Extrapair paternity in chestnut-sided warblers is correlated with consistent vocal performance

Bruce E. Byers
Department of Biology, University of Massachusetts, Amherst, MA 01003, USA

The elaborateness of many bird songs is commonly presumed to have evolved under the influence of sexual selection by female mate choice. Thus, aspects of acoustic diversity, such as song repertoire size, are seen as likely targets of female choice. In many songbird species with song repertoires, however, the repertoires are small. In such species, female choice might be based on song features other than, or in addition to, song diversity. To investigate this conjecture, I assessed singing and paternity in a population of chestnut-sided warblers (Dendroica pensylvanica), a species in which song repertoires are of modest size. Twenty-two song traits were evaluated to determine which ones best predicted male extrapair reproductive success. The candidate traits encompassed measures of song diversity (e.g., song repertoire size), gross-scale song performance (e.g., singing rate), and fine-scale song performance (e.g., variability among songs in a bout). Regression analysis revealed that the best predictor of extrapair success was singing with little variability. In particular, the most successful males sang with consistent pitch and timing, as well as high pitch. The greater extrapair success of males with more consistent vocal performance may be due to female preference for such performance, which could be an indicator of male quality. Key words: birdsong, extrapair paternity, mate choice, songbird, vocal performance, warbler. [Behav Ecol 18:130–136 (2007)]

Darwin, impressed by birds’ production of “vocal and instrumental music of the most varied kinds,” proposed that the songs of songbirds evolved in response to sexual selection exerted, at least in part, by female mate choice (Darwin 1871). Subsequent investigation of vocal variation in songbird populations has revealed that songs vary along many dimensions (Wiley 2000; Gil and Gahr 2002; Podos et al. 2004) and thus incorporate numerous features that might be targets of female choice (Searcy and Yasukawa 1996; MacDougal-Shackleton 1997). These variable features are diverse but fall into 2 broad categories: indicators of diversity and indicators of performance.

Measurements of song diversity generally attempt to capture some aspect of the elaborateness of singing. Diversity can be measured by indices derived from those used to measure species diversity (e.g., Trainer and Peltz 1996), indices of dissimilarity among songs (e.g., Podos et al. 1992), or indices of the frequency of transitions between different sounds (e.g., Gil and Slater 2000). Diversity is most commonly measured, however, by a simple count of repertoire size (the number of different songs or song elements used by an individual).

Singing performance is usually gauged by measures of the quantity of song output. Performance measures that have been shown to vary among males include singing rate (e.g., Hofstad et al. 2002), song duration (e.g., Kempenaers et al. 1997), ratio of sound to silence in song bouts (e.g., Poesel et al. 2001), and total amount of singing (e.g., Otter et al. 1997).

In addition to variation in gross-scale performance traits that reflect the sheer quantity of vocal output, song performance also varies at a finer scale. Fine-scale traits such as trill rate (e.g., Podos 1997), frequency bandwidth (e.g., Trillo and Vehrencamp 2005), and pitch (e.g., Osiejuk et al. 2005) can vary among males. An additional form of fine-scale performance variation is variability in the consistency of singing. For example, males may differ in ability to maintain consistent pitch across consecutive syllables in a song (Lambrechts 1997).

A key prediction of the hypothesis that song evolution has been influenced by female mate choice is that female songbirds choose reproductive partners on the basis of differences in singing among males. Indirect empirical support for this prediction has emerged from the findings of 2 complementary investigative approaches. In one of these approaches, behavioral responses of captive females are assayed for differential response to variation in broadcast, tape-recorded male songs (e.g., Riebel and Slater 1998; Nolan and Hill 2004; Hosoi et al. 2005). In a second approach, investigators search, in natural populations, for correlations between variable features of male singing and potential indicators of male reproductive success that are presumed to reflect the outcome of female choice (e.g., Buchanan and Catchpole 1997; Lampe and Espmark 2003; Reid et al. 2004).

Investigators applying these approaches to a variety of songbird species have implicated as potential targets of female choice an assortment of performance and diversity traits. Song diversity, however, has generally received more attention than song performance, especially fine-scale performance. This tendency is understandable, given that the diversity and complexity of birdsong are compelling features that demand evolutionary explanation. Nonetheless, considerable evidence suggests that song performance can also influence female mate choice. It is thus difficult to make a priori predictions about the song features most likely to be targets of female choice in a species. Accordingly, prudence dictates that investigations of singing and reproductive success should simultaneously consider multiple song traits. Such consideration is especially warranted in studies of species that have small song repertoires and thus may be less likely than species with larger repertoires to have been subject to selection favoring increased song diversity.

With this imperative in mind, I compared male extrapair reproductive success with singing in a population of chestnut-sided warblers (Dendroica pensylvanica), a songbird species in...
which males have modest song repertoires and use only a portion of the repertoire for most singing. To maximize the chances of identifying the signal features most important to female choice, I included in the comparison a large pool of song traits. The candidate traits included multiple aspects of song variability, gross-scale performance, and fine-scale performance.

METHODS

Study population

I measured singing and reproductive success in a color-banded population of chestnut-sided warblers in Berkshire County, MA (42°40′N, 73°3′W) in 1999 and 2000. The study included 36 different males, 3 of which were present in both years.

Chestnut-sided warblers are socially monogamous, sexually dimorphic long-distance migrants that breed in northern North America. Males use repertoires of 2–12 song types that incorporate 6–28 syllable types (Byers 1995). As in some other members of the family Parulidae, the songs in a chestnut-sided warbler’s repertoire fall into 2 distinct categories (Lein 1978). In chestnut-sided warblers, the 2 song categories are designated accentuated-ending (AE) and unaccentuated-ending (UE). Each male generally has one preferred song type in each category; the preferred types account for the vast majority of a male’s singing. The other song types in a repertoire are used only rarely.

Males use AE and UE songs in different contexts (Byers 1996). One aspect of the behavioral distinction between the categories is that AE songs are not used at dawn. Instead, dawn singing by mated males is a distinctive performance that consists of rapidly repeated, exclusively UE songs interspersed with bursts of characteristic chip notes (Byers 1995). Later, after the dawn bout has concluded, singing is quite different. The pace is more leisurely, AE songs predominate, and the chip bursts diagnostic of the dawn bout are not used.

In the analyses reported here, dawn songs were analyzed separately from songs sung later in the day. Songs uttered in association with the unmistakable dawn bout behavior described above were classified as dawn songs. All other songs were classified as daytime songs, with the restriction that the daytime song category could include no song uttered fewer than 40 min after the first song of the day. This restriction ensured that no dawn songs were inadvertently misclassified as daytime songs, because the duration of the longest dawn bouts in this population is less than 40 min (Kroodsma et al. 1989). Separate analysis of dawn and daytime singing was intended to guard against conflation of the effects of qualitatively different singing behaviors that might have different functions and stimulate different responses by receivers (Spector 1992; Staicer et al. 1996).

Reproductive success

I used number of genetic offspring outside a male’s own nest as a measure of male extrapair reproductive success. Extrapair paternity was common in the study population; 47% of offspring had extrapair fathers, and 60% of the broods hatched during the study period contained extrapair young. The proportion of extrapair offspring did not differ between the 2 years of the study (χ² = 0.90, degree of freedom = 1, P = 0.34). The genetic father was known for all but 6 of the 95 offspring fledged in the study area during the study period (for details, see Byers et al. 2004). Total reproductive success varied considerably among the males in the population (0–7 offspring/year), and differences among males in extrapair success (0–6 offspring/year) contributed significantly to this variability.

Song sample

The singing of males resident in the study area was recorded daily, 25 May–25 June 1999 and 23 May–27 June 2000. These recording periods began after females had settled on the territories of their social mates and concluded when eggs had hatched on several territories. The recording period thus coincided with the time span in which all known extrapair fertilizations occurred (Byers et al. 2004).

Recordings were made on analog cassette tapes using a Telinga parabolic reflector, a Sennheiser MKH 62 microphone, and either a Sony TC-D5 Pro or Marantz PDMD 222 recorder. As a precaution to detect recordings compromised by weak batteries or other causes of faulty tape speed, each recordist periodically recorded a 440 Hz tone from a pitch pipe. If later playback of a recorded pitch pipe tone did not yield a 440 Hz signal (as determined by spectrographic analysis), the affected recording was discarded.

Sampling began before dawn with the first songs of the day and continued until approximately 0700 h eastern standard time. Recordists followed a protocol in which a focal male’s songs were recorded for 5 min, after which the recordist proceeded to the next adjacent territory (skipping any territory containing a silent male) to begin another 5-min sample. In this fashion, multiple 5-min samples were obtained each day for most males in the study area. The locations at which recording began each day were systematically rotated such that each male was the first bird recorded a roughly equal number of times across a season and sampling effort was roughly equivalent for each male. Approximately 16 000 songs were recorded.

Song measurements

From the full song sample, I selected song bouts for analysis. A bout was defined as a continuous recording of at least 10 but no more than 20 repetitions of a single song type. (An instance of chestnut-sided warbler singing typically includes only a single song type; type switches are rare.) Bouts were intended to serve as representative samples of spontaneous, undisturbed singing. The 10-repetition minimum was specified so that sporadic singing, such as typically occurred in the context of male–male fights or chases, would not be included. The specified 20-repetition maximum was an arbitrary limit imposed to ensure that the measurement task could be completed within a reasonable time.

For each male, 10 bouts were selected, based primarily on recording quality but subject to the constraint that 5 bouts were from dawn singing, 5 bouts were from daytime singing, and each bout in a category was recorded on a different day. For the 3 males who were recorded in both 1999 and 2000, bouts were selected from only one year’s recordings (chosen by coin flip). The song samples of 4 males did not include 10 bouts of sufficiently high quality, and only 7 or 8 bouts were analyzed for these males. Overall, 350 bouts were selected, containing a total of 5811 songs.

For each song in a bout, an observer measured the song’s highest frequency, lowest frequency, frequency of peak power, start time, and end time. Additional variables were derived from the ones directly measured. These variables were frequency bandwidth, song duration, duration of the pause between songs, and time between the beginnings of successive songs (a measure of singing rate). In addition, the number of syllables in each song was counted. Most of the measured variables differed significantly between dawn and daytime singing, providing additional justification for separate analysis of the 2 singing periods (paired t-tests with Bonferroni-adjusted P values; P < 0.05 for 6 of 8 variables).
Because the sampling protocol favored sample breadth (number of males sampled per day) over sample depth (duration for which each male was sampled), each day’s combined sample from a particular male was of relatively short duration. Such samples precluded accurate estimation of cumulative performance variables such as total time spent singing, total numbers of songs uttered, duration of the dawn bout, or proportion of AE versus UE songs. With regard to performance variables, then, the study focused on those that might be assessed over relatively short listening periods, at the expense of those that would require longer term assessment.

Song measurements were made on digitized versions of the recorded songs (sampling rate 24 kHz), using Signal 4.0 and RTSD 2.1 sound analysis software (Engineering Design, Berkeley, CA). Frequency of peak power was measured with the automated MEASURE command in Signal. The remaining time and frequency measurements were made by manual placement of screen cursors on spectrograph displays generated by RTSD (time measurements: 64 point fast Fourier transform [FFT], resolution 2.6 ms; frequency measurements: 512 point FFT, resolution 47 Hz).

Song and syllable repertoire sizes were estimated for both AE and UE songs by inspection of the entire recorded song sample. First, an observer compiled a collection of printed spectrograms containing multiple examples of each song type used by each male. Next, a second, naïve observer was instructed to sort unlabeled copies of the spectrograms into groups, such that each group contained identical songs. Each resulting group was deemed to represent a song type, and the number of different types used by each male was deemed to be his song repertoire. Finally, one printed spectrogram of each song type from each male was cut to yield spectrograms of the song’s individual syllables, and the sorting procedure repeated to yield syllable repertoire sizes.

**Statistical analysis**

I used regression analysis to determine the song traits that best predicted extrapair reproductive success. Because the regression’s response variable (number of extrapair offspring) was a counted variable with a small number of discrete values, an ordinary least-squares linear regression approach would be likely to yield biased parameter estimates (Long 1997) and was therefore not suitable for my data set. It was necessary to instead choose a regression model appropriate for a counted response variable. I chose a Poisson regression model, which is suitable for count data provided that certain assumptions are met (Cameron and Trivedi 1998). The key assumption is that the response variable is equidispersed (variance equal to the mean). This assumption can be checked with a postregression likelihood ratio test of the null hypothesis that the overdispersion parameter \( \varphi \) equals zero (Cameron and Trivedi 1990). If this null hypothesis is rejected, the response variable is probably not equidispersed and the Poisson model is not appropriate.

The 25 candidate predictor variables included those described above as well as calculated estimators (coefficients of variation [CVs]) of the within-bout variability in time and frequency measurements. These variability measurements were intended to assess the consistency with which the songs in a bout were performed. Additional candidate predictor variables included calendar year and 2 measures of male body size (weight and tarsus length). None of the candidate song and size variables differed between the 2 years of the study (pooled \( t \) tests with song bout as the unit of analysis; all \( P > 0.4 \)).

Because each bird contributed only one data record to the regression analysis, candidate song performance variables were calculated as per-bird averages. For example, measured song durations were averaged across the 10–20 songs in a bout, and the per-bout averages were themselves averaged across all bouts of a given individual to yield a single number characterizing song duration for that bird. Similarly, coefficients of variation for measured song variables were calculated for each bout, so that each CV measured consistency of a trait across a bout. The per-bout CVs were then averaged across the bouts of an individual.

Collapsing song performance measures into per-bird averages had the potential to obscure variation that would arise if song traits differed at different times of day. In order to reduce the impact on the analysis of any such time-of-day effects, the song-sampling protocol was designed to ensure that the distribution of times of day was roughly equivalent in samples from different birds. In addition, the one known time-of-day effect (the distinctiveness of the dawn bout) was accounted for by performing separate analyses of dawn and daytime singing, as noted above. Nonetheless, it remained possible that, within each time category, between-bird differences in measured song traits arose due to differences in the time of day of recordings rather than due to differences intrinsic to the birds. Such an effect, however, seems unlikely, as there was no significant statistical association between time of day (measured as minutes after the first song of the day) and any of the song performance measures for either dawn songs or daytime songs (analyses of variance with bird as factor and time of day as covariate; for bird, all \( P < 0.02 \); for time of day, all \( P > 0.13 \)).

Two regression models (one for dawn singing and one for daytime singing) were developed using a model-building strategy advocated by Hosmer and Lemeshow (2000). Candidate predictor variables were first screened by univariate regressions on the response variable. Any variable whose univariate regression indicated that the variable was a very poor predictor of the response variable (i.e., the regression yielded a likelihood ratio chi-square test with a \( P \) value higher than 0.30) was excluded from the model at this step. Next, a multivariate model containing all the remaining variables was fit. Each variable in the fitted model was examined, and those whose inclusion did not add considerable (\( P < 0.15 \)) predictive value to the model were removed. A new model containing the reduced variable set was then fit. The new model was compared with the old model with a likelihood ratio test to determine whether the new model was a significant improvement over the old one. In addition, the coefficients of the variables in the new model were compared with their values in the old model, to determine if any coefficients had changed markedly. (Such changes can indicate that a removed variable, even if not by itself a good predictor, provides an important adjustment of the effect of another variable). If coefficients had changed, the previously removed variables were tested one at a time to see if re-adding one or more of them to the model would improve the stability of coefficients. If any previously removed variables were added back, a model was fit for the now-augmented collection of variables. Once again, the newly fitted model was compared with the previous one and its component variables examined, beginning another round of “deleting, refitting, and verifying” as described above (but with progressively more stringent \( P \) value cutoffs for removal at the deleting step). This iterative process continued until only significant (\( P < 0.05 \)) predictors remained in the model. As a last check of the final model, a series of additional follow-up models was fit. Each of these models included the variables from final model plus one of the previously removed variables. These follow-up models provided confirmation that including a previously removed variable would not significantly improve the final model. All statistical analyses were performed with Stata 8.0.
RESULTS

Model fitting

In both regressions, reproductive success was best predicted by a reduced model containing only 2 or 3 variables (Table 1). Both models had highly significant predictive power relative to the null model (likelihood ratio test, \( P < 0.005 \) for both models). In each model, no excluded variable significantly improved the model (Table 2). None of the variables remaining in the models were significantly correlated with any of the excluded variables.

Regression diagnostics showed that the poisson model was appropriate for both regressions. Goodness of fit chi-square values, which compare a fitted model with the saturated model, were low and insignificant for both models, indicating good model performance (\( P > 0.5 \) for both models). In addition, for both models, the overdispersion parameter \( \alpha \) was not significantly different from 0, indicating that the response variable was equidispersed, as assumed under a Poisson regression model (\( P > 0.5 \) for both models).

Best predictors

Almost all the variables in the final models were measures of variability in song performance. For all these measures of variability, regression coefficients were negative, indicating an inverse relationship between song performance variability and male success. Overall, more consistent (less variable) singing was associated with greater extrapair reproductive success.

For dawn singing, the best predictor variables were the coefficient of variation of peak-power frequency and the coefficient of variation of lowest frequency. Thus, the most successful males were those that, during the dawn bout, tended to sing with little variation in frequency (pitch) among the songs in a bout. For later, daytime singing, performance variability was again important in predicting success but, in this case, the important

---

Table 1
Best-fit models from 2 regressions of song traits on extrapair reproductive success (\( N = 36 \) for both models)

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>Coefficient</th>
<th>( P ) (2 tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dawn singing</td>
<td>CV peak-power frequency</td>
<td>-0.709</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>CV lowest frequency</td>
<td>-0.477</td>
<td>0.016</td>
</tr>
<tr>
<td>Daytime singing</td>
<td>CV peak-power frequency</td>
<td>-0.202</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>CV intersong interval</td>
<td>-0.597</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>Peak-power frequency</td>
<td>1.96</td>
<td>0.008</td>
</tr>
</tbody>
</table>

---

Table 2
Candidate variables excluded from 2 regressions of song traits on extrapair reproductive success

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dawn singing</th>
<th></th>
<th></th>
<th>Daytime singing</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Coefficient</td>
<td>( P ) (2 tailed)</td>
<td></td>
<td>Coefficient</td>
<td>( P ) (2 tailed)</td>
</tr>
<tr>
<td>CV lowest frequency</td>
<td></td>
<td>-0.23</td>
<td>0.12</td>
<td></td>
<td></td>
<td>-0.33</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
<td>0.76</td>
<td>0.47</td>
<td></td>
<td></td>
<td>-0.46</td>
</tr>
<tr>
<td>CV duration</td>
<td></td>
<td>0.14</td>
<td>0.43</td>
<td></td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>Intersong interval</td>
<td></td>
<td>-0.01</td>
<td>0.92</td>
<td></td>
<td></td>
<td>0.17</td>
</tr>
<tr>
<td>CV intersong interval</td>
<td></td>
<td>0.05</td>
<td>0.77</td>
<td></td>
<td></td>
<td>—</td>
</tr>
<tr>
<td>Lowest frequency</td>
<td></td>
<td>-1.94</td>
<td>0.27</td>
<td></td>
<td></td>
<td>1.11</td>
</tr>
<tr>
<td>Highest frequency</td>
<td></td>
<td>0.03</td>
<td>0.97</td>
<td></td>
<td></td>
<td>0.73</td>
</tr>
<tr>
<td>CV highest frequency</td>
<td></td>
<td>-0.36</td>
<td>0.21</td>
<td></td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>Bandwidth</td>
<td></td>
<td>0.28</td>
<td>0.64</td>
<td></td>
<td></td>
<td>0.49</td>
</tr>
<tr>
<td>CV bandwidth</td>
<td></td>
<td>-0.32</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>Number of syllables in song</td>
<td></td>
<td>0.12</td>
<td>0.45</td>
<td></td>
<td></td>
<td>-0.06</td>
</tr>
<tr>
<td>CV number of syllables in song</td>
<td></td>
<td>-0.11</td>
<td>0.51</td>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>Peak-power frequency</td>
<td></td>
<td>-1.21</td>
<td>0.31</td>
<td></td>
<td></td>
<td>1.30</td>
</tr>
<tr>
<td>AE song repertoire</td>
<td></td>
<td>0.19</td>
<td>0.78</td>
<td></td>
<td></td>
<td>0.68</td>
</tr>
<tr>
<td>AE syllable repertoire</td>
<td></td>
<td>-0.45</td>
<td>0.83</td>
<td></td>
<td></td>
<td>-0.20</td>
</tr>
<tr>
<td>UE song repertoire</td>
<td></td>
<td>-0.64</td>
<td>0.22</td>
<td></td>
<td></td>
<td>-0.02</td>
</tr>
<tr>
<td>UE syllable repertoire</td>
<td></td>
<td>0.10</td>
<td>0.32</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>Total song repertoire</td>
<td></td>
<td>0.19</td>
<td>0.78</td>
<td></td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>Total syllable repertoire</td>
<td></td>
<td>0.11</td>
<td>0.34</td>
<td></td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td>-0.55</td>
<td>0.64</td>
<td></td>
<td></td>
<td>0.98</td>
</tr>
<tr>
<td>Tarsus length</td>
<td></td>
<td>0.55</td>
<td>0.64</td>
<td></td>
<td></td>
<td>0.98</td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td>0.08</td>
<td>0.89</td>
<td></td>
<td></td>
<td>-0.02</td>
</tr>
<tr>
<td>Between onset interval</td>
<td></td>
<td>0.18</td>
<td>0.55</td>
<td></td>
<td></td>
<td>0.77</td>
</tr>
<tr>
<td>CV between onset interval</td>
<td></td>
<td>-0.53</td>
<td>0.29</td>
<td></td>
<td></td>
<td>-0.14</td>
</tr>
</tbody>
</table>

Parameter estimates derived by entering each variable by itself into one of the final models shown in Table 1.
variability encompassed timing as well as pitch. The coefficient of variation of peak-power frequency was again a good predictor, but the daytime singing model also included the coefficient of variation of intersong pause. The model for daytime singing also included peak-power frequency; this variable had a positive regression coefficient. Thus, the most successful males were those that, during postdawn singing, tended to sing high-pitched songs whose pitch varied little over the course of a bout and that were delivered with little variation in the length of the silent period between songs in a bout.

**Paternity gainers versus paternity losers**

In pairwise comparisons of the genetic and social fathers of particular extrapair offspring, the singing of the genetic fathers tended to be less variable and (for daytime songs) higher pitched than the singing of the cuckolded social fathers. Seventeen different males from the study population were known to have fathered extrapair offspring in the study area, and 3 of these males fathered extrapair young in 2 different nests. The study population thus contained 20 instances in which it was possible to compare the singing of males that fathered extrapair young with the males attending the nests that contained those offspring. Comparisons of males’ values of each of the 5 predictor variables revealed that the song performance of paternity winners was superior to that of paternity losers in 65–85% of cases (Figure 1). For all 5 variables, the difference between paternity winners and paternity losers was statistically significant (paired t-tests, see Figure 1 for results). Every paternity winner was superior to the corresponding paternity loser in at least 2 of the 5 predictor variables. All but 2 paternity winners were superior in three or more variables.

**DISCUSSION**

In my study population, male extrapair reproductive success was associated with precise song performance: consistent,
invariant repetition of high-pitched songs. The most successful males sang with consistent pitch at dawn and with consistent pitch and timing, as well as high pitch, later in the day. In contrast, variation among males in song diversity (e.g., song and syllable repertoire size) and gross-scale performance features (e.g., singing rate, song duration) did not influence variation in extrapair reproductive success.

These correlative results cannot by themselves demonstrate that female chestnut-sided warblers use song variation to assess potential extrapair partners. Correlations between song traits and reproductive success might, for example, reflect female preference for nonvocal signals such as plumage traits that are themselves correlated with song traits. Alternatively, song traits might determine the outcome of male–male contests whose results in turn influence female mate choice. Nonetheless, my results show that to the extent that females do use songs in extrapair mate choice, they prefer consistent singers and are not influenced by song diversity or gross-scale performance features.

A preference for males whose singing is characterized by consistent performance of high-pitched songs might indicate that the preference evolved because its expression conferred a fitness advantage on females. Such advantage could arise if consistent performance were a reliable indicator of male quality. One condition under which that sort of reliability can evolve is when production of a signal is directly dependent on physical qualities of the signaler that cannot be faked (Maynard Smith and Harper 2003). Vocal performance traits in general have the potential to meet this criterion because a male’s ability to perform physically challenging vocal tasks is limited by morphological and physiological constraints, and variation among males in performance of such tasks reflects phenotypic variation in ability to approach the limits imposed by these constraints (Podos and Nowicki 2004). Ability to perform consistently may be an example of a vocal task whose expression depends on phenotypic quality.

In songbirds, performance of songs requires precise control of rapid and repeated contraction of dozens of syrinxal, respiratory, and other muscles (Suthers and Zollinger 2004). These efforts are doubtless difficult to sustain in stereotyped fashion over an extended bout of singing, and fatigue may lead to increasing variation between song renditions. The degree to which a bird can sustain consistency may thus reflect levels of stamina and neuromuscular coordination that are intrinsic to his phenotype. Highly consistent singing may be possible only for individuals with superior phenotypes.

There is some evidence that consistent vocal performance can be an indicator of male quality. Lambrechts and Dhondt (1997) found that variability in the duration of between-song pauses was whose social mates had larger clutches tended to sing bouts in which variability in the duration of between-song pauses was low (Poesel et al. 2001).

Although the chestnut-sided warbler data presented here represent, to my knowledge, the only demonstration of a connection between extrapair reproductive success and consistent song performance, other song traits have been linked to extrapair success in other species. Both diversity and performance have been implicated. In 2 populations of great reed warblers (Acrocephalus arundinaceus), the fathers of extrapair offspring had larger repertoires than the males they had cuckolded (Hasselquist et al. 1996; Leisler et al. 2000). A similar pairwise comparison of paternity winners and losers in blue tits found that the winners tended to sing songs of longer duration (Kempenaers et al. 1997). In duskies warblers (Phylloscopus fuscatus), extrapair success was associated with ability to sing rapidly modulated syllables at high amplitude but was not associated with repertoire size (Forstmeier et al. 2002).

That even the relatively few studies of song and extrapair paternity have implicated a variety of song traits as possible targets of female choice suggests that we should expect variety among species in the song traits used by females to assess potential extrapair mates. However, we should remain alert to the possibility that broad patterns will emerge as the number of studied species increases and as investigators perform more definitive tests of the relationship between song and mate choice. Perhaps some kinds of song traits will be revealed to be especially well suited for revealing male quality and therefore widely used in assessment of extrapair mates. Further investigation may also reveal whether the song traits important in extrapair mate choice are generally the same ones that are important in social mate choice.

I thank Elijah Goodwin, Lauren Tuori, and Erin Welby for outstanding assistance in the field and in the laboratory. Don Kroodsma and 2 anonymous reviewers provided helpful comments on the manuscript. This research was supported by a Faculty Research Grant from the University of Massachusetts, Amherst.

REFERENCES


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.