are qualitatively the same using uncorrected

vocal deviation (not shown), or using alternative measures of pitch (Supplementary Material Table S.1.1).

To test for the effect of playback treatments, I constructed separate models for each response variable, testing whether response to playback depended on playback treatment (high, medium, or low vocal deviation; high or low trill consistency). I analyzed protocols 1 and 2 separately, and included both vocal deviation (a within-male effect) and trill consistency (a between-male effect) in each model. Male identity was nested within playback stimulus set, and both male and stimulus set were random effects in all models. For all response variables, I included mean responses during playback presentation and during the post-playback observation, with a “phase” term denoting playback vs. post-playback. I also included a categorical “order” term in models. I investigated alternate methods for expressing the treatment terms (i.e., predicting playback response as a function of the measurements of the stimulus song, or as a function of the relative measure of the stimulus song vs. the focal male’s own songs; see Supplementary Table S2). These models gave qualitatively similar results, as did models including other potential covariates (e.g., male body size, female presence, and date; not shown).

Residuals for song rate were normally distributed. Proportion of time within 5 m of the speaker and rate of flights across the speaker both had many zero values, so I divided each of these variables into two analyses. I analyzed whether the bird approached within 5 m of the speaker and whether the bird flew across the speaker, assuming a binary error distribution and a logit link function. Among the birds that did approach to within 5 m of the speaker, the proportion of time spent within 5 m of the speaker was uniformly distributed, and parametric statistics are robust to this violation of assumptions (Sokal and Rohlf 1995). I therefore analyzed the proportion of time within 5 m with parametric statistics. Among the birds that did fly across

the speaker, the distribution of flight count was still strongly skewed, and this response was analyzed with a negative binomial distribution.

For consistency and for vocal deviation, I performed separate Kaplan-Meier survival analysis in SPSS 19 to determine whether the latency to approach to within 5 m of the speaker differed between treatments (Botero and Vehrencamp 2007). This test is useful for data such as approach latency, where many individuals did not reach the critical distance from the speaker and therefore an ordinary paired test would be biased (Botero and Vehrencamp 2007). This test does not allow for within-individual comparisons, so each playback trial was treated as an independent event.

I corrected for multiple testing by running false discovery rate (FDR) correction in R version 2.9.2 (R [Development](#) Core Team 2009; Benjamini and Hochberg 1995) on all the effect tests. All other statistics were conducted in SAS 9.2.

## RESULTS

### *Vocal deviation in house wrens*

As in Podos (1997), plotting frequency bandwidth as a function of trill rate created a triangular distribution of trills (Figure 1.1), and the upper bound regression was statistically significant ( $r^2 = 0.97$ ,  $F_{1,8} = 293.20$ ,  $p < 0.0001$ ; prediction expression used in calculating vocal deviation: frequency bandwidth =  $-168.50 * \text{trill rate} + 6019$  Hz, grey points in Figure 1.1). This line excluded several bins that did not contain a large number of trills, and therefore did not contain high-performance examples. However, a line including those bins was still statistically significant ( $r^2 = 0.41$ ,  $F_{1,15} = 10.44$ ,  $p = 0.006$ , prediction expression: frequency bandwidth =  $-97.84 * \text{trill rate} + 4614.42$ , grey and white points in Figure 1.1).

*Pitch-vocal deviation relationship:*

Vocal deviation was significantly correlated with pitch: 93% of the variation in vocal deviation across 4645 trills was explained by a model including only the trill type, pitch, and their interaction ( $r^2 = 0.93$ ,  $F_{15,4629} = 3898.85$ ,  $p < 0.0001$ ; each predictor,  $p < 0.0001$ , Figure 1.2). Higher pitches corresponded to lower vocal deviations, a pattern that was driven primarily by the effect of pitch on frequency bandwidth: when I measured the effect of pitch on trill rate and frequency bandwidth in separate models for each trill type, the correlation was always stronger, and the effect more significant, on frequency bandwidth than on trill rate (Supplementary Material Table S.1.1). Trills sung at higher pitches had broader frequency bandwidths. Results are similar when I estimated pitch as the mean low, rather than high, frequency in the trill (Table S.1.1), because the high and low frequencies in the trill are tightly correlated (separate models for each trill type predicting high frequency as a function of low frequency, all  $r^2 > 0.45$ , all  $p < 0.0001$ ). The frequency bandwidth increases with increasing high frequency not by mathematical necessity, but because the highest frequency in the trill rises faster than the low frequency; that is, on average, a 1-kHz increase in the low frequency of a trill corresponded to a 1.3 kHz increase in the high frequency (modeled across all trill types).

*Playback responses:*

Song rate, flight rate, and proportion of time within 5 m of the speaker did not differ significantly across playback treatments after correction for multiple testing (Figure 1.3, Table 1.1), though for both experimental protocols, males tended to fly across the speaker and approached less strongly to the high-consistency and low-deviation treatments (Figure 1.3). Males approached to within 5 m of the speaker equally quickly regardless of the playback treatments (Figure 1.4, vocal deviation protocol 1, Log-rank test  $\chi^2_2 = 0.73$ ,  $p = 0.70$ ; vocal

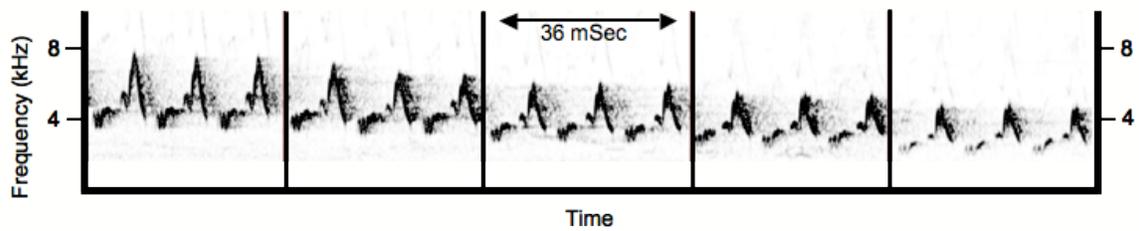


Figure 1.2. Frequency bandwidth is higher for trills with higher pitches. This spectrogram is concatenated from 5 songs sung by the same male, ordered according to pitch. Higher pitched, and thus broader bandwidth, trills are generally positioned earlier in the song, as illustrated in this concatenation (see supplementary material Figures S.1.1, S.1.2, Table S.1.1).

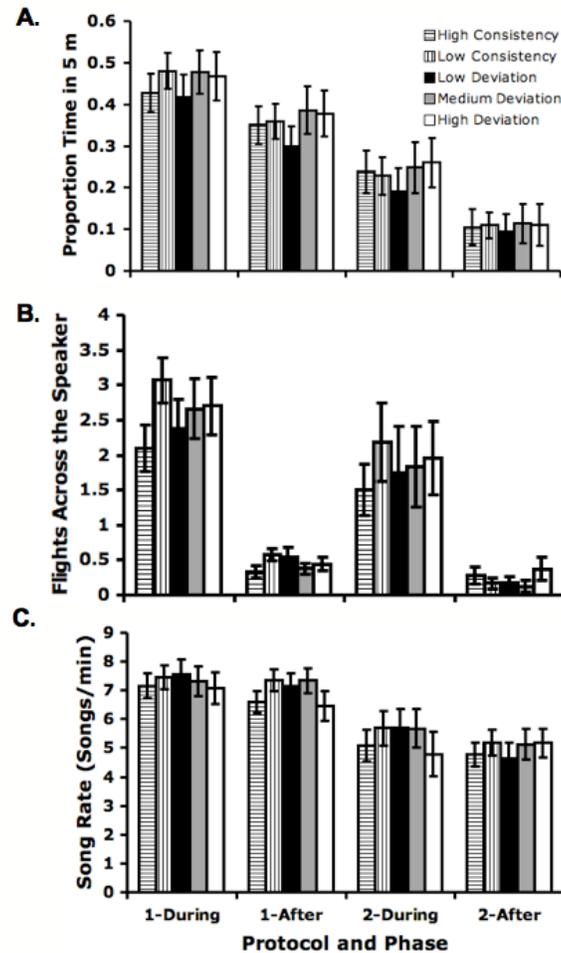


Figure 1.3. Responses to playback as a function of playback treatment, playback phase (during vs. after), and protocol. A) Mean proportion of time spent within 5 m of the speaker B) Mean number of flights across the speaker C) Mean song rate. Note that proportion of time within 5 m of the speaker and number of flights across the speaker each had a large number of zero values, so statistical tests were in two parts (binary response, and a continuous response from among the responders; see main text for details.) Protocol 1: n = 300 treatment-phases, 50 males; Protocol 2: n = 144 treatment-phases, 24 males. High and low consistency treatments are horizontal and vertical lines, respectively; low, medium, and high deviation treatments are black, grey, and white. Comparisons were among consistency treatments, and among deviation treatments; I found no significant effect of playback treatments.

Table 1.1. Least squares means estimates ( $\pm$  SE) of birds' responses to playback as a function of categorical playback treatment (mean for normal data, mean log odds for binary responses, and mean the negative binomial distribution for count data). Effects that remained significant after correcting for multiple testing are in bold.

Protocol	Response variable	Predictor variable	Performance Difficulty of Stimulus Treatment			F <sub>df</sub> (p)
			High	Med.	Low	
1	Approach (binary)	Voc Dev	1.09 ± 0.40	1.14 ± 0.40	1.36 ± 0.41	F <sub>2,293</sub> = 0.25 (0.78)
		Consist	0.91 ± 0.46		1.48 ± 0.48	F <sub>1,41.92</sub> = 0.73 (0.40)
		Phase Order				F <sub>1,293</sub> = 2.53 (0.11) F <sub>2,293</sub> = 0.48 (0.62)
	Approach (normal)	Voc Dev	0.48 ± 0.05	0.6 ± 0.05	0.57 ± 0.05	F <sub>2,164.8</sub> = 3.83 (0.02)
		Consist	0.56 ± 0.05		0.54 ± 0.05	F <sub>1,24</sub> = 0.12 (0.74)
		Phase Order				<b>F<sub>1,164</sub> = 7.88 (0.006)</b> F <sub>2,164.7</sub> = 3.89 (0.02)
	Flights (binary)	Voc Dev	-0.13 ± 0.34	0.03 ± 0.34	0.12 ± 0.34	F <sub>2,293</sub> = 0.26 (0.77)
		Consist	-0.38 ± 0.38		0.4 ± 0.38	F <sub>1,37.26</sub> = 2.08 (0.16)
		Phase Order				<b>F<sub>1,293</sub> = 36.59 (0.0001)</b> F <sub>2,293</sub> = 2.56 (0.08)
	Flights (count)	Voc Dev	0.66 ± 0.13	0.63 ± 0.12	0.66 ± 0.12	F <sub>2,86.78</sub> = 0.03 (0.97)
		Consist	0.52 ± 0.13		0.78 ± 0.12	F <sub>1,19.26</sub> = 3.24 (0.09)
		Phase Order				<b>F<sub>1,144</sub> = 75.55 (0.0001)</b> F <sub>2,86.41</sub> = 0.05 (0.95)
	Song	Voc Dev	7.34 ± 0.46	7.31 ± 0.46	6.79 ± 0.46	F <sub>2,245</sub> = 1.37 (0.26)
		Consist	6.88 ± 0.51		7.42 ± 0.51	F <sub>1,24</sub> = 0.76 (0.39)
		Phase Order				F <sub>1,245</sub> = 1.19 (0.28) F <sub>2,245</sub> = 0.97 (0.38)

Table 1.1 Continued

Approach (binary)	Voc Dev	-0.53 ± 0.63	-0.32 ± 0.62	-0.02 ± 0.62	F <sub>2,137</sub> = 0.43 (0.65)
	Consist	-0.39 ± 0.74		-0.19 ± 0.76	F <sub>1,15.81</sub> = 0.04 (0.85)
	<b>Phase</b>				<b>F<sub>1,137</sub> = 11.16 (0.001)</b>
	Order				F <sub>2,137</sub> = 0.88 (0.42)
Approach (normal)	Voc Dev	0.31 ± 0.07	0.32 ± 0.07	0.33 ± 0.07	F <sub>2,46.14</sub> = 0.05 (0.95)
	Consist	0.34 ± 0.08		0.31 ± 0.08	F <sub>1,4.64</sub> = 0.11 (0.76)
	<b>Phase</b>				<b>F<sub>1,45.87</sub> = 14.29 (0.0005)</b>
	Order				F <sub>2,45.93</sub> = 4.26 (0.02)
2 Flights (binary)	Voc Dev	-1.15 ± 0.6	-1.63 ± 0.62	-0.66 ± 0.59	F <sub>2,137</sub> = 1.43 (0.24)
	Consist	-1.04 ± 0.65		-1.25 ± 0.66	F <sub>1,9.34</sub> = 0.07 (0.80)
	<b>Phase</b>				<b>F<sub>1,137</sub> = 23.07 (0.0001)</b>
	Order				F <sub>2,137</sub> = 1.12 (0.33)
Flights (count)	Voc Dev	0.43 ± 0.25	0.55 ± 0.25	0.37 ± 0.23	F <sub>2,41</sub> = 0.31 (0.74)
	Consist	0.24 ± 0.27		0.65 ± 0.27	F <sub>1,16.65</sub> = 1.45 (0.24)
	<b>Phase</b>				<b>F<sub>1,41</sub> = 19.31 (0.0001)</b>
	Order				F <sub>2,41</sub> = 3.8 (0.03)
Song	Voc Dev	5.18 ± 0.63	5.4 ± 0.63	4.98 ± 0.63	F <sub>2,115</sub> = 0.51 (0.60)
	Consist	4.93 ± 0.7		5.44 ± 0.7	F <sub>1,11</sub> = 0.46 (0.51)
	<b>Phase</b>				F <sub>1,115</sub> = 1.38 (0.24)
	Order				F <sub>2,115</sub> = 3.21 (0.04)

Note that for vocal deviation, the high performance challenge stimulus corresponds to a low residual vocal deviation. Least squares means (LSM) estimates are from models including a random effect of male identity and fixed effects of playback phase (during vs. after playback) and playback order, and they reflect the mean log odds of approaching to within 5 m of the speaker and of flying across the speaker at least once for binary responses. They reflect the scaling factor of the negative binomial distribution for the count data.

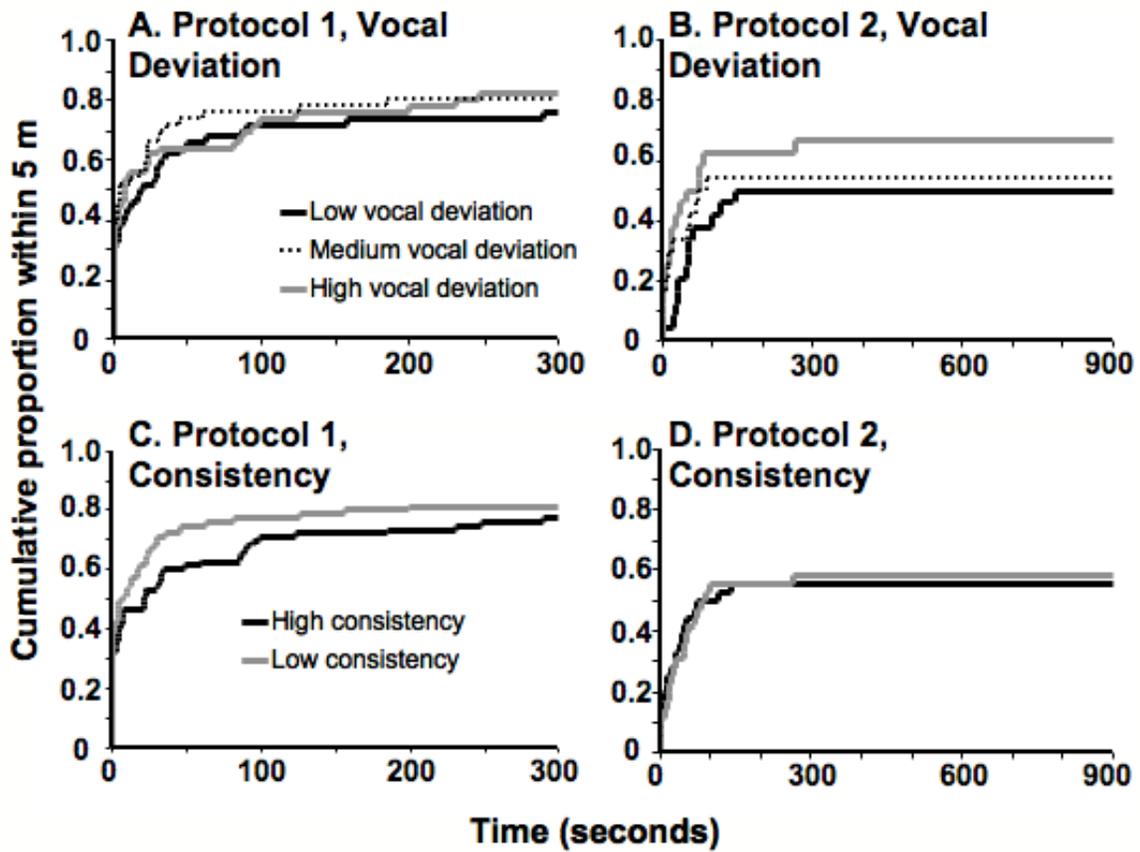


Figure 1.4. Latency to approach within 5 m of the speaker did not differ with playback treatments. Kaplan-Meier survival curves showing cumulative proportion of individuals that had approached to within 5 m of the speaker for the different vocal deviation levels (A, protocol 1, and B, protocol 2) and different consistency levels (C, protocol 1, and D, protocol 2).

deviation protocol 2: Log-rank test,  $\chi^2_2 = 2.23$ ,  $p = 0.33$ ; Consistency, protocol 1: Log-rank test  $\chi^2_1 = 1.29$ ,  $p = 0.26$ ; Consistency, protocol 2: log-rank test,  $\chi^2_1 = 0.00$ ,  $p = 0.98$ ). I found no evidence for an interaction between vocal deviation and trill consistency treatments (not shown). Analyses that excluded individuals that came within 5 m of the speaker before the playback began or that excluded individuals that perched on their nest boxes during playback did not reveal different results (not shown).

## DISCUSSION

Male house wrens did not respond differently to simulated territorial intruders singing trills with different vocal deviation and trill consistency scores. This study used large sample sizes and revealed no strong trends in male response, suggesting that this result is robust and not a consequence of low statistical power. Previous playbacks on males that made large-scale manipulations on vocal deviation and trill consistency typically found that males respond about half as strongly to one playback treatment compared to the other (Cramer and Price 2007; de Kort et al. 2009a,b; Sewall et al. 2010; DuBois et al. 2011). While several comparisons between treatments in this study were significant before correction for multiple testing, the magnitude of the differences in males' responses to different playbacks was quite small, suggesting that it is not biologically relevant (Figure 1.3). To the best of my knowledge, no published study that has made large modifications to song acoustic structure has failed to find an effect of vocal deviation and trill consistency on receiver responses. These results are therefore novel and suggest that, despite their logical mechanical explanations and support in many species so far, these signal components are not universal indicators of male quality in a male-male communication context. I am currently testing whether these signal elements may have a role in male-female

communication in house wrens (unpublished), as females are more discriminating of song quality than males are in some species (e.g., Danner et al. 2011).

The structure of house wren syllables and songs may make vocal deviation and trill consistency imperfect measures of performance difficulty. Many house wren syllables (e.g., Figure 1.2) involve both up- and down-sweeps in frequency, such that the syllable begins and ends at approximately the same frequency. Syllables with this pattern may be easier to perform with low vocal deviation songs compared to trills involving uni-directional frequency sweeps: for unidirectional sweeps, the bird must rapidly re-configure its vocal tract to be at the correct frequency for the start of the next syllable (Podos et al. 2009). Moreover, some house wren syllable types involve multiple changes in direction of frequency, so that the total frequency modulation in the note is only poorly captured by the difference between high and low frequencies.

The fact that house wrens produce the same or similar trill types at a range of pitches may negate the signal value of both trill consistency and vocal deviation. For trill consistency, selection to be able to produce syllables at a range of pitches may be stronger than selective pressure to repeat notes consistently. That is, producing a syllable at a range of pitches may be critically important to house wren communication, but necessarily creates inconsistency in syllable structure as I measured it. While most variation in pitch occurs between rather than within trills, consistency in general may not be under strong selection in house wrens. For vocal deviation, the strong correlation between pitch and frequency bandwidth might make assessing vocal deviation computationally intensive: a listening bird would have to simultaneously assess trill rate, frequency bandwidth, and pitch to derive a meaningful measure of the singer's ability.

Both sound perception and production occur on non-linear scales (Hurly et al. 1992, Goller and Suthers 1996, Cynx 2004, Fletcher et al. 2006). While all studies of vocal deviation to my knowledge, including this one, compute frequency bandwidth as the difference between high and low frequencies (“additive” frequency bandwidth), the ratio of the two may be a more relevant measure for production difficulty, and for perception (e.g., Hurly et al. 1992, Cynx 2004). From a production perspective, the activity level of the syringeal muscle that controls fundamental frequency increases exponentially, not linearly, with increasing frequency (Goller and Suthers 1996). “Downstream” from the syrinx, the three-dimensional volume of the upper vocal tract is non-linearly related to its resonance properties, such that a given change in volume has a larger effect on resonant frequency when that change occurs at a high pitch (Fletcher et al. 2006). Singing the same bandwidth at high and low pitches may therefore represent different performance challenges, and thus may have different value in conveying information about male quality. If the difficulty of producing a certain frequency bandwidth does indeed depend on the pitch of the notes, the interaction between absolute pitch and vocal deviation may also represent a novel dimension of performance challenge.

I did not specifically design my playback experiment to control for or assess differences in response to trills sung at different pitches. However, the high performance playback stimuli (playbacks with low vocal deviation) also were high performance with respect to pitch (i.e. lower vocal deviation than expected given the pitch at which they were sung; Supplementary Table S.1.3). The playback stimuli therefore appear completely adequate to test for a difference in response to stimuli with different vocal deviations based on additive frequency bandwidth. Unfortunately, I did not consider the frequency ratio as I was creating playback stimuli (an anonymous reviewer pointed out the logic and perceptual importance of the frequency ratio), and

the playback stimuli did not cover the full natural range of frequency ratios (Supplementary Table S.1.3). I re-analyzed playback responses using frequency ratio and a “multiplicative” vocal deviation based on frequency ratio (see Supplementary Table S.1.2), and found weak trends for males to respond more strongly to stimuli with higher multiplicative vocal deviation (i.e., “easier” songs; Supplementary Table S.1.2) for three of ten statistical tests. It is unclear whether these trends occur by chance, or whether they reflect a potentially important signal that my playback stimuli were insufficient to test. Future studies of vocal deviation should consider absolute pitch and frequency ratio to determine whether additive frequency bandwidth or frequency ratio is more important for signaling.

I conducted the second playback protocol to eliminate possible “ceiling effects” due to proximity to the nest. Some previous studies suggest that playbacks conducted where males are most motivated to defend (i.e., the center of a territory) elicit highly aggressive responses regardless of the playback stimuli, while playbacks at the edge of a territory are more likely to elicit differential responses (Stoddard et al. 1991). The fact that house wrens responded less strongly but still did not discriminate among treatments in the second protocol, combined with a number of studies that do find discrimination among stimuli presented at the center of the territory (e.g., Brenowitz 1981; Falls and Brooks 1975; Brunton et al. 2008), suggests that my results are not a consequence of a ceiling effect and that I would have detected differential responses if they were present.

House wrens are a notoriously aggressive species (e.g., Belles-Isles and Picman 1987). I have not observed many close territorial encounters, but anecdotally, territorial fights begin with the males approaching a territorial border and singing; males then begin to chase each other, pausing and singing between chases. Finally, they may engage in physical combat; the only

fighting pair I observed ceased singing when they began to physically fight. While song seems important in the early stages of the fight, males appear to vocalize less frequently as the fight becomes more intense, perhaps suggesting that other measures of male quality (e.g., flying speed and maneuverability, body size) are more important than vocal ability in mediating territorial aggression.

In conclusion, I find no evidence that house wren males attend to differences in vocal deviation or in trill consistency when assessing territorial rivals. The complexities of house wren trills, notably the way that house wrens sing individual trill types at a range of pitches, may detract from the signal value of single performance measurements such as vocal deviation and trill consistency.

#### ACKNOWLEDGEMENTS AND FUNDING

This work was supported by a National Science Foundation Graduate Research Fellowship, the Animal Behavior Society, the American Ornithologists' Union, the Cornell University Department of Neurobiology and Behavior, Cornell Sigma Xi, and the Cornell Lab of Ornithology and its donors.

Thanks to Sandra Vehrencamp, Emma Greig, Jenelle Dowling, Petra Deane, Sara deLeon, and Julian Kapoor for comments on the manuscript; to Cornell Research Ponds for access to the field site; to the Cornell Statistical Consulting Unit for statistical advice; and to Paulo Llambias and Taza Schaming for sharing the study system.

## REFERENCES

- Ballentine B. 2009. The ability to perform physically challenging songs predicts age and size in male swamp sparrows, *Melospiza georgiana*. *Anim Behav.* 77:973-978.
- Ballentine B, Hyman J, Nowicki S. 2004. Vocal performance influences female response to male bird song: an experimental test. *Behav Ecol.* 15:163-168.
- Beebe MD. 2004. Variation in vocal performance in the songs of a wood-warbler: evidence for the function of distinct singing modes. *Ethology.* 110:531-542.
- Belles-Isles J-C, Picman J. 1987. Suspected adult intraspecific killing by house wrens. *Wilson Bull.* 99:497-498.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B.* 57:289-300.
- Botero C, Vehrencamp SL. 2007. Responses of male tropical mockingbirds (*Mimus gilvus*) to variation in within-song and between-song versatility. *Auk.* 124:185-196.
- Botero CA, Rossman RJ, Caro LM, Stenzler LM, Lovette IJ, de Kort SR, Vehrencamp SL. 2009. Performance variability is related to age, dominance, and reproductive success in the tropical mockingbird. *Anim Behav.* 77:701-706.
- Bradbury JW, Vehrencamp SL. 2000. Economic models of animal communication. *Anim Behav.* 59:259-268.
- Bradbury JW, Vehrencamp SL. 2011. *Principles of Animal Communication*, 2nd Ed. Sunderland, MA: Sinauer Associates.
- Brenowitz EA. 1981. The effect of stimulus presentation sequence on the response of red-winged blackbirds in playback studies. *Auk.* 98:355-360.
- Brunton DH, Evans B, Cope T, Ji W. 2008. A test of the dear enemy hypothesis in female New Zealand bellbirds (*Anthornis melanura*): female neighbors as threats. *Behav Ecol.* 19:791-798.
- Byers BE. 2007. Extrapair paternity in chestnut-sided warblers is correlated with consistent vocal performance. *Behav Ecol.* 18:130-136.

- Byers J, Hebets EA, Podos J. 2010. Female mate choice based upon male motor performance. *Anim Behav.* 79:771-778.
- Cardoso GC, Atwell JW, Ketterson ED. 2007. Inferring performance in the songs of dark-eyed juncos (*Junco hyemalis*). *Behav Ecol.* 18:1051-1057.
- Cardoso GC, Atwell JW, Ketterson ED, Price TD. 2009. Song types, song performance, and the use of repertoires in dark-eyed juncos (*Junco hyemalis*). *Behav Ecol.* 20:901-907.
- Cardoso GC, Atwell JW, Hu Y, Ketterson ED, Price TD. 2012. No correlation between three selected trade-offs in birdsong performance and male quality for a species with song repertoires. *Ethology.* 118:584-593.
- Cardoso GC, Hu Y. 2011. Birdsong performance and the evolution of simple (rather than elaborate) sexual signals. *Am Nat.* 178:679-686.
- Caro SP, Sewall KB, Salvante KG, Sockman KW. 2010. Female Lincoln's sparrows modulate their behavior in response to variation in male song quality. *Behav Ecol.* 21: 562-569.
- Catchpole CK, Slater PJB. 2008. *Birdsong: biological themes and variations*, 2nd ed. New York: Cambridge University Press.
- Christensen R, Kleindorfer S, Robertson J. 2006. Song is a reliable signal of bill morphology in Darwin's small tree finch *Camarhynchus parvulus*, and vocal performance predicts male pairing success. *J Avian Biol.* 37:6:617-624.
- Collins S. 2004. Vocal fighting and flirting: the functions of birdsong. In: Marler P, Slabbekoorn H, editors. *Nature's Music: the science of birdsong*. Boston: Elsevier Academic Press. p. 39-79.
- Cortopassi KA, Bradbury JW. 2000. The comparison of harmonically rich sounds using spectrographic cross-correlation and principal coordinates analysis. *Bioacoustics.* 11:89-127.
- Cramer ERA, Hall ML, deKort SR, Lovette IJ, Vehrencamp SL. 2011. Infrequent extra-pair paternity in the banded wren, a synchronously breeding tropical passerine. *Condor.* 113:637-645.
- Cramer ERA, Price JJ. 2007. Red-winged blackbirds *Agelaius phoeniceus* respond differently to song types with different performance levels. *J Avian Biol.* 38:122-127.

- Cynx J. 2004. Are songbirds Pythagoreans? Absolute and relative pitch perception. In: Marler PR, Slabbekoorn H, editors. *Nature's Music: The Science of Birdsong*. San Diego, California: Elsevier. p. 218.
- Danner JE, Danner RM, Bonier F, Martin PR, Small TW, Moore IT. 2011. Female, but not male, tropical sparrows respond more strongly to the local song dialect: implications for population divergence. *Am Nat.* 178:53-63.
- deKort SR, Eldermire ERB, Valderrama S, Botero CA, Vehrencamp SL. 2009a. Trill consistency is an age-related assessment signal in banded wrens. *Proc R Soc Lond B.* 269:2525-2531.
- deKort SR, Eldermire ERB, Cramer ERA, Vehrencamp SL. 2009b. The deterrent effect of bird song in territory defense. *Behav Ecol.* 20:200-206.
- Drăgănoiu TI, Nagle L, Kreutzer M. 2002. Directional female preference for an exaggerated male trait in canary (*Serinus canaria*) song. *Proc R Soc Lond B.* 269:2525-2531.
- DuBois AL, Nowicki S, Searcy WA. 2009. Swamp sparrows modulate vocal performance in an aggressive context. *Biol Lett.* 5:163-165.
- DuBois AL, Nowicki S, Searcy WA. 2011. Discrimination of vocal performance by male swamp sparrows. *Behav Ecol Sociobiol.* 65:717-726.
- Falls JB, Brooks RJ. 1975. Individual recognition by song in white-throated sparrows. II. Effects of location. *Can J Zool.* 53:1412-1420.
- Fletcher F, Riede T, Suthers RA. 2006. Model for vocalization by a bird with distensible vocal cavity and open beak. *J Acoust Soc Am.* 119:1005-1011.
- Goller F, Suthers RA. 1996. Role of syringeal muscles in controlling the phonology of bird song. *J Neurophysiol.* 76:287-300.
- Greene E, Lyon BE, Muehter VR, Ratcliffe LM, Oliver SJ, Boag PT. 2000. Disruptive sexual selection for plumage coloration in a passerine bird. *Nature.* 407:1000-1003.
- Hall ML, Illes AE, Vehrencamp SL. 2006. Overlapping signals in banded wrens: long-term effects of prior experience on males and females. *Behav Ecol.* 17:260-269.

- Hoese WJ, Podos J, Boetticher NC, Nowicki S. 2000. Vocal tract function in birdsong production: experimental manipulation of beak movements. *J Exp Biol.* 203:1845-1855.
- Hurly TA, Ratcliffe L, Weary DM, Weisman R. 1992. White-throated sparrows (*Zonotrichia albicollis*) can perceive pitch change in conspecific song by using the frequency ratio independent of the frequency difference. *J Comp Psych.* 106:388-391.
- Illes AE, Hall ML, Vehrencamp SL. 2006. Vocal performance influences male receiver response in the banded wren. *Proc R Soc Lond B.* 273:1907-1912.
- Jarvis ED. 2004. Brains and birdsong. In: Marler P, Slabbekoorn H, editors. *Nature's Music: the science of birdsong.* Boston: Elsevier Academic Press. p. 226-271.
- Kunc HP, Amrhein V, Naguib M. 2006. Vocal interactions in nightingales, *Luscinia megarhynchos*: more aggressive males have higher pairing success. *Anim Behav.* 72:25-30.
- LaBarbera K, Llambias PE, Cramer ERA, Schaming TD, Lovette IJ. 2010. Synchrony does not explain extrapair paternity rate variation in northern or southern house wrens. *Behav Ecol.* 21:773-780.
- Langemann U, Tavares JP, Peake TM, McGregor PK. 2000. Response of great tits to escalating patterns of playback. *Behaviour.* 137:451-471.
- Llambias PE. 2009. Why monogamy? Comparing house wren social mating systems in two hemispheres. [dissertation]. Ithaca, NY: Cornell University; 119 p.
- Logue DM, Forstmeier W. 2008. Constrained performance in a communication network: implications for the function of song type matching and for the evolution of multiple ornaments. *Am Nat.* 172:34-41.
- Podos J. 1997. A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution.* 51:537-551.
- Podos J. 2001. Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature.* 409:185-188.
- Podos J, Lahti DC, Moseley DL. 2009. Sensorimotor learning in songbirds. *Adv Stud Behav.* 40:159-195.

Riede T, Suthers R, Fletcher NH, Blevins WE. 2006. Songbirds tune their vocal tract to the fundamental frequency of their song. PNAS. 103:5543-5548.

Rivera-Gutierrez HF, Pinxten R, Eens M. 2011. Songs differing in consistency elicit differential aggressive response in territorial birds. Biol Lett. 7:339-342.

Sakata JT, Vehrencamp SL. 2012. Integrating perspectives on vocal performance and consistency. J Exp Biol. 215:201-209.

Searcy WA, Nowicki S. 2000. Male-male competition and female choice in the evolution of vocal signaling. In: Espmark Y, Amundsen T, Rosenqvist G, editors. Animal Signals: Signalling and Signal Design in Animal Communication. Trondheim, Norway: Tapir Academic Press. p. 301-315.

Searcy WA, Nowicki S. 2005. The Evolution of Animal Communication. Princeton, NJ: Princeton University Press.

Sewall KB, Dankoski EC, Sockman KW. 2010. Song environment affects singing effort and vasotocin immunoreactivity in the forebrain of male Lincoln's sparrows. Horm Behav. 58:544-553.

Sibley DA. 2000. The Sibley Guide to Birds. New York: Knopf.

Sockman KW. 2009. Annual variation in vocal performance and its relationship with bill morphology in Lincoln's sparrows, *Melospiza lincolnii*. Anim Behav. 77:663-671.

Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research. New York: Freeman.

Stoddard PK, Beecher MD, Horning CL, Campbell SE. 1991. Recognition of individual neighbors by song in the song sparrow, a species with song repertoires. Behav Ecol Sociobiol. 29:211-215.

Suthers RA. 2001. Peripheral vocal mechanisms in birds: are songbirds special? Neth J Zool. 51:217-242.

Trillo PA, Vehrencamp SL. 2005. Song types and their structural features are associated with specific contexts in the banded wren. Anim Behav. 70:921-935.

Vallet E, Beme I, Kreutzer M. 1998. Two-note syllables in canary song elicit high levels of sexual display. *Anim Behav.* 55:291-297.

Vehrencamp SL. 2000. Handicap, index, and conventional signal elements of bird song. In: Espmark Y, Amundsen T, Rosenqvist G, editors. *Animal Signals: Signalling and Signal Design in Animal Communication*. Trondheim, Norway: Tapir Academic Press. p. 277-300.

Wegrzyn E, Leniowski K, Osiejuk TS. 2010. Whistle duration and consistency reflect philopatry and harem size in great reed warblers. *Anim Behav.* 79:1363-1372.

## APPENDIX 1.

### SUPPLEMENTARY MATERIALS TO CHAPTER 1

#### Supplementary material 1. Pitch-trill placement relationship

I defined pitch as the mean of the high frequency for all syllables within a trill, using measurement procedures in the main text. For eight males and 809 trills of the eight common trill types, I categorized the trill as: the first trill in the song following the introductory notes, a middle trill, or the last trill in the song (Figure S.1.1). I then tested for an effect of trill order on trill pitch by constructing a model with male identity as a random effect and trill type and trill position as predictors. The random effect of male identity explained 10% of the variation in the model, but trill type and trill position also explained significant portions of the variation (whole model  $r^2 = 0.65$ ; trill type  $F_{7,795.2} = 30.13$ ,  $p < 0.0001$ ; trill position  $F_{2,794.1} = 385.84$ ,  $p < 0.0001$ ; Figure S.1.2). The effect was similarly strong when pitch was expressed as the mean low frequency in the trill ( $r^2 = 0.69$ , trill type  $F_{7,795.2} = 57.45$ ,  $p < 0.0001$ ; trill position  $F_{2,794.1} = 404.84$ ,  $p < 0.0001$ ).

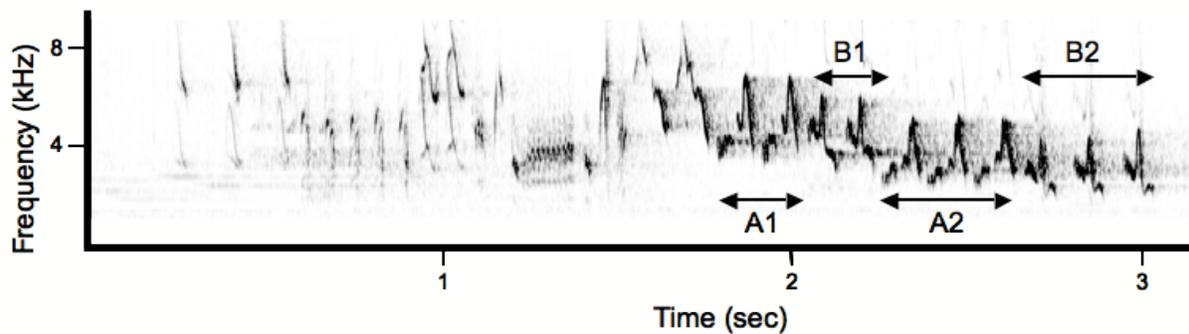


Figure S.1.1. Spectrogram example of a song with four trills belonging to two trill types, labeled A and B. Trill A1 would be categorized “first”, B1 and A2 as “middle”, and B2 as “last” for this analysis. Note how the pitch of the trills decreases over the course of the song.

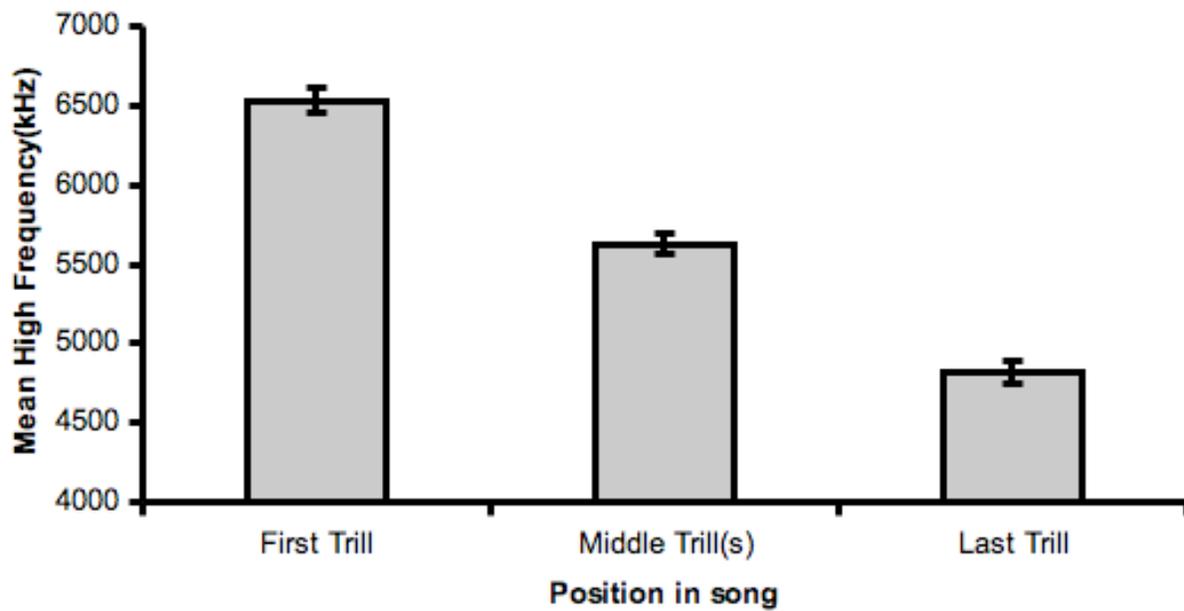


Figure S.1.2. Mean  $\pm$  SE high frequency in the trill as a function of position within the song, measured for 8 males and 809 trills. Values are the least squared means and standard errors from a mixed model including male identity as a random effect and trill type as a covariate.

## Supplementary material 2. Evaluation of different measures of pitch.

To determine whether high or low frequency in the song was a better measure of pitch for the pitch-performance relationship, I constructed models expressing vocal deviation as a function of trill type, high or low frequency, and the interaction between the pitch measure and trill type. To determine whether the relationship between vocal deviation and pitch was driven by a pitch relationship with frequency bandwidth or with trill rate (the two factors that contribute to variation in vocal deviation) I constructed additional models. In these models, I used either frequency bandwidth or trill rate as the response variable and trill type, a pitch measure (either high or low frequency), and the interaction of the two predictors. For all tests,  $n = 4645$  songs of the eight common trill types, recorded from 61 males during playback trials. The relationship between vocal deviation and pitch is stronger (note higher  $r^2$  and larger  $F$ ) when I measured pitch as the high frequency in the song, so I used high frequency in my main analyses. Pitch had a stronger effect on frequency bandwidth than it did on trill rate, as indicated by the higher  $F$  test scores for the pitch-bandwidth relationship than for the pitch-trill rate relationship. Note that the very high  $r^2$  values for the trill rate-pitch relationships are driven primarily by the large effect of trill type on trill rate. While high frequency was a superior pitch measure, the relationships between pitch and vocal deviation, frequency bandwidth, and trill rate were highly statistically significant regardless of which measure I used for pitch.

Table S.1.1. High and low frequencies are similar predictors of vocal deviation, frequency bandwidth, and trill rate.

Response variable	Predictor	Whole model: $r^2$ , F, p	Individual effect	Individual effect F	df	p
Vocal Deviation	High Frequency	0.93, 3898.85, 0	Trill Type	3574.92	7, 4629	<0.0001
			Pitch	2943.54	1, 4629	<0.0001
			Interaction	187.92	7, 4629	<0.0001
	Low Frequency	0.84, 1603.69, 0	Trill Type	2006.99	7, 4629	<0.0001
			Pitch	463.43	1, 4629	<0.0001
			Interaction	138.93	7, 4629	<0.0001
Frequency Bandwidth	High Frequency	0.85, 1773.26, 0	Trill Type	311.40	7, 4629	<0.0001
			Pitch	2407.15	1, 4629	<0.0001
			Interaction	123.01	7, 4629	<0.0001
	Low Frequency	0.68, 664.53, 0	Trill Type	992.25	7, 4629	<0.0001
			Pitch	320.87	1, 4629	<0.0001
			Interaction	100.88	7, 4629	<0.0001
Trill Rate	High Frequency	0.99, 20592.44, 0	Trill Type	26497.80	7, 4629	<0.0001
			Pitch	194.69	1, 4629	<0.0001
			Interaction	92.34	7, 4629	<0.0001
	Low Frequency	0.99, 21027.29, 0	Trill Type	35611.62	7, 4629	<0.0001
			Pitch	204.68	1, 4629	<0.0001
			Interaction	105.32	7, 4629	<0.0001

Supplementary 3. Other approaches to analyzing response to playback.

In the main text, I analyzed playback response as a function of the categorical playback treatment, but expressing the playback stimulus quality as a continuous measure can also be informative (e.g., Illes et al. 2006). Moreover, the relative quality of the focal male vs. the playback stimulus may be a more powerful approach to detecting a pattern (e.g., Collins 2004). Here, I present alternate analyses of playback responses as a function of the stimulus song parameters, and as a function of stimulus song measures relative to the same song measures of the focal male. I also present data on “multiplicative vocal deviation,” under the following rationale (pointed out to us by an anonymous reviewer, whom I thank).

*Multiplicative vocal deviation--*. Avian sound perception may, like human sound perception, be more sensitive to the ratio rather than the difference between frequencies (e.g., Hurly et al. 1992). Analogously, the relationship of fundamental sound frequency with tension in the syringeal muscles and with the volume of the upper vocal tract is also non-linear (Goller and Suthers 1996, Fletcher et al. 2006). The salient feature in my playback stimuli may therefore have been the frequency ratio-trill rate relationship, rather than the frequency bandwidth-trill rate relationship. To evaluate whether males responded differently to playback stimuli depending on the frequency ratio-trill rate relationship, I therefore calculated a performance relating the frequency ratio (high frequency/low frequency) to trill rate. Following Podos’ (1997) protocol for additive bandwidth, I created bins by trill rate, chose the maximum frequency ratio for each bin, and performed upper-bound regression across these highest-performance trills. As with the additive bandwidth calculations, several bins did not have apparent high-performance exemplars, and I excluded 6 of 17 bins (regression line: frequency ratio =  $3.19 - 0.051 * \text{trill rate (Hz)}$ ;  $r^2 = 0.82$ ,  $F_{1,9} = 42.27$ ,  $p < 0.0001$ ). The upper-bound regression was, however, significant and

negative even when all bins were included (equation: frequency ratio =  $2.78 - 0.034 * \text{trill rate}$ ;  $r^2 = 0.27$ ,  $F_{1,15} = 5.60$ ,  $p = 0.03$ ). I calculated orthogonal deviation from this line and called this measure “multiplicative vocal deviation.” Multiplicative vocal deviation was not consistently correlated with pitch: five syllables types showed significant positive correlations ( $r^2 = 0.02 - 0.59$ ) while the other three showed significant negative correlations ( $r^2 = 0.02 - 0.13$ ). I therefore present multiplicative vocal deviation without a pitch correction. This multiplicative vocal deviation was highly correlated with additive vocal deviation ( $r^2 = 0.63$ ,  $F_{1,4790} = 201.87$ ,  $p < 0.0001$ ) in a model controlling for male identity (random effect), year, and trill type. It was also highly correlated with residual additive vocal deviation (i.e., additive vocal deviation after correcting for pitch) in a model with the same other parameters ( $r^2 = 0.73$ ,  $F_{1,4780} = 2117.34$ ,  $p < 0.0001$ ).

*Statistical analysis.*-- I examined playback responses both as a function of the absolute measurements of the stimulus songs and as a function of the measurements of the stimulus song relative to the focal male. As with the main analyses with categorical treatment variables, I included a random effect of bird identity nested within a random effect of stimulus source. I also included playback phase and order as fixed effects, and analyzed data from each experimental protocol separately. For tests examining the relative song measurements of the focal male and the playback stimulus, I expressed playback treatments as the log ratio of the focal bird’s mean song measures (residual vocal deviation, trill consistency, and multiplicative vocal deviation) to the stimulus song’s measure. Song measures differ substantially among syllable types, so an ideal comparison would be within-syllable-type only. However, all birds did not sing all trill types in response to playback, so doing this analysis would reduce the sample size by eight individuals. To avoid this reduction in power, I calculated the mean vocal deviation and mean

trill consistency for all individuals across all syllable types, and calculated the log ratio using these means. The analyses of playback response using the reduced data set with only matching trill types gave the same results (not shown), and the matched-type log ratios were tightly correlated with the unmatched-all-types log ratios ( $r^2 = 0.80$ ,  $F_{1,394} = 1545.69$ ,  $p < 0.0001$ ). I re-scaled the residual vocal deviation and multiplicative vocal deviation scores to be all positive values so that I could use the log ratio, which is the preferred statistical method for this analysis.

I also investigated models with other covariates, and with squared vocal deviation terms (to allow for non-linear relationships between a focal male's response to playback and his quality relative to that of the stimulus; Collins 2004). These analyses did not reveal significant treatment effects after correcting for multiple testing, and are not shown.

*Results*—No treatment effects were robust to correction for multiple testing, and the different analytical approaches (categorical vs. continuous, relative vs. absolute) generally reveal similar patterns (Table S.1.2, Table 1.1). The signs of the relationships between playback response and absolute vs. relative stimulus measures typically differ. This pattern makes sense in light of my calculations of the ratio between the focal male and the stimulus songs: a higher absolute measure for the stimulus song should correspond to a lower focal male/stimulus song ratio. Note that the parameter estimates of the absolute and relative song measurements are not directly comparable, since the relative scores are re-scaled log ratios while the absolute measures are song measurements.

The trends apparent in the categorical analysis in the main text are also reflected in these analyses: males tended to spend more time close to higher vocal deviation stimuli and to fly across the less consistent stimuli more in all analyses in Protocol 1, though the effect sizes for these differences would be small.

Multiplicative vocal deviation showed trends towards significance for several measures, with males tending to spend a lower proportion of time within 5 m of the speaker for lower multiplicative vocal deviation (more challenging) stimuli in Protocol 1, and being less likely to fly across the speaker, and flying across the speaker less often in response to lower multiplicative vocal deviation stimuli in Protocol 2 (Table S.1.2). These patterns appeared to be driven by the frequency ratio, rather than the trill rate (not shown). Since the stimuli were not designed to capture the full range of frequency ratios (Table S.1.3), it is difficult to interpret these trends: the fact that trends appear in so many response variables for multiplicative vocal deviation is suggestive that this measure could be important, but I lack the data to test the idea thoroughly.

Table S.1.2. Relationships between male response to playback and characteristics of the playback stimulus. No stimulus measurements remained significant after correction for multiple testing. All models included a random effect of male identity and stimulus set, as well as fixed effects of order, phase (during vs. after playback presentation), and a stimulus measurement.

\*Effects of order and phase were generally similar for all models of the same response variable and stimulus measure, so I report only a summary of these effects for all six models.

\*\*Consistency is the spectrogram cross-correlation score of the playback stimulus; Voc. Dev. is the residual vocal deviation (controlling for pitch and trill type), calculated with additive frequency bandwidth. Mult. Voc. Dev. is deviation from an analogous performance limit to vocal deviation, but using the ratio of high:low frequency rather than the difference between high and low frequency. Rel. Mult. Voc. Dev. is the multiplicative vocal deviation of the focal male relative to the playback stimulus, calculated as described in the main text for relative additive residual vocal deviation (i.e., multiplicative vocal deviations were re-scaled to be positive, and I took the log of the ratio). For binary response variables, I modeled the probability that the response would occur.

Protocol	Response Effects of Phase and Order*	Stimulus measurement**	Coefficient ± SE	F <sub>df</sub> (p)
1	Approach (binary)	Consistency	1.328 ± 1.897	F <sub>1,62.1</sub> = 0.49 (0.49)
		Voc. Dev.	0.043 ± 0.055	F <sub>1,295</sub> = 0.61 (0.44)
		Mult. Voc. Dev.	0.22 ± 0.921	F <sub>1,295</sub> = 0.06 (0.81)
		Rel. Consistency	-2.800 ± 3.351	F <sub>1,69.6</sub> = 0.70 (0.41)
		Rel. Voc. Dev.	-0.970 ± 1.145	F <sub>1,295</sub> = 0.72 (0.49)
	Phase P > 0.11	Rel. Mult. Voc. Dev.	0.073 ± 1.039	F <sub>1,295</sub> = 0.00 (0.94)
	Approach (normal)	Consistency	0.023 ± 0.21	F <sub>1,50.05</sub> = 0.01 (0.91)
		Voc. Dev.	0.012 ± 0.006	F <sub>1,174.1</sub> = 3.8 (0.053)
		Mult. Voc. Dev.	0.166 ± 0.106	F <sub>1,187.7</sub> = 2.46 (0.12)
		Rel. Consistency	-0.108 ± 0.371	F <sub>1,61.4</sub> = 0.09 (0.77)
		Rel. Voc. Dev.	-0.250 ± 0.125	F <sub>1,176.9</sub> = 3.96 (0.05)
	Phase P < 0.008	Rel. Mult. Voc. Dev.	-0.162 ± 0.105	F <sub>1,194.8</sub> = 2.39 (0.12)
	Flights (binary)	Consistency	-1.499 ± 1.571	F <sub>1,60.7</sub> = 0.91 (0.34)
		Voc. Dev.	0.034 ± 0.05	F <sub>1,295</sub> = 0.46 (0.50)
		Mult. Voc. Dev.	0.509 ± 0.832	F <sub>1,295</sub> = 0.38 (0.54)
		Rel. Consistency	2.584 ± 2.782	F <sub>1,67.7</sub> = 0.86 (0.36)
		Rel. Voc. Dev.	-0.734 ± 1.026	F <sub>1,295</sub> = 0.51 (0.48)
	Phase P < 0.001	Rel. Mult. Voc. Dev.	-0.33 ± 0.866	F <sub>1,295</sub> = 0.15 (0.70)
	Flights (count)	Consistency	-0.645 ± 0.412	F <sub>1,29.4</sub> = 2.45 (0.13)
		Voc. Dev.	-0.003 ± 0.017	F <sub>1,90.4</sub> = 0.03 (0.87)
Mult. Voc. Dev.		-0.168 ± 0.299	F <sub>1,119</sub> = 0.32 (0.57)	
Rel. Consistency		0.954 ± 0.728	F <sub>1,32.7</sub> = 1.72 (0.20)	
Rel. Voc. Dev.		0.091 ± 0.347	F <sub>1,90.7</sub> = 0.07 (0.79)	
Phase P < 0.001	Rel. Mult. Voc. Dev.	0.213 ± 0.303	F <sub>1,125.3</sub> = 0.49 (0.48)	
Songs	Consistency	-1.529 ± 1.771	F <sub>1,42.7</sub> = 0.74 (0.39)	
	Voc. Dev.	-0.08 ± 0.053	F <sub>1,251.3</sub> = 2.27 (0.13)	
	Mult. Voc. Dev.	-1.419 ± 0.902	F <sub>1,264.7</sub> = 2.47 (0.12)	
	Rel. Consistency	4.017 ± 3.095	F <sub>1,49.5</sub> = 1.68 (0.20)	
	Rel. Voc. Dev.	1.407 ± 1.102	F <sub>1,255.7</sub> = 1.63 (0.20)	
Phase P > 0.25	Rel. Mult. Voc. Dev.	1.149 ± 0.961	F <sub>1,279.2</sub> = 1.43 (0.23)	
Order P > 0.30				

Table S.1.2 Continued

	Consistency	$0.935 \pm 2.765$	$F_{1,24.5} = 0.11 (0.74)$
Approach (binary)	Voc. Dev.	$0.101 \pm 0.114$	$F_{1,139} = 0.79 (0.38)$
	Mult. Voc. Dev.	$2.23 \pm 1.413$	$F_{1,139} = 2.49 (0.12)$
	Rel. Consistency	$-1.706 \pm 4.628$	$F_{1,28} = 0.14 (0.72)$
	Phase P < 0.002	Rel. Voc. Dev.	$-2.560 \pm 2.536$
Order P > 0.43	Rel. Mult. Voc. Dev.	$-4.143 \pm 2.156$	$F_{1,139} = 3.69 (0.06)$
	Consistency	$0.099 \pm 0.266$	$F_{1,10.7} = 0.14 (0.72)$
Approach (normal)	Voc. Dev.	$0.004 \pm 0.012$	$F_{1,45.5} = 0.14 (0.71)$
	Mult. Voc. Dev.	$0.109 \pm 0.152$	$F_{1,49.5} = 0.52 (0.48)$
	Rel. Consistency	$-0.132 \pm 0.457$	$F_{1,15.84} = 0.08 (0.78)$
	Phase P < 0.001	Rel. Voc. Dev.	$-0.127 \pm 0.269$
Order P < 0.02	Rel. Mult. Voc. Dev.	$-0.134 \pm 0.247$	$F_{1,50.6} = 0.30 (0.59)$
	Consistency	$1.758 \pm 2.34$	$F_{1,12.7} = 0.56 (0.47)$
2 Flights (binary)	Voc. Dev.	$0.131 \pm 0.117$	$F_{1,139} = 1.26 (0.26)$
	Mult. Voc. Dev.	$2.27 \pm 1.438$	$F_{1,139} = 2.49 (0.12)$
	Rel. Consistency	$-2.415 \pm 4.030$	$F_{1,16.4} = 0.36 (0.56)$
	Phase P < 0.001	Rel. Voc. Dev.	$-3.199 \pm 2.591$
Order P > 0.30	Rel. Mult. Voc. Dev.	$-4.099 \pm 2.23$	$F_{1,139} = 3.38 (0.07)$
	Consistency	$-1.463 \pm 0.932$	$F_{1,18.4} = 2.47 (0.13)$
Flights (count)	Voc. Dev.	$0 \pm 0.044$	$F_{1,43} = 0 (0.99)$
	Mult. Voc. Dev.	Model did not converge	
	Rel. Consistency	$3.127 \pm 1.528$	$F_{1,19.2} = 4.19 (0.054)$
	Phase P < 0.001	Rel. Voc. Dev.	$-0.330 \pm 1.002$
Order P < 0.09	Rel. Mult. Voc. Dev.	$-0.921 \pm 0.928$	$F_{1,31.2} = 0.98 (0.33)$
	Consistency	$-1.328 \pm 2.028$	$F_{1,17.6} = 0.43 (0.52)$
Songs	Voc. Dev.	$-0.050 \pm 0.088$	$F_{1,117.5} = 0.33 (0.57)$
	Mult. Voc. Dev.	$-0.436 \pm 1.062$	$F_{1,123.5} = 0.17 (0.68)$
	Rel. Consistency	$2.730 \pm 3.467$	$F_{1,22.26} = 0.62 (0.44)$
	Phase P > 0.23	Rel. Voc. Dev.	$1.046 \pm 1.940$
Order P < 0.05	Rel. Mult. Voc. Dev.	$0.57 \pm 1.567$	$F_{1,124.9} = 0.13 (0.72)$

Supplementary material 4. Description of playback stimulus quality relative to the natural range of variation in songs.

Here, I present a comparison of the song variables I specifically set out to manipulate (vocal deviation, trill rate, frequency bandwidth) as well as pitch (expressed as high frequency), since it was an important covariate with vocal deviation. I also present values for the residual vocal deviation correcting for pitch, frequency ratio, and multiplicative vocal deviation (see Supplementary Material 3), to evaluate the playback stimuli on these parameters, which I did not intentionally manipulate. While the playback stimuli showed good separation and covered most of the natural range for multiplicative vocal deviation, this effect was largely driven by the trill rate manipulation: for many trill types, the frequency ratio of the playback stimuli did not cover the full natural range of frequency ratios. Playback stimuli covered the natural range for the other song parameters, which I had intended to manipulate; note that syllable types A and T were not included in many playback stimuli, so comparisons of playback stimuli to natural songs for these syllable types suffered from low statistical power.

The length of the playback stimuli did differ slightly because the trill rate of the different treatments differed, but I decided that it was better to allow the stimulus lengths to differ rather than to have the stimuli differ by the number of note repetitions (e.g., Illes et al. 2006).

Table S.1.3. Descriptive statistics and comparison of acoustic measurements of natural songs and playback stimuli. N trills are given in the first row of the table for that category; sample sizes were the same for all acoustic measurements reported here.

These measurements only refer to the manipulated trills in the playback stimuli (see main text).

\* Unless otherwise noted, for comparisons relevant to vocal deviation, the three playback stimulus classes differed significantly from each other, and the low- and high-deviation stimuli differed significantly from the natural songs. Where noted, rows without a letter in common are statistically different from each other (all comparisons are within syllable types only). For consistency comparisons, high consistency stimuli were not re-measured because they, by definition, had perfect consistency (equal to 1) and any difference from that would be due to measurement error. Statistics for the consistency comparison refer to a comparison of low to natural consistencies.

\*\* Dev. = Vocal Deviation playback stimulus class. Cons. = Trill Consistency playback class

Acoustic measure	Syllable Type and ANOVA results	Trill context (n)**	Mean $\pm$ SD (95% confidence Interval)	Range	Grouping*
Vocal Deviation (No pitch correction)	2D $F_{3,1181} = 408.62,$ $P < 0.0001$	Natural (1023)	6.07 $\pm$ 1.23 (6.00, 6.15)	1.81, 9.46	
		Low Dev. (54)	1.86 $\pm$ 1.47 (1.46, 2.26)	-1.84, 5.03	
		Medium Dev. (54)	6.11 $\pm$ 0.96 (5.84, 6.38)	3.52, 8.17	
		High Dev. (54)	10.11 $\pm$ 0.95 (9.85, 10.37)	7.41, 12.00	
	2U $F_{3,1004} = 65.99,$ $P < 0.0001$	Natural (948)	11.31 $\pm$ 1.74 (11.20, 11.42)	5.87, 16.64	
		Low Dev. (20)	7.95 $\pm$ 0.95 (7.51, 8.40)	6.24, 9.75	
		Medium Dev. (20)	11.72 $\pm$ 0.53 (11.47, 11.97)	10.85, 12.39	
		High Dev. (20)	15.46 $\pm$ 1.01 (14.98, 15.93)	13.44, 17.13	
	A $F_{3,974} = 14.83,$ $P < 0.0001$	Natural (966)	10.81 $\pm$ 2.82 (10.64, 10.99)	1.24, 15.51	
		Low Dev. (4)	2.13 $\pm$ 0.81 (0.83, 3.43)	1.36, 2.98	
		Medium Dev. (4)	9.97 $\pm$ 1.74 (7.2, 12.74)	8.37, 11.51	
		High Dev. (4)	14.25 $\pm$ 0.73 (13.09, 15.4)	13.62, 14.97	
	B $F_{3,348} = 55.45,$ $P < 0.0001$	Natural (262)	4.74 $\pm$ 2.01 (4.49, 4.98)	-0.97, 11.89	
		Low Dev. (30)	1.74 $\pm$ 1.22 (1.29, 2.20)	-0.44, 4.08	
		Medium Dev. (30)	5.19 $\pm$ 0.99 (4.82, 5.56)	3.61, 7.91	
		High Dev. (30)	7.84 $\pm$ 1.37 (7.33, 8.35)	5.88, 11.09	
	Q $F_{3,678} = 24.39,$ $P < 0.0001$	Natural (652)	12.58 $\pm$ 2.40 (12.40, 12.77)	4.58, 18.14	
		Low Dev. (10)	8.00 $\pm$ 1.80 (6.71, 9.29)	5.89, 10.97	
		Medium Dev. (10)	12.79 $\pm$ 1.34 (11.83, 13.75)	10.88, 14.84	
		High Dev. (10)	17.07 $\pm$ 0.84 (16.47, 17.67)	15.68, 18.18	
S $F_{3,432} = 118.34,$ $P < 0.0001$	Natural (370)	2.53 $\pm$ 1.62 (2.37, 2.70)	-2.49, 6.42		
	Low Dev. (22)	-1.27 $\pm$ 1.08 (-1.74, -0.79)	-3.19, 1.02		
	Medium Dev. (22)	2.46 $\pm$ 0.59 (2.20, 2.72)	1.18, 3.35		
	High Dev. (22)	7.36 $\pm$ 0.74 (7.03, 7.69)	6.22, 8.60		

Table S.1.3 Continued

Residual Vocal Deviation (After pitch correction)	T $F_{3,106} = 5.87,$ $P = 0.001$	Natural (104)	$6.01 \pm 1.81$ (5.66, 6.36)	1.52, 9.34	a
		Low Dev. (2)	$1.21 \pm 0.01$ (1.13, 1.30)	1.21, 1.22	b
		Medium Dev. (2)	$4.72 \pm 0.05$ (4.31, 5.12)	4.68, 4.75	ab
		High Dev. (2)	$7.90 \pm 0.03$ (7.64, 8.17)	7.88, 7.93	a
	V $F_{3,341} = 31.28,$ $P < 0.0001$	Natural (320)	$4.22 \pm 1.60$ (4.05, 4.40)	-0.65, 9.43	
		Low Dev. (9)	$1.56 \pm 1.32$ (0.55, 2.57)	-0.12, 3.23	
		Medium Dev. (9)	$5.14 \pm 1.9$ (3.55, 6.72)	2.38, 6.97	
		High Dev. (9)	$8.9 \pm 1.98$ (7.25, 10.55)	5.96, 10.80	
	2D $F_{3,1181} = 432.27,$ $P < 0.0001$	Natural	$0 \pm 1.07$ (-0.07, 0.07)	-3.14, 3.09	
		Low Dev.	$-3.98 \pm 1.47$ (-4.38, -3.58)	-7.48, -1.51	
		Medium Dev.	$-0.16 \pm 0.77$ (-0.37, 0.05)	-2.35, 1.23	
		High Dev.	$3.39 \pm 0.78$ (3.17, 3.60)	1.06, 5.11	
	2U $F_{3,1004} = 62.57,$ $P < 0.0001$	Natural	$0.00 \pm 1.05$ (-0.07, 0.07)	-3.14, 3.30	
		Low Dev.	$-1.82 \pm 1.38$ (-2.47, -1.18)	-4.04, 0.46	
		Medium Dev.	$0.18 \pm 0.76$ (-0.18, 0.53)	-0.81, 2.31	
		High Dev.	$2.64 \pm 0.75$ (2.29, 2.99)	1.125, 4.07	
	A $F_{3,974} = 16.67,$ $P < 0.0001$	Natural	$0.00 \pm 1.16$ (-0.07, 0.07)	-4.54, 5.87	
		Low Dev.	$-3.76 \pm 1.29$ (-5.82, -1.71)	-4.92, -2.52	
		Medium Dev.	$-0.73 \pm 0.54$ (-1.59, 0.13)	-1.24, -0.04	
		High Dev.	$1.42 \pm 0.36$ (0.85, 2.00)	1.07, 1.87	
B $F_{3,348} = 86.91,$ $P < 0.0001$	Natural	$0.00 \pm 0.75$ (-0.09, 0.09)	-2.41, 2.32		
	Low Dev.	$-1.62 \pm 0.74$ (-1.89, -1.34)	-2.96, 0.14		
	Medium Dev.	$-0.12 \pm 0.55$ (-0.32, 0.09)	-1.27, 0.68		
	High Dev.	$1.41 \pm 0.66$ (1.17, 1.66)	-0.17, 2.66		
Q $F_{3,678} = 39.41,$ $P < 0.0001$	Natural	$0.00 \pm 0.75$ (-0.06, 0.06)	-3.24, 2.07		
	Low Dev.	$-2.27 \pm 1.52$ (-3.36, -1.18)	-4.12, 1.441		
	Medium Dev.	$0.37 \pm 0.76$ (-0.17, 0.92)	-1.06, 1.97		
	High Dev.	$1.28 \pm 0.87$ (0.66, 1.90)	-0.03, 2.27		

Table S.1.3 Continued

Multi- plicative Vocal Deviation	S $F_{3,432} = 120.15,$ $P < 0.0001$	Natural	$0.00 \pm 1.36$ (-0.13, 0.13)	-4.12, 4.63	
		Low Dev.	$-2.48 \pm 0.67$ (-2.77, -2.18)	-4.01, -1.28	
		Medium Dev.	$0.43 \pm 1.14$ (-0.07, 0.94)	-1.23, 3.42	
		High Dev.	$4.22 \pm 0.72$ (3.90, 4.53)	2.97, 5.63	
	T $F_{3,106} = 16.23,$ $P < 0.0001$	Natural	$0.00 \pm 1.36$ (-0.26, 0.26)	-2.96, 2.30	a
		Low Dev.	$-5.93 \pm 0.38$ (-9.38, -2.47)	-6.20, -5.66	b
		Medium Dev.	$-2.17 \pm 0.15$ (-3.47, -0.86)	-2.27, -2.06	c
		High Dev.	$2.13 \pm 0.11$ (1.16, 3.09)	2.05, 2.21	d
	V $F_{3,341} = 38.58,$ $P < 0.0001$	Natural	$0.00 \pm 1.24$ (-0.14, 0.14)	-3.55, 5.49	
		Low Dev.	$-2.87 \pm 0.96$ (-3.61, -2.13)	-4.28, -1.85	
		Medium Dev.	$0.59 \pm 1.06$ (-0.29, 1.47)	-1.55, 1.92	
		High Dev.	$3.59 \pm 2.02$ (1.91, 5.28)	1.26, 7.11	
	2D $F_{3,1181} = 33.68,$ $P < 0.0001$	Natural	$0.41 \pm 0.19$ (0.4, 0.42)	-0.11, 0.95	a
		Low Dev.	$0.18 \pm 0.19$ (0.13, 0.23)	-0.18, 0.59	b
		Medium Dev.	$0.35 \pm 0.12$ (0.32, 0.38)	0.05, 0.58	c
		High Dev.	$0.52 \pm 0.12$ (0.49, 0.55)	0.25, 0.81	d
2U $F_{3,1004} = 30.63,$ $P < 0.0001$	Natural	$0.59 \pm 0.17$ (0.58, 0.60)	0.06, 0.99		
	Low Dev.	$0.40 \pm 0.23$ (0.29, 0.51)	0.05, 0.75		
	Medium Dev.	$0.63 \pm 0.11$ (0.58, 0.68)	0.48, 0.90		
	High Dev.	$0.89 \pm 0.10$ (0.84, 0.94)	0.65, 1.06		
A $F_{3,974} = 3.70,$ $P = 0.01$	Natural	$0.75 \pm 0.13$ (0.74, 0.76)	0.34, 1.27	a	
	Low Dev.	$0.56 \pm 0.16$ (0.30, 0.82)	0.41, 0.71	b	
	Medium Dev.	$0.68 \pm 0.05$ (0.60, 0.76)	0.64, 0.75	ab	
	High Dev.	$0.84 \pm 0.04$ (0.78, 0.90)	0.80, 0.89	a	
B $F_{3,348} = 50.37,$ $P < 0.0001$	Natural	$0.33 \pm 0.12$ (0.32, 0.35)	-0.02, 0.78		
	Low Dev.	$0.15 \pm 0.09$ (0.12, 0.18)	-0.02, 0.36		
	Medium Dev.	$0.32 \pm 0.08$ (0.29, 0.35)	0.15, 0.44		
	High Dev.	$0.50 \pm 0.10$ (0.47, 0.54)	0.25, 0.71		

Table S.1.3 Continued

Multi- plicative Vocal Deviation	Q $F_{3,678} = 27.52,$ $P < 0.0001$	Natural	$0.74 \pm 0.10$ (0.74, 0.75)	0.40, 1.09	
		Low Dev.	$0.52 \pm 0.18$ (0.39, 0.65)	0.20, 0.87	
		Medium Dev.	$0.78 \pm 0.07$ (0.73, 0.83)	0.61, 0.88	
		High Dev.	$0.91 \pm 0.09$ (0.84, 0.97)	0.74, 1.02	
	S $F_{3,432} = 35.02,$ $P < 0.0001$	Natural	$0.35 \pm 0.13$ (0.34, 0.36)	-0.12, 0.76	a
		Low Dev.	$0.26 \pm 0.09$ (0.22, 0.30)	0.07, 0.39	b
		Medium Dev.	$0.42 \pm 0.13$ (0.36, 0.47)	0.20, 0.69	c
		High Dev.	$0.62 \pm 0.11$ (0.57, 0.66)	0.43, 0.79	d
	T $F_{3,106} = 16.45,$ $P < 0.0001$	Natural	$0.88 \pm 0.08$ (0.86, 0.89)	0.67, 1.03	a
		Low Dev.	$0.50 \pm 0.03$ (0.26, 0.74)	0.48, 0.52	c
		Medium Dev.	$0.73 \pm 0.01$ (0.63, 0.83)	0.72, 0.74	b
		High Dev.	$0.97 \pm 0.00$ (0.94, 1.01)	0.97, 0.98	a
V $F_{3,341} = 25.82,$ $P < 0.0001$	Natural	$0.64 \pm 0.11$ (0.63, 0.67)	0.31, 1.02		
	Low Dev.	$0.42 \pm 0.10$ (0.34, 0.50)	0.27, 0.52		
	Medium Dev.	$0.70 \pm 0.08$ (0.63, 0.76)	0.54, 0.80		
	High Dev.	$0.87 \pm 0.14$ (0.75, 0.99)	0.69, 1.11		
Frequency Bandwidth (kHz)	2D $F_{3,1181} = 72.28,$ $P < 0.0001$	Natural	$2755 \pm 202$ (2743, 2767)	2063, 3402	
		Low Dev.	$3008 \pm 150$ (2967, 3049)	2656, 3404	
		Medium Dev.	$2751 \pm 146$ (2711, 2791)	2391, 3186	
		High Dev.	$2450 \pm 192$ (2397, 2502)	2121, 2949	
	2U $F_{3,1004} = 42.30,$ $P < 0.0001$	Natural	$2880 \pm 274$ (2862, 2897)	2001, 3744	
		Low Dev.	$3286 \pm 163$ (3210, 3363)	2977, 3574	
		Medium Dev.	$2797 \pm 83$ (2758, 2836)	2686, 2934	
		High Dev.	$2344 \pm 166$ (2267, 2422)	2063, 2672	
	A $F_{3,974} = 11.51$ $P < 0.0001$	Natural	$2882 \pm 443$ (2854, 2910)	2105, 4383	a
		Low Dev.	$4098 \pm 200$ (3780, 4416)	3897, 4273	b
		Medium Dev.	$3044 \pm 314$ (2544, 3545)	2766, 3328	a
		High Dev.	$2457 \pm 112$ (2280, 2635)	2344, 2555	a

Table S.1.3 Continued

Frequency Bandwidth (kHz)	B $F_{3,348} = 28.54,$ $P < 0.0001$	Natural	3922 ± 339 (3881, 3963)	2754, 4762	
		Low Dev.	4266 ± 194 (4194, 4339)	3912, 4576	
		Medium Dev.	3846 ± 164 (3785, 3908)	3403, 4125	
		High Dev.	3533 ± 232 (3447, 3620)	2977, 3867	
	Q $F_{3,678} = 16.85,$ $P < 0.0001$	Natural	2521 ± 332 (2495, 2546)	1594, 3359	
		Low Dev.	3005 ± 280 (2805, 3206)	2578, 3398	
		Medium Dev.	2509 ± 222 (2350, 2667)	2191, 2844	
		High Dev.	1964 ± 159 (1851, 2078)	1805, 2231	
	S $F_{3,432} = 27.06,$ $P < 0.0001$	Natural	2259 ± 270 (2232, 2287)	1522, 3052	
		Low Dev.	2605 ± 169 (2529, 2680)	2329, 2865	
		Medium Dev.	2257 ± 107 (2209, 2304)	2075, 2542	
		High Dev.	1910 ± 138 (1849, 1972)	1674, 2203	
T $F_{3,106} = 6.02,$ $P = 0.008$	Natural	2558 ± 249 (2510, 2607)	1914, 3191	b	
	Low Dev.	3151 ± 2 (3137, 3166)	3150, 3152	a	
	Medium Dev.	2993 ± 13 (2874, 3112)	2984, 3002	a	
	High Dev.	2734 ± 5 (2689, 2779)	2731, 2738	ab	
V $F_{3,341} = 17.01,$ $P < 0.0001$	Natural	2904 ± 256 (2876, 2933)	2074, 3504		
	Low Dev.	3165 ± 238 (2982, 3347)	2906, 3497		
	Medium Dev.	2726 ± 309 (2468, 2984)	2409, 3176		
	High Dev.	2334 ± 347 (2044, 2624)	2004, 2848		
Frequency ratio (high frequency: low frequency)	2D $F_{3,1181} = 5.99,$ $P = 0.0005$	Natural	2.11 ± 0.19 (2.09, 2.18)	1.53, 2.67	c
		Low Dev.	2.20 ± 0.16 (2.16, 2.24)	1.84, 2.49	a
		Medium Dev.	2.17 ± 0.13 (2.13, 2.20)	1.90, 2.48	ab
		High Dev.	2.11 ± 0.13 (2.08, 2.15)	1.83, 2.40	bc
	2U $F_{3,1004} = 19.94,$ $P < 0.0001$	Natural	2.23 ± 0.17 (2.22, 2.25)	1.84, 2.77	
		Low Dev.	2.38 ± 0.24 (2.27, 2.49)	2.02, 2.71	
		Medium Dev.	2.19 ± 0.11 (2.14, 2.24)	1.91, 2.33	
		High Dev.	1.98 ± 0.11 (1.93, 2.03)	1.80, 2.22	

Table S.1.3 Continued

Frequency ratio (high frequency: low frequency)	A $F_{3,974} = 1.38,$ $P = 0.25$	Natural	$2.05 \pm 0.14$ (2.04, 2.05)	1.47, 2.47	
		Low Dev.	$2.16 \pm 0.18$ (1.87, 2.45)	1.99, 2.33	
		Medium Dev.	$2.12 \pm 0.05$ (2.04, 2.21)	2.05, 2.16	
		High Dev.	$2.00 \pm 0.03$ (1.95, 2.06)	1.96, 2.04	
	B $F_{3,348} = 28.98$ $P < 0.001$	Natural	$2.47 \pm 0.12$ (2.46, 2.48)	2.04, 2.81	
		Low Dev.	$2.60 \pm 0.09$ (2.57, 2.63)	2.43, 2.75	
		Medium Dev.	$2.48 \pm 0.08$ (2.45, 2.51)	2.36, 2.65	
		High Dev.	$2.34 \pm 0.10$ (2.30, 2.37)	2.14, 2.60	
	Q $F_{3,678} = 10.46,$ $P < 0.0001$	Natural	$2.03 \pm 0.10$ (2.03, 2.04)	1.69, 2.38	b
		Low Dev.	$2.17 \pm 0.19$ (2.04, 2.31)	1.82, 2.51	a
		Medium Dev.	$2.01 \pm 0.07$ (1.96, 2.05)	1.91, 2.17	bc
		High Dev.	$1.93 \pm 0.09$ (1.87, 1.99)	1.83, 2.11	c
	S $F_{3,432} = 8.73,$ $P < 0.0001$	Natural	$1.84 \pm 0.13$ (1.82, 1.85)	1.48, 2.27	a
		Low Dev.	$1.84 \pm 0.09$ (1.80, 1.88)	1.72, 2.04	ab
		Medium Dev.	$1.76 \pm 0.12$ (1.71, 1.82)	1.50, 2.00	bc
		High Dev.	$1.71 \pm 0.11$ (1.66, 1.76)	1.53, 1.93	c
T $F_{3,106} = 15.79,$ $P < 0.0001$	Natural	$1.58 \pm 0.08$ (1.56, 1.59)	1.39, 1.81	b	
	Low Dev.	$1.89 \pm 0.03$ (1.64, 2.13)	1.87, 1.91	a	
	Medium Dev.	$1.79 \pm 0.02$ (1.64, 1.95)	1.78, 1.80	a	
	High Dev.	$1.63 \pm 0.00$ (1.60, 1.66)	1.63, 1.63	b	
V $F_{3,341} = 12.73,$ $P < 0.0001$	Natural	$1.82 \pm 0.11$ (1.81, 1.83)	1.46, 2.13	b	
	Low Dev.	$1.99 \pm 0.12$ (1.90, 2.08)	1.87, 2.16	a	
	Medium Dev.	$1.76 \pm 0.08$ (1.70, 1.83)	1.66, 1.91	bc	
	High Dev.	$1.66 \pm 0.14$ (1.54, 1.78)	1.44, 1.85	c	
Trill rate (syllables/sec)	2D $F_{3,1181} = 702.11,$ $P < 0.0001$	Natural	$13.3 \pm 0.5$ (13.3, 13.3)	11.5, 15.6	
		Low Dev.	$16.0 \pm 1.1$ (15.7, 16.3)	14.4, 17.4	
		Medium Dev.	$13.3 \pm 0.4$ (13.2, 13.4)	12.9, 15.0	
		High Dev.	$11.1 \pm 0.4$ (11.0, 11.2)	10.5, 11.8	

Table S.1.3 Continued

Trill rate (syllables/ sec)	2U $F_{3,1004} = 129.56,$ $P < 0.0001$	Natural	$7.3 \pm 0.3$ (7.3, 7.3)	6.0, 8.2	
		Low Dev.	$8.3 \pm 0.2$ (8.2, 8.3)	7.9, 8.6	
		Medium Dev.	$7.4 \pm 0.1$ (7.3, 7.5)	7.2, 7.6	
		High Dev.	$6.4 \pm 0.1$ (6.3, 6.4)	6.3, 6.5	
	A $F_{3,974} = 18.36,$ $P < 0.0001$	Natural	$7.1 \pm 0.5$ (7.8, 7.8)	6.8, 10.3	
		Low Dev.	$9.3 \pm 0.4$ (8.7, 9.9)	8.8, 9.6	
		Medium Dev.	$7.7 \pm 0.1$ (7.5, 7.9)	7.5, 7.8	
		High Dev.	$6.9 \pm 0.1$ (6.8, 7)	6.8, 7.0	
	B $F_{3,348} = 176.96,$ $P < 0.0001$	Natural	$7.7 \pm 0.3$ (7.7, 7.7)	6.9, 8.7	
		Low Dev.	$8.7 \pm 0.5$ (8.5, 8.8)	7.9, 9.6	
		Medium Dev.	$7.7 \pm 0.1$ (7.7, 7.7)	7.6, 7.9	
		High Dev.	$6.9 \pm 0.2$ (6.8, 7.0)	6.5, 7.3	
	Q $F_{3,678} = 35.86,$ $P < 0.0001$	Natural	$8.2 \pm 0.6$ (8.1, 8.2)	6.7, 11.4	
		Low Dev.	$9.9 \pm 0.3$ (9.7, 10.1)	9.3, 10.4	
		Medium Dev.	$8.0 \pm 0.2$ (7.9, 8.2)	7.8, 8.3	
		High Dev.	$7.0 \pm 0.6$ (6.6, 7.4)	6.6, 8.1	
	S $F_{3,432} = 183.20,$ $P < 0.0001$	Natural	$19.8 \pm 0.7$ (19.7, 19.9)	16.2, 21.5	
		Low Dev.	$21.5 \pm 0.4$ (21.3, 21.7)	20.8, 22.4	
		Medium Dev.	$19.9 \pm 0.3$ (19.7, 20)	19.4, 20.4	
		High Dev.	$17 \pm 0.3$ (16.9, 17.2)	16.3, 17.7	
T $F_{3,106} = 8.86,$ $P < 0.0001$	Natural	$14.5 \pm 1.0$ (14.3, 14.7)	12.7, 16.3	ab	
	Low Dev.	$15.8 \pm 0.0$ (15.8, 15.8)	15.8, 15.8	a	
	Medium Dev.	$13.2 \pm 0.1$ (12.1, 14.4)	13.2, 13.3	bc	
	High Dev.	$11.6 \pm 0.0$ (11.6, 11.6)	11.6, 11.6	c	
V $F_{3,341} = 25.60,$ $P < 0.0001$	Natural	$14.3 \pm 0.6$ (14.2, 14.3)	12.7, 16.4		
	Low Dev.	$15.4 \pm 0.3$ (15.2, 15.6)	15.1, 15.9		
	Medium Dev.	$14.4 \pm 0.1$ (14.3, 14.5)	14.2, 14.7		
	High Dev.	$13.0 \pm 0.2$ (12.8, 13.1)	12.8, 13.3		

Table S.1.3 Continued

Pitch (High frequency, kHz)	2D $F_{3,1181} = 25.40$ , $P < 0.0001$	Natural	$5331 \pm 627$ (5292, 5369)	4135, 7668	a
		Low Dev.	$5563 \pm 457$ (5439, 5688)	4688, 6766	b
		Medium Dev.	$5130 \pm 340$ (5037, 5222)	4373, 6029	c
		High Dev.	$4675 \pm 311$ (4590, 4760)	3984, 5472	d
	2U $F_{3,1004} = 16.67$ , $P < 0.0001$	Natural	$5240 \pm 453$ (5212, 5269)	4152, 7418	
		Low Dev.	$5742 \pm 483$ (5515, 5968)	5175, 7017	
		Medium Dev.	$5164 \pm 248$ (5047, 5280)	4760, 5762	
		High Dev.	$4747 \pm 274$ (4619, 4875)	4195, 5172	
	A $F_{3,974} = 5.74$ , $P = 0.0007$	Natural	$5707 \pm 1025$ (5642, 5772)	4110, 8820	a
		Low Dev.	$7669 \pm 194$ (7361, 7977)	7488, 7845	b
		Medium Dev.	$5751 \pm 506$ (4945, 6557)	5203, 6188	a
		High Dev.	$4905 \pm 157$ (4655, 5155)	4746, 5063	a
	B $F_{3,348} = 17.09$ , $P < 0.0001$	Natural	$6595 \pm 457$ (6540, 6651)	5250, 7570	
		Low Dev.	$6933 \pm 260$ (6836, 7030)	6363, 7371	
		Medium Dev.	$6455 \pm 253$ (6361, 6550)	5801, 6879	
		High Dev.	$6181 \pm 327$ (6059, 6303)	5334, 6797	
	Q $F_{3,678} = 10.19$ , $P < 0.0001$	Natural	$4972 \pm 635$ (4923, 5021)	3805, 6969	
		Low Dev.	$5614 \pm 543$ (5225, 6002)	4758, 6633	
		Medium Dev.	$5018 \pm 531$ (4639, 5398)	4371, 5953	
		High Dev.	$4080 \pm 216$ (3925, 4234)	3773, 4305	
S $F_{3,432} = 16.51$ , $P < 0.0001$	Natural	$5007 \pm 566$ (4949, 5065)	3986, 6434	a	
	Low Dev.	$5741 \pm 513$ (5514, 5969)	4576, 6723	b	
	Medium Dev.	$5287 \pm 547$ (5044, 5529)	4316, 6527	c	
	High Dev.	$4670 \pm 535$ (4433, 4907)	3762, 5431	d	
T $F_{3,106} = 0.98$ , $P = 0.40$	Natural	$7022 \pm 332$ (6957, 7086)	6203, 7973	a	
	Low Dev.	$6707 \pm 104$ (5768, 7645)	6633, 6781	a	
	Medium Dev.	$6778 \pm 53$ (6302, 7254)	6741, 6816	a	
	High Dev.	$7086 \pm 38$ (6744, 7429)	7059, 7113	a	

Table S.1.3 Continued

Pitch (High frequency, kHz)	V $F_{3,341} = 3.20$ , P = 0.02	Natural	6481 ± 510 (6425, 6537)	5285, 7934	a
		Low Dev.	6376 ± 289 (6154, 6598)	5888, 6891	ab
		Medium Dev.	6319 ± 654 (5772, 6867)	5550, 7652	ab
		High Dev.	5941 ± 645 (5402, 6481)	4875, 6750	b
Consistency	2D $F_{1,1102} = 224.47$ , P < 0.0001	Natural (1023)	0.84 ± 0.07 (0.83, 0.84)	0.37, 0.93	
		Low Cons. (81)	0.71 ± 0.11 (0.68, 0.73)	0.44, 1.07	
	2U $F_{1,973} = 308.35$ , P < 0.0001	Natural (948)	0.82 ± 0.05 (0.82, 0.83)	0.56, 0.94	
		Low Cons. (27)	0.63 ± 0.14 (0.57, 0.68)	0.42, 0.87	
	A $F_{1,970} = 0.50$ , P = 0.48	Natural (966)	0.80 ± 0.09 (0.80, 0.81)	0.27, 0.93	
		Low Cons. (6)	0.78 ± 0.05 (0.73, 0.83)	0.70, 0.84	
	B $F_{1,299} = 17.17$ , P < 0.0001	Natural (262)	0.82 ± 0.07 (0.81, 0.82)	0.51, 0.92	
		Low Cons. (39)	0.77 ± 0.08 (0.74, 0.79)	0.51, 0.87	
	Q $F_{1,662} = 73.40$ , P < 0.0001	Natural (652)	0.81 ± 0.08 (0.80, 0.82)	0.41, 1.00	
		Low Cons. (12)	0.61 ± 0.16 (0.51, 0.71)	0.36, 0.85	
	S $F_{1,401} = 65.85$ , P < 0.0001	Natural (370)	0.80 ± 0.07 (0.80, 0.81)	0.58, 0.92	
		Low Cons. (33)	0.69 ± 0.12 (0.65, 0.73)	0.47, 0.88	
T $F_{1,105} = 4.00$ , P = 0.05	Natural (104)	0.73 ± 0.11 (0.70, 0.75)	0.35, 0.89		
	Low Cons. (3)	0.60 ± 0.10 (0.34, 0.86)	0.52, 0.72		
V $F_{1,330} = 90.14$ , P < 0.0001	Natural (320)	0.79 ± 0.09 (0.78, 0.80)	0.26, 0.93		
	Low Cons. (12)	0.52 ± 0.14 (0.43, 0.61)	0.37, 0.81		

## CHAPTER 2

### VOCAL DEVIATION AND TRILL CONSISTENCY DO NOT INDICATE MALE QUALITY IN HOUSE WRENS

EMILY CRAMER

#### ABSTRACT

Physically challenging signals are likely to honestly indicate signaler quality. In trilled bird song two physically challenging parameters are vocal deviation (the speed of sound frequency modulation) and trill consistency (how precisely syllables are repeated). They are signals in most bird species tested so far, but differences in selective pressures and song acoustic structure could prevent them from being universal. In particular, there may be opposing selection between song complexity and song performance difficulty, such that in species where song complexity is strongly selected, there may not be strong selection on performance-based traits. I tested whether vocal deviation and trill consistency are signals of male quality in house wrens (*Troglodytes aedon*), a species with complex song structure. Males' singing ability did not correlate with several measures of male quality, except that older males and males that sang at higher rates during playback sang with higher trill consistency. Moreover, song did not relate to polygyny, extra-pair paternity, or annual reproductive success. I conclude that vocal deviation and trill consistency do not signal male quality in this species.

#### INTRODUCTION

In species with traditional sex roles, intrasexual selection favors male traits that enhance their ability to out-compete other males for mating opportunities, and intersexual selection favors

traits that make males more attractive to females (Andersson 1994). Sexual signals are generally thought to be honest signals of male quality, because signal receivers should rapidly evolve to disregard dishonest signals (Searcy and Nowicki 2005, Bradbury and Vehrencamp 2011). For a signal to honestly indicate male quality, there must be a cost of or constraint on signal production that makes it not economical or impossible for low quality males to produce high quality signals (Grafen 1990, Maynard Smith and Harper 1995, reviewed in Searcy and Nowicki 2005, Bradbury and Vehrencamp 2011). Signals that incorporate challenging motor displays may be particularly likely to be costly or constrained, and therefore to be honest signals (Byers et al. 2010). Signal complexity may also be under strong selection (e.g., Catchpole and Slater 2008), and there may be divergent selection pressures such that species that are selected to have more complex songs are not under selection for performance-based signals, while species with strong selection on performance-based signals may not be under strong selection for signal complexity (Cardoso and Hu 2011).

Birds' songs are elaborate signals that probably represent a substantial motor challenge because they involve highly coordinated movements incorporating the respiratory system, the vocal organ (the syrinx), and the upper vocal tract (Suthers 2001, Riede et al. 2006). As such, they have been heavily studied with regard to honest signaling (Vehrencamp 2000, Catchpole and Slater 2008). Vocal deviation and consistency are two aspects of song that have recently received a great deal of attention as potential honest signals of male quality because they appear to represent particularly challenging motor displays.

Vocal deviation is a measure of how quickly the bird modifies sound frequency in a trill, or a series of repeated syllables (Podos 1997, 2001). In trill production, a bird cannot simultaneously maximize frequency bandwidth and trill rate (Podos 1997): a broad frequency

bandwidth requires a large-magnitude change in the volume of the oropharyngeal cavity (Reide et al. 2006) and in beak gape (Hoese et al. 2000), while a high trill rate requires rapid repetition of those changes. Due to mechanical constraints, then, there is an upper limit on the combination of frequency bandwidth and trill rate (Podos 1997), and deviation from this performance limit is thought to reflect trill difficulty. Trills with low vocal deviation from the performance limit combine a relatively broad frequency bandwidth with a relatively fast trill syllable repetition rate and therefore require rapid frequency modulation. Trills with greater vocal deviation relative to the performance limit have relatively narrow frequency bandwidths given their syllable repetition rates, require slower frequency modulation, and are putatively less physically challenging to produce (Podos 1997). A growing number of studies suggest that females prefer males with lower deviation, more challenging trills (Vallet et al. 1998, Draganoui et al. 2002, Ballentine et al. 2004, Christensen et al. 2006, Caro et al. 2010, Cramer et al. 2011). Males discriminate among rivals with different vocal deviation characteristics (Illes et al. 2006, Cramer and Price 2007, de Kort et al. 2009b, Sewall et al. 2010, duBois et al. 2011) and sing with different vocal deviation levels when in different motivational states (e.g., during dawn chorus vs. daytime song, Beebee 2004; and during playback vs. during solo singing, duBois et al. 2009). Vocal deviation correlates with male quality in some species, further suggesting that it is an honest signal (Ballentine 2009, Sockman et al. 2009).

Vocal deviation is based on frequency bandwidth, or the difference between the high and low frequencies. However, song production and perception occur on logarithmic, not linear, scales (Hurly et al. 1992, Goller and Suthers 1996, Cynx 2004, Fletcher et al. 2006), so it may be more appropriate to investigate the ratio of, rather than the difference in, frequencies. This fact has not been addressed in the vocal deviation literature (but see Cramer in review), so here I

investigate both the standard measure of vocal deviation and a measure I call “multiplicative vocal deviation,” which focuses on the tradeoff between frequency ratio and trill rate instead of between frequency bandwidth and trill rate.

Consistency is a measure of how precisely a sound is reproduced each time the bird repeats it, measured at the level of either whole songs or individual, repeated notes. Producing consistent songs and trills might require an especially high degree of integration across multiple brain regions, including the direct motor control of respiratory, syringeal, and vocal tract muscles (Suthers 2001, Jarvis 2004, Byers 2007, Sakata and Vehrencamp 2012). As with vocal deviation, a growing body of literature suggests that consistency may be an important signal of male quality in birds (Byers 2007, Botero et al. 2009, de Kort et al. 2009a, Wegrzyn et al. 2010, Cramer et al. 2011, Rivera-Gutierrez et al. 2011, Rivera-Gutierrez et al. 2012).

While these two song parameters have attracted substantial attention, behavioral ecologists recognize that other aspects of song can also affect the difficulty of song production (Podos et al. 2009, Cardoso et al. 2007), and that sexual selection can favor different traits in different taxa (e.g., Cardoso and Hu 2011). Further study is therefore needed, especially in species with complex song structure, to determine how widely applicable vocal deviation and trill consistency are as salient features of song influencing decision-making by conspecifics. I studied whether vocal deviation, multiplicative vocal deviation, and trill consistency affect male mating success in the house wren (*Troglodytes aedon*). In this species, song is the most probable target of sexual selection: house wrens are dull-colored and sexually monomorphic (though males are slightly larger than females; Johnson 1998). However, males sing elaborate songs that typically begin with relatively low-amplitude introductory notes and end with a series of trills, each composed of a different syllable type. Within each trill, mean pitch is generally fairly

constant, but each succeeding trill usually occurs at a lower mean pitch than the one before (Cramer in review). This flexibility in pitch, combined with the fact that males alter the frequency bandwidth of the trills depending on their pitch, theoretically allows me to disentangle frequency ratio and bandwidth, because the measures can vary more-or-less independently. Males sing at a high rate, especially when they are establishing a new territory and attempting to attract a new mate (Johnson and Kermott 1991). Males sing much more frequently, and with more elaborate song structure, than females (Platt and Ficken 1987, Johnson and Kermott 1990). Females are more likely to visit a nest box from which male song is played than a silent nest box (Johnson and Searcy 1996), though they may value territory characteristics over male or song characteristics in their final mating decisions (Eckerle and Thompson 2006). Fine-scale acoustic differences in song between males, such as vocal deviation and trill consistency, have not been studied in this species (but see Cramer in review).

Variation in reproductive success is higher for male than for female house wrens, largely because some males pair with two females simultaneously, while other males have lower social pairing success (Whittingham and Dunn 2005). Extra-pair (EP) paternity is moderately common, accounting for approximately 15% of offspring (Johnson et al. 2009b), though it does not contribute strongly to variation in male reproductive success in a Wisconsin population of house wrens (Whittingham and Dunn 2005). Since both polygyny and EP paternity can be affected by female preferences for certain male traits and by male competitive ability (e.g., Lightbody and Weatherhead 1987, Lifjeld et al. 1994), I discuss polygyny and EP paternity as variation in “mating success” to avoid making assumptions about the mechanisms driving success.

In a series of steps, I tested the hypotheses that vocal deviation, multiplicative vocal deviation, and trill consistency are honest indicators of male quality that affect mating success

and reproductive success in house wrens. 1) For song characteristics to carry information about an individual, they must reflect consistent differences in singing ability. If these are “reliable” measurements of underlying singing ability, among-male variation in vocal deviation and trill consistency should be greater than within-male variation. Moreover, singing ability in one song type should predict singing ability in other song types (e.g., Ballentine et al. 2004, Cardoso et al. 2009). 2) If underlying singing ability signals male quality, song measures and male phenotypic quality should be correlated. Specifically, male quality measures should negatively correlate with vocal deviation and multiplicative vocal deviation (since lower vocal deviation indicates a more challenging song) and positively correlate with trill consistency. 3) If songs signal quality and influence decisions related to mating, singing ability should relate to mating success. Males with higher success getting matings should sing with lower vocal deviation and multiplicative vocal deviation and with higher trill consistency. 4) For song characteristics to be under selection, they should affect reproductive success. I tested the prediction that males with higher annual reproductive success sing with lower vocal deviation and higher trill consistency. I further tested for relationships between male quality measures and male mating success, and between both of these measures and reproductive success, to assess the value of these measures.

## METHODS

### *Field procedures*

We studied house wrens nesting in boxes at two partially-wooded sites in Ithaca, NY (see Llambias 2009, LaBarbera et al. 2010 for details on study sites and field procedures). I captured, banded, and bled most breeding adults and offspring between April and August 2008-2011. For adults, I measured wing chord (Avinet wing rule, 0.5 mm accuracy), tarsus length (SPI Dial

calipers, 0.1 mm accuracy), and weight (to the nearest 0.1 g with a Pesola spring scale). I monitored all breeding attempts on the field sites and banded chicks at approximately 7 days of age. To prevent premature fledging, I did not continue to count offspring after banding; for reproductive success estimates, I assumed that all banded chicks fledged unless I saw obvious signs of depredation or nestling starvation. Annual reproductive success for each male was the sum of the number of chicks fledged from all his nests, accounting for gains or losses due to extra-pair paternity. Some nests were involved in brood size manipulations, and these males were not included in analyses of total reproductive success.

### *Song measurements*

We conducted playbacks in 2009 and 2010 (see details in Cramer in review), and I used songs from focal males recorded during and immediately preceding those playback trials in analyses. Details on song measurements are given in Cramer (in review). Briefly, I isolated individual songs from each playback recording in Syrinx PC (John Burt). Using only trills that could be assigned to one of the eight common syllable types in the population, I measured frequency range and trill rate for each trill in RavenPro 1.4 and measured trill consistency via cross correlation with SoundXT (Bradbury and Cortopassi 2000; mean  $\pm$  SE, range:  $62.7 \pm 4.14$ , 8-193 trills per male per year, 59 males with 14 males measured in two years, 4580 trills). All recordings were made with a Marantz PMD 690 and Sennheiser ME 67 or MKH 816. I estimated the performance limit of frequency bandwidth and trill rate (i.e., the upper-bound regression of frequency bandwidth regressed on trill rate; Podos 1997), and calculated vocal deviation as the orthogonal distance from each song to this performance limit (Cramer in review). For multiplicative vocal deviation, I conducted the same analysis, but using the frequency ratio in

place of the frequency bandwidth (Cramer in review). All analyses controlled for syllable type, and analyses of standard vocal deviation controlled for the pitch of the trill (measured as the maximum note frequency) by using the residual of the regression of vocal deviation on pitch (Cramer in review). Controlling for pitch should make my results more comparable to other studies' results by removing the confounding factor of pitch. Results presented in the paper were qualitatively unchanged if I analyzed raw, rather than residual, vocal deviation; I use "residual vocal deviation" for clarity to refer to the vocal deviation score based on the frequency bandwidth of the songs.

Five hundred thirty-nine songs were recorded before playback, and the remaining 4041 were recorded during or immediately after playback. Residual vocal deviation does not differ between the pre-playback and the during/post-playback time periods, while trill consistency increases slightly but significantly from pre-playback to during/post playback (unpublished results). To maintain high statistical power, I included all songs in the analyses, but results were qualitatively unchanged if I instead restricted analyses to only songs recorded during/post playback, which should equalize motivational state across males and allow for a more accurate between-male comparison.

### *Male phenotypic quality*

In total, I captured 125 males a total of 253 times over four breeding seasons, and individuals were caught at varying stages of breeding. I measured male size, body condition, age, health, and aggressiveness (since male competitive ability may influence mating success; Lightbody and Weatherhead 1987, Lifjeld et al. 1994). While I do not know if these measures

are salient aspects of male quality in house wrens, I collectively call them “male quality” measures.

Wing and tail measures increased with age, which makes it difficult to combine size measures using principal components analysis: it is not appropriate to include data from multiple captures for only some individuals, but applying a value based on one year’s data to multiple years is also inappropriate. I therefore investigated relationships between song and tarsus, wing, and tail measures individually, using the average measure from within a year if a male was captured more than once in that year. For body condition, I used the standardized residual of a regression of weight on tarsus, controlling for date and time of day captured. I did not re-measure males specifically to calculate measurement repeatability, but measurement repeatability calculated across repeated captures of the same male was highly statistically significant within years (*sensu* Lessells and Boag 1989,  $r > 0.68$ ,  $F > 5.18$ ,  $p < 0.0001$  for tarsus, wing, tail, and weight,  $n = 112-114$  measurements on 50-51 males for tarsus, tail, and weight;  $n = 62$  measures and 28 males for wing). Measurements were also repeatable between years ( $r > 0.46$ ,  $F > 2.67$ ,  $p < 0.002$ ,  $n = 78-87$  measures of 34-37 males).

We could assign age only for a subset of individuals (84 male-years) that had been banded on-site in a previous year. I categorized males as second-year if they had been banded as nestlings the previous year (*i.e.*, this was their first breeding season) and after-second-year if they had been banded as adults in a previous season. I did not make finer-scale age assignments among after-second-year males that were present multiple years.

In 2009, I used two ecoimmunology techniques to assess male health (see Cramer et al. in review for details). Briefly, I followed procedures in Millet et al. (2007) to measure the bactericidal capacity of whole blood samples collected from the brachial vein after ethanol

sterilization. I also took blood smears and had the ratio of heterophiles:lymphocytes counted by the Animal Health Diagnostic Center at Cornell University Veterinary College (Ots et al. 1998). Scores for the bactericidal assay are thought to increase with improved innate immunity (Millett et al. 2007), while scores for the heterophile:lymphocyte ratio increase in response to stress (Ots et al. 1998).

I derived aggression scores from playback experiments conducted in 2008, 2009, and 2010 as part of other studies (Cramer accepted and Cramer in review). Briefly, for each playback experiment, I presented males with on-off bouts of playback, with a single stimulus song repeated consecutively within each on-bout, and with each male hearing a different stimulus set. In 2008, a single song stimulus was repeated for all six on-off bouts (Cramer accepted). In 2009 and 2010, each on-bout had a different manipulation of a single song, but I found no evidence that males responded differently depending on how the stimulus had been manipulated for that on-bout (Cramer in review). Each year's experiment had other unique attributes (e.g., speaker brand and distance to the nest box) that could affect responses to playback, so I always included a year/experiment variable in analyses. For all playback experiments, I measured responses in terms of the number of songs the focal male gave, how much time he spent within 5 m of the speaker, and how many times he flew within a 2 m ring across the speaker. I calculated the mean song rate across all on-off bouts for each male; song rate was not strongly correlated with either flights across the speaker or time spent close to the speaker. The latter two variables each had a large number of zero values and were highly correlated (Spearman rank correlation  $\rho = 0.53$ ,  $p < 0.001$ ), so I combined them. Because they were on different scales (0-1 for proportion of time close to the speaker, and approximately 0-10 for number of flights across the speaker), I first normalized each of these response variables to 1 within a year. I then averaged across both

response variables and across all on- and off-bouts of playback to get a single score for each male. Largely similar results were obtained if I instead used mean rate of flights across the speaker and mean proportion of time within 5 m of the speaker as independent response variables, but these could not be transformed for normality, so I preferred the combined approach score response. Neither song rate ( $r = 0.132$ ,  $F_{18,22} = 1.30$ ,  $p = 0.27$ ) nor approach response ( $r = 0.000$ ,  $F_{18,22} = 1.0$ ,  $p = 0.49$ ) were repeatable across experiments (sensu Lessells and Boag 1987; calculated with 19 males exposed to 2 or 3 playback experiments each, and a total of 41 playbacks).

I included only a single capture event per year for analyses of how male quality relates to song and mating success, using the capture closer in time to the time of song recording (45 captures were on the same day as song recording and 26 banded males were recorded on average  $26 \pm 3.6$  days before capture). Fourteen males were recorded in two years, and two banded males were recorded and monitored for breeding, but not captured in the year of song recording. For analyses of how body condition relates to mating success, I included the first capture event for each male; results were not changed if I instead randomly chose a capture event to include. For the ecoimmunology measures, more data were available from the first captures, so I used only first captures of males captured twice.

#### *Paternity analysis and male mating success*

I followed the protocol of LaBarbera et al. (2010) and genotyped all individuals using a panel of 7 microsatellite markers. I conducted paternity analysis using Cervus 3.0, including the social mother as a known parent (Kalinowski et al. 2007), and located mismatching loci using GenoPed (Z. Zhang). I re-genotyped mismatching loci to confirm allele calls. To most

conservatively estimate EP paternity, I attributed a chick to EP paternity if it had more than one mismatch with its social father. In assigning EP sires, I allowed EP fathers to have a single null-allele mismatch (see Dakin and Avise 2004) with his putative offspring. Three nests were provisioned by males that I had not observed advertising at the box, and that were not the genetic sires of the young in the nest. In this case, I cannot distinguish rapid mate-switching during the female's fertile period from cuckoldry, and I excluded these males from analyses of extra-pair success. Two males were excluded from analysis of reproductive success and maintaining WP success because a majority of their offspring had two null-allele mismatches. Males that were extra-pair sires of chicks in these five nests were included in analysis of success in cuckolding other individuals, and these chicks were included towards their annual reproductive success.

I considered a male polygynous if he attracted a female to a second nest box while his primary female still had an active nest (i.e., simultaneously polygyny, as in Soukup and Thompson 1997; results remain qualitatively unchanged if I also consider males polygynous if they had different females for each brood). Soukup and Thompson (1997) found a higher rate of EP paternity in the secondary nests of polygynous males, perhaps indicating a reduced efficacy in mate-guarding of the second female. In this study, however, I found no evidence that polygynous males were more likely to be cuckolded in their secondary than in their primary nests, nor that the rate differed from monogamous males ( $n = 98$  males, 179 nests, modeled the likelihood of being cuckolded as a function of nest type with year as a covariate and male ID as a random effect, effect of nest type:  $F_{2,173} = 0.58$ ,  $p = 0.56$ ). For monogamous males, 37% (45/121) of nests contained at least one EP offspring; 45% (13/29) of polygynous males' primary nests and 31% (9/29) of polygynous males' secondary nests contained EP offspring. Thus, I found it unlikely that WP paternity merely indicated mate-guarding efficacy.

I considered a male to have been cuckolded if one or more offspring in his social nest was sired by an EP male. Results of analyses were unchanged if I excluded nests where the male was captured or a playback was conducted during the putative fertile period of his mate (i.e., from two weeks before the first egg was laid until the day the last egg was laid). I also noted whether or not each male that bred on-site was an extra-pair sire of at least one chick in another nest. For this measure, I excluded males that deserted the study site immediately after capture. For both WP paternity loss and EP paternity gain, I treated each year as a separate measurement for males present in multiple years.

I also conducted pair-wise tests of EP sires to the WP males they cuckolded in cases where the EP sire could be identified. For ease of discussion, I consider success in attracting a secondary female, success in maintaining WP paternity, and success in gaining EP paternity in other nests as having higher mating success. I also consider the EP sires more successful than the WP males they cuckolded in paired comparisons.

### *Statistical analysis*

I performed two analyses to assess whether my song measures reflected underlying singing ability. First, I asked whether a male's singing ability in one syllable type predicted his singing ability in other syllable types. For this analysis, I assessed pair-wise correlations for residual vocal deviation, multiplicative vocal deviation, and trill consistency between all pairs of syllable types within males, using the mean vocal deviation and trill consistency for that syllable type for that male. For the 14 males with song measurements in both years, I randomly chose one year's values to be included. Second, I ran models with and without random effects of male identity (sensu Nakagawa and Schielzeth 2010). In one model, I assessed repeatability across all

trills measured by using either residual vocal deviation, multiplicative vocal deviation, or trill consistency as the response measure, with syllable type and year as fixed effects. To investigate repeatability between years, I ran the same model, but only used syllable types that were measured in two years for single males (a total of 14 males were recorded in both years, encompassing 1607 trills).

I assessed whether residual vocal deviation, multiplicative vocal deviation, and trill consistency were associated with male quality measures by fitting general linear models with the song measure as the response variable; year, syllable type, and a single male quality measure as fixed effects; and male identity as a random effect. Because the heterophile:lymphocyte ratio and the bactericidal assay were analyzed in a single year and a single time per male, those models did not include year effects or male identity terms.

Next, I assessed whether any of the male or song quality measures was related to mating success. I constructed a separate model for each measure of mating success (polygynous vs. monogamous, cuckolded vs. not cuckolded, and EP sire vs. not an EP sire in other nest) and for each male quality or song measure. Data were missing from different variables for different males, so constructing separate models allowed me to maximize sample size for each analysis. Each model used the song or male quality measure as a response variable and included as predictors a measure of mating success, a fixed effect of syllable type (for song analyses only) and year (except for health measures), and a random effect of male identity (except for health measures). I use the same process to test whether song quality, male quality, and male mating success related to annual reproductive success, using reproductive success as a predictor variable because it could not be transformed for normality to allow it to be a response variable. Using reproductive success as a predictor rather than a response variable also allowed me to control for

syllable type in analyses of song. While this approach reverses the logical dependent and independent variable, the goal of the analysis is to measure the association between the two variables, and the strength of the association should be unaffected by which variable I use for the dependent vs. independent variable.

For paired comparisons of EP males to the WP males they cuckolded, I conducted paired t-tests for each male quality measure. For song measures, because I had many measurements for each male, I constructed models to predict residual vocal deviation, multiplicative vocal deviation, or trill consistency, with syllable type and role (EP vs. WP) as fixed effects, and random effects of male identity and a grouping variable to associate EP males with the WP males they cuckolded.

Residual vocal deviation, multiplicative vocal deviation, trill consistency, size, body condition, and song rate in response to playback approached normal distributions, and transformations did not improve the distribution. The ratio of heterophiles:lymphocytes was log-transformed, and approach scores were arc-sine square-root transformed for normality. The percent bacteria killed was strongly skewed and could not be transformed for normality. All tests were performed in JMP 7.0 except for analyses with a categorical response variable and a random predictor variable (e.g., whether male age differed between different levels of attractiveness; whether reproductive success related to categorical mating success variables). These tests were performed in SAS 9.2, using PROC GLIMMIX with a binary error distribution.

To correct for multiple testing, I used false discovery rate (Benjamini and Hochberg 1995), implemented in R version 2.9.2 (R [Development](#) Core Team 2009). I conducted table-wise corrections. P-values listed in the tables are un-corrected, but I note in the text whenever statistical significance changed following correction for multiple testing.

## RESULTS

### *Paternity*

Across all years, 13.7% (118/863) of offspring in 37.7% (69/183) of nests were extra-pair young. Single pair-wise mismatches consistent with being null alleles occurred in over 40 chicks for each parental sex. Sixteen of 863 offspring had single mismatches with their social mothers that could not be attributed to null alleles, and 9/863 offspring had single non-null mismatches with their social fathers. Intra-specific brood parasitism has not been reported in this species (Soukup and Thompson 1997, Johnson et al. 2002), so I allowed these single mismatches with the putative parents. For five males, the chick with the single mismatch was the only chick in that year with any mismatches, and so this paternity call determined whether the male was considered cuckolded or not. Changing the cuckoldry status of these males did not substantially affect the outcome of other tests. The other four males with chicks with single mismatches had other chicks that could more definitively be attributed to extra-pair paternity.

The likelihood of maintaining full paternity in his own nest was not related to a male's likelihood of gaining EP offspring in other nests ( $n = 131$  nests with WP success data, 98 males, effect of EP sire status  $F_{1,126} = 0.26$ ,  $p = 0.61$ ). For 39.1% (27/69) of nests, the social father of the nest was not cuckolded and was an EP sire in another nest, whereas for 37.1% (23/62) of nests, the social father was cuckolded but was also an EP sire elsewhere.

### *Reliability of song measures*

Vocal deviation and trill consistency weakly, but reliably, indicated underlying singing ability. In general, a male's residual vocal deviation and trill consistency in each syllable type

weakly but consistently predicted his residual vocal deviation and trill consistency in all other syllable types (Figure 2.1). Twenty-four of 28 pairwise correlations were positive for residual vocal deviation, and 26 of 28 were positive for trill consistency. Obtaining this proportion of positive correlations by chance is highly unlikely (binomial test,  $p < 0.001$  for both song measures). For multiplicative vocal deviation, 8 of the 28 correlations were negative, and only two of the positive correlations were significant (none of the negative correlations were; data not shown). Pairwise correlation coefficients were significantly higher on average for residual vocal deviation than for multiplicative vocal deviation (paired t-test, mean for residual vocal deviation 0.154, mean for multiplicative vocal deviation 0.099,  $t_{27} = 2.54$ ,  $p = 0.02$ )

Male identity explained a small to moderate amount of variation in song measurements. In the full models including all songs and syllable type as a covariate, male identity explained 19.6% of the variation in residual vocal deviation, 13.4% of the variation in multiplicative vocal deviation, and 24.4% of the variation in trill consistency. When I only included song types that I had measured from the same male in both years, male identity explained 21.7, 8.9, and 16.8% of the variation in song measures, respectively. Repeatability was highly statistically significant for residual vocal deviation and trill consistency ( $p < 0.0001$ ).

#### *Correlations between song quality and male quality*

Few correlations between song quality and male quality were statistically significant (Table 2.1). Residual vocal deviation and multiplicative vocal deviation correlated with tail length, such that males with longer tails sang lower-deviation (i.e., more challenging) songs (Table 2.1), though these results were not robust to correction for multiple testing. Trill consistency correlated with age, size, and vocal response to playback: males with more consistent trills were older, larger,

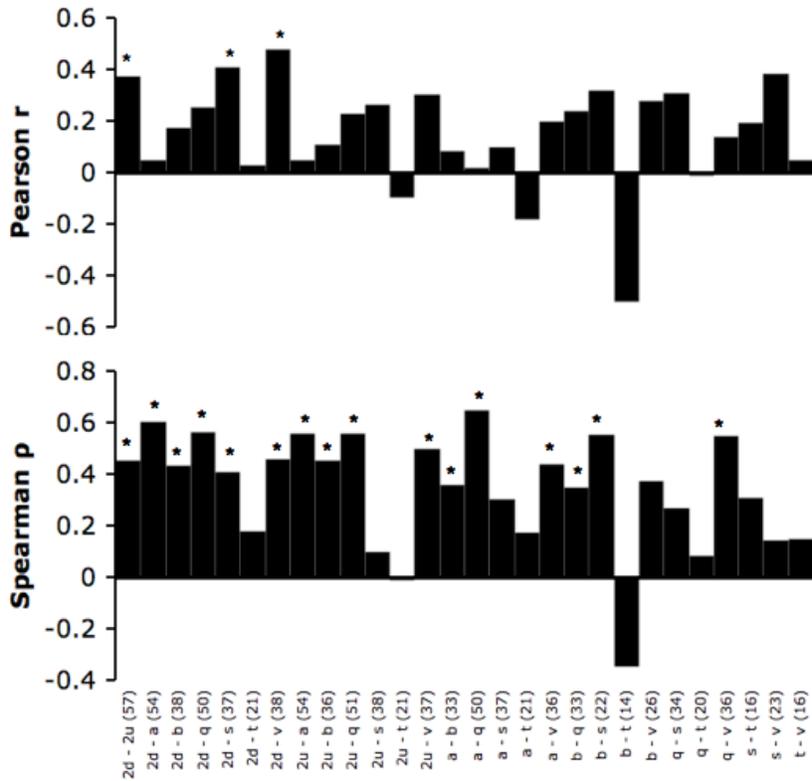


Figure 2.1. Pair-wise correlations of (a) residual vocal deviation and (b) trill consistency measurements across males and across the eight syllable types common in the population. I calculated the mean residual vocal deviation or trill consistency for each type for each male, and performed correlations on these values. A positive correlation therefore indicates that males that on average sing one syllable type well also sing the other syllable type relatively well on average. The pair of syllable types correlated and number of males included is listed on the x-axis. Sample sizes vary because I did not measure all syllable types for all males. Bars marked with an asterisk were statistically significant correlations before correcting for multiple testing ( $p < 0.05$ ). I used Spearman correlation for trill consistency because it was not normally distributed. The probability of getting 24/28 (residual vocal deviation) and 26/28 (trill consistency) positive correlations by chance is very low ( $p < 0.001$ ).

Table 2.1. Estimated effects of putative male quality measures on vocal deviation and trill consistency. Effects that remained statistically significant after correction for multiple testing are in bold.

Male Quality Measure	N males (trills)	Song Measurement	Estimate <sup>†</sup> ± SE (95% CI)	t <sub>df</sub> (p)
Tarsus	58 (4450)	Resid Voc Dev*	0.173 ± 0.145 (-0.1154, 0.460)	t <sub>101.7</sub> = 1.20 (0.23)
		Mult Voc Dev	-0.212 ± 0.258 (-0.726, 0.301)	t <sub>80.3</sub> = -0.82 (0.41)
		Consistency	0.003 ± 0.011 (-0.020, 0.026)	t <sub>121.6</sub> = 0.26 (0.79)
Wing	58 (4450)	Resid Voc Dev*	-0.024 ± 0.035 (-0.093, 0.045)	t <sub>271.5</sub> = -0.68 (0.50)
		Mult Voc Dev	-0.023 ± 0.067 (-0.155, 0.109)	t <sub>167.5</sub> = -0.34 (0.73)
		<b>Consistency</b>	<b>0.017 ± 0.003 (0.012, 0.023)</b>	<b>t<sub>413.8</sub> = 6.54 (.0001)</b>
Tail	58 (4450)	Resid Voc Dev*	-0.07 ± 0.037 (-0.148, -0.001)	t <sub>88.75</sub> = -2.01 (0.05)
		Mult Voc Dev	-0.153 ± 0.065 (-0.284, -0.023)	t <sub>78.01</sub> = -2.35 (0.02)
		Consistency	0.005 ± 0.003 (-0.001, 0.011)	t <sub>100.2</sub> = 1.80 (0.08)
Condition	58 (4450)	Resid Voc Dev*	-0.023 ± 0.113 (-0.246, 0.200)	t <sub>128.3</sub> = -0.20 (0.84)
		Mult Voc Dev	-0.125 ± 0.206 (-0.535, 0.285)	t <sub>94.3</sub> = -0.61 (0.55)
		Consistency	-0.006 ± 0.009 (-0.023, 0.012)	t <sub>162.2</sub> = -0.63 (0.53)
Age	34 (2667)	Resid Voc Dev*	0.029 ± 0.063 (-0.154, 0.095)	t <sub>505.1</sub> = -0.46 (0.64)
		Mult Voc Dev	0.125 ± 0.123 (-0.115, 0.366)	t <sub>354.7</sub> = 1.02 (0.31)
		<b>Consistency*</b>	<b>0.019 ± 0.004 (0.011, 0.028)</b>	<b>t<sub>367.1</sub> = 4.52 (0.0001)</b>
Heterophile : lymphocyte ratio	43 (2763)	Resid Voc Dev	-0.082 ± 0.095 (-0.275, 0.110)	t <sub>39.8</sub> = -0.86 (0.39)
		Mult Voc Dev	0.051 ± 0.168 (-0.290, 0.391)	t <sub>38.96</sub> = 0.30 (0.77)
		Consistency	0.007 ± 0.007 (-0.008, 0.021)	t <sub>39.3</sub> = 0.90 (0.37)
Bactericidal capacity	42 (2687)	Resid Voc Dev	0.074 ± 0.294 (-0.519, 0.668)	t <sub>40.54</sub> = 0.25 (0.80)
		Mult Voc Dev	0.658 ± 0.523 (-0.399, 1.716)	t <sub>39.60</sub> = 1.26 (0.22)
		Consistency	-0.008 ± 0.021 (-0.051, 0.035)	t <sub>38.92</sub> = -0.37 (0.71)
Song Rate in Playback	58 (4541)	Resid Voc Dev	0.001 ± 0.014 (-0.027, 0.029)	t <sub>879.4</sub> = 0.07 (0.95)
		Mult Voc Dev	-0.041 ± 0.028 (-0.096, 0.014)	t <sub>499.2</sub> = 1.47 (0.14)
		<b>Consistency</b>	<b>0.007 ± 0.001 (0.005, 0.009)</b>	<b>t<sub>1100</sub> = 7.38 (0.0001)</b>
Approach score	58 (4541)	Resid Voc Dev	-0.032 ± 0.188 (-0.401, 0.337)	t <sub>1087</sub> = -0.17 (0.86)
		Mult Voc Dev	-0.379 ± 0.377 (-1.120, 0.362)	t <sub>605</sub> = 1.00 (0.32)
		Consistency	0.024 ± 0.014 (-0.003, 0.050)	t <sub>1501</sub> = 1.75 (0.08)

\*Models with a significant (p < 0.05) year.

<sup>†</sup>For 1 unit change in predictor, size and direction of change in the y variable.

and sang more to playback (Table 2.1). While those relationships remained statistically significant following correction for multiple testing, the tendency for males with more consistent trills to approach playback more closely was not significant after correction (corrected  $p = 0.28$ ).

More detailed analyses showed that the apparent effect of wing and tail length on trill consistency is likely driven by age effects on both size and consistency: older males are larger and also sing with higher consistency than second-year males. After-second year males had longer wings (least squares mean  $\pm$  SE wing for after-second-year males  $51.2 \pm 0.17$  mm, for second year males  $49.7 \pm 0.36$ ,  $F_{1,63.33} = 15.63$ ,  $p = 0.0002$ ,  $n = 80$  observations) and tended to have longer tails (least squares mean  $\pm$  SE tail for after-second-year males  $43.8 \pm 0.18$  mm, for second year males  $43.1 \pm 0.35$ ,  $F_{1,48.22} = 3.79$ ,  $p = 0.06$ ,  $n = 80$  observations) than second-year males. When I simultaneously assessed the effect of age and wing chord on trill consistency, age had an effect (least squares mean after-second-year consistency  $0.809 \pm 0.005$ , second-year consistency  $0.765 \pm 0.012$ ,  $F_{1,63.67} = 13.34$ ,  $p = 0.0005$ ,  $n = 2537$  songs,  $n = 32$  males, with 2 males appearing in both years), while wing chord did not (estimated effect  $\pm$  SE of wing on consistency  $-0.003 \pm 0.004$ ,  $F_{1,35.7} = 0.26$ ,  $p = 0.62$ ). Controlling for age also made the trend between tail length and trill consistency not significant, and changed the sign of the trend (effect of tail:  $-0.004 \pm 0.004$ ,  $F_{1,37.87} = 1.09$ ,  $p = 0.30$ ), while the age effect remained significant in this model ( $p < 0.0001$ ). Age effects did not appear to drive the trend between tail length and vocal deviation and multiplicative vocal deviation: controlling for age in those models did not substantially change the parameter estimates, though it did reduce the significance of the tail-song relationships ( $p = 0.11$  for multiplicative vocal deviation,  $p = 0.21$  for residual vocal deviation).

The original, fully-parameterized model examining trill consistency and age showed a significant year effect (Table 2.1), but this year effect probably did not drive the increase in consistency with age. If the apparent age effect were due to systematic differences in consistency measurements between years rather than due to actual age effects, the difference in years should be apparent when testing only same-age birds. I re-tested the year effect only within after-second year birds, and found no effect of year in this reduced model ( $r^2 = 0.002$ ,  $n = 35$  males,  $F_{1,33} = 0.068$ ,  $p = 0.80$ ; sample sizes are too small to powerfully test year effects within the second-year age class). Rather, the year effect is likely due in part to a different composition of age classes across years: in 2009 (LSM trill consistency  $0.798 \pm 0.006$ ), 18.2% (4/22) known-age birds were second-year, while in 2010 (trill consistency  $0.778 \pm 0.007$ ), only 5.6% (1/18) known-age birds was second-year. The higher proportion of second-year birds singing less consistent trills in 2009 could have caused the lower overall trill consistency in that year.

Although my sample sizes were too small for strong longitudinal analyses of within-individual changes in trill consistency, I investigated these within-individual patterns to see if they were consistent with the between-individual patterns described above. Because syllable type affects trill consistency, for each male separately, I excluded types that I only measured in one year. For the two birds recorded both as second-years (in 2009) and after-second-years (in 2010), trill consistency increased with age, consistent with the between-individual analysis (LSM SY  $0.755 \pm 0.007$ ; ASY  $0.781 \pm 0.008$ ;  $F_{1,280} = 7.48$ ,  $p = 0.007$ ,  $n = 289$ ). Four males were recorded as after-second-year males in both 2009 and 2010, and their trill consistency was slightly but significantly lower in 2010 than 2009, the opposite pattern from the increase in consistency from the second-year to after-second year transition (mean  $\pm$  SE 2009  $0.831 \pm 0.015$ , 2010  $0.812 \pm 0.015$ ,  $F_{1,564,2} = 9.40$ ,  $p = 0.002$ ,  $n = 573$  trills).

Most measures of male quality were not strongly inter-correlated in simple regressions, with the following exceptions. Wing chord correlated positively with body condition, tarsus, and tail. The correlation between tarsus and tail only approached significance ( $p = 0.07$ ). Age only affected wing and tail measures (see above).

*Song and male quality: relation to male mating success*

Song quality related to two of the four measures of mating success, but in the opposite direction from predicted. Males that lost paternity in their social nests had lower residual vocal deviation and higher trill consistency—both putatively “better” song characteristics—than males that were not cuckolded (Table 2.2, Figure 2.2). Similarly, males that gained EP offspring in other nests on site had higher residual vocal deviation and multiplicative vocal deviation than males that succeeded in siring EP offspring on site, although the males that gained EP paternity did have higher trill consistency (Table 2.2, Figure 2.2). Vocal deviation and trill consistency did not differ between polygynous and monogamous males or between EP males and the WP males they cuckolded in paired comparisons (Table 2.2, 2.3).

Measures of male quality were largely un-related to mating success. Males that approached playback more were more likely to be polygynous (Table 2.2), but the effect was not robust to correction for multiple testing (adjusted  $p = 0.29$ ). EP males did not differ from the WP males they cuckolded, except that WP males sang at a higher song rate in response to playback (Table 2.3). Again, this relationship was not robust to correction for multiple testing (adjusted  $p = 0.55$ ).

The effect of age on maintaining WP success could not be estimated in the random effects model because the model did not converge, but a chi-squared test treating each

Table 2.2. Estimated associations between mating success and male or song quality measures. Effects that remained significant after correction for multiple testing are in bold. Parameter estimates are the difference from the reference category to the grand mean across all individuals, with the mating success variable treated as a predictor. Reference categories were as follows: was monogamous, was cuckolded, and was not an EP sire. Therefore, positive scores indicate that more successful males had higher male/song quality scores (note that a higher score for vocal deviation indicates lower song quality). Age was a categorical outcome variable, and the probability of being after-second year was modeled with logistic regression. Estimates for age indicate the change in log-odds of being after-second year with changing from polygynous to monogamous or from being an EP sire in other nests to not being one. H:L refers to the ratio of heterophiles:lymphocytes.

Mating success metric	Quality or Song trait	N males (obs.)	Estimate ± SE (95% CI)	T <sub>df</sub> (p)
Polygyny	Tarsus	120 (164)	0.024 ± 0.021 (-0.018, 0.067)	t <sub>49.93</sub> = 1.15 (0.25)
	Wing	120 (164)	0.116 ± 0.112 (-0.107, 0.338)	t <sub>76.23</sub> = 1.04 (0.30)
	Tail	120 (163)	0.054 ± 0.092 (-0.130, 0.238)	t <sub>50.55</sub> = 0.59 (0.56)
	Condition	120 (164)	-0.032 ± 0.039 (-0.110, 0.047)	t <sub>89.87</sub> = -0.80 (0.42)
	Age	56 (79)	0.76 ± 0.88 (-0.995, 2.514)	t <sub>74</sub> = 0.86 (0.39)
	H:L	39	-0.009 ± 0.145 (-0.303, 0.285)	t <sub>37</sub> = -0.06 (0.95)
	Bactericidal assay	39	0.026 ± 0.047 (-0.068, 0.121)	t <sub>37</sub> = 0.57 (0.58)
	Song rate in playback	64 (85)	0.136 ± 0.322 (-0.506, 0.777)	t <sub>76.0</sub> = 0.42 (0.67)
	Approach Score	64 (85)	0.064 ± 0.032 (0.000, 0.128)	t <sub>73.9</sub> = 1.99 (0.0502)
	Vocal Deviation	50 (3907)	-0.040 ± 0.032 (-0.102, 0.022)	t <sub>1588</sub> = -1.27 (0.20)
	Mult Voc Dev	50 (3907)	-0.095 ± 0.064 (-0.220, 0.029)	t <sub>1026</sub> = -1.50 (0.13)
Consistency	50 (3907)	-0.002 ± 0.002 (-0.006, 0.002)	t <sub>1985</sub> = -0.87 (0.38)	
Maintaining WP success †	Tarsus	96 (127)	0.017 ± 0.028 (-0.039, 0.073)	t <sub>50.18</sub> = 0.61 (0.54)
	Wing	96 (127)	-0.129 ± 0.110 (-0.348, 0.089)	t <sub>75.56</sub> = -1.18 (0.24)
	Tail	96 (127)	-0.107 ± 0.117 (-0.340, 0.127)	t <sub>61.23</sub> = -0.91 (0.37)
	Condition	96 (127)	-0.036 ± 0.043 (-0.121, 0.048)	t <sub>114.2</sub> = -0.86 (0.39)
	H:L	34	0.065 ± 0.143 (-0.356, 0.226)	t <sub>32</sub> = -0.45 (0.65)
	Bactericidal assay	34	-0.024 ± 0.042 (-0.110, 0.062)	t <sub>32</sub> = -0.57 (0.57)
	Song rate in playback	54 (69)	-0.341 ± 0.320 (-0.980, 0.298)	t <sub>62.1</sub> = -1.07 (0.29)
	Approach Score	54 (69)	-0.043 ± 0.033 (-0.108, 0.023)	t <sub>60.8</sub> = -1.30 (0.20)
	<b>Vocal Deviation</b>	<b>39 (3037)</b>	<b>0.245 ± 0.057 (0.132, 0.359)</b>	<b>t<sub>146.9</sub> = 4.28 (0.0001)</b>
	Mult Voc Dev	39 (3037)	0.201 ± 0.105 (-0.007, 0.409)	t <sub>103.5</sub> = 1.91 (0.06)
<b>Consistency</b>	<b>39 (3037)</b>	<b>-0.017 ± 0.004 (-0.025, -0.009)</b>	<b>t<sub>190.1</sub> = -4.06 (0.0001)</b>	

Table 2.2 Continued

	Tarsus	124 (168)	$-0.001 \pm 0.021$ (-0.042, 0.041)	$t_{59.56} = 0.04$ (0.97)
	Wing	124 (168)	$0.087 \pm 0.101$ (-0.113, 0.287)	$t_{96.33} = 0.87$ (0.39)
	Tail	124 (167)	$0.026 \pm 0.092$ (-0.158, 0.210)	$t_{65.33} = 0.28$ (0.78)
	Condition	124 (168)	$0.030 \pm 0.035$ (-0.040, 0.100)	$t_{116.5} = 0.85$ (0.40)
Success	Age	55 (78)	$0.254 \pm 0.79$ (-1.318, 1.827)	$t_{73} = 0.32$ (0.75)
Siring EP	H:L	39	$0.076 \pm 0.126$ (-0.179, 0.330)	$t_{39} = 0.6$ (0.55)
Offspring in	Bactericidal assay	40	$0.023 \pm 0.037$ (-0.052, 0.097)	$t_{39} = 0.62$ (0.54)
Other Nests	Song rate in playback	63 (84)	$0.023 \pm 0.295$ (-0.564, 0.610)	$t_{79.9} = 0.08$ (0.94)
	Approach Score	63 (84)	$0.019 \pm 0.030$ (-0.040, 0.079)	$t_{80} = 0.65$ (0.52)
	<b>Vocal Deviation</b>	<b>49 (3858)</b>	<b><math>0.121 \pm 0.032</math> (0.057, 0.184)</b>	<b><math>t_{1433} = 3.74</math> (0.0002)</b>
	<b>Mult Voc Dev</b>	<b>49 (3858)</b>	<b><math>0.201 \pm 0.064</math> (0.075, 0.327)</b>	<b><math>t_{746.9} = 3.14</math> (0.002)</b>
	<b>Consistency</b>	<b>49 (3858)</b>	<b><math>0.008 \pm 0.002</math> (0.003, 0.012)</b>	<b><math>t_{1772} = 3.48</math> (0.0005)</b>

†The model including age did not converge.

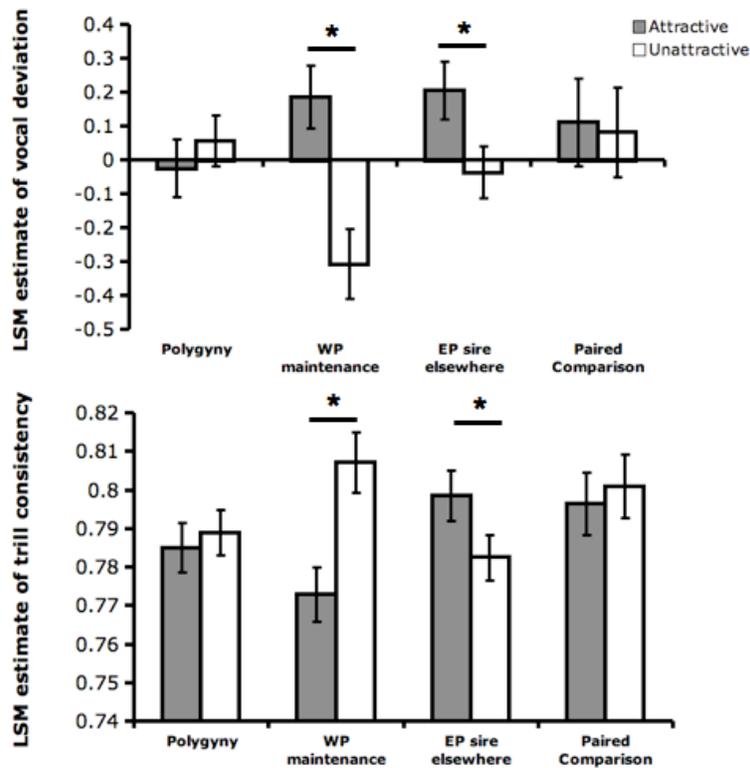


Figure 2.2. Comparisons of the least squares mean  $\pm$  SE of song quality (a, residual vocal deviation; b, trill consistency) for males that differ in their apparent attractiveness to females. I compared males that were vs. were not successful in simultaneously socially pairing with two females; males that were vs. were not successful in maintaining WP paternity within their social broods; males that were vs. were not successful in gaining EP offspring in other broods on site; and, in a paired comparison, EP males to the WP males they cuckolded (“attractive” vs. “unattractive”, in grey and white, respectively). Song quality differed significantly between cuckolded and uncuckolded males, as well as between males that did and did not gain EP offspring in other nests. Note that lower scores correspond to more physically challenging (“better”) songs for residual vocal deviation, but higher scores are better for trill consistency.

Table 2.3. Paired comparisons of male quality and song quality measures, between EP males and the WP males they cuckolded. Test statistics refer to paired tests. For male phenotype measures, I used paired t-tests. For song measures, I constructed a model with male identity and a grouping variable as random effects, with fixed effects of syllable type, year, and role (EP or WP). No effects were statistically significant after correcting for multiple testing.

Male trait	Mean for WP male	Mean for EP male	t <sub>df</sub> (p)	N pairs (males)
Tarsus	16.84	16.82	t <sub>68</sub> = 0.23 (0.82)	69 (74)
Wing	50.69	50.55	t <sub>68</sub> = 0.72 (0.48)	69 (74)
Tail	43.6	43.2	t <sub>68</sub> = 1.32 (0.19)	69 (74)
Condition	-0.026	0.007	t <sub>68</sub> = -0.45 (0.66)	69 (74)
Heterophile : Lymphocyte*	-0.476	-0.548	t <sub>14</sub> = 0.41 (0.69)	15 (23)
Bactericidal assay	0.889	0.876	t <sub>15</sub> = 0.20 (0.85)	16 (25)
Residual vocal deviation	0.082	0.112	t <sub>199.6</sub> = 0.38 (0.71)	18 (25)
Mult Voc Dev	0.593	0.596	t <sub>164.5</sub> = 0.33 (0.74)	18 (25)
Consistency	0.801	0.796	t <sub>202.1</sub> = -0.89 (0.37)	18 (25)
Song Rate in Playback	7.130	5.897	t <sub>27</sub> = 2.06 (0.05)	28 (31)
Approach Score	0.271	0.246	t <sub>27</sub> = 0.45 (0.66)	28 (31)

\*values log-transformed for normality

\*\* 2300 observations were included (for males present more than once in the comparisons, all trills measured for that male for that year were duplicated)

observation as an independent measure shows no relationship between them: 45% (26/57) of after-second-year birds and 42% (3/7) second-year birds were cuckolded ( $\chi^2_1 = 0.019$ ,  $p = 0.89$ ). For paired comparisons of EP and WP sires, I had age data on both males for only 19 pairs; in 16 pairs, both males were the same age, in two pairs, a younger male cuckolded an older male, and in the final pair, the older male cuckolded the younger pair.

*Song quality, male quality, and male mating success: relation to annual reproductive success*

Although song measures and the majority of male quality measures did not correlate with reproductive success, males that approached playback more strongly had higher reproductive success (Table 2.4). This relationship became non-significant (adjusted  $p = 0.11$ ) with correction for multiple testing. While mating success based on EP paternity did not correlate with reproductive success, polygyny did: males that were polygynous had higher reproductive success (Table 2.4).

## DISCUSSION

These results do not support the hypothesis that house wrens use low vocal deviation and high trill consistency as honest signals of high male quality. Although songs with low vocal deviation and high trill consistency are thought to be physically challenging to produce (Podos 1997, Sakata and Vehrencamp 2012) and therefore should be honest signals of male quality that affect mating success and reproductive success (Byers et al. 2010), I found no support for this hypothesis despite reasonable sample sizes and extensive recording. Residual vocal deviation (i.e., controlled for pitch) and trill consistency did appear to reflect some underlying singing ability, since song measures were repeatable and predictive across syllable types (Figure 2.1).

Table 2.4. Relationships between annual reproductive success and song quality, male quality, and male mating success. Effects that were significant after correcting for multiple testing are in bold. Estimated coefficients from logistic model.

Quality or Song trait	N males (obs.)	Estimate $\pm$ SE (95% CI)	T <sub>df</sub> (p)
Vocal Deviation	36 (2470)	-0.009 $\pm$ 0.018 (-0.045, 0.026)	t <sub>118.4</sub> = -0.52 (0.60)
Mult Voc Dev	36 (2470)	-0.025 $\pm$ 0.35 (-0.095, 0.044)	t <sub>87.12</sub> = 0.73 (0.47)
Consistency	36 (2470)	0.002 $\pm$ 0.001 (-0.001, 0.004)	t <sub>161.2</sub> = 1.23 (0.22)
Tarsus	100 (122)	0.010 $\pm$ 0.008 (-0.006, 0.025)	t <sub>32.96</sub> = 1.25 (0.22)
Wing	100 (122)	0.019 $\pm$ 0.037 (-0.055, 0.093)	t <sub>116.9</sub> = 0.50 (0.62)
Tail	100 (122)	-0.003 $\pm$ 0.031 (-0.067, 0.060)	t <sub>36.19</sub> = -0.10 (0.92)
Condition	100 (122)	-0.018 $\pm$ 0.013 (-0.043, 0.008)	t <sub>115.6</sub> = -1.37 (0.17)
Age	40 (53)	0.159 $\pm$ 0.138 (-0.120, 0.437)	t <sub>39</sub> = 1.15 (0.26)
Heterophile : Lymphocyte	33	-0.0054 $\pm$ 0.048 (-0.104, 0.094)	t <sub>31</sub> = -0.11 (0.91)
Bactericidal assay	33	0.011 $\pm$ 0.015 (-0.020, 0.043)	t <sub>31</sub> = 0.72 (0.47)
Song rate during playback	51 (58)	0.118 $\pm$ 0.086 (-0.054, 0.291)	t <sub>54</sub> = 1.38 (0.17)
Approach response to playback	51 (58)	0.023 $\pm$ 0.009 (0.005, 0.042)	t <sub>49.7</sub> = 2.53 (0.015)
Siring EP offspring elsewhere	100 (122)	-0.165 $\pm$ 0.057 (-0.885, 0.555)	t <sub>1</sub> = -2.91 (0.21)
Maintaining WP paternity	86 (105)	-0.058 $\pm$ 0.064 (-0.186, 0.070)	t <sub>83.9</sub> = 0.90 (0.37)
<b>Polygyny</b>	<b>100 (122)</b>	<b>0.241 <math>\pm</math> 0.074 (0.092, 0.389)</b>	<b>t<sub>58.1</sub> = -3.25 (0.002)</b>

Age, polygyny, and the EP paternity measures were modeled with logistic regression, so estimates reflect how a one-unit change in reproductive success changes the log odds of the mating success variable. A positive score indicates that increasing reproductive success corresponds to an increased probability of being after-second-year, of siring EP offspring elsewhere, of maintaining WP paternity, and of being polygynous.

However, these measures of song quality did not correlate with body condition or health (Table 2.1), suggesting that they do not carry information about male quality. Moreover, I found no evidence that males with “better” songs had higher mating success (Tables 2.2 and 2.3) or higher reproductive success (Table 2.4). In most analyses, there was no relationship between mating success and singing ability; in three of the four analyses where mating success was significantly related to singing ability, less successful males had better songs, the opposite of what I had predicted. I therefore conclude that there is little, if any, evidence that residual vocal deviation and trill consistency affect mating interactions in house wrens. In concert with a playback study on territorial male house wrens (Cramer in review), where I found no effect of the stimulus residual vocal deviation and trill consistency on males’ responses to playback, I conclude that neither song measure is a signal of male quality in house wrens.

These results relating age and song rate to trill consistency are highly consistent with the current literature. Males increased their trill consistency in the transition from their first to later breeding seasons (age effect, Table 2.1), as seen in other species (reviewed by Sakata and Vehrencamp 2012). Moreover, though my sample size is limited, consistency declined slightly but significantly over time within the after-second-year age class, an effect also reported in great tits (*Parus major*; Rivera-Gutierrez et al. 2012). Males are thought to increase their consistency over time due to increased opportunity to practice their vocal output (reviewed in Sakata and Vehrencamp 2012). In this study, males with more consistent songs sang at higher rates in response to playback. If song rate during playback reflects a male’s overall song rate, perhaps these males have simply practiced their songs more and therefore have higher trill consistency. Thus, trill consistency might honestly indicate male age, or the extent to which he has practiced singing, in house wrens. However, I found no evidence that older males or males with more

consistent songs have higher success maintaining WP paternity, attracting a polygynous female, or fledging more offspring (although more consistent singers had higher success gaining EP offspring in other nests), suggesting that this potential signal is not biologically relevant for this species. The sample size for age effects was, however, somewhat limited.

Cramer (in review) found no evidence that territorial males responded differently to songs that differed in residual vocal deviation or trill consistency characteristics. Assuming that male-male competition has a strong effect on mating success, it is perhaps unsurprising that mating success does not relate strongly to song characteristics. If these song parameters do not influence male-male competition, and male-male competition drives polygyny and EP mating, then the song parameters would be unlikely to relate to polygyny and EP mating. If female choice had a strong effect on mating success, and if females preferred low-deviation and/or high-consistency songs, there could be a relationship between song parameters and mating success in the absence of a relation between song parameters and male-male competition. I attempted to directly assess female preferences, but found that wild-caught females did not acclimate sufficiently well to captivity to conduct choice trials (unpublished).

Previous work in house wrens suggests that mate-guarding has an important effect on patterns of EP paternity (Soukup and Thompson 1997, Brylawski and Whittingham 2004). In light of that result, I was intrigued by the possibility that males approaching playback more strongly are more likely to be polygynous and to have higher reproductive success. Though neither of these relationships was robust to correction for multiple testing, the question of whether more aggressive males are able to defend larger or better territories, thereby attracting secondary females and increasing their reproductive success, appears worthy of further investigation in house wrens.

The negative relationships between song quality and EP success are intriguing but difficult to interpret. Males that lost paternity within their own nests sang “better” songs (lower residual vocal deviation and higher trill consistency), and males that gained EP offspring in other nests sang with “worse” residual vocal deviation than males that failed to gain EP offspring elsewhere. However, males that gained EP offspring in other nests did have higher trill consistency than males that did not, a relationship in the predicted direction. While I found no evidence that either song measure relates to characteristics of the males themselves, perhaps song quality trades off with an un-measured aspect of quality, and that quality affects EP mating success. In that case, it is not clear why higher trill consistency should confer an advantage in gaining EP offspring in other nests but a disadvantage in maintaining WP paternity within a male’s own nest. Perhaps different male characteristics are important in a mate-guarding context than in a mate-seeking context.

I found that annual reproductive success correlates positively with polygyny, but not with EP paternity. This result is consistent with Whittingham and Dunn (2005)’s result, that polygyny has a stronger effect on variation in male reproductive success than EP paternity. These results differ from previous work in house wrens, though, in that polygynous males did not lose more WP paternity in their secondary nests than in their primary nests (c.f., Soukup and Thompson 1997). Moreover, in an Illinois population of house wrens, older males tend to be less likely to be cuckolded and are significantly more likely to be polygynous than younger males (Soukup and Thompson 1997). Although my sample sizes were similar to Soukup and Thompson (1997), I found no evidence that females prefer older males in my New York population. There may be intraspecific variation in extra-pair mating behaviors (e.g., Johnsen and Lifjeld 2003).

I investigated several potential measures of male quality, none of which related to my measures of mating success or reproductive success. Previous work in house wrens shows that EP offspring are not healthier than their WP half-siblings (Forsman et al. 2008), so it is perhaps unsurprising that I found no relationship between health measures and success in siring EP offspring or maintaining WP paternity. Moreover, body condition is generally not different between EP and WP males across many species of birds (Akçay and Roughgarden 2007), perhaps suggesting that the typical measures of body condition are not meaningful in birds, or that body condition is not relevant to EP mating decisions.

The challenge involved in song production may be non-linearly related to the frequency bandwidth, and birds perceive frequencies in a non-linear fashion. If a ratio-based measure of vocal deviation is more informative than a difference-based measure, then I expected to see a stronger relationship between male quality or success and the ratio-based measure. However, since neither measure of vocal deviation was significantly correlated with male quality, it is not interesting to test for the strength of the relationships. The strength of the relationships between song quality and mating success were similar for the two measures, and if anything, multiplicative vocal deviation generally had less explanatory power than residual vocal deviation. Multiplicative vocal deviation also was less repeatable than residual vocal deviation. Despite the non-linear nature of sound production, continuing to use frequency bandwidths based on the difference between high and low frequencies may be appropriate.

I agree with the logic of Byers et al. (2010) that physically challenging aspects of song production seem like the most likely candidates to be honest signals of male quality. However, sexual selection can promote different signal properties in different lineages (Price and Lanyon 2004, Cardoso and Hu 2011), and I found no evidence that either vocal deviation or trill

consistency is a signal element promoted by sexual selection in house wrens. Perhaps complexities of house wren song structure complicate the interpretation of these particular parameters for listening birds (as I argue in Cramer in review), and other song parameters are the target for sexual selection in house wrens. If sexual selection cannot explain song elaboration in this species, then it is not clear from this study which characteristics of a male make him more likely to succeed reproductively.

## REFERENCES

- Akçay E, Roughgarden J. 2007. Extra-pair paternity in birds: review of the genetic benefits. *Evol Ecol Res.* 9:855-868.
- Andersson M. 1994. *Sexual Selection*. Princeton, NJ: Princeton University Press.
- Ballentine B. 2009. The ability to perform physically challenging songs predicts age and size in male swamp sparrows, *Melospiza georgiana*. *Anim Behav.* 77:973-978.
- Ballentine B, Hyman J, Nowicki S. 2004. Vocal performance influences female response to male bird song: an experimental test. *Behavioral Ecology* 15:163-168.
- Beebee MD. 2004. Variation in vocal performance in the songs of a wood-warbler: evidence for the function of distinct singing modes. *Ethology* 110:531-542.
- Botero CA, Rossman RJ, Caro LM, Stenzler LM, Lovette IJ, de Kort SR, Vehrencamp SL. 2009. Performance variability is related to age, dominance, and reproductive success in the tropical mockingbird. *Anim Behav.* 77:701-706.
- Bradbury JW, Vehrencamp SL. 2011. *Principles of Animal Communication*, 2nd Ed. Sunderland, MA: Sinauer Associates.
- Brylawski AMZ, Whittingham LA. 2004. An experimental study of mate guarding and paternity in house wrens. *Anim Behav.* 68:1417-1424.
- Byers BE. 2007. Extrapair paternity in chestnut-sided warblers is correlated with consistent vocal performance. *Behav Ecol.* 18:130-136.
- Byers J, Hebets EA, Podos J. 2010. Female mate choice based upon male motor performance. *Anim Behav.* 79:771-778.
- Cardoso GC, Atwell JW, Ketterson ED. 2007. Inferring performance in the songs of dark-eyed juncos (*Junco hyemalis*). *Behav Ecol.* 18:1051-1057.
- Cardoso GC, Atwell JW, Ketterson ED, Price TD. 2009. Song types, song performance, and the use of repertoires in dark-eyed juncos (*Junco hyemalis*). *Behav Ecol.* 20:901-907.

Cardoso GC, Hu Y. 2011. Birdsong performance and the evolution of simple (rather than elaborate sexual signals). *Am Nat* 178:679-686.

Caro SP, Sewall KB, Salvante KG, Sockman KW. 2010. Female Lincoln's sparrows modulate their behavior in response to variation in male song quality. *Behav Ecol*. 21: 562-569.

Catchpole CK, Slater PJB. 2008. *Birdsong: biological themes and variations*, 2nd ed. Cambridge University Press, NY, p.

Christensen R, Kleindorfer S, Robertson J. 2006. Song is a reliable signal of bill morphology in Darwin's small tree finch *Camarhynchus parvulus*, and vocal performance predicts male pairing success. *J Avian Biol*. 37:6:617-624.

Cramer ERA, Hall ML, deKort SR, Lovette IJ, Vehrencamp SL. 2011. Infrequent extra-pair paternity in the banded wren, a synchronously breeding tropical passerine. *Condor*. 113:637-645.

Cramer ERA, Price JJ. 2007. Red-winged blackbirds *Agelaius phoeniceus* respond differently to song types with different performance levels. *J Avian Biol*. 38:122-127.

Cynx J. 2004. Are songbirds Pythagoreans? Absolute and relative pitch perception. In: Marler PR, Slabbekoorn H, editors. *Nature's Music: The Science of Birdsong*. San Diego, California: Elsevier. p. 218.

Dakin EE, Avise JC. 2004. Microsatellite null alleles in parentage analysis. *Heredity*. 93:504-509.

De Kort SR, Eldermire ERB, Valderrama S, Botero CA, Vehrencamp SL. 2009. Trill consistency is an age-related assessment signal in banded wrens. *Proc R Soc Lond B*. 269:2525-2531.

deKort SR, Eldermire ERB, Cramer ERA, Vehrencamp SL. 2009. The deterrent effect of bird song in territory defense. *Behav Ecol*. 20:200-206.

Drăgănoui TI, Nagle L, Kreutzer M. 2002. Directional female preference for an exaggerated male trait in canary (*Serinus canaria*) song. *Proc R Soc Lond B*. 269:2525-2531.

duBois AL, Nowicki S, Searcy WA. 2009. Swamp sparrows modulate vocal performance in an aggressive context. *Biol Lett*. 5:163-165.

duBois AL, Nowicki S, Searcy WA. 2011. Discrimination of vocal performance by male swamp sparrows. *Behav Ecol Sociobiol.* 65:717-726.

Eckerle KP, Thompson CF. 2006. Mate choice in house wrens: nest cavities trump male characteristics. *Behaviour.* 143:253-271.

Fletcher F, Riede T, Suthers RA. 2006. Model for vocalization by a bird with distensible vocal cavity and open beak. *J Acoust Soc Am.* 119:1005-1011.

Goller F, Suthers RA. 1996. Role of syringeal muscles in controlling the phonology of bird song. *J Neurophysiol.* 76:287-300.

Grafen A. 1990. Biological signals as handicaps. *J Theor Biol.* 144:517-546.

Hoese WJ, Podos J, Boetticher NC, Nowicki S. 2000. Vocal tract function in birdsong production: experimental manipulation of beak movements. *J Exp Biol.* 203:1845-1855.

Hurly TA, Ratcliffe L, Weary DM, Weisman R. 1992. White-throated sparrows (*Zonotrichia albicollis*) can perceive pitch change in conspecific song by using the frequency ratio independent of the frequency difference. *J Comp Psych.* 106:388-391.

Illes AE, Hall ML, Vehrencamp SL. 2006. Vocal performance influences male receiver response in the banded wren. *Proc R Soc Lond B.* 273:1907-1912.

Jarvis ED. 2004. Brains and birdsong. In: Marler P, Slabbekoorn H, editors. *Nature's Music: the science of birdsong.* Boston: Elsevier Academic Press. p. 226-271.

Johnsen A, Lifjeld JT. 2003. Ecological constraints on extra-pair paternity in the bluethroat. *Oecologia.* 136:476-483.

Johnson LS, Hicks BG, Masters BS. 2002. Increased cuckoldry as a cost of breeding late for male house wrens (*Troglodytes aedon*). *Behav Ecol.* 13:670-675.

Johnson LS, Kermott LH. 1990. Structure and context of female song in a north-temperate population of house wrens. *J Field Ornithol.* 61:273-284.

Johnson LS, Kermott LH. 1991. The function of song in male house wrens (*Troglodytes aedon*). *Behaviour.* 116:190-209.

Johnson LS, Searcy WA. 1996. Female attraction to male song in house wrens (*Troglodytes aedon*). Behaviour. 133:357-366.

Johnson LS, Thompson CF, Sakaluk SK, Neuhauser M, Johnson BG, Soukup SS, Forsythe SJ, Masters BS. 2009. Extra-pair young in house wren broods are more likely to be male than female. Proc R Soc Lond B. 276:2285-2289.

Johnson LS. 1998. House Wren (*Troglodytes aedon*). The Birds of North America Online (A. Poole, Ed) Ithaca: Cornell Lab of Ornithology; retrieved from the Birds of North America Online: <http://bna.birds.cornell.edu/bna/species/380> . doi:10.2173/bna.380

Kalinowski ST, Taper ML, Marshall TC. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol. 16:1099-1106.

LaBarbera K, Llambías PE, Cramer ERA, Schaming TD, Lovette IJ. 2010. Synchrony does not explain extrapair paternity rate variation in northern or southern house wrens. Behav Ecol. 21:773-780.

Lessells CM, Boag PT. 1987. Unrepeatable repeatabilities: a common mistake. Auk. 104:116-121.

Lifjeld JT, Dunn PO, Westneat DR. 1994. Sexual selection by sperm competition in birds: male-male competition or female choice? J Avian Biol. 25:244-250.

Lightbody JP, Weatherhead PJ. 1987. Polygyny in the yellow-headed blackbird: female choice versus male competition. Anim Behav. 35:1670-1684.

Llambías PE. 2009. Why monogamy? Comparing house wren social mating systems in two hemispheres. [dissertation]. Ithaca, NY: Cornell University; 119 p.

Maynard Smith J, Harper DGC. 1995. Animal signals: Models and terminology. J Theor Biol. 77:305-311.

Millet S, Bennett J, Lee KA, Hau M, Klasing KC. 2007. Quantifying and comparing constitutive immunity across avian species. Dev Comp Immunol. 31:188-201.

Nakagawa S, Schielzeth H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol Rev.* 85:935-956.

Ots I, Murumägi A, Hõrak P. 1998. Haematological health state indices of reproducing great tits: methodology and sources of natural variation. *Funct Ecol.* 12:700-707.

Platt ME, Ficken MS. 1987. Organization of singing in house wrens. *J Field Ornithol.* 58:190-197.

Podos J. 1997. A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution.* 51:537-551.

Podos J. 2001. Correlated evolution of morphology and vocal signal structure in Darwin's finches. *Nature.* 409:185-188.

Podos, J., D.C. Lahti, and D.L. Moseley. 2009. Sensorimotor learning in songbirds. *Adv Stud Behav.* 40:159-195.

Riede T, Suthers R, Fletcher NH, Blevins WE. 2006. Songbirds tune their vocal tract to the fundamental frequency of their song. *PNAS.* 103:5543-5548.

Rivera-Gutierrez HF, Pinxten R, Eens M. 2012. Tuning and fading voices in songbirds: age-dependent changes in two acoustic traits across the life span. *Anim Behav.* 83:2012.

Rivera-Gutierrez HF, Pinxten R, Eens M. 2011. Songs differing in consistency elicit differential aggressive response in territorial birds. *Biol Lett.* 7:339-342.

Sakata JT, Vehrencamp SL. 2012. Integrating perspectives on vocal performance and consistency. *J Exp Biol.* 215:201-209.

Searcy WA, Nowicki S. 2005. *The Evolution of Animal Communication.* Princeton, NJ: Princeton University Press.

Sewall KB, Dankoski EC, Sockman KW. 2010. Song environment affects singing effort and vasotocin immunoreactivity in the forebrain of male Lincoln's sparrows. *Horm Behav.* 58:544-553.

- Sockman KW. 2009. Annual variation in vocal performance and its relationship with bill morphology in Lincoln's sparrows, *Melospiza lincolnii*. *Anim Behav.* 77:663-671.
- Soukup SS, Thompson CF. 1997. Social mating system affects the frequency of extra-pair paternity in house wrens. *Anim Behav.* 54:1089-1105.
- Suthers RA. 2001. Peripheral vocal mechanisms in birds: are songbirds special? *Neth J Zool.* 51:217-242.
- Vallet E, Beme I, Kreutzer M. 1998. Two-note syllables in canary song elicit high levels of sexual display. *Anim Behav.* 55:291-297.
- Vehrencamp SL. 2000. Handicap, index, and conventional signal elements of bird song. In: Espmark Y, Amundsen T, Rosenqvist G, editors. *Animal Signals: Signalling and Signal Design in Animal Communication*. Trondheim, Norway: Tapir Academic Press. p. 277-300.
- Wegrzyn E, Leniowski K, Osiejuk TS. 2010. Whistle duration and consistency reflect philopatry and harem size in great reed warblers. *Anim Behav* 79:1363-1372.
- Whittingham LA, Dunn PO. 2005. Effects of extra-pair and within-pair reproductive success on the opportunity for selection in birds. *Behav Ecol.* 16:138-144.

## CHAPTER 3

### SPERM LENGTH VARIATION IN HOUSE WRENS *TROGLODYTES AEDON*

EMILY R. A. CRAMER, TERJE LASKEMOEN, ODDMUND KLEVEN, AND JAN T. LIFJELD

#### ABSTRACT

It is well documented that sperm size and structure varies considerably among avian species, but we know much less about the extent of intraspecific variation in sperm morphometry and its possible co-variation with somatic traits like body size and condition. Here, we investigate patterns of sperm length variation and co-variation in a population of House Wrens (*Troglodytes aedon*). Total sperm length showed considerable between-male variation, with high repeatability between seasons indicative of a strong genetic basis for this trait. However, we also detected a seasonal increase in the flagellum:head length ratio, which might indicate phenotypic plasticity or adjustment in the relative size of sperm components. The variation in total sperm length within an ejaculate sample was higher for males sampled very early in the season, which may reflect more heterogeneity in the size of seminiferous tubules when testes are growing. None of the studied sperm morphometry traits correlated significantly with any measures of male body size or physiological condition. Further studies are needed to reveal if the observed individual variation in sperm morphology plays any functional or adaptive role.

#### INTRODUCTION

Sperm competition can occur if a female copulates with more than one male in a reproductive bout, and it consists of competition among sperm from different males (sperm competition, *sensu stricto*) and/or cryptic female preference for certain sperm traits (Birkhead and Pizzari 2002). Comparative studies suggest that sperm competition influences the evolution of reproductive traits. For instance, species with stronger sperm competition have larger testes relative to body size (Møller 1991) and produce more sperm (e.g., Tuttle et al. 1996). Sperm competition also acts on the morphology of sperm cells. Across passerine birds, the level of multiple mating by females correlates with several aspects of sperm morphology (detailed below) that likely affect sperm function, such as longevity or swimming speed (Kleven et al. 2009a; Lüpold et al. 2009b; but see Immler and Birkhead 2007). Morphological traits such as sperm total length and the length of different sperm components can vary considerably among and within individuals of the same species, particularly in species with low levels of sperm competition (Birkhead et al 2005; Calhim et al. 2007; Immler et al. 2008; Kleven et al. 2008; Lüpold et al. 2009b; Lifjeld et al. 2010; Lüpold et al. 2011; Schmoll and Kleven 2011; Lifjeld et al 2012). However, the causes of variation in sperm morphology between males are poorly understood, and the possibility of variation in sperm morphology between different ejaculates by the same male has rarely been explored in wild birds.

Several hypotheses address possible causes of sperm morphology variation within a species. Optimal sperm morphology may depend on male characteristics and on the likelihood of experiencing sperm competition. For instance, if producing high-quality sperm is costly (Dewsbury 1982), investment in sperm production may depend on a male's condition or health status, a hypothesis supported by the finding that Great Tits (*Parus major*) exposed to a more stressful environment produce sperm with reduced motility (Helfenstein et al. 2010b). Males

with low copulatory access to females may also benefit by investing heavily in sperm quality to maximize chances of achieving fertilizations following copulations, as occurs in subordinate male domestic chickens (*Gallus gallus domesticus*, Froman et al. 2002). Furthermore, captive Gouldian Finches (*Erythrura gouldiae*) show within-male plasticity in sperm morphology when the competitive environment changes (Immler et al. 2010).

At a mechanistic level, sperm morphology likely depends on the seminiferous tubules of the testes (Aire 2007a,b; Lüpold et al. 2009c). Because testes size changes seasonally (e.g., Calhim and Birkhead 2007), sperm morphology may vary within males across the breeding season. If individual males differ in testes size, these differences could also cause between-male variation in sperm morphology.

In this study on House Wrens (*Troglodytes aedon*), we assessed three levels of variability in sperm morphology: between males, within males but between ejaculate samples, and within ejaculate samples. We tested whether between-male variation in sperm morphology is related to male condition, health (using two standard ecoimmunology assays), access to females, date, and breeding stage. For a small number of individuals, we investigated variation within males between breeding seasons. For these analyses, we studied three measures of sperm morphology that either correlate with sperm function or that show evidence of selection in interspecific comparative studies: total sperm length (TSL), flagellum:head ratio (F:H), and midpiece:TSL ratio (M:TSL). Different studies have revealed different functional correlates of these measures. TSL likely mediates compatibility with the female reproductive tract (Briskie et al. 1997; Kleven et al. 2009a) and may relate to sperm longevity (Helfenstein et al. 2008, 2010a; but see Kleven et al. 2009b), but longer TSL typically does not correlate with faster swimming speed in birds (e.g., Birkhead et al. 2005; Helfenstein et al. 2008; Kleven et al. 2009a; Laskemoen et al. 2010; but see

Lüpold et al. 2009a; Mossman et al. 2009). Instead, swimming speed may increase with F:H and/or M:TSL. The midpiece, which consists of a single large, fused mitochondrion that powers swimming (Koehler 1995; Cardullo and Baltz 1991; Froman and Kirby 2005), and the flagellum, which supports the midpiece and extends beyond it to form the tail, both contribute to forward motion while the head of the sperm produces drag proportional to its surface area (Humphries et al. 2008). Higher F:H and M:TSL therefore represent a greater investment in propulsion relative to resistance to motion (Humphries et al. 2008; Lüpold et al. 2009a; Mossman et al. 2009; Helfenstein et al. 2010a; Immler et al. 2010; Laskemoen et al. 2010; but see Kleven et al. 2009a; Lüpold et al. 2009b). We also investigated variability in TSL within an ejaculate sample ( $CV_{wm}$ ), which is associated with the risk of sperm competition across species:  $CV_{wm}$  is lower in species with stronger sperm competition (Calhim et al. 2007; Kleven et al. 2008; Lüpold et al. 2009b; Lifjeld et al. 2010), although the strength of this association varies with phylogeny (Lifjeld et al. 2010).

## METHODS

### *Study system and field methods*

We studied House Wrens breeding in nest boxes in forest/pond edges near Ithaca, New York, USA (42°31'N, 76°28'W), from April-July 2009-2011. Individuals typically produce two broods per year, and newly-arriving males continue to establish territories throughout the breeding season. Extra-pair paternity occurs in approximately 30-50% of broods and accounts for 12-25% of chicks in our population, and about 25% of males are polygynous (LaBarbera et al. 2010 and unpublished data).

We captured and banded all males. We measured wing chord, tarsus length, tail length, and mass (to the nearest 0.1 g using a Pesola spring scale); took a blood sample from the brachial vein after swabbing with ethanol for sterilization (Millet et al. 2007); made a blood smear (2009 only); and took a semen sample. Semen samples were collected in microcapillary tubes using cloacal massage (e.g. Kleven et al. 2008). In 2009, semen was stored in 20-50  $\mu$ L 1X phosphate buffered saline (PBS) in the field and fixed by adding 100  $\mu$ L 5% formalin later in the laboratory. In 2010 and 2011, we mixed semen samples with 25  $\mu$ L 1X PBS and then immediately added 25  $\mu$ L 5% formalin in the field. We collected a total of 105 sperm samples from 79 different individuals (59 individuals sampled once, nine individuals sampled twice in 2009, 13 males sampled in two consecutive years, and two males sampled in all three years; 48 samples collected in 2009, 39 in 2010, and 18 in 2011).

All breeding attempts on the site were monitored to assess the effect of breeding stage on sperm morphology; we expressed breeding stage as number of days before or after the first egg date. Most nests were found during the egg laying stage, so we calculated the first egg date by counting backwards. For six nests, we estimated first egg dates based on timing of hatching or chick size. Fifteen males advertised for females on-site but did not progress to nesting, and we assigned these males the mean value of days before the first egg date calculated across other advertising males. We noted whether or not a male became polygynous by attracting a secondary female while his primary female still had an active nest.

Males in their first breeding year cannot be distinguished morphologically from older males, but 32 males' ages were known from their banding history.

### *Male condition and health*

We defined body size as the first principal component of wing chord, tail length, and tarsus length measurements, calculated separately for each year. We defined body condition as body weight in analyses that controlled for size as a covariate; this approach is preferred to using the residual of weight on size as a measure of condition (e.g., García-Berthou 2001).

In 2009, we attempted to manipulate body condition for an unrelated experiment by assigning males to one of three wing-clipping treatment groups (4-mm, 2-mm, or control 0 mm; Tieleman et al. 2008). We recaptured most males later in the season to re-measure body condition and assess the effectiveness of the manipulation, and found in a larger sample of males that the treatment had no measurable effect on male condition or health (data not shown).

For 2009 captures, we estimated health using two standard ecoimmunology measures: the ratio of heterophils:lymphocytes (H:L; Ots et al. 1998) and the bactericidal capacity of the blood (Millet et al. 2007). To assess H:L, 100 white blood cells were counted on blood smears stained with a Modified Wright's Stain (Hematek Stain Pak, Siemens Diagnostics) at the Cornell University Veterinary College Animal Health Diagnostic Center. For six samples, fewer than 100 cells could be counted due to blood smear quality.

For the bactericidal assay (Millet et al. 2007), we combined either 10  $\mu$ L blood (males captured in pre-nestling stage) or 5  $\mu$ L blood (nestling stage) with 10  $\mu$ L bacteria (*E. coli* strain ATCC 8739, American Type Culture Collection), 2  $\mu$ L 200 mM L-glutamine, and 88  $\mu$ L (pre-nestling) or 93  $\mu$ L (nestling) pre-heated CO<sub>2</sub>-independent growth medium (42 °C). One positive control (using the appropriate quantity of 1X PBS in place of blood) and one negative control (using PBS for both blood and bacteria) were run daily, and all assays were performed on the day of capture. These mixtures were incubated at 42 °C for 30 minutes, then two 50  $\mu$ L aliquots from each tube were plated individually onto standard, antibiotic-free Luria plates. Plates were

incubated overnight at 37 °C and the number of bacterial colonies was counted the following day. “Bactericidal activity” was the difference between the mean positive control count and the mean count for the individual’s blood, divided by the positive control.

### *Sperm measurements*

Approximately 10 µL of fixed semen was streaked onto a microscope slide, dried overnight, and gently rinsed with distilled water (Laskemoen et al. 2008). Re-sampling preliminary measurements of 70 sperm from a single male showed that 30 sperm gave a precise estimate of standard deviation in TSL, and increasing the sample size beyond 30 sperm did not greatly improve precision. We therefore photographed 30 morphologically normal sperm cells per sample under a light microscope (2009: 320x magnification, Leica Microsystems DM6000B DFC420 Leica digital camera, Heerbrugg, Switzerland; 2010 and 2011: 400x magnification, Zeiss Axiover 200M with AxioCam MRm, Carl Zeiss Inc). For two samples, only 29 sperm were available. Morphologically normal sperm were those without clear breaks within or between sections. We then measured the length of the head (including acrosome), midpiece, and tail of the sperm cells using the on-screen cursor line tool in the Leica Application Suite (version 2.6.0 R1; 2009 samples) or the line tool in ImageJ with a custom plug-in (NIH, 2010 and 2011 samples). Flagellum length was calculated by adding midpiece and tail, and the TSL was considered the sum of flagellum and head. Variability in TSL within ejaculate samples was calculated as the coefficient of variation across the 30 measured spermatozoa per sample ( $CV_{wm}$ , Lifjeld et al. 2010). Measurement accuracy was high (repeatability > 0.96 [sensu Lessells and Boag 1987] for measurements of mean TSL, F:H, M:TSL and  $CV_{wm}$  across 8 males,  $F_{7,8} > 52$ ,  $p < 0.001$  for all measures).

For males captured in 2009 and 2010, the 2009 samples were re-measured on the Zeiss microscope system so that they could be directly compared to the 2010 measurements, and to assess microscope effects on measurements. For 11 re-measured samples, the Zeiss system gave TSL values approximately 1.5  $\mu\text{m}$  longer than the Leica system (statistical model fit with male identity as a random factor, microscope effect  $F_{1,10} = 49.50$ ,  $p < 0.0001$ ) and F:H scores lower by 0.16 on average ( $F_{1,10} = 26.17$ ,  $p = 0.0005$ ). The microscopes did not significantly affect M:TSL ( $F_{1,10} = 0.42$ ,  $p = 0.5$ ) or  $CV_{\text{wm}}$  ( $F_{1,10} = 0.06$ ,  $p = 0.8$ ).

Many of our samples, particularly from 2009, either lacked sperm or had a high percentage of sperm that had broken, usually between the head and midpiece. Damaged sperm were likely due in part to the length of time semen samples were carried in PBS in 2009 before adding formalin, which kills and fixes the sperm cells; we therefore did not quantify the proportion of morphologically abnormal cells, as these were likely an artifact of collection method (Humphreys 1972). We avoided including obviously damaged cells in our sample, and measurements from samples that did ( $n = 8$ ) and did not ( $n = 40$  in 2009) have a high proportion of damaged cells did not differ in TSL, F:H, M:TSL, or  $CV_{\text{wm}}$  (all  $p > 0.3$ , all  $r^2 < 0.03$ ).

### *Statistical analysis*

We first assessed seasonal effects by constructing a set of models for each sperm measure and choosing among the models based on AIC scores. We included a year term in all models to account for the effects of the microscope and measurement software, as well as potential true yearly variation. We then tested models including all possible combinations of year with date, date squared (to allow for non-linear relationships), days to first egg date, days to first egg date squared, and pair-wise interaction terms. For each male that was sampled more than once, we

randomly selected a single sample to include, and we used average values for the sample for TSL, F:H, and M:TSL.

We tested for correlations between male quality measures (male age, polygyny status, condition, health, and wing-clipping treatment) and sperm traits. We constructed a separate model for each male quality-sperm morphometry combination, and included as covariate(s) the seasonal effect(s) from the best-fitting model chosen above. Since we only assessed ecoimmunology measures and conducted wing-clipping in 2009, only a sub-set of sperm samples had these associated data. Rather than use the same subset of males as we analyzed for seasonal effects (which, by chance, often excluded the capture event with health and treatment data), we considered only the data set with health measures and randomly chose a duplicate sample to exclude from this restricted data set.

To examine year-to-year changes in sperm, we calculated repeatability across years for the 15 males sampled in consecutive years by fitting models with or without a random effect of male identity (i.e., following the procedure of Nakagawa and Schielzeth 2010 for calculating repeatability). For this analysis, we used measurements of the 2009 samples taken using the same Zeiss microscope system that was used in 2010 and 2011. These models included a year effect but no date effect because too few re-measurements of 2009 samples were taken on the Zeiss system to allow for an accurate estimate of the sperm by date relationships. For the two males sampled in all three years, one pair of consecutive years was randomly selected.

$CV_{wm}$  was log-transformed for normality. Bactericidal capacity could not be transformed, so we used nonparametric tests. All other response variables were normally distributed, and all tests were two-tailed. Date and days to the first egg date were centered (Schielzeth 2010). To control for multiple testing, we used false discovery rate correction (FDR, Benjamini and

Hochberg 1995) implemented in R version 2.9.2 (R [Development](#) Core Team 2009); this method maintains higher statistical power than more traditional Bonferroni correction, but still controls experiment-wide Type I error (Verhoeven et al. 2005). All other statistical tests were conducted in JMP 7.0 and SAS 9.0 (SAS Institute, Cary, NC, USA). To eliminate possible inter-observer differences, one person took all sperm and morphology measurements.

## RESULTS

### *Size and structure of House Wren sperm*

Across all males ( $n = 79$  samples), mean ( $\pm$  SE) TSL was  $77.8 \pm 0.42 \mu\text{m}$ , with most of the length consisting of the flagellum (midpiece:  $50.4 \pm 0.41 \mu\text{m}$  and tail:  $14.8 \pm 0.39 \mu\text{m}$ ). Mean head length was  $12.6 \pm 0.07 \mu\text{m}$ . The length of each sperm component was repeatable within ejaculate samples (head  $R_M = 0.44$ , midpiece  $R_M = 0.57$ , and tail  $R_M = 0.50$ ), as was TSL ( $R_M = 0.70$ ), F:H ratio ( $R_M = 0.48$ ), and M:TSL ratio ( $R_M = 0.49$ ;  $p < 0.001$ ,  $n = 2368$  sperm and 79 samples for all analyses, calculated following Nakagawa and Schielzeth 2010). These repeatability values imply consistent, important individual differences among males in sperm size and the length of the components.

The between-male coefficient of variation ( $CV_{\text{bm}}$ ) in mean TSL was 4.0% in 2009, 4.7% in 2010, and 5.0% in 2011. The within-male coefficient of variation in sperm length ( $CV_{\text{wm}}$ ), calculated based on a single ejaculate sample per male, ranged from 1.7% to 8.1%, with a mean of  $3.3 \pm 0.2\%$  in 2009,  $2.8 \pm 0.2\%$  in 2010, and  $2.4 \pm 0.3\%$  in 2011.  $CV_{\text{wm}}$  correlated positively with variation in each of the components (with CV of head,  $r^2 = 0.13$ ,  $p = 0.001$ ; with CV midpiece,  $r^2 = 0.32$ ,  $p < 0.0001$ ; with CV tail,  $r^2 = 0.20$ ,  $p < 0.0001$ ;  $n = 79$  males), so we used  $CV_{\text{wm}}$ , based on the TSL, in all analyses.

### *Seasonal and annual variation in sperm morphometry*

Between-male variation in F:H and  $CV_{wm}$  was partly due to date effects (Figure 3.1): F:H increased linearly with date (best model  $F_{3,75} = 4.93$ ,  $p = 0.004$ ; date effect  $F_{1,75} = 11.57$ ,  $p = 0.001$ , year effect  $F_{2,75} = 2.67$ ,  $p = 0.08$ ). This increase corresponded with a seasonal increase in the length of the midpiece ( $F_{1,75} = 16.76$ ,  $p = 0.0001$ ) and to a tendency for head length to decrease seasonally ( $F_{1,75} = 3.06$ ,  $p = 0.09$ ); the length of the tail did not change ( $F_{1,75} = 2.04$ ,  $p = 0.16$ ).  $CV_{wm}$  decreased initially and then remained fairly constant (best model  $F_{4,74} = 6.74$ ,  $p = 0.0001$ ; effect of year  $F_{1,74} = 1.85$ ,  $p = 0.16$ , date  $F_{1,74} = 5.44$ ,  $p = 0.02$ ; date squared  $F_{1,74} = 7.77$ ,  $p = 0.007$ ). This quadratic relationship was driven largely by samples on the extreme ends of the season; the quadratic term was not significant ( $p > 0.6$ ) when samples from before 3 May (day 123; the earliest first egg dates were between 8 and 13 May, 2009-2011) and after July 14 (day 195; the last eggs were laid between 15-19 July, 2009-2011) were excluded.

There were no consistent seasonal trends with TSL or M:TSL. The best model for TSL included a marginally non-significant year by date interaction (whole model  $F_{5,73} = 2.77$ ,  $p = 0.02$ ; date effect  $F_{1,73} = 23.45$ ,  $p = 0.17$ ; year effect  $F_{2,73} = 4.31$ ,  $p = 0.02$ ; interaction  $F_{2,73} = 2.49$ ,  $p = 0.09$ ; note that the dates were centered and scaled, and the year effect may reflect variation between microscopes). The best model for M:TSL included only a non-significant year term ( $F_{2,76} = 0.16$ ,  $p = 0.85$ ).

Notably, breeding stage was not a factor in any of the models with low AIC scores. Excluding males that failed to attract females did not qualitatively change the models (data not shown). After correcting for multiple testing across all four best models, the linear term for

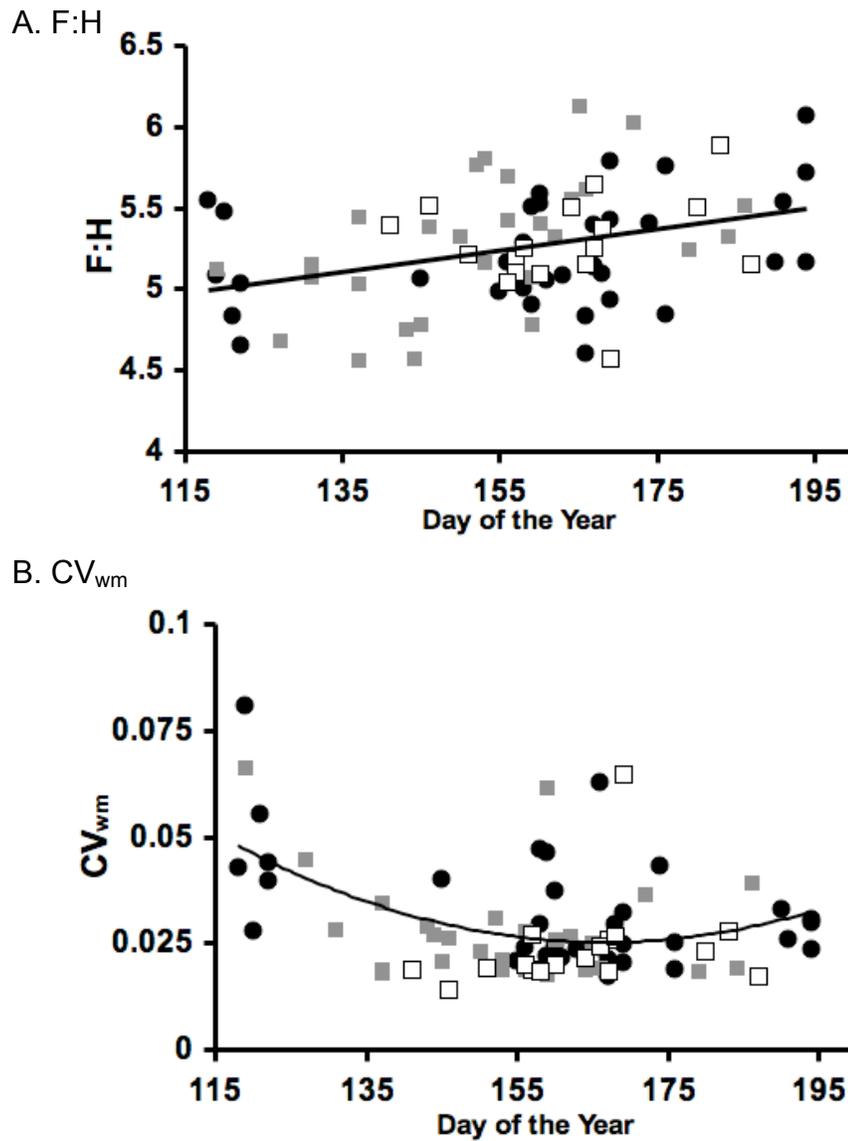


Figure 3.1. Seasonal patterns in (A) flagellum:head ratio (F:H) and (B) within-sample variability in total sperm length ( $CV_{wm}$ ). Years are indicated by different colors (2009, black; 2010, grey; 2011, white), and regression lines represent the best model (chosen by AIC) relating sperm traits to date. Date was significantly, linearly related to F:H, and significantly quadratically related to  $CV_{wm}$ . ( $p < 0.05$ ).  $CV_{wm}$  was log-transformed for statistical testing, but untransformed values are shown.

$CV_{wm}$  and the year effect for TSL became marginally non-significant (both  $p = 0.051$ ), but other significance levels did not change qualitatively.

For samples of the same individual taken in consecutive years, repeatability was high and significant for TSL (Figure 3.2,  $R_M = 0.84$ ,  $p = 0.004$ ) and M:TSL ( $R_M = 0.82$ ,  $p = 0.006$ ), moderate but non-significant for F:H ( $R_M = 0.56$ ,  $p = 0.11$ ) and low and not significant for  $CV_{wm}$  ( $R = 0.13$ ,  $p = 0.75$ ;  $n = 15$  males measured in two consecutive seasons on the same microscope for all tests, Nakagawa and Schielzeth 2010). Significance levels of repeatability scores did not change qualitatively after controlling for multiple tests.

#### *Sperm morphometry and male phenotype*

Before correcting for multiple testing, second-year males had higher M:TSL than older males, and polygynous males tended to have higher F:H than socially monogamous males (Table 3.1). TSL tended to correlate positively with body condition and bactericidal capacity of the blood, but also correlated positively with the ratio of heterophils:lymphocytes, which increases with stress (Ots et al. 1998). However, none of these relationships was robust to correction for multiple testing (Tables 3.1 and 3.2). For individuals sampled in consecutive years, body condition was moderately, but not significantly, repeatable ( $R_M = 0.47$ ,  $p = 0.21$ ). Results using raw sperm measurements (i.e., not corrected for date effects) were similar (data not shown).

## DISCUSSION

We found large and consistent differences among individuals in sperm morphology that were unrelated to breeding stage, age, or several measures of male health and condition. Differences in the flagellum:head ratio (F:H) were partly explained by date of sampling, as the F:H increased

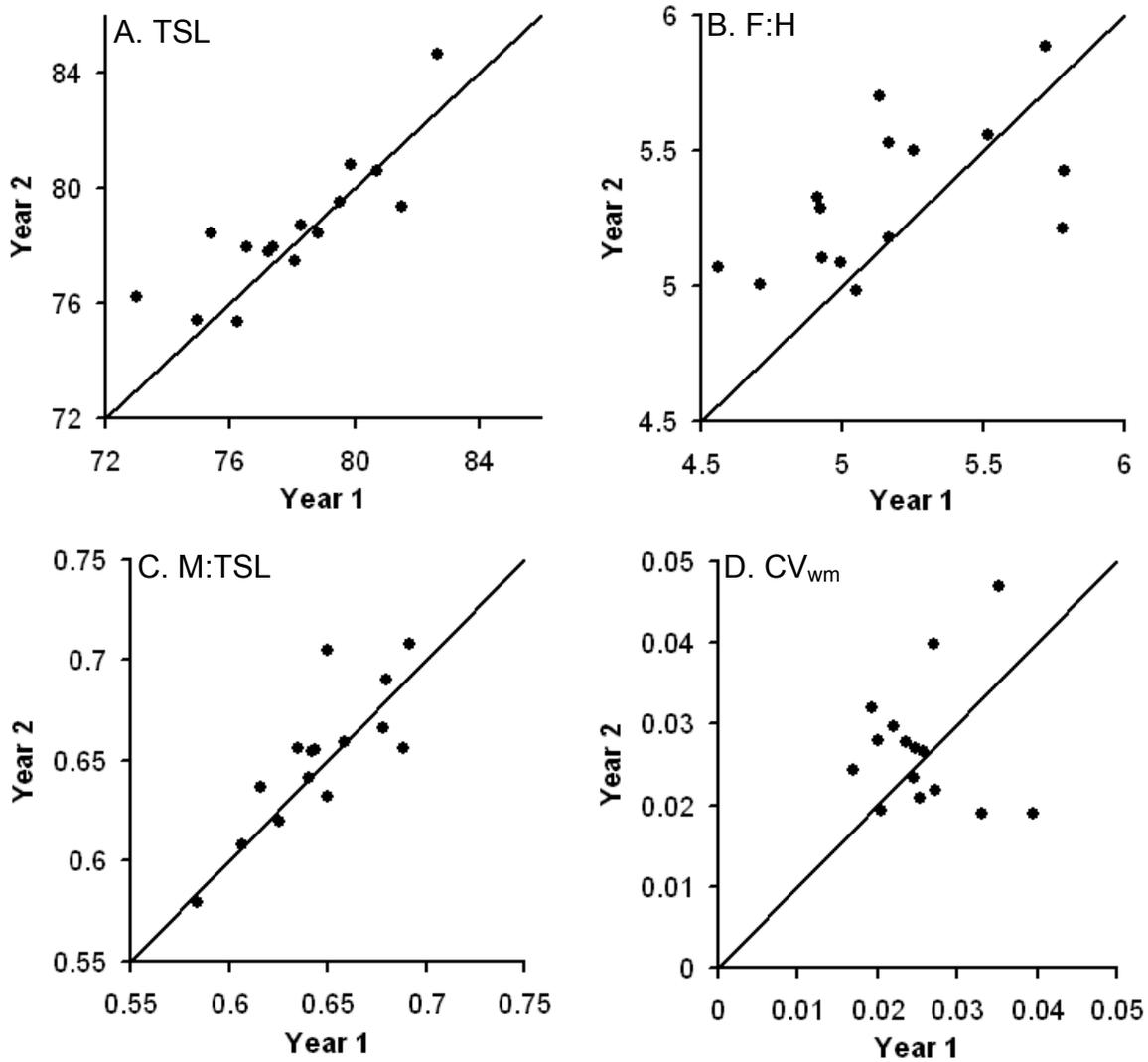


Figure 3.2. Repeatability of sperm morphology between two years for 15 males in (A) total sperm length (TSL), (B) flagellum:head ratio (F:H), (C) midpiece:TSL ratio (M:TSL), and (D) variability in TSL ( $CV_{wm}$ ). The lines represent unity. Note that measurements used in this analysis were taken on the same microscopes, and that untransformed values are shown.

Table 3.1. Least squared mean estimates and standard error of sperm morphometry as a function of male age, polygyny status, and experimental groups (wing clipping; 2009 only). In each model, we corrected for date based on the best-fitting seasonal effects model. Results using uncorrected sperm morphometry are similar. P values are not corrected for multiple comparisons; after correction with false discovery rate (Benjamini and Hochberg 1995), none of the relationships approached significance.

Sperm trait	Factor	Test statistic ( <i>p</i> )	Level of factor	n	Least Squared Mean ± SE
Total length	Age <sup>a</sup>	F <sub>1,25</sub> = 0.70 (0.41)	ASY	24	78.66 ± 0.72
			SY	8	77.42 ± 1.33
	Social mates	F <sub>1,56</sub> = 0.01 (0.93)	Monogamous	51	77.92 ± 0.58
			Polygynous	12	78.02 ± 1.17
	Experimental group	F <sub>2,22</sub> = 0.62 (0.44)	Big clip	7	76.04 ± 1.29
			Little clip	10	78.71 ± 1.11
No clip			9	76.50 ± 1.05	
Flagellum: head ratio	Age	F <sub>1,27</sub> = 0.32 (0.57)	ASY	24	5.33 ± 0.08
			SY	8	5.23 ± 0.15
	Social mates	F <sub>1,58</sub> = 2.77 (0.10)	Monogamous	51	5.17 ± 0.05
			Polygynous	12	5.34 ± 0.10
	Experimental group	F <sub>2,22</sub> = 0.03 (0.97)	Big clip	7	5.31 ± 0.16
			Little clip	10	5.26 ± 0.13
No clip			9	5.28 ± 0.13	
Midpiece: total ratio	Age	F <sub>2,28</sub> = 6.87 (0.01)	ASY	24	0.64 ± 0.01
			SY	8	0.68 ± 0.02
	Social mates	F <sub>1,59</sub> = 0.02 (0.90)	Monogamous	51	0.65 ± 0.01
			Polygynous	12	0.65 ± 0.01
	Experimental group	F <sub>2,23</sub> = 0.10 (0.91)	Big clip	7	0.65 ± 0.01
			Little clip	9	0.65 ± 0.01
No clip			9	0.65 ± 0.01	
Sperm length variability <sup>b</sup>	Age	F <sub>1,26</sub> = 0.78 (0.38)	ASY	24	-3.69 ± 0.08
			SY	8	-3.57 ± 0.12
	Social mates	F <sub>1,57</sub> = 0.05 (0.83)	Monogamous	51	-3.68 ± 0.06
			Polygynous	12	-3.71 ± 0.11
	Experimental group	F <sub>2,21</sub> = 0.28 (0.76)	Big clip	7	-3.72 ± -0.18
			Little clip	10	-3.60 ± 0.13
No clip			9	-3.73 ± 0.13	

<sup>a</sup>Ages: ASY (After Second Year) includes males who had bred on the site in a previous year. SY (Second Year) includes males banded as nestlings in the previous year; this was therefore their first breeding season.

<sup>b</sup>Log-transformed values are reported.

Table 3.2. Linear correlations between sperm morphometry and male size, condition and health. We corrected for seasonal effects for each sperm trait based on the best-fitting seasonal model. Estimates and test statistics are for the effect of the health/condition feature alone. P values are not corrected for multiple comparisons; after correction with false discovery rate (Benjamini and Hochberg 1995), none of the relationships approached significance.

Sperm trait	Feature	Estimate $\pm$ SE	t	P	n
Total length	Structural size <sup>a</sup>	0.21 $\pm$ 0.30	0.68	0.50	79
	Body condition	1.42 $\pm$ 0.78	1.84	0.07	79
	H:L <sup>c</sup>	2.17 $\pm$ 0.79	2.75	0.01	27
	Percent bacteria killed, early <sup>d</sup>	8.72 $\pm$ 4.46	1.95	0.07	17
	Percent bacteria killed, late	-0.16 $\pm$ 2.74	-0.06	0.96	17
Flagellum: head ratio	Structural size	0.01 $\pm$ 0.03	0.45	0.66	79
	Body condition	-0.04 $\pm$ 0.07	-0.53	0.60	79
	H:L	0.01 $\pm$ 0.08	0.15	0.88	27
	Percent bacteria killed, early	-0.46 $\pm$ 0.27	-1.71	0.11	17
	Percent bacteria killed, late	-0.06 $\pm$ 0.26	-0.24	0.81	17
Midpiece: total ratio	Structural size	-0.003 $\pm$ 0.003	-0.88	0.38	79
	Body condition	0.005 $\pm$ 0.01	0.55	0.58	79
	H:L	-0.01 $\pm$ 0.01	-1.12	0.27	27
	Percent bacteria killed, early	0.004 $\pm$ 0.07	0.06	0.95	17
	Percent bacteria killed, late	0.005 $\pm$ 0.02	0.29	0.77	17
Sperm length variability	Structural size	0.02 $\pm$ 0.03	0.58	0.57	79
	Body condition	-0.04 $\pm$ 0.07	-0.55	0.58	79
	H:L	-0.03 $\pm$ 0.00	1.54	0.13	27
	Percent bacteria killed, early	0.10 $\pm$ 0.46	0.21	0.84	17
	Percent bacteria killed, late	0.34 $\pm$ 0.30	1.15	0.27	17

<sup>a</sup> Structural size is the first principal component of wing chord, tarsus, and tail length.

<sup>b</sup> Body condition is weight with structural size included as a covariate.

<sup>c</sup> Heterophil:Lymphocyte ratio (H:L) was log-transformed for normality. 2009 data only.

<sup>d</sup> Early (mostly pre-incubation) bactericidal assays were run using 10  $\mu$ L blood; late (nestling stage) bactericidal assays were run using 5  $\mu$ L. Neither could be normalized, so Spearman rank correlation ( $\rho^2$ ) was performed separately for each. Note that there is overlap in the timing of early and late samples between males due to re-nesting and second broods. 2009 data only.

slightly across the season (Figure 3.1). Date also affected the level of variability in sperm length within ejaculate samples, with the lowest variability in the middle of the breeding season, perhaps corresponding to seasonal changes in testicular size and structure.

Between years, total sperm length (TSL) and the midpiece:TSL ratio (M:TSL) were highly consistent within individual males, and the F:H also tended to be consistent between years (Figure 3.2). The repeatability of sperm morphology between years is striking given that the testes of birds regress and re-grow between breeding seasons (Aire 2007a), and is consistent with a strong genetic basis for sperm morphology (Birkhead et al. 2005; Mossman et al. 2009).

Given the consistent sperm morphology within males, it is perhaps unsurprising that sperm morphology did not relate to our measures of male body condition or health (Tables 1 and 2). However, correlating potentially energy-limited traits with condition is problematic, because investing in a costly trait can reduce body condition, as suggested for sperm energetic content in Lake Whitefish (*Coregonus clupeaformis*; Burness et al. 2008). Furthermore, sperm cell production and maturation requires approximately two weeks in birds (Aire 2007b), so the lack of a correlation with current body condition is only imperfect evidence that sperm quality is not condition-dependent. We did not test whether other ejaculate characteristics, such as sperm number or the proportion of normal sperm, are related to body condition.

Previous studies have reported no age effects on sperm morphology (Laskemoen et al. 2008; Møller et al. 2009; Laskemoen et al. 2010; Rowe et al. 2010). Although the M:TSL tended to decrease with age in House Wrens (Table 1), after correction for multiple testing, no sperm trait differed significantly between age classes. Lack of an age effect is consistent with the high between-year repeatability we observed in TSL and M:TSL (Figure 3.2). Sperm also did not

differ between socially monogamous and polygynous males, suggesting that attractiveness to females is not related to sperm morphology.

F:H increased with sample date, and variability in TSL within ejaculate samples ( $CV_{wm}$ ) decreased with date initially and then remained constant (Figure 3.1). Date was not correlated with M:TSL, and date correlations with TSL were not consistent across years. These seasonal patterns could be driven by between-male differences in sperm morphology coupled with a bias in the timing of capture (e.g., males that were captured late in the season tended to be individuals with high F:H) or by within-male changes in sperm traits across the season. While we cannot eliminate the possibility that seasonal patterns were driven by between-male differences, we find the explanation unlikely, as the date of sperm sampling was not related to any measure of male quality (data not shown), and male quality was not related to sperm morphology. Further studies are needed to determine whether seasonal changes in F:H are driven by within-male changes.

At a mechanistic level, variation in sperm morphology within ejaculates could arise from conflicts among the haploid sperm (Parker and Begon 1993) or from errors in the sperm production process (Cohen 1967; Knudsen 2009). We suggest that changes in the level of variability in sperm morphology could also arise if the testes produce highly variable sperm when these organs are not in full breeding condition. The testes of passerine birds undergo dramatic seasonal changes, reaching a maximum size at the peak of breeding (Calhim and Birkhead 2007). Testes that are not in full breeding condition may have more-variable seminiferous tubules and therefore may produce more-variable sperm. This hypothesis would predict highly variable sperm at the extremes of the season, when the testes are not in full breeding condition, as we observed in the  $CV_{wm}$  measure.

Although the seasonal variation in sperm morphology may be simply a non-adaptive by-product of seasonal testicular growth and regression, the increase in F:H could also have adaptive consequences. It could allow sperm to perform more successfully in an environment with changing levels of sperm competition, or it could parallel seasonal changes in the female reproductive tract. In some populations of House Wrens (Johnson et al. 2002) though not in ours (LaBarbera et al. 2010), the incidence of extra-pair paternity, and putatively sperm competition, increases across the season. F:H correlates with swimming speed (Helfenstein et al. 2010a; Lüpold et al. 2009a; Mossman et al. 2009; Immler et al. 2010; Laskemoen et al. 2010; but see Kleven et al. 2009a; Lüpold et al. 2009b), and may therefore be important in sperm competition. Seasonal change in sperm motility has been shown in Great Tits (Helfenstein et al. 2010b), and captive Gouldian Finches increase the length of their sperm midpiece when exposed to higher social competition (Immler et al. 2010). Further work is needed to assess potential fitness consequences of seasonal variation in sperm morphology.

#### ACKNOWLEDGEMENTS

We thank Cornell University Research Ponds for access to the field site, field assistants (Eileen McIver, Katie Baird, Noelle Chaine, Natalie Koscal, and Carly Hodes) and others (Paulo Llambías, Taza Schaming, Katie LaBarbera, Kim Bostwick, Charles Dardia, and Irby Lovette and the Evolutionary Biology Lab) for field and logistical support; Elaina Tuttle, Stephen Pruett-Jones, Emma Greig, anonymous reviewers, and the Cornell Behavior Journal Club for feedback on the manuscript and data interpretation; Sandy Vehrencamp for support throughout; and Bob Doran and the Paula Cohen lab for microscope assistance in the USA. Funding was provided by grants from the Animal Behavior Society, American Ornithologists' Union, Cornell Department

of Neurobiology and Behavior, Cornell University Sigma Xi Chapter, as well as a donation to the Cornell Lab of Ornithology from the Kramer family and a National Science Foundation (USA) Graduate Research Fellowship and Nordic Research Opportunity Fellowship. TL, OK and JTL were supported by a grant from the Research Council of Norway. This study was approved by the Cornell University Institutional Animal Care and Use Committee, and complied with the current laws in Norway and the USA.

## REFERENCES

- Aire TA. 2007a. Anatomy of the testis and male reproductive tract. In: Jamieson BGM (ed) Reproductive Biology and Phylogeny of Birds, Vol 6A. Science Publishers, Enfield, New Hampshire, pp 37-114.
- Aire TA. 2007b. Spermatogenesis and testicular cycles. In: Jamieson BGM (ed) Reproductive Biology and Phylogeny of Birds, Vol 6A. Science Publishers, Enfield, New Hampshire, pp 279-347.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 57:289-300.
- Birkhead TR, Pellatt EJ, Brekke P, Yeates R, Castillo-Juarez H. 2005. Genetic effects on sperm design in the zebra finch. *Nature*. 434:383-387.
- Birkhead TR, Pizzari T. 2002. Postcopulatory sexual selection. *Nature Reviews*. 3:262-273.
- Briskie JV, Montgomerie R, Birkhead TR. 1997 The evolution of sperm size in birds. *Evolution*. 51:937-945.
- Burness G, Schulte-Hostedde AI, Montgomerie R. 2008. Body condition influences sperm energetics in lake whitefish (*Coregonus clupeaformis*). *Can J Fish Aquat Sci*. 65:615-620.
- Calhim S, Birkhead T. 2007. Testes size in birds: quality versus quantity--assumptions, errors, and estimates. *Behav Ecol*. 18:271-275.
- Calhim S, Immler S, Birkhead TR. 2007. Postcopulatory sexual selection is associated with reduced variation in sperm morphology. *PLoS One*. 2 5:e413.
- Cardullo RA, Baltz JM. 1991. Metabolic regulation in mammalian sperm: mitochondrial volume determines sperm length and flagellar beat frequency. *Cell Motil Cytoskel*. 19:180-188.
- Cohen J. 1967. Correlation between sperm "redundancy" and chiasma frequency. *Nature*. 215:862-863.
- Dewsbury DA. 1982. Ejaculate cost and male choice. *Am Nat*. 119:601-610.

- Froman DP, Kirby JD. 2005. Sperm mobility: phenotype in roosters (*Gallus domesticus*) determined by mitochondrial function. *Biol Reprod.* 72:562-567.
- Froman DP, Pizzari T, Feltmann AJ, Castillo-Juarez H, Birkhead TR. 2002. Sperm mobility: mechanisms of fertilizing efficiency, genetic variation and phenotypic relationship with male status in the domestic fowl, *Gallus gallus domesticus*. *Proc R Soc Lond B.* 269:607-612.
- García-Berthou E. 2001. On the misuse of residuals in ecology: testing regression residuals versus the analysis of covariance. *J Anim Ecol.* 70:708-711.
- Helfenstein F, Losdat S, Møller AP, Blount JD, Richner H. 2010b. Sperm of colourful males are better protected against oxidative stress. *Ecol Lett.* 13:213-222.
- Helfenstein F, Podevin M, Richner H. 2010a. Sperm morphology, swimming velocity, and longevity in the house sparrow *Passer domesticus*. *Behav Ecol Sociobiol.* 64:557-565.
- Helfenstein F, Szep T, Nagy Z, Kempnaers B, Wagner RH. 2008. Between-male variation in sperm size, velocity and longevity in sand martins *Riparia riparia*. *J Avian Biol.* 39:647-652.
- Humphreys PN. 1972. Brief observations on the semen and spermatozoa of certain passerine and non-passerine birds. *J Reprod Fert.* 29:327-336.
- Humphries S, Evans JP, Simmons LW. 2008. Sperm competition: linking form to function. *BMC Evol Biol.* 8:319.
- Immler S, Birkhead TR. 2007. Sperm competition and sperm midpiece size: no consistent pattern in passerine birds. *Proc R Soc Lond B.* 274:561-568.
- Immler S, Calhim S, Birkhead TR. 2008. Increased postcopulatory sexual selection reduces the intramale variation in sperm design. *Evolution.* 62:1538-1543.
- Immler S, Pryke SR, Birkhead TR, Griffith SC. 2010. Pronounced within-individual plasticity in sperm morphometry across social environments. *Evolution.* 64:1634-1643.
- Johnson LS, Hicks BG, Masters BS. 2002. Increased cuckoldry as a cost of breeding late for male house wrens (*Troglodytes aedon*). *Behav Ecol.* 13:670-675.

Kleven O, Fossøy F, Laskemoen T, Robertson RJ, Rudolfsen G, Lifjeld JT. 2009a. Comparative evidence for the evolution of sperm swimming speed by sperm competition and female sperm storage duration in passerine birds. *Evolution*. 63:2466-2473.

Kleven O, Laskemoen T, Fossøy F, Robertson RJ, Lifjeld JT. 2008. Intraspecific variation in sperm length is negatively related to sperm competition in passerine birds. *Evolution*. 62:494-499.

Kleven O, Laskemoen T, Lifjeld JT. 2009b. Sperm length in sand martins *Riparia riparia*: a comment on Helfenstein et al. *J Avian Biol* 40:241-242.

Knudsen J. 2009. Sperm production and variance in sperm quality. Master's Thesis, Queen's University.

Koehler LD. 1995. Diversity of avian spermatozoa ultrastructure with emphasis on the members of the order Passeriformes. *Mem Mus Natn Hist Nat*. 166:437-444.

LaBarbera K, Llambías PE, Cramer ERA, Schaming TD, Lovette IJ. 2010. Synchrony does not explain extrapair paternity rate variation in northern or southern house wrens. *Behav Ecol*. 21:773-780.

Laskemoen T, Kleven O, Fossøy F, Robertson RJ, Rudolfsen G, Lifjeld JT. 2010. Sperm quantity and quality effects on fertilization success in a highly promiscuous passerine, the tree swallow *Tachycineta bicolor*. *Behav Ecol Sociobiol*. 64:1473-1483.

Laskemoen T, Fossøy F, Rudolfsen G, Lifjeld JT. 2008. Age-related variation in primary sexual characters in a passerine with male age-related fertilization success, the bluethroat *Luscinia svecica*. *J Avian Biol*. 39:322-328.

Lessells CM, Boag PT. 1987. Unrepeatable repeatabilities: a common mistake. *Auk*. 104:116-121.

Lifjeld JT, Laskemoen T, Kleven O, Pedersen ATM, Lampe HM, Rudolfsen G, Schmoll T, Slagsvold T. 2012. No evidence for pre-copulatory sexual selection on sperm length in a passerine bird. *PLoS One*. 7:e32611.

Lifjeld JT, Laskemoen T, Kleven O, Albrecht T, Robertson RJ. 2010. Sperm length variation as a predictor of extrapair paternity in passerine birds. *PLoS One*. 5:e13456.

- Lüpold S, Calhim S, Immler S, Birkhead TR. 2009a. Sperm morphology and sperm velocity in passerine birds. *Proc R Soc Lond B*. 276:1175-1181.
- Lüpold S, Linz GM, Birkhead TR. 2009b. Sperm design and variation in the New World blackbirds (Icteridae). *Behav Ecol Sociobiol*. 63:899-909.
- Lüpold S, Linz GM, Rivers JW, Westneat DF, Birkhead TR. 2009c. Sperm competition selects beyond relative testes size in birds. *Evolution*. 63:391-402.
- Lüpold S, Westneat DF, Birkhead TR. 2011. Geographical variation in sperm morphology in the red-winged blackbird (*Agelaius phoeniceus*). *Evol Ecol*. 25:373-390.
- Millet S, Bennett J, Lee KA, Hau M, Klasing KC. 2007. Quantifying and comparing constitutive immunity across avian species. *Dev Comp Immunol*. 31:188-201.
- Møller AP. 1991. Sperm competition, sperm depletion, paternal care, and relative testis size in birds. *Am Nat*. 137:882-906.
- Møller AO, Mousseau TA, Rudolfson G, Balbontín J, Marzal A, Hermosell I, De Lope F. 2009. Senescent sperm performance in old male birds. *J Evol Biol*. 22:334-34.
- Mossman J, Slate J, Humphries S, Birkhead TR. 2009. Sperm morphology and velocity are genetically codetermined in the zebra finch. *Evolution*. 63:2730-2737.
- Nakagawa S, Schielzeth H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol Rev*. 85:935-956.
- Ots I, Murumägi A, Hõrak P. 1998. Haematological health state indices of reproducing great tits: methodology and sources of natural variation. *Funct Ecol*. 12:700-707.
- Parker GA, Begon ME. 1993. Sperm competition games: sperm size and number under gametic control. *Proc R Soc Lond B*. 253:255-262.
- R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rowe M, Swaddle JP, Pruett-Jones S, Webster MS. 2010. Plumage coloration, ejaculate quality and reproductive phenotype in the red-backed fairy wren. *Anim Behav*. 79:1239-1246.

Schiele H. 2010. Simple means to improve the interpretability of regression coefficients. *Methods Ecol Evol.* 1:103-113.

Schmoll T, Kleven O. 2011. Sperm dimensions differ between two coal tit *Parus ater* populations. *J Ornithol.* 152:515-520.

Tieleman BI, Dijkstra TH, Klasing KC, Visser GH, Williams JB. 2008. Effects of experimentally increased costs of activity during reproduction on parental investment and self-maintenance in tropical house wrens. *Behav Ecol.* 19:949-959.

Tuttle EM, Pruett-Jones S, Webster MS. 1996. Cloacal protuberances and extreme sperm production in Australian fairy-wrens. *Proc R Soc Lond B.* 263:1359-1364.

Verhoeven KJF, Simonsen KL, McIntyre LM. 2005. Implementing false discovery rate control: increasing your power. *Oikos.* 108:643-647.

## CHAPTER 4

### SPERM MORPHOLOGY DOES NOT RELATE TO EXTRA-PAIR PATERNITY SUCCESS IN HOUSE WRENS

EMILY R. A. CRAMER, TERJE LASKEMOEN, ODDMUND KLEVEN, KATIE  
LABARBERA, IRBY J. LOVETTE, AND JAN T. LIFJELD

#### ABSTRACT

The potential evolutionary importance of sperm competition in internally fertilizing vertebrates has recently become a topic of great interest, yet relatively little is known about what characteristics confer competitive advantages in sperm competition. In this intraspecific study, we investigated several measures of sperm morphology that have been identified as potentially contributing to sperm success in interspecific studies. We tested whether males with differing success in sperm competition via extra-pair paternity had different sperm morphology. We found no evidence for this pattern, and moreover, found no correlation between sperm morphology and annual reproductive success after accounting for extra-pair paternity. Males may use behavioral strategies to improve the success of extra-pair copulations, and these may have a larger effect than sperm morphology does on the outcome of sperm competition. Selection on house wrens sperm may be relatively weak despite moderate levels of sperm competition, and within-species variability in sperm morphology may be relatively low, making it difficult to detect elements of sperm morphology that contribute to success in sperm competition.

#### INTRODUCTION

When females copulate with multiple males, sperm from those males can compete, and females can exert cryptic choice for certain sperm traits (Parker 1970, Eberhard 1996). While it is clear in some species that sperm characteristics can affect fertilization success (e.g., Birkhead et al. 1999), for most internally-fertilizing vertebrates and passerines in particular, we know relatively little about how between-male variation in sperm characteristics affects the outcome of sperm competition (*sensu lato*). Furthering this knowledge is important for understanding how sperm competition works, and for understanding whether sperm traits correlate with other male phenotypes in a way that could confound studies of pre-copulatory sexual selection (e.g., Sheldon 1994, Andersson and Simmons 2006).

Interspecific studies have identified several sperm characteristics that are likely to be important in sperm competition in passerines. Species with more frequent multiple mating have larger testes (putatively allowing them to produce a higher number of sperm; Møller and Briskie 1995, Lifjeld et al. 2010), faster-swimming sperm (Kleven et al. 2009), longer sperm cells (Kleven et al. 2009, Lüpold et al. 2009a, Lifjeld et al. 2010, but see Immler and Birkhead 2007), and lower variability in sperm length, likely reflecting stronger stabilizing selection for optimal sperm traits in these species (Calhim et al. 2007, Immler et al. 2008, Kleven et al. 2008, Lifjeld et al. 2010). Thus, increased sperm number, length, and swimming speed may improve a male's sperm competitive ability.

Several intraspecific studies in domestic birds have demonstrated that sperm motility affects fertilization success: when females were artificially inseminated with a mixture of high and low motility sperm, the high-motility sperm sired a disproportionate number of offspring (Birkhead et al. 1999, Donoghue et al. 1999, Denk et al. 2005). Work from both inter- and intra-specific studies shows that swimming speed may be increased by increasing the relative length of

the sperm midpiece (a single, fused mitochondrion that wraps around the flagellum; Lüpold et al. 2009a, Laskemoen et al. 2010) or by increasing the relative length of the flagellum (Lüpold et al. 2009a, Mossman et al. 2009, Helfenstein et al. 2010a, Immler et al. 2010; although other papers find no relationship: Kleven et al. 2009a, Lüpold et al. 2009b, Laskemoen et al. 2010, Lifjeld et al. 2012). This pattern makes mechanical sense because the flagellum and midpiece are thought to contribute to forward motion while the head of the sperm (which contains the DNA) produces only drag (Humphries et al. 2008). Consistent with swimming speed being important in sperm competition, passerines species with higher sperm competition have relatively longer midpieces (Lüpold et al. 2009a; though swimming speed did not correlate with the strength of sperm competition in that study).

Taken together, these studies suggest that variation in sperm morphology and swimming speed within a species could affect male sperm competitive ability. To our knowledge, only one paper has tested this prediction in a wild bird: Laskemoen et al. (2010) found little effect of sperm morphology on sperm competition in tree swallows (*Tachycineta bicolor*), although males with larger cloacal protuberances (likely indicating higher sperm production) and with sperm with longer midpieces had higher fertilization success overall. Here, we examined how sperm characteristics relate to fertilization success in another passerine, the house wren (*Troglodytes aedon*). House wrens have moderate level of extra-pair (EP) paternity (e.g., Forsman et al. 2008, Labarbera et al. 2010), generating the potential for substantial sperm competition between males. We assumed that EP sires have an inherent disadvantage in sperm competition (as is commonly assumed to be the case, due to predicted higher rates of within-pair [WP] copulations, e.g., Birkhead et al. 1987). Differences in the morphology of sperm between EP and WP males could therefore reflect sperm characteristics that improve the competitive ability of EP males' sperm

and allow them to overcome this disadvantage and achieve fertilizations. Based on results from the above interspecific studies (also see Laskemoen et al. 2010), we predicted that EP males should have longer sperm, sperm length closer to the population mean length, and/or a longer relative midpiece or flagellum length. This prediction assumes that longer midpieces and/or flagella promote faster swimming, and that faster swimming offers an advantage in sperm competition. Because low variability in sperm morphology within an ejaculate sample could reflect more consistent (and therefore better) sperm production capacities, we further predicted that EP males would have lower variation in sperm morphology, regardless of their mean sperm length.

## METHODS

### *Field methods and study system*

House wrens are migratory, double-brooded cavity-nesting passerines with approximately 15-20% of offspring sired by extra-pair males on our two field sites in Ithaca, NY (lat 42°31'N, 76°28'W; Labarbera et al. 2010 and see paternity results below). The sites are 3 km apart, with about 75 nest boxes on each site. For details on the study sites, see Llambías (2009). Details on sperm sampling are given in Cramer et al. (accepted), and details on paternity analysis and field methods are in Cramer (in review). Briefly, we captured, banded, and bled all individuals between April and August, 2009-2011, and collected an ejaculate sample from males using cloacal massage (e.g., Kleven et al. 2008). Nestlings were banded at approximately day 8, and all banded nestlings were assumed to fledge unless there were signs of nest depredation.

We measured the head, midpiece, and tail length of 30 morphologically normal sperm cells per ejaculate sample using light microscopy (2009: 320x magnification, Lieca

Microsystems DM6000B DFC420 Leica digital camera, Heerbrugg, Switzerland; 2010 and 2011: 400x magnification, Zeiss Axiovert 200M with AxioCam MRm, Carl Zeiss Inc) and the on-screen cursor line tool in the Leica Application Suite (version 2.6.0 R1; 2009 samples) or the line tool in ImageJ with a custom plug-in (NIH, 2010 and 2011 samples; see further details in Cramer et al. accepted). From the measurements, we calculated total sperm length, flagellum : head ratio, midpiece : total sperm length ratio, and variability (estimated by the coefficient of variation in the total length of the sperm cells). To test the hypothesis concerning the divergence of sperm length from the population mean, we took the absolute value of the difference of sperm length from the mean sperm length of the year (to control for microscope effects). Hereafter, we call this variable “length extremeness.”

Details on paternity analysis are in Chapter 2; here, we report only on the sub-set of individuals for which sperm data were available ( $n = 57$  males sampled in a single year only, 12 males sampled in two years, and 2 males sampled in 3 years, for a total of 87 male-years; for more information about within- and between-season variation in sperm morphology, see Cramer et al. in press). Briefly, we used the loci and genotyping conditions described in LaBarbera et al. (2010), compared the genotypes of offspring and their putative parents in Cervus 3.0 (Kalinowski et al. 2007), located mismatches between social parents and offspring using GenoPed (Z. Zhang), and confirmed those mis-matches by re-genotyping. To conservatively estimate extra-pair paternity, we attributed a chick to extra-pair paternity if it had more than one “non-null” mismatch with its social father (null allele rates were estimated in Cervus 3.0 and are reported for the population in LaBarbera et al. 2010). We also allowed single null-allele mismatches between the candidate extra-pair father and the offspring. Most assigned EP fathers were territorial neighbors rearing their own broods.

We categorized males as having been cuckolded if one or more of his social offspring was sired by another male. We categorized males as having failed to sire EP offspring in another nest on the study site only if he stayed on the study site long enough to attract a female and failed to sire EP offspring (i.e., we excluded males that left the study site soon after capture). The “not cuckolded” category could include males with successful sperm that outcompeted that of all EP males, as well as males whose females did not mate multiply. Conversely, the “not EP sire on-site” category could include males with very unsuccessful sperm that copulated with EP female(s) without achieving fertilizations as well as males that failed to attract EP females for copulation. We therefore created two additional categories using males whose sperm was known to have been in competition: males whose sperm was known to compete be successful (i.e., males that gained EP paternity in other nests without losing any WP paternity in their own broods), and those whose sperm was only known to be less successful (i.e., males that lost WP paternity without gaining EP offspring in other nests). We call this the “restricted” data set.

### *Statistical analysis*

We compared sperm traits for males that were cuckolded to males that were not cuckolded, and for males that were extra-pair sires on-site to males that were not, using unpaired tests assuming unequal variances. Ten males had paternity and sperm data in more than one year, and we randomly chose which year to include for each male (final sample sizes: 29 males maintained complete WP paternity vs. 27 males were cuckolded; 22 males sired EP offspring on the study site vs. 49 did not; 10 males that were only known to succeed in both categories vs. 22 males that were only known to have failed in both categories). Some males had two sperm samples collected in the same year, and we randomly chose one to include in the analysis.

We conducted paired t-tests to compare the sperm morphology of EP males to the WP males they cuckolded. Because of incomplete sperm sampling, the final data set for paired comparisons included 25 unique pairs of EP and WP sires, encompassing 34 males. Seven males appeared twice in this data set, and single males appeared three, four, and five times in the data set; the males that appeared four and five times were sampled in two different years. Four males were both EP and WP males, including one instance of reciprocal cuckoldry, where each member of the pair cuckolded the other.

Males were sampled at various points during the season, and sample date is correlated with various aspects of sperm morphology (Cramer et al. accepted). While it seems likely that this seasonal effect reflects within-male changes in sperm morphology rather than between-male differences in sperm morphology coupled with a bias in capture timing, we lacked the data to test the mechanism thoroughly (Cramer et al. accepted). We find it more conservative to compare actual sperm measurements rather than performing a statistical control for date, when potential within-male changes in sperm morphology are uncertain. Moreover, the date of sperm sampling did not differ significantly for any comparison (maintaining own paternity,  $t_{51.78} = 0.57$ ,  $p = 0.57$ ; gaining EP paternity in other nests,  $t_{32.89} = 1.32$ ,  $p = 0.20$ ; paired test: mean difference  $9.4 \pm 5.55$  days;  $t_{24} = -1.69$ ,  $p = 0.10$ ), suggesting that date effects should not bias our results. Analyses attempting to control for date using residuals from the population-wide correlations between date and sperm morphology gave similar results to those presented here (not shown).

Our a priori predictions concerned sperm total length, length extremeness, within-male variability in sperm morphology, and the relative investment in the flagellum : head ratio and the midpiece : total sperm length ratio. For completeness, we also tested for differences between groups in the length of individual sperm components.

Only sperm samples taken in the year of the EP event were used. All statistical tests were performed in JMP 7.0 (SAS Institute, Cary, NC). We used non-parametric tests for unpaired tests of variability in sperm length, the midpiece : total sperm length ratio, and midpiece length because they were not normally distributed. Reproductive success could not be transformed, so we also used nonparametric correlations to test for associations with reproductive success.

## RESULTS

Sperm characteristics from males that were cuckolded did not differ from those of males that were not cuckolded, and males that sired EP offspring in other nests on-site did not differ from those that did not sire EP offspring elsewhere (Table 4.1). While there was a trend for males that maintained complete WP paternity to have a larger midpiece : total sperm length ratio, the pattern was not apparent in the restricted data set (i.e., including only males whose sperm was known to have been in competition with the sperm of other males). Males that failed to gain EP paternity in other nests tended to have longer absolute midpiece lengths, but again this trend was not evident in the restricted data set. Moreover, sperm measures did not differ between EP males and the WP males they cuckolded in paired comparisons (Table 4.2). Sperm measures did not correlate with annual genetic reproductive success (Table 4.3).

## DISCUSSION

Our prediction that sperm morphology would differ between males with relatively successful and relatively unsuccessful sperm was not supported. Sperm morphology did not differ between males that succeeded in sperm competition (that is, males that gained extra-pair paternity in other nests and maintained full within-pair paternity in their own broods) and males

Table 4.1. Comparisons of males with relatively successful or unsuccessful performance in sperm competition. We defined success within “Own” nests as maintaining complete paternity within their own social broods. Success in “Other” nests was siring offspring in the nest of another male on site. Some males gained EP offspring in other nests without losing WP paternity (successful in “Both” nests), and others lost WP paternity without gaining EP offspring in other nests (unsuccessful in “Both” nests).

Sperm trait	Nest Type	Mean $\pm$ SE (95% Confidence Interval)		Test statistic (p)
		Successful males	Unsuccessful males	
Total Sperm Length	Own	77.19 $\pm$ 0.68 (75.79, 78.59)	78.36 $\pm$ 0.71 (76.90, 79.82)	t <sub>54.00</sub> = 1.19 (0.24)
	Other	77.48 $\pm$ 0.70 (76.02, 78.93)	78.06 $\pm$ 0.57 (76.92, 79.21)	t <sub>48.83</sub> = -0.65 (0.52)
	Both	77.94 $\pm$ 1.12 (75.41, 80.47)	78.81 $\pm$ 0.83 (77.09, 80.53)	t <sub>19.09</sub> = 0.63 (0.54)
Length Extremeness	Own	1.45 $\pm$ 0.14 (1.17, 1.73)	1.59 $\pm$ 0.13 (1.32, 1.86)	t <sub>53.72</sub> = 0.72 (0.47)
	Other	1.45 $\pm$ 0.14 (1.16, 1.74)	1.58 $\pm$ 0.11 (1.37, 1.79)	t <sub>44.81</sub> = -0.73 (0.47)
	Both	1.47 $\pm$ 0.22 (0.98, 1.97)	1.61 $\pm$ 0.16 (1.28, 1.95)	t <sub>19.05</sub> = 0.52 (0.61)
Variability in Total Length	Own	2.98 $\pm$ 0.23 (2.52, 3.45)	2.96 $\pm$ 0.24 (2.46, 3.46)	Z = 0.54 (0.59)
	Other	2.88 $\pm$ 0.31 (2.22, 3.53)	2.95 $\pm$ 0.17 (2.61, 3.30)	Z = -0.96 (0.34)
	Both	3.03 $\pm$ 0.47 (1.97, 4.10)	2.89 $\pm$ 0.24 (2.39, 3.39)	Z = -0.1 (0.92)
Flagellum : Head	Own	5.19 $\pm$ 0.07 (5.05, 5.33)	5.21 $\pm$ 0.07 (5.07, 5.35)	t <sub>53.96</sub> = 0.20 (0.85)
	Other	5.20 $\pm$ 0.08 (5.03, 5.37)	5.21 $\pm$ 0.05 (5.11, 5.31)	t <sub>36.58</sub> = -0.10 (0.92)
	Both	5.29 $\pm$ 0.15 (4.95, 5.62)	5.25 $\pm$ 0.08 (5.09, 5.41)	t <sub>14.04</sub> = -0.21 (0.84)
Midpiece : Total Length	Own	0.66 $\pm$ 0.01 (0.64, 0.67)	0.64 $\pm$ 0.01 (0.62, 0.65)	Z = 1.93 (0.053)
	Other	0.64 $\pm$ 0.01 (0.62, 0.66)	0.65 $\pm$ 0.01 (0.64, 0.66)	Z = -1.25 (0.21)
	Both	0.64 $\pm$ 0.02 (0.61, 0.68)	0.64 $\pm$ 0.01 (0.63, 0.66)	Z = 0.14 (0.89)
Head Length	Own	12.52 $\pm$ 0.14 (12.24, 12.80)	12.67 $\pm$ 0.12 (12.43, 12.90)	t <sub>51.95</sub> = 0.80 (0.43)
	Other	12.56 $\pm$ 0.16 (12.23, 12.89)	12.62 $\pm$ 0.08 (12.45, 12.79)	t <sub>32.94</sub> = -0.31 (0.76)
	Both	12.48 $\pm$ 0.31 (11.79, 13.17)	12.65 $\pm$ 0.14 (12.37, 12.94)	t <sub>12.83</sub> = 0.52 (0.61)
Midpiece Length	Own	50.60 $\pm$ 0.66 (49.25, 51.96)	49.97 $\pm$ 0.72 (48.51, 51.44)	Z = 0.52 (0.60)
	Other	49.39 $\pm$ 0.75 (47.83, 50.95)	50.90 $\pm$ 0.50 (49.91, 51.90)	Z = -1.81 (0.07)
	Both	50.12 $\pm$ 1.28 (47.23, 53.02)	50.73 $\pm$ 0.82 (49.02, 52.44)	Z = -0.43 (0.67)
Flagellum Length	Own	64.67 $\pm$ 0.63 (63.38, 65.96)	65.69 $\pm$ 0.69 (64.28, 67.11)	t <sub>53.82</sub> = 1.10 (0.28)
	Other	64.92 $\pm$ 0.66 (63.55, 66.28)	65.45 $\pm$ 0.55 (64.34, 66.55)	t <sub>49.88</sub> = -0.62 (0.54)
	Both	65.46 $\pm$ 1.01 (63.17, 67.75)	66.16 $\pm$ 0.79 (64.51, 67.81)	t <sub>20.15</sub> = 0.54 (0.59)

Table 4.2. Mean values and paired comparisons of EP males to the WP males they cuckolded (n = 25 unique pairs, 34 males)

Sperm trait	EP mean	WP mean	Mean Difference $\pm$ SE (95% CI)	Test Statistic (p)
Total Sperm Length	78.16	78.13	0.03 $\pm$ 0.87 (-1.77, 1.83)	$t_{24} = 0.04$ (0.97)
Length Extremeness	1.39	1.18	0.21 $\pm$ 0.17 (-0.13, 0.56)	$t_{24} = 1.26$ (0.22)
Variability in Total Length	2.79	2.73	0.06 $\pm$ 0.23 (-0.42, 0.54)	$Z = -2$ (0.96)
Flagellum : Head	5.35	5.28	0.07 $\pm$ 0.11 (-0.15, 0.29)	$t_{24} = 0.69$ (0.50)
Midpiece : Total Length	0.64	0.63	0.01 $\pm$ 0.01 (-0.02, 0.03)	$t_{24} = 0.74$ (0.47)
Head Length	12.39	12.52	-0.13 $\pm$ 0.24 (-0.62, 0.37)	$t_{24} = -0.53$ (0.60)
Midpiece Length	49.79	49.18	0.61 $\pm$ 0.89 (-1.24, 2.46)	$t_{24} = 0.68$ (0.50)
Flagellum Length	65.77	65.62	0.16 $\pm$ 0.77 (-1.43, 1.75)	$t_{24} = 0.2$ (0.84)

Table 4.3. Correlations between annual genetic reproductive success and sperm traits (n = 52 males)

Trait	Spearman's $\rho$	p
Total Sperm Length	-0.03	0.84
Length Extremeness	0.03	0.84
Variability in Total Length	0.1	0.47
Flagellum : Head	-0.08	0.56
Midpiece : Total Length	-0.04	0.76
Head Length	0.11	0.42
Midpiece Length	-0.12	0.41
Flagellum Length	-0.03	0.84

that had lower success (i.e., males that failed to gain EP paternity in other nests and lost paternity in their own nests). Sperm morphology did not correlate with genetic reproductive success that includes gains and losses due to EP paternity. Using patterns of EP paternity may not be the most powerful way to detect which characteristics are important in sperm competition (see below). However, for socially monogamous species, sperm competition only occurs through EP copulations, so this type of comparison should reveal sperm characteristics that influence fertilization success.

We assessed EP behavior only by the paternity of nestlings rather than by directly observing copulations. Males with the least successful sperm—those that copulated with females but failed to fertilize her eggs—were therefore not identified, and comparing successful males against these individuals would provide a better test of which sperm traits confer an advantage. Moreover, we cannot be completely sure of which males faced sperm competition: males that maintained complete WP paternity may have effectively guarded their mates and prevented them from engaging in EP copulations, and males that failed to gain EP fertilizations in other nests may not have inseminated any EP females if females rejected them as EP partners. We could partially overcome this issue by restricting the data set to only include males that were known to face sperm competition because they sired offspring in mixed-paternity broods. For this comparison, we defined successful males as gaining EP paternity in other broods without losing WP paternity, and unsuccessful males as losing WP paternity without gaining EP paternity. While this comparison necessarily has reduced sample size and statistical power, the mean sperm values for successful and unsuccessful males were still very similar, suggesting that any effects of sperm morphology on EP fertilization success, if present, must be quite small.

Two of our measures—the flagellum : head ratio and the midpiece : total sperm length ratio—correlate with sperm swimming speed in other species (Lüpold et al. 2009a, Mossman et al. 2009, Helfenstein et al. 2010a, Laskemoen et al. 2010, Immler et al. 2010), and swimming speed is important in sperm competition (e.g., Birkhead et al. 1999, Donoghue et al. 1999, Denk et al. 2005, Snook 2005). However, correlations between sperm swimming speed and sperm morphology differ among species and studies (e.g., Kleven et al. 2009a, Lüpold et al. 2009b, Helfenstein et al. 2010a, Laskemoen et al. 2010, Lifjeld et al. 2012), so the flagellum : head ratio and the midpiece : total sperm length ratio may not correlate with sperm swimming performance in house wrens. If that is the case, and if sperm swimming speed rather than morphology itself is the target of sperm competition, we would not expect to find a difference in sperm morphology between EP and WP males.

Compatibility with the female reproductive tract is likely important in sperm competition, but the mechanisms determining compatibility are unknown (Bakst et al. 1994). Sperm total length correlates strongly with the length of sperm storage tubules (specialized storage tubules at the utero-vaginal junction, which are the source of the sperm that achieve fertilization) across species, suggesting that sperm length may play a role in compatibility (Bakst et al. 1994, Briskie et al. 1997). Individual conspecific females have sperm storage tubules of different lengths (Briskie and Montgomerie 1993). From a very simplistic perspective, if the match between the length of sperm and sperm storage tubules is important, the sperm length that matches best may vary among females. Even with more complex mechanisms, variation among females in which male traits confer a fertilization advantage with that female would make it difficult to detect which male traits are generally most successful.

The results of this study contrast with interspecific patterns, where some aspect of sperm morphology typically correlates with the level of sperm competition (although the particular patterns often depend on the taxa studied; e.g., Immler and Birkhead 2007, Kleven et al. 2008, Lüpold et al. 2009a). Patterns that are apparent at large taxonomic scales may not be detected at smaller taxonomic scales because the variation within the smaller taxonomic scale is typically lower (e.g., Read and Weary 1992). Two other studies investigating sperm morphology and fertilization success in birds also found no effect of morphology on fertilization success (Denk et al. 2005, Laskemoen et al. 2010), although the ratio of midpiece : total sperm length was significantly related to reproductive success in multivariate, but not simple, analyses in Laskemoen et al. (2010). House wren sperm morphology may be under relatively weak selection, despite a moderate degree of multiple mating by females. Across passerine species, between-male variability in sperm length strongly predicts the level of extra-pair paternity (Calhim et al. 2007, Kleven et al. 2008, Lifjeld et al. 2010). Based on the regression line in Lifjeld et al. (2010) and a mean proportion of EP offspring of 13.7% across four years in this population, the between-male variability in sperm length is expected to be 2.82%, while the observed between-male variability is 4.0-5.0% (Cramer et al. accepted). Assuming that lower variability in sperm length in species with higher sperm competition reflects stronger stabilizing selection (e.g., Lifjeld et al. 2010), this relatively high level of between-male variability may indicate relaxed selection on sperm morphology in house wrens. Why house wren sperm morphology might be under relatively weak selection remains unclear.

Successful EP males may use behavioral strategies that enhance fertilization success irrespective of their sperm morphology, which would make it difficult to detect effects of sperm morphology on reproductive success. The last male to copulate often sires a disproportionate

number of offspring (Birkhead and Møller 1992), and copulating at the peak of female fertility may also improve fertilization success. Males could strategically invest large numbers of sperm into EP copulations, as occurs in domestic fowl (Pizzari et al. 2003). Studies on passerine copulation are rare, but in several species, EP males do not appear to time their copulations better than WP males (e.g., Johnsen et al. 2012), nor do EP males physiologically control the number of sperm ejaculated in a strategic manner (Birkhead and Fletcher 1995). Relatively little is known about copulation behavior in house wrens (Brylawski and Whittingham 2004), making it difficult to speculate on whether males follow these potential strategies.

Sperm traits may correlate with phenotypes involved in pre-copulatory mate choice (e.g., as predicted by Sheldon 1994), making it necessary to disentangle the effects of pre- and post-copulatory selection to fully understand sexual selection within a species. While several studies in birds have found no correlation between sperm traits and features involved in mate choice (e.g., Birkhead et al. 1997, Lifjeld et al. 2012), some studies have found negative correlations between putatively advantageous sperm and pre-mating traits (Rowe et al. 2010), and other studies have found positive correlations (Peters et al. 2004, Helfenstein et al. 2010b). Factors affecting pre-copulatory sexual selection are also poorly understood in house wrens (Johnson and Searcy 1996, Eckerle and Thompson 2006, Cramer in review), so possible correlations between traits important in pre- and post-copulatory sexual selection remain a valuable line of future inquiry.

In conclusion, sperm morphology does not appear to have a strong effect on sperm competition in house wrens, despite strong evidence that sperm characteristics relate to sperm competition at a between-species level (e.g., Kleven et al. 2009, Lüpold et al. 2009a, Lifjeld et al. 2010). Sperm traits could correlate with other aspects of phenotype that affect mating success;

males may use behavioral strategies to enhance reproductive success, and within-species variation may be too slight to detect effects on reproductive success without very large sample sizes. More work in more species is needed to understand how sperm traits affect sperm competition, and in particular how those traits relate to pre-mating phenotypes.

#### ACKNOWLEDGEMENTS

We thank the Paula Cohen and Bob Doran for access to their microscope, and Dan Fergus, Kelly Zamudio, Sigal Balshine, Kern Reeve, Wes Hochachka, Jon Lambert, Emma Greig, the NBB Behavior reading group, and the Webster and Lovette labs for thought-provoking discussions. This research was supported by NSF Graduate Research Fellowship and the Nordic Research Opportunity Fellowship and grants from the American Ornithologists' Union, Animal Behavior Society, Cornell University Department of Neurobiology and Behavior, and Cornell Lab of Ornithology to ERAC; an Einhorn Discovery Grant to KL.

## REFERENCES

- Andersson M, Simmons LW. 2006. Sexual selection and mate choice. *Trends Ecol Evol.* 21:296-302.
- Bakst MR, Wishart G, Brillard JP. 1994. Oviductal sperm selection, transport, and storage in poultry. *Poult Sci Res.* 5:117-143.
- Birkhead TR, Atkin L, Møller AP. 1987. Copulation behaviour of birds. *Behaviour.* 101:101-138.
- Birkhead TR, Fletcher F. 1995. Depletion determines sperm numbers in male zebra finches. *Anim Behav.* 49:451-456.
- Birkhead TR, Buchanan KL, DeVoogd TJ, Pellatt EJ, Szekely CK, Catchpole CK. 1997. Song, sperm quality and testes asymmetry in the sedge warbler. *Anim Behav.* 53:965-971.
- Birkhead TR, Fletcher F. 1995. Male phenotype and ejaculate quality in the zebra finch *Taeniopygia guttata*. *Proc R Soc Lond B.* 262:329-334.
- Birkhead TR, Martinez JG, Burke T, Froman DP. 1999. Sperm mobility determines the outcome of sperm competition in the domestic fowl. *Proc R Soc Lond B.* 266:1759-1764.
- Birkhead TR, Møller AP. 1992. Sperm competition in birds: evolutionary causes and consequences. Academic Press, San Diego.
- Briskie JV, Montgomerie R. 1993. Patterns of sperm storage in relation to sperm competition in passerine birds. *Condor.* 95:442-454.
- Briskie JV, Montgomerie R, Birkhead TR. 1997. The evolution of sperm size in birds. *Evolution.* 51:937-945.
- Brylawski AMZ, Whittingham LA. 2004. An experimental study of mate guarding and paternity in house wrens. *Anim Behav.* 68:1417-1424.
- Calhim S, Immler S, Birkhead TR. 2007. Postcopulatory sexual selection is associated with reduced variation in sperm morphology. *PLOS ONE* 2 5:e413.

Cramer ERA, Laskemoen T, Kleven O, and Lifjeld JT. Sperm length variation in house wrens *Troglodytes aedon*. Accepted at J Ornith.

Denk AG, Holzmann A, Peters A, Vermeirssen ELM, Kempnaers B. 2005. Paternity in mallards: effects of sperm quality and female sperm selection for inbreeding avoidance. *Behav Ecol*. 16:825-833.

Donoghue AM, Sonstegard TS, King LM, Smith EJ, Burt DW. 1999. Turkey sperm mobility influences paternity in the context of competitive fertilization. *Biol Reprod*. 61:422-427.

Eberhard WG. 1996. *Female control: sexual selection by cryptic female choice*. Princeton University Press, Princeton.

Eckerle KP, Thompson CF. 2006. Mate choice in house wrens: nest cavities trump male characteristics. *Behaviour* 143:253-271.

Forsman AM, Vogel LA, Sakaluk SK, Johnson BG, Masters BS, Johnson LS, Thompson CF. 2008. Female house wrens (*Troglodytes aedon*) increase the size, but not immunocompetence, of their offspring through extra-pair mating. *Mol Ecol*. 17:3697-3706.

Helfenstein F, Losdat S, Møller AP, Blount JD, Richner H. 2010. Sperm of colourful males are better protected against oxidative stress. *Ecol Lett*. 13:213-222.

Helfenstein F, Podelvin M, Richner H. 2010. Sperm morphology, swimming velocity, and longevity in the house sparrow *Passer domesticus*. *Behav Ecol Sociobiol*. 64:557-565.

Humphries S, Evans JP, Simmons LW. 2008. Sperm competition: linking form to function. *BMC Evol Biol*. 8:319.

Immler S, Birkhead TR. 2007. Sperm competition and sperm midpiece size: no consistent pattern in passerine birds. *Proc R Soc Lond B*. 274:561-568.

Immler S, S. Calhim, and T.R. Birkhead. 2008. Increased postcopulatory sexual selection reduces the intramale variation in sperm design. *Evolution*. 62:1538-1543.

Immler S, Pryke SR, Birkhead TR, Griffith SC. 2010. Pronounced within-individual plasticity in sperm morphometry across social environments. *Evolution*. 64:1634-1643.

Johnsen A, Carter KL, Delhey K, Lifjeld JT, Robertson RJ, Kempenaers B. 2012. Laying-order effects on sperm numbers and on paternity: comparing three passerine birds with different life histories. *Behav Ecol Sociobiol.* 66:181-190.

Johnson LS, Searcy WA. 1996. Female attraction to male song in house wrens (*Troglodytes aedon*). *Behaviour.* 133:357-366.

Kalinowski ST, Taper ML, Marshall TC. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol Ecol.* 16:1099-1106.

Kleven O, Fossøy F, TLaskemoen T, Robertson RJ, Rudolfsen G, Lifjeld JT. 2009. Comparative evidence for the evolution of sperm swimming speed by sperm competition and female sperm storage duration in passerine birds. *Evolution.* 63:2466-2473.

Kleven O, Laskemoen T, Fossøy F, Robertson RJ, Lifjeld JT. 2008. Intraspecific variation in sperm length is negatively related to sperm competition in passerine birds. *Evolution.* 62:494-499.

LaBarbera K, Llambías PE, Cramer ERA, Schaming TD, Lovette IJ. 2010. Synchrony does not explain extrapair paternity rate variation in northern or southern house wrens. *Behav Ecol.* 21:773-780.

Laskemoen T, Kleven O, Fossøy F, Lifjeld JT. 2007. Intraspecific variation in sperm length in two passerine species, the Bluethroat *Luscinia svecica* and the Willow Warbler *Phylloscopus trochilus*. *Ornis Fenn.* 84:131-139.

Laskemoen T, Kleven O, Fossøy F, Robertson RJ, Rudolfsen G, Lifjeld JT. 2010. Sperm quantity and quality effects on fertilization success in a highly promiscuous passerine, the tree swallow *Tachycineta bicolor*. *Behav Ecol. Sociobiol.* 64:1473-1483.

Lifjeld JT, Laskemoen L, Kleven O, Albrecht T, Robertson RJ. 2010. Sperm length variation as a predictor of extrapair paternity in passerine birds. *PLOS ONE.* 5:e13456.

Lifjeld JT, Laskemoen T, Kleven O, Pedersen ATM, Lampe HM, Rudolfsen G, Schmoll T, Slagsvold T. 2012. No evidence for pre-copulatory sexual selection on sperm length in a passerine bird. *PLOS ONE.* 7:e32611.

- Llambías PE. 2009. Why monogamy? Comparing house wren social mating systems in two hemispheres. [dissertation]. Ithaca, NY: Cornell University; 119 p.
- Lüpold S, Calhim S, Immler S, Birkhead TR. 2009. Sperm morphology and sperm velocity in passerine birds. *Proc R Soc Lond B*. 276:1175-1181.
- Lüpold S, Linz GM, Birkhead TR. 2009. Sperm design and variation in the New World blackbirds (Icteridae). *Behav Ecol Sociobiol*. 63:899-909.
- Møller AP, Briskie JV. 1995. Extra-pair paternity, sperm competition and the evolution of testis size in birds. *Behav Ecol Sociobiol*. 36:357-365.
- Mossman J, Slate J, Humphries S, Birkhead TR. 2009. Sperm morphology and velocity are genetically codetermined in the zebra finch. *Evolution* 63:2730-2737.
- Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects. *Biol Rev*. 45:525-567.
- Peters A, Denk AG, Delhey K, Kempenaers B. 2004. Carotenoid-based bill colour as an indicator of immunocompetence and sperm performance in male mallards. *J Evol Biol*. 17:1111-1120.
- Pizzari T, Cornwallis CK, Lovlie H, Jakobsson S, Birkhead TR. 2003. Sophisticated sperm allocation in male fowl. *Nature*. 426:70-74.
- Read AF, Weary DM. 1992. The evolution of bird song: comparative analyses. *Phil Trans Biol Sci*. 338:165-187.
- Rowe M, Swaddle JP, Pruett-Jones S, Webster MS. 2010. Plumage coloration, ejaculate quality and reproductive phenotype in the red-backed fairy wren. *Anim Behav*. 79:1239-1246.
- Sheldon BC, Birkhead TR. 1994. Reproductive anatomy of the chaffinch in relation to sperm competition. *Condor*. 96:1099-1103.
- Snook RR. 2005. Sperm in competition: not playing by the numbers. *Trends Ecol Evol*. 20:46-53.

## CHAPTER 5

### ARE ANDROGENS RELATED TO AGGRESSION IN HOUSE WRENS?

EMILY CRAMER

#### ABSTRACT

Elevated circulating testosterone levels are hypothesized to allow male animals to direct resources into territorial and mating behaviors at the expense of reducing paternal care of offspring. For this hypothesis to apply, testosterone must facilitate territorial/mating behaviors and have antagonistic effects on paternal care, but this pattern has only been supported in some, not all, species. I tested whether androgens correlate with aggressive behaviors in male house wrens (*Troglodytes aedon*), a double-brooded species where paternal and aggressive behaviors overlap temporally. House wrens may therefore benefit from having a hormonal mechanism that allows males to rapidly change behavioral states. However, I found no evidence that androgens (testosterone and 5 $\alpha$ -dihydrotestosterone) relate to aggression in house wrens: androgens did not increase in response to playback, and endogenous circulating androgens were not correlated with how aggressively males responded to those playbacks. Moreover, androgen levels were low during the pre-breeding stage of the second brood, when many males establish new territories and attract new mates. This study adds to a growing body of literature suggesting that the relationship between circulating androgens and aggressive behavior is more complex than originally thought.

## INTRODUCTION

During the reproductive period, males of many species face a trade-off between engaging in territorial/mate-attraction behaviors and investing in offspring care (Trivers 1972). At the proximate level, investment into mating versus paternal care may be influenced by hormonal mechanisms such as circulating levels of testosterone (Hau 2007). In male birds, increasing circulating testosterone is typically thought to increase aggressive/mate attraction behavior and decrease paternal investment. A male could therefore change his testosterone levels to alter his relative investment into each option (e.g., Wingfield et al. 1990, Hau 2007, Ketterson et al. 2009). For instance, he could strategically increase investment into aggression by temporarily increasing circulating testosterone during male-male interactions, when it is crucial to be aggressive but when paternal behavior is not immediately important (an aspect of Wingfield et al. [1990]’s Challenge Hypothesis). It is well-documented in some bird species that testosterone can have such “pleiotropic” effects on multiple different behaviors (e.g., dark-eyed juncos, *Junco hyemalis*; Ketterson et al. 2009). Alternatively, testosterone’s effects on aggressive and mating behaviors could evolve independently from its effects on paternal behaviors (the “evolutionary potential hypothesis” in Hau 2007, or “phenotypic independence” in Ketterson et al. 2009). Before testing whether testosterone mediates alternative investment strategies, then, it is important to test whether testosterone is related to the behaviors of interest in a given species. In this study, I tested whether testosterone relates to aggression in house wrens (*Troglodytes aedon*).

The regulation of aggression and paternal care by testosterone may depend on a species’ ecological or life history traits (reviewed in Hirschenhauser and Oliveira 2006, Lynn 2008, Goymann 2009). Territorial aggression and paternal care are less likely to depend on circulating

testosterone levels in single-brooded bird species (Goymann et al. 2007, Landys et al. 2007) and in bird species with short breeding seasons (Wingfield and Hunt 2002) compared to double-brooded species with longer breeding seasons. In a single-brooded, short-season species, if testosterone increases to facilitate aggression but the increase greatly decreases paternal care, the cost of reduced paternal care could out-weigh the benefit of the aggression. Selection would then act against a single hormonal mechanism that antagonistically affects both behaviors (Hau 2007, Ketterson et al. 2009). The relative costs and benefits may differ in double-brooded species with longer breeding seasons, thereby selecting for a single hormonally-mediated mechanism to regulate both territorial aggression and paternal care (reviewed in Lynn 2008, Goymann 2009). That is, double-brooded, long-season species may benefit more from having a hormonal mechanism to rapidly and strategically alter investment into different components of reproductive success (as argued by Peters 2002).

The house wren is a species where a fast-acting hormonal switch between aggression and parental care may be particularly likely, as male house wrens must simultaneously exhibit parental and aggressive behaviors. House wrens are cavity nesters and are double-brooded; males usually provide nestlings with a substantial amount of food (Johnson et al. 1992, 1993); and polygyny is common (about 25% in the study population, comparable to the 25-40% rate reported in another box-nesting population, Johnson et al. 1993). Territory defense during the nestling period is crucial to nest success, because other prospecting males frequently destroy eggs or nestlings to take over nest cavities (Johnson and Kermott 1993). Successful territory defense may improve a male's chance of being able to breed on that same territory in the second brood (as hypothesized by Drilling and Thompson 1991). Additionally, polygynous males simultaneously invest in non-aggressive and territorial/mate attraction behaviors, because they

advertise for secondary females while attending the nests of their incubating primary females (Johnson and Kermott 1991, Ziolkowski et al. 1997). Switching territories and mates between broods is common (Drilling and Thompson 1991), so males may benefit by exhibiting the same suite of territorial and mating behaviors over a protracted proportion of the breeding season.

House wrens are also an interesting study system because they are polygynous and have high paternal investment. Most work on testosterone and aggression in birds has focused on monogamous species with biparental care or on polygynous species without paternal care. Relatively few polygynous species with paternal care have been studied (but see European starlings *Sturnus vulgaris*, e.g., Gwinner et al. 2002; pied flycatchers *Ficedula hypoleuca*, e.g., Silverin 1993; and yellow-headed blackbirds *Xanthocephalus xanthocephalus*, Beletsky et al. 1990).

I tested the hypothesis that testosterone is associated with aggression in house wrens. Either of two patterns would be consistent with this hypothesis, following the logic of Wingfield et al. (1990). Males might display physiological maximum levels of testosterone throughout the time period where territory defense and mate attraction occur. In this case, circulating testosterone is not expected to increase in response to territorial intrusions, and individual variation in aggressiveness would likely depend on factors other than circulating testosterone. Endogenous testosterone is therefore not predicted to correlate with aggression. Alternatively, males might only express physiological maximum testosterone levels during intense encounters with other males. In this case, testosterone should increase during territorial intrusions, and endogenous testosterone should correlate with individual aggressiveness. The latter scenario is more in accordance with the hypothesis that testosterone allows males to flexibly switch between paternal care and aggression within a day. In either case, testosterone is expected to be at high

levels when males are primarily engaged in territorial defense, that is, before the first brood and during the pre-incubation stage of the second brood.

## METHODS

### *Field methods*

I studied a migratory population of house wrens nesting in nest boxes in edge habitat between forest and bogs or fields in Ithaca, New York, USA (42°31'N, 76°28'W; see Llambías 2009 for details on the study sites). I monitored the pairing status and nesting success of all males on the study sites in April-August 2008-2010. I described males' breeding stages using a continuous variable (days before or after his mate laid the first egg in the clutch) or categorically. I captured and color banded all individuals and took blood samples (up to 100 µl) into heparinized microcapillary tubes by venipuncture of the brachial vein. In 2009, the venipuncture site was sterilized with ethanol for a separate project on immunology. Blood samples were stored on ice for 1 - 6.5 h ( $3.75 \pm 0.2$  mean  $\pm$  SE) and centrifuged to separate plasma from red blood cells. Plasma was stored at -20° C until analysis.

I measured tarsus length, wing chord, and tail length, and I combined these measures into a single measure of body size using principal components analysis. The first principal component had an eigenvalue of 1.61 and explained 54% of the variation in size measures. All three measures loaded positively (eigenvectors: 0.40, 0.67, and 0.63, respectively). I then calculated body condition as the standardized residuals from a regression of body weight on size, controlling for date and time of capture as covariates.

I mist-netted 67 males using playback, and four males were netted before playback began. For netted males, the duration of playback depended partly on netting conditions (e.g.,

shade and wind) and partly on the response of the bird, because highly aggressive individuals usually flew into the net more quickly. Playback ended when the bird was caught. I noted playback duration for all individuals ( $294 \pm 34$  s) and the time from capture until blood collection for 30 individuals ( $407 \pm 32$ , range 120 – 840 s), so I could assess the effects of handling time. I trapped 11 males at the box as they were feeding nestlings. Seven of the 67 males netted with playback were also feeding nestlings (playback duration:  $589 \pm 196$  s). Animals were released at the site of capture immediately after processing. All protocols were approved by Cornell University's IACUC committee (Protocol 2007-0123), and appropriate state and federal permits were obtained (Federal banding under Dr. Sandra Vehrencamp, 20954; New York state permit 1231).

#### *Playback procedures*

For 22 of the males in 2008, I conducted a long playback during which I quantified aggression before capture. To mimic song bouts and decrease the chance of habituation to playback, each playback stimulus consisted of a single song repeated for two minutes at a rate of 1 song/8 s, followed by three min of silence. This 5-min sequence was repeated 6 times (30 min total) before I attempted to capture the male. Two males were exposed to 5 rounds of playback and silence (25 min total), and one male was exposed to a seventh round (35 min total). After a 15-min pause during which the net was unfurled, I played single songs from the playback stimulus at approximately 2-min intervals until the male was captured or until 45 min had passed, at which time the trial was terminated. Each playback stimulus was presented to only one male. Recordings were from distant males in the population and so were presumed to be unfamiliar to the focal male. Playback stimuli were constructed using Syrinx PC (J. Burt) and, after high-pass

filtering at 1 kHz, were stored as .WAV files on an iPod. I considered playback duration to last from the beginning of playback to the time the last song ended, including silent periods.

I played stimuli through an Anchor Minivox speaker tied to a tree 10-15 m from the active nest box and 1-1.5 m above the ground. Each location had many perches with cover both near and far from the speaker. Throughout playback, I noted the distance of the male to the speaker using a 5 m ring of flagging tape as a reference. I also assessed flights across the speaker, defined as a flight that passed within a 2 m horizontal radius of the speaker. Spoken observations and males' vocalizations were recorded using a Marantz PMD 690 recorder and Sennheiser ME67 shotgun microphone. From these recordings, I measured song rate in response to playback and extracted my spoken observations of other behavioral responses. I combined aggressive responses using a principal components analysis including song rate, rate of flights across the speaker, and proportion of time spent within 5 m of the speaker. The first principle component had an eigenvalue of 1.91 and explained 64% of the variance in aggressive behaviors. All three responses loaded positively (eigenvectors: 0.58, 0.61, and 0.53, respectively; loading coefficients: 0.81, 0.84, and 0.73).

All males were in early stages of breeding, either advertising for a female or mated to a female that had not finished laying the full clutch of eggs. One male was advertising for a secondary female and had a primary female incubating.

#### *Testosterone analysis*

I measured androgen concentration using an Extended Range Salivary Testosterone Assay (Salimetrics, Catalog # 1-2402, State College, PA) following the manufacturers' instructions. This kit has been validated for use with plasma from multiple passerine species by Washburn et

al. (2007). To validate the kit in house wrens, I demonstrated parallelism by diluting wren plasma in assay buffer. Recovery of testosterone for three dilutions of pooled wren plasma spiked with testosterone ranged from 85-107%. The intra- and inter-assay coefficients of variation were 12.3 and 11.4%, respectively.

According to the manufacturer's information based on salivary hormones, this assay has 36.4% cross-reactivity with 5 $\alpha$ -dihydrotestosterone (DHT) and 21.0% cross-reactivity with 19-nortestosterone. Cross-reactivity is less than 1.5% with other steroids. Because of the cross-reactivity with DHT, I refer to the Salimetrics kit assay results as T/DHT.

Based on preliminary results, I diluted samples by a factor of 6-15 (males) or 2-3 (females) to fall on the standard curve. For constructing the standard curve, the lowest-concentration testosterone provided with the kit was 6.1 pg/mL; six wells for four individuals (including one female) had concentrations lower than this value in the diluted samples. Because I could not reliably estimate the testosterone concentration below the standard curve, I assigned these wells a concentration of 6.1 pg/mL.

Samples were assayed after the end of each breeding season, which meant that samples were not randomized across the four assay plates run between 2008 and 2011. However, given a coefficient of variation of 11.4%, and based on statistics specifically investigating plate-to-plate differences (not shown), this lack of randomization should not drive the patterns I report. Within plates, samples were randomized. Males for which I assessed aggression were run in triplicate on one assay plate, to ensure comparability.

### *Statistical analysis*

I analyzed data in a series of steps, largely because missing data limited sample sizes for some

comparisons. I first confirmed that T/DHT concentration was not affected by methodological factors by testing for correlations with handling time and method of capture (netting vs. at nest box). Because these factors were not significant (as also found for handling time in other birds by Schwabl et al. 2005; Horton et al. 2010), I did not include them in further models, although I did test their effect again in the final model. I therefore coded males captured at the nest box as not having been exposed to playback, although results were similar when these individuals were instead excluded from analyses involving playback duration.

I then tested for effects that could relate to aggressive behaviors by examining, in a single model, the relationships between T/DHT concentration and: playback duration, breeding stage, male size, male body condition, male age, date, and year. The year term encompasses true annual variation as well as some plate-to-plate variability, as two plates were run in 2008 and one each in 2009 and 2011. For 11 males that advertised but did not breed on site, I estimated the time before the first egg date as the mean value for other males in the advertising stage (-11 days); results were qualitatively unchanged if these males were instead excluded from analyses. Male identity was included as a random factor; 70 individuals were included, with eight sampled twice (once in each of two years) and two sampled three times. I included the significant predictors from this model as covariates in subsequent models. T/DHT concentration was log-transformed for normality.

To test for a relationship between T/DHT and aggression, I compared the first principle component of aggressive response to T/DHT while controlling for body size and capture date (from the above analysis). To determine whether individual response variables (flights across the speaker, proportion of time within 5 m, and song rate) showed different patterns with respect to T/DHT, I also performed the same analysis using each response variable separately. Aggressive

responses had normally distributed error. All statistics were run in JMP 9.0 and all tests were two-tailed.

## RESULTS

### *Correlations with season and playback duration*

Capture date was the only significant predictor of T/DHT levels in a full model relating T/DHT to date, year, time of day, breeding stage, playback duration, size, body condition, and age, although the effects of body size and year approached significance (Table 5.1). I therefore included capture date and size as covariates in the model relating aggressiveness to T/DHT (see below; aggressiveness was only assessed in a single year). Multicollinearity due to correlations among predictors was not problematic (variance inflation factor < 3.5 for all predictors). T/DHT levels were also not associated with breeding stage when I analyzed it as a categorical predictor (i.e., pre-laying, laying, incubating, or nestling stages). The relationship between T/DHT and date was log-linear in males, showing no evidence of a second peak when the wren population begins the second clutch (Figure 5.1). The random effect of individual did not explain variation in any model. Because the full model included a relatively large number of parameters, which could have reduced statistical power for each parameter, I also assessed the influence of breeding stage, time, playback duration, and age individually in models that included only year, date, and one of these predictors at a time. These predictors remained non-significant (not shown).

As expected, the 11 females captured had lower T/DHT concentrations than males (effect of sex:  $F_{1,98} = 57.46$ ,  $p < 0.0001$ ,  $n = 100$  samples; mean  $\pm$  SE T/DHT for females,  $0.04 \pm 0.02$  ng/mL).

Table 5.1. Parameter estimates, 95% confidence intervals, and effect tests from a general linear mixed model relating plasma T/DHT concentration (pg/ml) to male phenotype, breeding stage, capture date and time, and playback duration. T/DHT concentration was log-transformed for analysis. Male identity was a random effect (n = 82 samples, 70 males). Significant effects are bolded. For categorical variables, I present least squared means estimates.

Term	Parameter estimate (95% confidence limits)	df	F	p	
Intercept	4.68 (3.86, 5.50)	1,71	129.28	<0.0001	
Size	0.06 (-0.01, 0.14)	1,71	2.69	0.11	
Condition	0.05 (-0.13, 0.24)	1,71	0.34	0.56	
<b>Date (days)</b>	<b>-0.013 (-0.017, -0.008)</b>	<b>1,71</b>	<b>27.38</b>	<b>&lt; 0.0001</b>	
Time (min)	0.0004 (-0.0004, 0.001)	1,71	0.99	0.32	
Breeding Stage (days)	-0.005 (-0.01, 0.003)	1,71	1.58	0.21	
Playback Duration (sec)	-0.00001 (-0.0001, 0.00001)	1,71	0.08	0.80	
Year	2008	2.72 (2.55, 2.90)	2,71	2.48	0.09
	2009	2.61 (2.38, 2.84)			
	2010	2.93 (2.73, 3.13)			
	ASY	2.74 (2.56, 2.92)			
Age	SY	2.87 (2.55, 3.19)	2,71	0.79	0.46
	UK	2.66 (2.50, 2.80)			

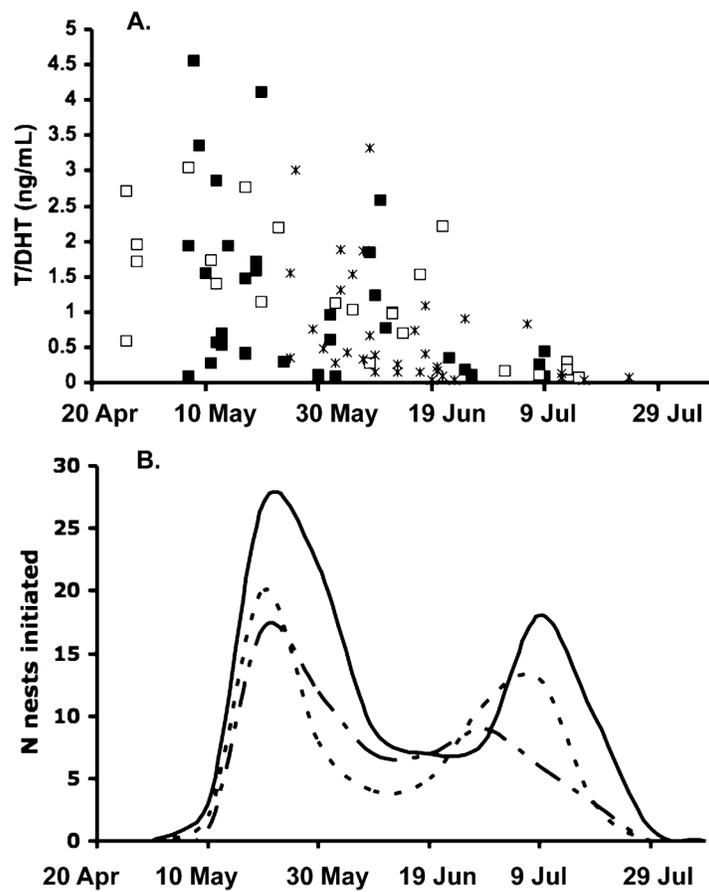


Figure 5.1. Seasonal patterns in A) circulating T/DHT and B) nest initiations. A: Plasma T/DHT concentration as a function of capture date, breeding stage, and playback treatment. Squares are males in early breeding (before pairing, or paired to a female that has not yet begun incubating), and crosses are “late-stage” males paired to females that are incubating or rearing nestlings. Dark squares represent males captured with long playbacks (18-90 min) and white squares are early-stage males captured with shorter playbacks (0-10 min, excluding handling time after capture). Late-stage males include 22 males captured with a mist net using playback and 11 males captured as they were entering the nest box to feed nestlings. B: Number of nests initiated across the breeding season in 2008 (solid), 2009 (dashed), and 2010 (dotted). Lines are smoothed across histograms in 10-day bins. I defined nest initiation as the date the female laid the first egg.

### *Correlations with aggressive behaviors*

Aggressive responses to playback, measured immediately before blood sampling, were not related to T/DHT levels (Table 5.2, Figure 5.2). Androgens tended to be negatively related to aggressive responses, but this apparent relationship was partly driven by a date confound, whereby aggressive responses tended to increase, and testosterone decreased, with date. Each of the aggressive behaviors analyzed by itself showed a similar lack of association with T/DHT (not shown).

## DISCUSSION

High circulating T/DHT appears unnecessary for territory establishment and aggression in house wrens, though it could have important organizational effects early in the season that allow males to later behave aggressively. T/DHT did not increase in response to playback and did not positively correlate with the expression of aggressive behavior. This pattern could arise if T/DHT is necessary for aggression but is expressed at the maximum physiological levels throughout the period when aggression is necessary. However, T/DHT levels were very low during the pre-mating stage for the second brood, when male house wrens perform extensive territorial defense and mate attraction behaviors. Eight of the 12 males in the pre-incubation stage for the second brood (i.e., captured after 19 June) were establishing new territories on the study site, and only one of those males had a T/DHT concentration above 1 ng/mL. The high variation in T/DHT early in the season also suggests that not all males were expressing their physiological maximum levels, or that males have different physiological maxima. Even males that are not expressing their physiological maximum androgen levels may fail to increase androgens during male-male interactions: black redstart (*Phoenicurus ochruros*) males increase circulating testosterone in

Table 5.2. Parameter estimates, effect tests, and 95% confidence limits relating aggressive behaviors to endogenous plasma T/DHT in male house wrens. Aggressive behaviors were measured during an approximately 30-minute playback immediately prior to capture. The full model includes covariates that affected T/DHT concentration in a larger dataset.

Model	Term	Parameter estimate (95% confidence limits)	df	F	p
Full	Intercept	-1.02 (-7.27, 5.23)	1,17	0.12	0.73
	T/DHT (pg/mL)	-0.0003 (-0.0002, 0.0003)	1,17	1.47	0.24
	Date (days)	0.01 (-0.03, 0.05)	1,17	0.23	0.64
	Size	-0.25 (-0.78, 0.29)	1,17	0.96	0.34
Simple	Intercept	0.51 (-0.29, 1.31)	1,20	1.8	0.20
	T/DHT (pg/mL)	-0.0004 (-0.0009, 0.00003)	1,20	3.79	0.07

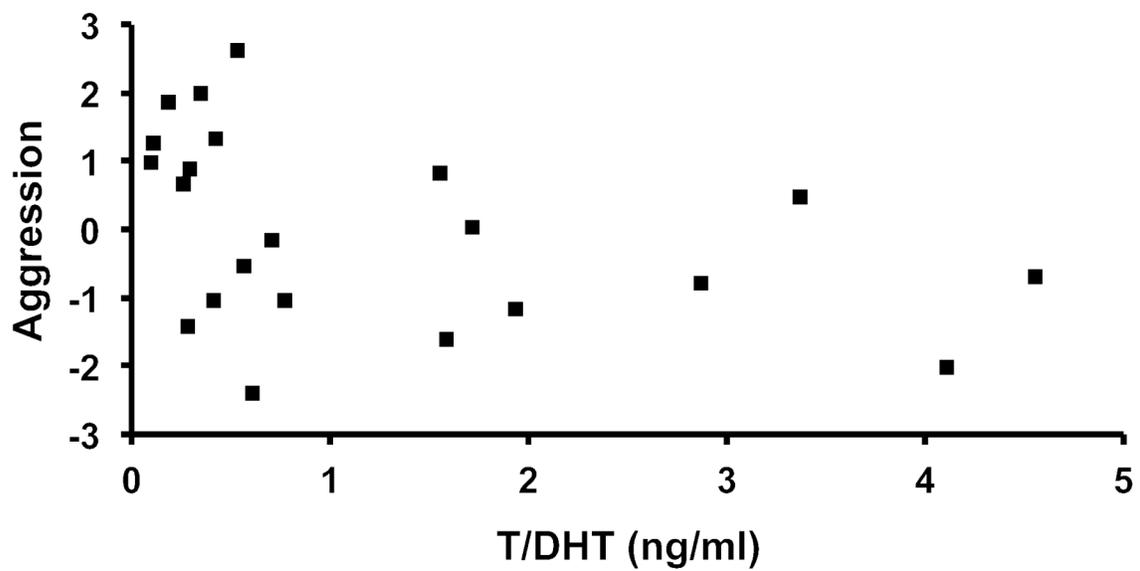


Figure 5.2. Aggressive behavior was not correlated with plasma T/DHT concentration. Aggression is a composite measure including song rate, proportion of time within 5 m of the speaker, and rate of flights across the speaker. The tendency towards a negative correlation was driven by a confounding date effect.

response to gonadotropin-releasing hormone injections but not in response to simulated territorial intrusions (Apfelbeck and Goymann 2011). Moreover, since T/DHT did not decrease in house wrens during the parental phase and did not differ between males captured with playback and those captured at the nest box while feeding nestlings, T/DHT seems unlikely to inhibit paternal care.

Endogenous plasma T/DHT levels were not correlated with aggressive behaviors in house wrens, and if anything, tended to be negatively related to aggression, as has been found in Siberian hamsters (*Phodopus sungorus*; Scotti et al. 2008). Many other studies have shown a lack of correlation between endogenous androgens and aggression during the breeding season in birds (e.g., Gwinner et al. 2002 in European starlings; Wiley and Goldizen 2003 in the buff-banded rail *Gallirallus philippensis*; Silverin et al. 2005 in pied flycatchers; van Duyse et al. 2004 in great tits *Parus major*; Schwabl et al. 2005 in European stonechats *Saxicola torquata*; and Scriba and Goymann 2010 in European robins *Erithacus rubecula*). Many studies have shown that individuals can respond very aggressively to conspecific stimuli during the non-breeding season, when testosterone levels are low (for instance in species holding year-round territories: Levin and Wingfield 1992; Hau et al. 2004; Schwabl et al. 2005). Since a large number of avian species appear to be “behaviorally insensitive” to testosterone (Lynn 2008), it is important to test the assumption that testosterone regulates aggression in any given species before testing further hypotheses on hormone function.

Several hypotheses address why behaviors do not relate to circulating testosterone in some species. Individuals may differ in their sensitivity to testosterone (Ketterson et al. 2009). Within an individual, tissues may differ in hormone receptors and in the levels of testosterone and its metabolites (reviewed in Soma et al. 2008). Since behaviors presumably depend on the

concentration of hormones and hormone receptors in the brain (e.g., Schlinger and Callard 1989, Silverin et al. 2005, Canoine et al. 2007, Charlier et al. 2011) rather than on circulating hormone levels, circulating testosterone may not necessarily be relevant for behaviors (reviewed in Soma et al. 2008). For instance, circulating levels of the testosterone-precursor dehydroepiandrosterone (DHEA) may regulate aggression in several bird species, perhaps via conversion to testosterone or estrogen in the brain (e.g., song sparrows, Soma and Wingfield 2001; spotted antbirds *Hylophylax naevioides*, Hau et al. 2004). In this study, I only measured testosterone and 5 $\alpha$ -dihydrotestosterone (5 $\alpha$ -dihydrotestosterone has been less-studied than other androgens but its seasonal pattern parallels that of testosterone in song sparrows; Wingfield and Hahn 1994). Therefore, it is possible that other androgens or testosterone metabolites could influence aggression in this species. I did not investigate variation in sensitivity to testosterone.

I used only a playback stimulus, rather than a playback stimulus in conjunction with a conspecific decoy, which is the preferred method based on the results of Wingfield and Wada (1989; reviewed in Goymann et al. 2007). That study observed a significant increase in testosterone only when a decoy was presented simultaneously with vocalizations. When males were presented with playback alone, the increase in testosterone was almost equal in magnitude to the decoy + vocalization treatment but was not statistically significant due to high variation within the treatment groups (see Figure 5 in Wingfield and Wada 1989). However, a number of studies (e.g., Horton et al. 2010, Scriba and Goymann 2011, Apfelbeck and Goymann 2011) report no increase in testosterone in response to simulated territorial intrusions with combined decoy and playback. Because house wrens responded strongly to playback alone, I conclude that playback should have been sufficient to elicit an increase in androgens. Anecdotally, I also captured one male that had been vigorously chasing another male for at least an hour, beginning

two hours prior to the capture time. This male's T/DHT levels were not elevated compared to other males captured near that day (July 8; 1.04 ng/mL; he was excluded from analyses).

It is possible that my playbacks were not long enough to cause an increase in androgen levels. Two hours of simulated territorial intrusions were necessary for spotted antbirds to show an increase in androgens (Wikelski et al. 1999). However, many other species increase testosterone almost immediately, with significant increases in testosterone following only 10 min of playback (Wingfield and Wada 1989). Such quick responses may be more biologically relevant than slower responses, because they are more likely to occur within the duration of a typical male-male interaction.

House wrens are a species where males simultaneously invest in paternal care and territorial defense. As such, they are a species where a hormonal switch between behavioral states (i.e., paternal and aggressive) might be particularly useful. For testosterone to act as such a switch, it must affect both aggression and paternal care, as it does in many species (reviewed in Lynn 2008). For instance, unpaired pied flycatcher males are more likely to attack a simulated intruder if they have higher endogenous testosterone (Silverin 1993), and experimentally elevated testosterone increases aggressiveness in white-crowned sparrows (*Zonotrichia leucophrys*; Moore 1984). However, aggression does not appear to depend on circulating T/DHT in house wrens. The diversity of effects of testosterone across species (reviewed in Lynn 2008, Goymann 2009) supports the hypothesis that the physiological effects of circulating testosterone on one behavior can evolve independently from its effects on other behaviors (Hau 2007, Ketterson et al. 2009). It further suggests that circulating androgens may not be the primary regulator of aggression in many bird species.

## ACKNOWLEDGEMENTS

I thank Sandra Vehrencamp, Mike Webster, Irby Lovette, Andy Bass, Elizabeth Adkins-Regan, Ned Place, and Sara Kaiser, as well as the Cornell Behavior journal club, the Bass lab, the Webster lab, anonymous reviewers, and the Cornell Statistical Consulting Unit for feedback on analysis and manuscript. Skip Nelson at Salimetrics gave me valuable advice on assay validation. This work was funded by an NSF predoctoral fellowship, the Animal Behavior Society, the American Ornithologists' Union, the Cornell University Department of Neurobiology and Behavior, Cornell Sigma Xi, and the Cornell Lab of Ornithology and its donors. Cornell Research Ponds kindly allowed me access to the field site, and I am deeply indebted to Paulo Llambias and Taza Schaming for sharing the study system. I have no conflict of interest.

## REFERENCES

- Apfelbeck B, Goymann W. 2011. Ignoring the challenge? Male black redstarts (*Phoenicurus ochruros*) do not increase testosterone levels during territorial conflicts but they do so in response to gonadotropin-releasing hormone. *Proc R Soc Lond B*. 278:3233-3242.
- Beletsky LD, Orians GH, Wingfield JC. 1990. Steroid hormones in relation to territoriality, breeding density, and parental behavior in male yellow-headed blackbirds. *Auk*. 107:60-68.
- Canoine V, Fusani L, Schlinger BA, Hau M. 2007. Low sex steroids, high steroid receptors: increasing the sensitivity of the nonreproductive brain. *J Neurobiol*. 67:57-67.
- Charlier TD, Newman AEM, Heimovics SA, Po KWL, Saldanha CJ, Soma KK. 2011. Rapid effects of aggressive interactions on aromatase activity and oestradiol in discrete brain regions of wild male white-crowned sparrows. *J Neuroendocrinol*. 23:742-753.
- Drilling NE, Thompson CF. 1991. Mate switching in multibrooded house wrens. *Auk*. 108:60-70.
- Goymann W. 2009. Social modulation of androgens in male birds. *Gen Comp Endocrinol*. 163:149-157.
- Goymann W, Landys MM, Wingfield JC. 2007. Distinguishing seasonal androgen responses from male-male androgen responsiveness—Revisiting the Challenge Hypothesis. *Horm Behav*. 51:463-476.
- Gwinner H, Van't Hof T, Zeman M. 2002. Hormonal and behavioral responses of starlings during a confrontation with males or females at nest boxes during the reproductive season. *Horm Behav*. 42:21-31.
- Hau M, Stoddard ST, Soma KK. 2004. Territorial aggression and hormones during the non-breeding season in a tropical bird. *Horm Behav*. 45:40-49.
- Hau M. 2007. Regulation of male traits by testosterone: implications for the evolution of vertebrate life histories. *BioEssays*. 29:133-144.
- Hirschenhauser K, Oliveira RF. 2006. Social modulation of androgens in male vertebrates: a meta-analysis of the challenge hypothesis. *Anim Behav*. 71:265-277.

- Hirschenhauser K, Winkler H, Oliveira RF. 2003. Comparative analysis of male androgen responsiveness to social environment in birds: the effects of mating system and paternal incubation. *Horm Behav.* 43:508-519.
- Horton BM, Yoon J, Ghalambor CK, Moore IT, Sillett TS. 2010. Seasonal and population variation in male testosterone levels in breeding orange-crowned warblers (*Vermivora celata*). *Gen Comp Endocrinol.* 168:333-339.
- Johnson LS, Kermott LH. 1991. The functions of song in male house wrens (*Troglodytes aedon*). *Behaviour.* 116:190-209.
- Johnson LS, Kermott LH. 1993. Why is reduced male parental assistance detrimental to the reproductive success of secondary female house wrens? *Anim Behav.* 46:1111-1120.
- Johnson LS, Kermott LH, Lein MR. 1993. The cost of polygyny in the house wren *Troglodytes aedon*. *J Anim Ecol.* 62:669-682.
- Johnson LS, Merkle MS, Kermott LH. 1992. Experimental evidence for importance of male parental care in monogamous house wrens. *Auk.* 109:662-664.
- Ketterson ED, Atwell JW, McGlothlin JW. 2009. Phenotypic integration and independence: hormones, performance, and response to environmental change. *Integr Comp Biol.* 49:365-379.
- Landys MM, Goymann W, Raess M, Slagsvold T. 2007. Hormonal responses to male-male social challenge in the blue tit *Cyanistes caeruleus*: single-broodedness as an explanatory variable. *Physiol Biochem Zool.* 80:228-240.
- Lessells CM, Boag PT. 1987. Unrepeatable repeatabilities: a common mistake. *Auk.* 104:116-121.
- Levin RN, Wingfield JC. 1992. The hormonal control of territorial aggression in tropical birds. *Ornis Scand.* 23:284-291.
- Llambías PE. 2009. Why monogamy? Comparing house wren social mating systems in two hemispheres. [dissertation]. Ithaca, NY: Cornell University; 119 p.
- Lynn SE. 2008. Behavioral insensitivity to testosterone: why and how does testosterone alter paternal and aggressive behavior in some avian species but not others? *Gen Comp Endocrinol.*

157:233-240.

Moore MC. 1984: Changes in territorial defense produced by changes in circulating levels of testosterone: a possible hormonal basis for mate-guarding behavior in white-crowned sparrows. *Behaviour*. 88:215-226.

Peters A. 2002. Testosterone and the trade-off between mating and paternal effort in extrapair-mating superb fairy-wrens. *Anim Behav*. 64:103-112.

Schlinger BA, Callard GV. 1989. Aromatase activity in quail brain: correlation with aggressiveness. *Endocrinology*. 124:437-443.

Schwabl H, Flinks H, Gwinner E. 2005. Testosterone, reproductive stage, and territorial behavior of male and female European stonechats *Saxicola torquata*. *Horm Behav*. 47:503-512.

Scotti ML, Belén J, Jackson JE, Demas GE. 2008. The role of androgens in the mediation of seasonal territorial aggression in male Siberian hamsters (*Phodopus sungorus*). *Physiol Behav*. 95:633-640.

Scriba MF, Goymann W. 2010. European robins (*Erithacus rubecula*) lack an increase in testosterone during simulated territorial intrusions. *J Ornithol*. 151:607-614.

Silverin B. 1993. Territorial aggressiveness and its relation to the endocrine system in the pied flycatcher. *Gen Comp Endocrinol*. 89:206-213.

Silverin B, Baillien M, Balthazart J. 2005. Territorial aggression, circulating levels of testosterone, and brain aromatase activity in free-living pied flycatchers. *Horm Behav*. 45:225-234.

Soma KK, Scotti MAL, Newman AEM, Charlier TD, Demas GE. 2008. Novel mechanisms for neuroendocrine regulation of aggression. *Front Neuroendocrinol*. 29:476-489.

Soma KK, Wingfield JC. 2001. Dehydroepiandrosterone in songbird plasma: seasonal regulation and relationship to territorial aggression. *Gen Comp Endocrinol*. 123:144-155.

Trivers RL. 1972: Parental investment and sexual selection. In: *Sexual selection and the Descent of Man 1871-1971* (Campbell, B., ed.). Heinemann, London, p. 136-179.

van Duyse E, Pinxten R, Darras VM, Arckens L, Eens M. 2004. Opposite changes in plasma testosterone and corticosterone levels following a simulated territorial challenge in male great tits. *Behaviour*. 141:451-467.

Washburn BE, Millspaugh JJ, Morris DL, Schultz JH, Faaborg J. 2007. Using a commercially available enzyme immunoassay to quantify testosterone in avian plasma. *Condor*. 109:181-186.

Wikelski M, Hau M, Wingfield JC. 1999. Social instability increases plasma testosterone in a year-round territorial neotropical bird. *Proc R Soc Lond B*. 266:551-556.

Wiley CJ, Goldizen AW. 2003. Testosterone is correlated with courtship but not aggression in the tropical buff-banded rail, *Gallirallus philippensis*. *Horm Behav*. 43:554-560.

Wingfield JC, Hahn TP. 1994. Testosterone and territorial behaviour in sedentary and migratory sparrows. *Anim Behav*. 47:77-89.

Wingfield JC, Hegner RE, Dufty AMJ, Ball GF. 1990. The "challenge hypothesis:" Theoretical implications for patterns of testosterone secretion, mating systems, and breeding strategies. *Am Nat*. 136:829-846.

Wingfield JC, Hunt KE. 2002. Arctic spring: hormone-behavior interactions in a severe environment. *Comp Biochem Physiol B*. 132:275-286.

Wingfield JC, Wada M. 1989. Changes in the plasma levels of testosterone during male-male interactions in the song sparrow, *Melospiza melodia*: time course and specificity of response. *J Comp Physiol A*. 166:189-194.

Ziolkowski DJ, Johnson LS, Hannam KM, Searcy WA. 1997. Coordination of female nest attentiveness with male song output in the cavity-nesting house wren *Troglodytes aedon*. *J Avian Biol*. 28:9-14.

## CONCLUSIONS

House wren songs are quite complex, and I only tested two physically challenging parameters. A whole slew of other parameters could be indicators of male quality and would be intriguing to investigate (e.g., amplitude, particularly at low frequencies; coordination of the left and right sides of the syrinx). It would also be interesting to investigate the signal value, perceptual consequences, and neurobiological underpinnings of singing the same trill type at multiple pitches. Do the same neurons in the brain code for the same syllables regardless of pitch, or does each slightly-different-pitched syllable get its own special subset of neurons? House wrens also produce ultra-high-frequency calls at periods of very high motivation, for instance, when courting a female that may choose to settle on the territory or in the midst of a physical fight. How well can other house wrens hear these high-pitched notes? Is it difficult to produce these high frequencies, and if so, do these act as indicators of male quality? Or, perhaps, we should acknowledge the possibility that song is generally fairly cheap to produce. Perhaps males just need to sound like a house wren, and the complexities of house wren song structure reflect an accumulation of random cultural mutations that signal nothing about male quality. Perhaps songs just serve to get other house wrens close enough for assessment by other signals. As a bioacoustician, I do find this possibility a bit unsatisfying, but I feel that it needs to be raised as a possibility.

Sperm competition should be a very real force in the lives of house wrens, given their moderate level of extra-pair paternity. It would be interesting to investigate sperm function (swimming speed, longevity) and how that relates to extra-pair paternity, to complete the work looking at sperm morphology; it would also be interesting to collect additional samples to determine whether the seasonal changes in sperm morphology reflect within-individual changes.

Perhaps most interesting would be to be able to get inside the female reproductive tract and really understand how some sperm get fertilizations and others don't. What is the role of the female? How do the sperm compete? At the very least, it would be good to know more about the copulations that do not lead to fertilizations.

The amount of work on androgens and behavior in birds is staggering, but most of the in-depth work has been conducted on only a handful of species. Leaders in the field recognize that there are species level differences in how testosterone acts, but most beginners in the field accept the original hypotheses about the functions of testosterone as dogma. Follow-up work in house wrens (e.g., experimental implantation and blocking of testosterone and other androgens; examining variation in the expression of receptors; examining variation in local concentrations of androgens) would be interesting to determine whether androgens truly have no influence on aggression, or whether measuring circulating levels of testosterone is just too gross a scale to work from. I would also love to do some experiments on the transition from pre-breeding to breeding levels of testosterone. If aggression is not testosterone dependent, is song? All the work on song centers in the brain is on a handful of species, mostly the same handful where work on aggression has been done. Can we really generalize about the mechanisms of action of testosterone from these few species to birds in general?

This work on house wrens has generated more questions than answers. I hope that pursuing those answers would help advance our understanding of sexual selection.