Electronic Supplementary Material 1: Reanalysis of Searcy et al. 2010

Methods

Searcy et al. [20] tested the response of swamp sparrow females to songs sung by males with differing developmental history. They used 10 females from the same Conneaut Marsh population as in the present study, captured as adults in 2006 (i.e., two years before our song recording survey). Their experimental design was similar to the one we used in our female preference experiment. The most substantive difference was that females were played just two exemplars, one exemplar in a morning trial, and the other in an afternoon trial, with the experiment repeated on a second day with the order of presentation reversed. In each trial, a female heard 18 repetitions of a song, twice what females in our experiment heard, and therefore the number of displays per song were approximately double those made in our experiment.

The stimuli used in the experiment were songs recorded in the laboratory from hand-reared males. Half of the stimuli were produced by males with superior growth histories and half by males with inferior growth histories as nestlings (the different growth histories were the result of different feeding regimes). Subjects were tutored with a set of 14 song types recorded from the Conneaut population. Tutor stimuli were not selected on the basis of their frequency in the population.

After songs had crystallized, recordings were made of males’ repertoires. Song performance was measured using the same deviation method that we used. In the original study, learning accuracy was assessed as the cross correlation between a male song syllable and the most similar tutor song syllable in the set. Ten stimuli were chosen from the superior group and ten from the inferior group and used as exemplars for the female preference experiment. Females responded more strongly to songs of males with superior growth histories than to songs of males with inferior growth histories, but variation in response was not explained by differences in learning accuracy [20]: the correlation between female response to test songs and the cross-correlation scores with tutor songs was near 0 (r² = 0.056, P = 0.81).

Analysis and Results

We carried out a DTW analysis of our entire dataset of recordings plus the stimulus songs used by Searcy et al., and clustered songs using UPGMA. DTW parameters were identical to those used in our main analysis. Global Silhouette Widths of the UPGMA tree suggested the presence of 68 clusters. In this analysis stimulus songs fell into two broad groups. Eleven exemplars (which we labelled “outliers”) were not clustered together with a syllable type with more than five members. The remaining group of nine exemplars (“typical” songs) were clustered with syllable types with 9 or more members. The d_c scores of the typical group overlapped with the low-d_c and some of the high-d_c stimuli in the principal experiments we report in this paper. The “typical” songs in this reanalysis were therefore less typical than the
low-$d_c$ stimuli in our principal study, and the “outliers” in the reanalysis more atypical of population syllable type categories than the high-$d_c$ stimuli in our principal study. Females gave more CSDs to typical songs as defined by our reanalysis (mean: 9.33) than to outliers (mean: 2.18).

We tested whether a song being typical or an outlier predicted the number of female displays for the Searcy et al. data, analyzing the data with a Generalized Linear Mixed Model, using MCMCglmm in R [25, 26], with number of CSDs as the dependent factor, and typicality (coded as a two-level factor “typical” vs “outlier”), song performance and song length as independent variables and subject as a random factor. In the original study, song length was the principal significant factor predicting the number of CSDs [20].

We found that typicality was the only significant factor predicting the number of CSDs in the model (typicality: $p<0.04$; 95% credible interval $-2.16$ to $-0.015$; vocal performance: $p>0.5$, CI $-0.226$ to $0.069$; song length: $p>0.5$, CI $-0.0021$ to $0.0027$). In a simpler model using only typicality, the effect was even clearer ($p<0.0025$, CI $-2.22$ to $-0.43$).
Electronic Supplementary Material 2: Errors in Discriminating Typical from Less Typical Version of Syllable Types

(a) Proportion of males singing rare songs

In our experiments, we labelled one group of stimuli ‘high-$d_c$’, and found that they were discriminated against by receivers. Here we report more details of our syllable comparison to examine more thoroughly how unusual these stimuli were.

The average $d_c$ score of the high-$d_c$ stimuli was 0.0602. Of the syllables sung by at least 5 males (i.e. at least 2.5% of males in our sample), 91 had a $d_c$ score of 0.0602 or higher. 89 of these 91 syllables belonged to syllable types sung by at least 10 males.

In addition, 66 syllables belonged to syllable types sung by 5 males or fewer (22 syllables belonged to types sung only by one male, 34 syllables were from types sung by two or fewer males, 94 syllables were from types sung by 10 or fewer males).

In total, therefore, 157 syllables had higher $d_c$ scores than our high-$d_c$ stimuli, or were members of rare types. This represented 24.9% of the total sample of syllables in our sample.

(b) Probability of a receiver confusing High $d_c$ with Low $d_c$ Stimuli

One mechanism for how receivers might have discriminated high-$d_c$ from low-$d_c$ stimuli is that they first learned a set of exemplars, and subsequently attempted to match each stimulus with one of those exemplars. The more similar a song is to one of the exemplars, the stronger the response it would generate from the receiver. Because low-$d_c$ songs are more common than high-$d_c$ songs, receivers using this mechanism would be more likely to learn songs similar to the low-$d_c$ stimuli than songs similar to the high-$d_c$ stimuli.

We investigated how likely it would be for a receiver employing this strategy to confuse our high-$d_c$ stimuli with our low-$d_c$ stimuli – that is the probability of a receiver finding that the high-$d_c$ stimulus was more similar than the low-$d_c$ stimulus was to a learned exemplar.

To estimate this probability, we used the syllable dissimilarities calculated from the DTW analysis as proxies for the perceived dissimilarities between syllables. We first investigated the scenario in which a receiver memorizes at random just one stimulus from each syllable type. We therefore compared each of the non-stimulus members with each pair of high-$d_c$ and low-$d_c$ stimuli of the same syllable-type. Averaged over all 17 stimulus pairs, 17.2% of syllables were more similar to the high-$d_c$ stimulus, and 82.8% were more similar to the low-$d_c$ stimulus.
We next conducted more complex simulations in which ‘receivers’ first sampled $n$ syllables from our sample of syllables (excluding the exemplars for the stimuli themselves). They were then presented with each pair of high-$d_c$ and low-$d_c$ stimuli. They assessed learning accuracy by matching the stimuli to the $n$ syllables in their memory. The lowest DTW dissimilarity for each stimulus was taken as an assessment of learning accuracy. We then tallied how frequently the high-$d_c$ stimulus was matched more closely to a memorized exemplar than the low-$d_c$ stimulus. We examined values of $n$ from 5 to 300, and in each case, simulated 1000 receivers. The outcome of this simulation is shown in Fig. S1. It demonstrates that if receivers sample only a small number of songs, they have approximately a 20% chance of confusing a low-$d_c$ with a high-$d_c$ stimulus, but even if they sample several hundred songs, there is still a greater than 10% chance of erroneously picking the high-$d_c$ stimulus. If preferences for more typical songs is part of a signaling system, there is therefore a substantial probability of mistaken discrimination using this exemplar-based mechanism.
Fig. S1

The outcome of simulations to investigate the probability that receivers would judge our high-$d_c$ stimuli ‘better learned’ than our low-$d_c$ stimuli. The y-axis shows the probability of a simulated receiver finding a closer match for a high-$d_c$ stimulus than for its low-$d_c$ complement after comparing them with memorized syllables. $n$ is the number of syllables that receivers memorize.