Animal Behaviour xxx (2016) e1–e16

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Forum

Birdsong performance studies: a contrary view

Donald Kroodsma*

Department of Biology, University of Massachusetts Amherst, Amherst, MA, U.S.A.

ARTICLE INFO

Article history:
Received 5 February 2016
Initial acceptance 29 March 2016
Final acceptance 30 June 2016
Available online xxx
MS. number: AF-16-00107R

Keywords:
birdsong
chipping sparrow
performance
scepticism
sexual selection
swamp sparrow

Birdsong biologists interested in sexual selection and honest signalling have repeatedly reported confirmation, over more than a decade, of the biological significance of a scatterplot between trill rate and frequency bandwidth. This 'performance hypothesis' proposes that the closer a song plots to an upper bound on the graph, the more difficult the song is to sing, and the more difficult the song the higher quality the singer, so that song quality honestly reveals male quality. In reviewing the confirming literature, however, I can find no support for this performance hypothesis. I will argue here that the scatter in the graph for songbirds is better explained by social factors and song learning. When songbirds learn their songs from each other, multiple males in a neighbourhood will sing the same song type. The need to conform to the local dialect of song types guides a male to learn a typical example of each song type for that population, not to take a memorized song and diminish or exaggerate it in trill rate or frequency bandwidth to honestly demonstrate his relative prowess. When data in this scatterplot are coded both by song type and by male, it is the song type and the need to conform that explains the variability, not the quality of different males. There is no consistent, reliable information in the song performance measures that can be used to evaluate a singing male.

Something in how a male songbird delivers his songs may convey something about his relative quality to those who listen, especially females, but identifying those somethings has proven challenging. In the study of birdsong repertoires and female choice, for example, it has been widely accepted that 'Females of many songbird species show a preference for mating with males that have larger song repertoires' (Nowicki, Hasselquist, Bensch, & Peters, 2000, page 2419), but in spite of a host of studies claiming to confirm that relationship, there is no strong evidence that males or females attend to the number of different songs that a male can sing (Byers & Kroodsma, 2009).

Another idea that has over the last decade gained much traction is the performance hypothesis developed by Podos, Peters, and Nowicki (2004) and Ballentine, Hyman, and Nowicki (2004), based on motor and performance constraints described by Podos (1996, 1997). Scatterplots of trill rates and frequency bandwidths show an inverse relationship: the more rapid the trill, the narrower the bandwidth (see Figs. 4, 5 and 11 for examples). Blank areas with no data beyond an upper bound suggest a motor constraint; that is, the birds cannot produce those combinations of trill rates and bandwidths (but see Figs. 4 and 6). The interesting hypothesis is that how close a song plots to the upper bound might reveal the difficulty of producing that song, so that songs near the upper bound honestly reveal a high-quality singer; both prospective mates and competing males might then use those high-performance songs to detect high-quality singers.

This hypothesis has 'been adopted widely in tests of song function' (Goodwin & Podos, 2015, page 1), is touted as 'a premiere illustration of how performance constraints shape the evolution of mating displays [with] sexual selection favoring high performance trills' (Wilson, Bitton, Podos, & Mennill, 2014, page 214), and has been repeatedly confirmed over the past decade. My careful scrutiny of those studies here, however, reveals that the hypothesis has become largely an assumption ('generally assumed'; Cardoso, Atwell, Ketterson, & Price, 2007, page 901) and never truly tested (Prum, 2010, 2012); furthermore, given how song performance measures are distributed among song types and among males, the hypothesis becomes biologically implausible, if not impossible. Here I review the confirming studies, beginning with a most recent paper on chipping sparrows, *Spizella passerina* (Goodwin & Podos, 2014), because it reveals especially clearly the methods used to confirm the hypothesis. I then proceed to the studies of swamp sparrows, *Melospiza georgiana*, before briefly reviewing other species.

The problems that plague this birdsong performance literature are pervasive in sexual selection studies (Prum, 2010, 2012), including the study of birdsong repertoires (Byers & Kroodsma,

* Correspondence: D. Kroodsma, Department of Biology, University of Massachusetts Amherst, Amherst, MA 01003-9297, U.S.A.
E-mail address: DonaldKroodsma@gmail.com.

http://dx.doi.org/10.1016/j.anbehav.2016.06.015
0003-3472/

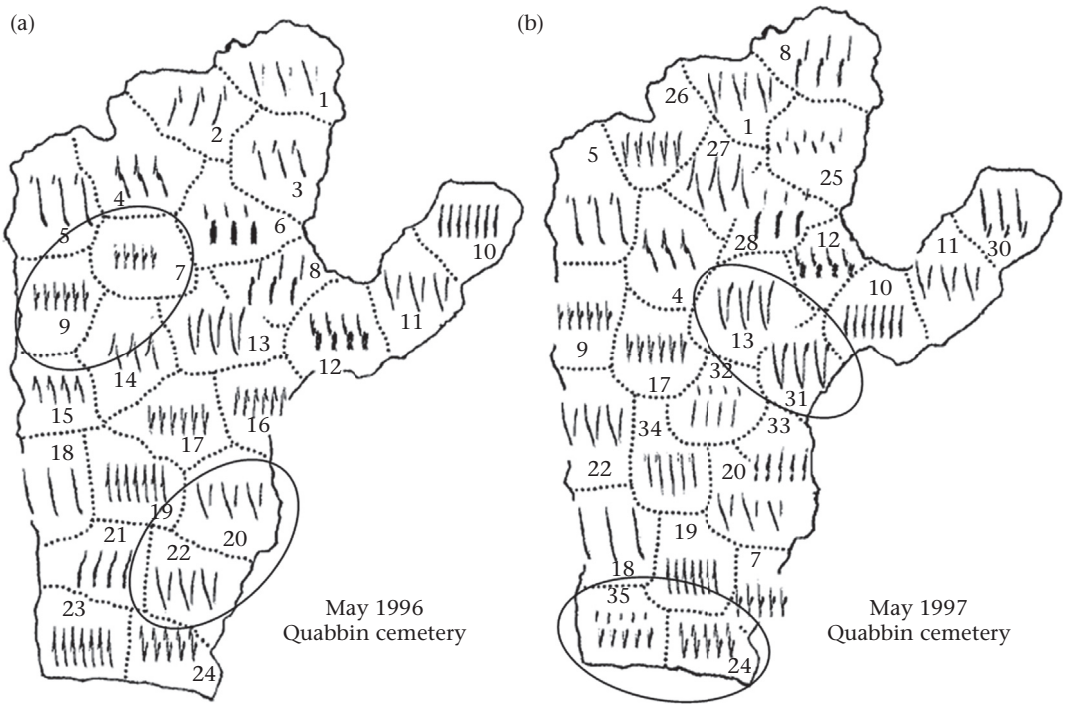

Figure 1. Reprinted with permission from Figure 2 in Liu and Kroodsma (2006). 'Yearling Chipping Sparrows imitate songs of an immediate neighbor, but the instability of territories results in only short-term song sharing among neighbors. (a) In May of 1996, 24 territorial males (numbered 1–24) were found in the Quabbin Cemetery, and a portion of each male's song type (0.35 sec) is illustrated. Males 7 and 9 share similar song types, as do males 20 and 22. (b) In May of 1997, 26 territorial males were found in the cemetery, 16 returning adults (at least two years old) from the previous year and 10 birds breeding there for the first time. Males 7 and 9 both returned but are no longer neighbors, and male 22 did not return. Male 35 did not learn the song of his father (male 5) or his father's close neighbors in 1996 or 1997, but instead appeared to learn from his immediate neighbor in 1997 (male 24). Male 31 also appeared to acquire his song from an immediate neighbor (male 13) in 1997, not from his father (male 10). Both 1997 yearlings (31 and 35) hatched late in the 1996 breeding season, and each most likely acquired his song during 1997, as a yearling'.

2009). By discussing these problems in considerable detail here, I would hope that future studies could be more successful in avoiding these problems and do a better job of revealing truths about the natural world.

CHIPPING SPARROW

I begin by illustrating how a chipping sparrow acquires his song, because the roots of implausibility for the performance hypothesis lie in the biological basis for song learning. Then I show how those songs are used during aggressive interactions among males, especially in lek-like arenas during the dawn chorus. These two known biological features of chipping sparrows are not referenced by Goodwin and Podos (2014), but seriously undermine their conclusions.

The Biology of Song Learning by Chipping Sparrows

A young chipping sparrow acquires his song by copying the song of an adult next to whom he settles, as illustrated by Liu and Kroodsma (2006; Fig. 1). The adult's song is copied whether the trill is delivered slowly (males 13 and 31, and males 20 and 22) or more rapidly (males 7 and 9, and males 24 and 35), during the social and aggressive interactions between the adult tutor and the youngster who is establishing his first territory. This conclusion is based on solid field evidence by colour banding 324 young chipping sparrows and following them during dispersal.

To further illustrate how a young male chipping sparrow learns rather precisely the song of his adult tutor, and especially the tutor's trill rate, I recorded chipping sparrows during early May (2015) when they first returned from migration, before postlearning dispersal might occur. I used a Sound Devices 722 digital recorder and a stereo Telinga microphone to record 67 different males in two populations, one on a golf course in Lewiston, Michigan, U.S.A., the other in a city park in Northampton, Massachusetts, U.S.A. Birds were not banded, but I recorded most of the birds in rapid succession by moving directly from one singer to the next, so that the previous and next singer could be heard while recording a given male. If songs of suspected neighbours were identical, and I could not distinguish their songs in sonagrams, I conservatively assumed they were the same male and discarded one of the recordings from the data set. Using Raven Pro software, I measured trill rates and frequency bandwidths for three high-quality songs for each male, and used the median value in analyses ('spectrogram window size' in Raven: 110 for temporal measures, 2050 for frequency; lower and upper frequencies measured as -24 dB down from maximum power; I believe these methods match those routinely used in performance studies).

Among these 67 males, I found 14 pairs of adjacent males with essentially identical songs (see Fig. 2), as one would expect based on how chipping sparrows learn their songs. As is clear in Fig. 2, song types and trill rates are determined by where and from whom a male learns his song and cannot reflect any measure of his quality, in the performance sense of Ballentine et al. (2004) and Podos et al. (2004). A male with a trill rate of 25 is not 'better' than a male with a trill rate of 7; instead, he simply learned his song from a male having a trill rate of about 25, whereas the other bird learned his song from a male having a trill rate of about 7. One might argue, if pressed, that a young male could innately know his relative singing ability and then choose to settle next to an adult

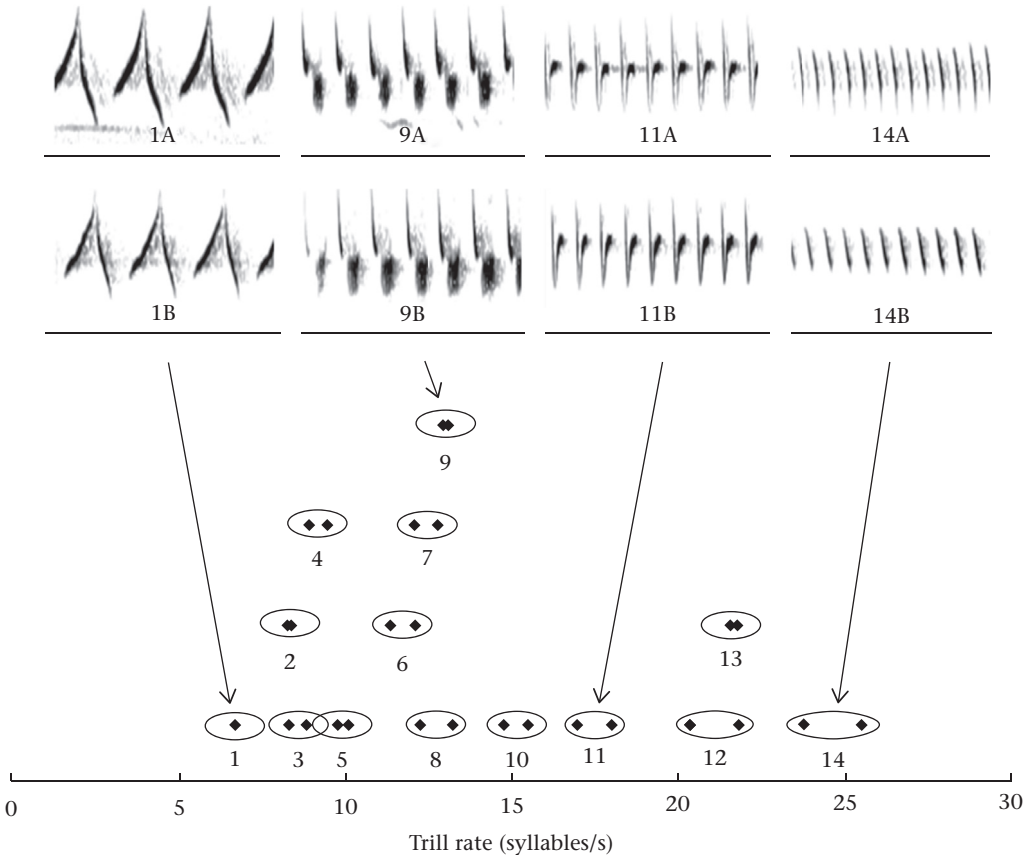

Figure 2. A few dozen different song types can occur within a chipping sparrow population (only four illustrated here: 1, 9, 11, 14), but neighbouring males (A and B) often have nearly identical songs, the result of a young male copying the song of a nearby adult singer (Liu & Kroodsma, 1999, 2006); all features of a male's song, including his trill rate as illustrated here (14 examples), are determined by that adult tutor. In the lower graph, each oval encircles the two data points (pairs 1 and 2 are identical) for trill rates from two neighbouring males with the same song types (data are distributed vertically for easier visibility). Each data point is the median of three measurements for a given male.

male whose song he can master, thus choosing a tutor male with an appropriate trill rate somewhere between 7 and 25. As shown in Fig. 2, neighbours often show nearly identical song types, a result of one bird learning from the other, and these matched pairs vary from slow to intermediate to fast trills, but there is no evidence for song learning in any songbird species or especially in chipping sparrows (Liu & Kroodsma, 1999, 2006) that a male is in any way limited in what naturally occurring trill rate he can learn.

The Biology of Song Use in Chipping Sparrows

Well before sunrise, during the dawn chorus, male chipping sparrows range widely over space, especially into neighbouring territories, but they can also display with other males in arenas far removed from their daytime centres of nesting activity. For example,

> *If territories are widely dispersed, it seems that the males still convene at a traditional location, sparring there even if some of the males don't own territories that border that place*
>
> (**Kroodsma, 2005, page 319)** (see also Liu, 2004).

One example suffices, from a location in eastern Missouri, U.S.A. (see Fig. 3). In that example, four males displayed simultaneously in a lek-like arena during the dawn chorus. As is typical of such gatherings, the males were on the ground (a paved road in this example) and aggressively chasing each other, all the while singing highly abbreviated songs at a rapid pace, up to 60 song fragments per minute instead of the four or so far longer songs per minute during daytime singing. Participation was not continuous, with one male arriving late, another leaving and then returning minutes later, the return sparking high-intensity calling before singing resumed. Before sunrise, the four males all dispersed, presumably to their daytime centres of activity where singing is typically resumed at a much slower rate high in the trees. Replacing those four males after sunrise were two other males, each now on his daytime centre of activity, each of which was presumably displaying elsewhere during the dawn chorus. Male chipping sparrows thus routinely intrude on the daytime activity centres (i.e. 'territories') of other males and display there competitively with lek-like behaviour.

The Trill Rate/Frequency Bandwidth Graph

The standard graph provided in studies of song performance is the scatterplot of frequency bandwidth versus trill rate (Fig. 4). The distance from a given plotted point to the upper-bound regression line (i.e. the deviation from the line) is then interpreted as a measure of a male's performance or proficiency on that particular song. A small deviation is considered a high-performance song, and a large deviation is considered a low-performance song. Because information on song type and individual males is not encoded in the data, however, the biological significance of the graph is obscured.

Consider, then, a graph of this sort that includes the information necessary to interpret it in a biological context (Fig. 5). Given how a

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

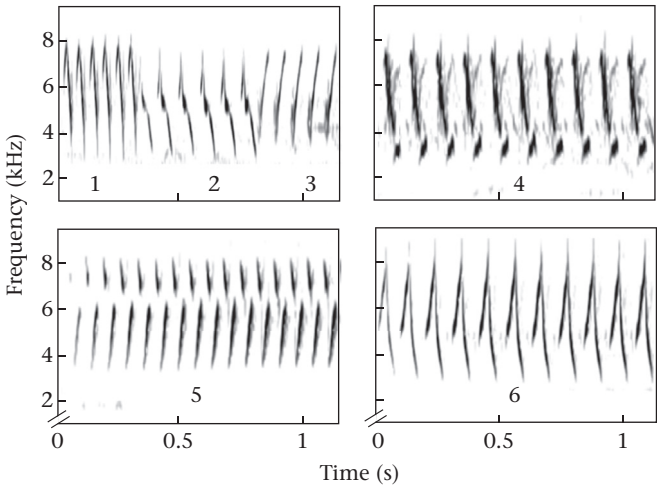

Figure 3. During the dawn chorus, chipping sparrow males can gather in lek-like, competitive singing arenas well away from their daytime territories. In this example, during the dawn chorus, four males (1–4) gathered in a lek-like arena on a paved road, displaying with brief songs in the dark, sight unseen (the half-second song fragment of bird 2 is a typical dawn song); individuality in their songs allowed each to be identified. After the dawn singers had dispersed to their daytime territories, two other males (5, 6), who presumably were elsewhere during the dawn chorus, sang at this location on their daytime territories; their songs were far longer, abbreviated to 1 s here only for illustration.

chipping sparrow learns his song from a neighbouring adult, it is clear from this figure that social factors and song learning explain not only (1) the variability in trill rates within a population (as in Fig. 2), but also (2) the variability in frequency bandwidth (Fig. 5), and therefore also (3) the scatter in the plot from Goodwin and Podos (Fig. 4). Scatter in the graph is explained not by trill rates or frequency bandwidths that reflect male quality, but instead by variation in the trill rate and frequency bandwidth of distinctive song types, reflecting the young bird's attempt to precisely imitate the song of his tutor neighbour.

A Focused Critique of Goodwin and Podos (2014)

The claims made by Goodwin and Podos (2014) are substantial, and novel (quotes below are from the title and abstract, with my edits in brackets):

> *Team of rivals: alliance formation [a cooperative fighting team] in territorial songbirds is predicted by vocal signal structure* [trill rate] … *Our results provide the first evidence that animals like chipping sparrows rely on precise assessments of mating signal features* [trill rates], *as well as relative comparisons of signal properties* [trill rates] *among multiple animals in communication networks, when deciding when and with whom to form temporary alliances* [cooperative fighting teams] *against a backdrop of competition and rivalry*
>
> **(Goodwin & Podos, 2014, page 1)**

These claims are made, however, by omitting reference to two ornithological facts about the subject animal that were published on the same population of chipping sparrows. (1) Trill rate reflects song learning from neighbours, not male quality (see above), and males cannot therefore assess one another based on trill rate, let alone precisely; and (2) published information on male behaviours would lead one to believe that these gatherings of singing males in small singing arenas are not cooperative but instead competitive, as they routinely involve much aggression among all birds involved.

A third major problem with Goodwin and Podos (2014) lies in their statistics. First, as pointed out by Akçay and Beecher (2015), the three tests supporting coalitions are simply done wrong; even when 'corrected' by Goodwin and Podos (2015), the one remaining significant test ($P = 0.03$) remains problematical. The reported statistically significant tests are gleaned from a much larger, unreported series of nonsignificant tests. The authors analysed data on (1) frequency bandwidths alone, (2) trill rates alone and (3) a combination of frequency bandwidths and trill rates. Even though only the bandwidth/rate combination makes any sense for the performance literature, the authors report only data on trill rates,

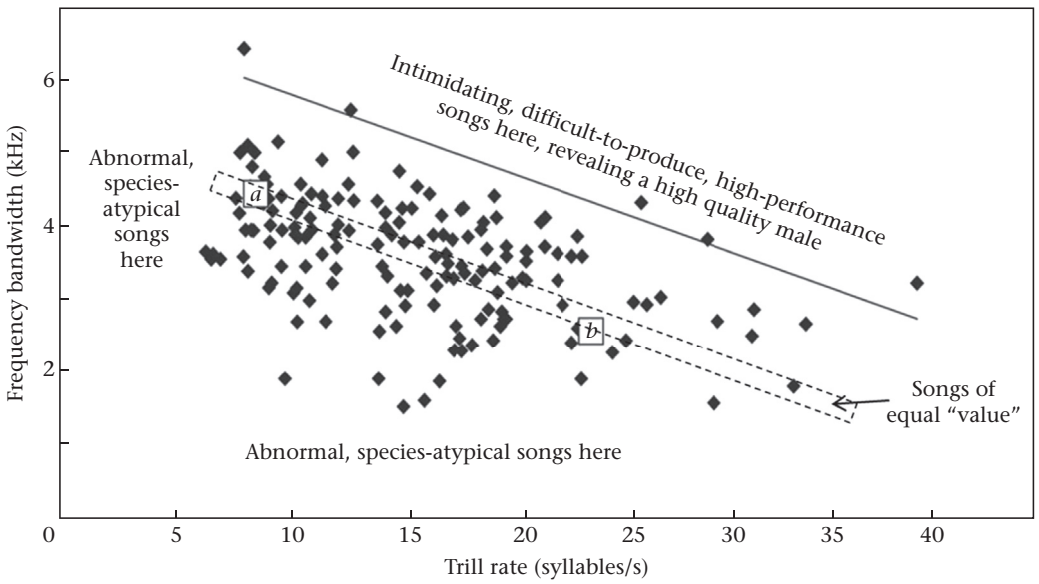

Figure 4. Reprinted with permission from Figure 1 in Goodwin and Podos (2014). 'Chipping sparrow songs show evidence of a vocal constraint … Biplot of trill rate and frequency bandwidth ($N = 160$ males) reveals a performance trade-off in vocal production'. Letters 'a' and 'b' refer to a portion of the original figure not illustrated here. Songs that plot within the dashed rectangle have similar deviations from the upper bound and would therefore be considered equally good, or proficient, in other performance studies. Data are replotted on expanded axes to show the open space below and to the left of the data points.

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

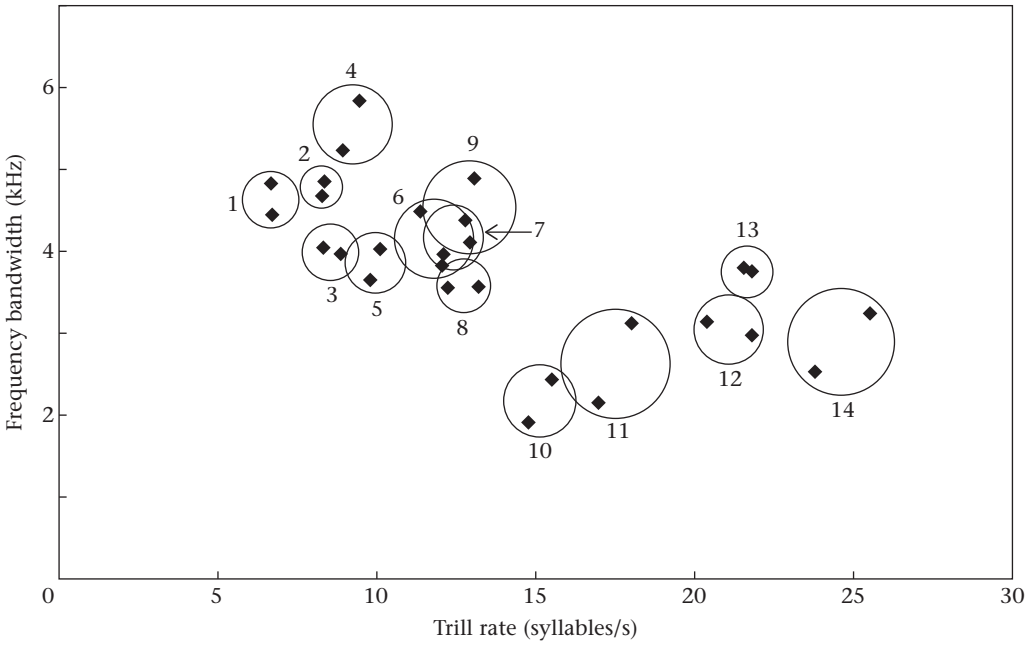

Figure 5. All aspects of a male chipping sparrow's song, including the trill rate and frequency bandwidth, are determined largely by the song that he imitates from an adult male, so that song types dictate the scatter in the plot, not relative male quality. Plotted data are from Fig. 2, and each of the 14 circles encompasses the songs of two neighbouring males.

presumably because the only statistically significant tests were found there. When undisclosed 'multiple comparisons' have been made, however, building a biologically significant story only on reported statistical tests that just barely reach $\alpha = 0.05$ is problematic, because any correction for multiple comparisons (e.g. Bonferroni) would render the reported tests nonsignificant; for example, the one remaining test that Goodwin and Podos (2015) claim to be significant, at $P = 0.03$, is nonsignificant when it is corrected ($\alpha = 0.05/3 = 0.017$) for the three admitted tests that were done. Include the undisclosed tests in the correction and there remains nothing on which to base their story; this general issue of 'undisclosed flexibility in data collection and analysis' is effectively addressed by Simmons, Nelson, and Simonsohn (2011, page 1359).

A fourth serious problem with Goodwin and Podos (2014, page 2) is revealed in this seemingly innocuous statement: 'We created stimuli by increasing or decreasing trill rate while ensuring the song was within the observed population range.' This oft-used method, with a single-minded focus on trill rate alone, dates back to Podos (1996) and simply creates highly abnormal, experimental song stimuli. Consider a song with a trill rate of 28, for example (see Fig. 4, or song 14 in Fig. 2); if three of four syllables are replaced with silence, thus reducing the trill rate to seven, the experimental trill rate is still 'within the observed population range' of trill rates and is therefore still considered normal, even though other dimensions of the song (e.g. ratio of syllable and intersyllable durations) are highly abnormal and unlike anything a chipping sparrow would ever hear or sing (for an expanded treatment of this issue, see Fig. 10 and accompanying text for an illustration of this problem with swamp sparrows). If trill rates of 21, 14 and 7 are created from a wild-type trill rate of 28, those songs become increasingly abnormal, yet only the declining 'performance value' of the song is considered relevant in performance studies. Using these four trill rates (7, 14, 21, 28) in any playback experiment, one would no doubt learn that 'aggressive behaviors … (approached the speaker more closely … spent more time within 2 m of the speaker … attacked the mount more often) … were significantly greater in response to fast trill rates' (Goodwin & Podos, 2014, page 3). I would suggest that responses to these experimental stimuli more likely reflect only how abnormal they are (or how much song stimulus is delivered; see also below, for Moseley, Lahti, & Podos, 2013), not how far these strange, experimental songs plot from an upper bound on the performance scale.

Still other issues have been addressed by Akçay and Beecher (2015). Those issues, together with the four major problems addressed above, lead me to conclude that these data do not constitute support for the performance hypothesis. Nevertheless, the flawed conclusions of Goodwin and Podos (2014; and of other performance studies in general, see ensuing discussion) are perpetuated when cited uncritically in support of subsequent studies of sexual selection and birdsong. For example, 'Neighbouring songbirds can even form alliances to expel common enemies, like … conspecific intruders' (Snijders, van der Eijk, van Rooij, de Goede, van Oers, et al., 2015, page 2).

SWAMP SPARROWS

The Biology of Song Learning by Swamp Sparrows

To illustrate the implications of song learning for swamp sparrows, I recorded birds at three locations during 2015 (Fig. 6). At each site, I used a stereo Telinga parabolic microphone, and either a Sound Devices 722 or Marantz PMD661 digital recorder. Birds were unbanded, but each male sang repeatedly over a few hours from the same predictable locations, and attributing each recording to a particular male was not difficult; if any doubts existed as to the origin of a song, it was discarded from the analyses. Songs were then analysed on Raven Pro 1.4 software (settings the same as for chipping sparrows), and the median of three examples of each song type from each bird was used in the analyses.

Several important points are revealed in these data (Fig. 6).

(1) Normal, wild-type swamp sparrow songs are restricted to a relatively limited set of all possible trill rates and frequency bandwidths (upper left subfigure in Fig. 6). Outside of this restricted area, all songs are, by definition, abnormal. For the

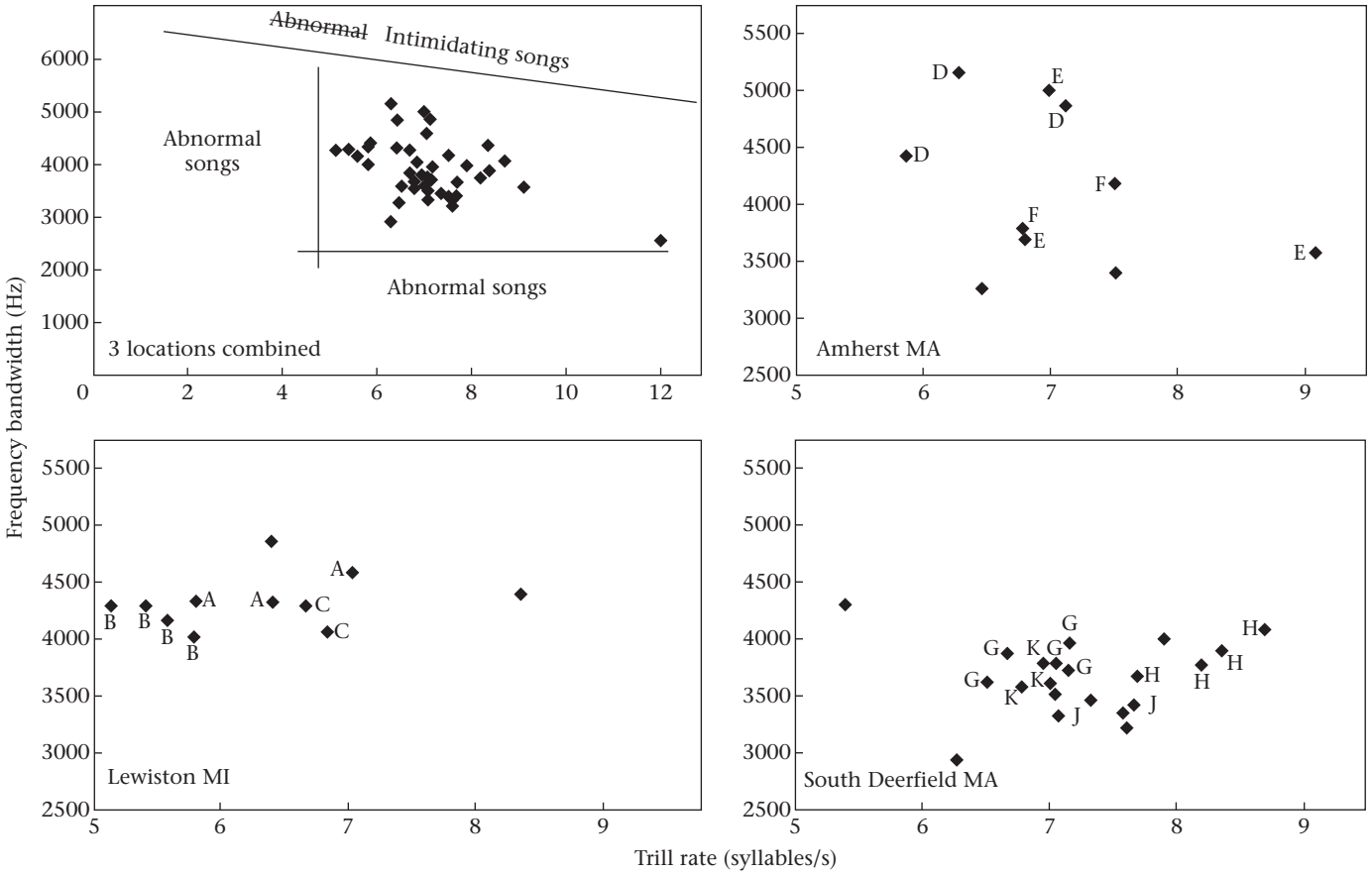

Figure 6. Scatterplots of trill rate and frequency bandwidth for swamp sparrows from three locations, revealing that scatter is largely due to song types (letters A–K; song types deemed unique to an individual are not labelled), leaving little if any information available about the quality of individual singers. In the upper left, data for all three locations are combined, and the axes meet at the origin (0,0); the graphs for the three separate locations are drawn to a different scale, expanded to better show the variation within locations (note that the outlying data point at 12 syllables/s is omitted from the lower-right graph; excluding it has no bearing on the conclusions to be drawn from this figure). The upper-bound line in the upper-left subfigure is from Ballentine et al. (2004); the bounds below and to the left of the data points are placed arbitrarily. In the three subplots, letters label different renditions of a given song type from different males; because of local dialects, song types were not shared across locations.

performance hypothesis, abnormal songs that approach the bound above the sea of data are considered supernormal and especially high performance, and so intimidating and threatening that listening males might well flee them (e.g. Illes, Hall, & Vehrencamp, 2006). Songs to the left and below the normal songs are just abnormal, although by definition the bounds to the left and below the data are also 'performance limits' for the birds. When bounded lines are added to the graph on all sides, they draw attention to the limits and demand explanations everywhere.

(2) Trill rates and frequency bandwidths can vary significantly by location, depending on the local dialect. Frequency bandwidths from Lewiston, Michigan, for example, are mostly above 4 kHz and those from South Deerfield, Massachusetts are below 4 kHz; trill rates are correspondingly slower at Lewiston. It is conceivable that some geographical differences might occur because of, for example, morphological differences in the birds, but it is difficult to come up with explanations other than 'local dialects' when the populations are only a few kilometres apart (see Fig. 8).

(3) Much of the scatter in the data is explained by song types (see also Fig. 7): Birds learn their songs (including trill rates and corresponding bandwidths) from one another, and as a result, many songs are shared within the population, so that songs judged to be of the same song type from different males tend to plot near one another (especially clear for Lewiston and South Deerfield).

A critical but untested feature of the performance hypothesis is that songs actually provide reliable, honest signals of male quality. If these scatterplots with the upper bound are at all relevant to how male and female swamp sparrows might assess a singer, then the performance measures should provide consistently reliable information about the singer. If no reliable information is provided, the relative performance of different males cannot be used as an honest signal of their relative quality.

As revealed in Fig. 8, measures for different males are broadly overlapping, and a given male might have both the 'best' and the 'worst' song in his repertoire. One might still argue that variation within song types could reveal differences in male quality, but then why would any male who is capable of singing such high-performance songs on one song type submit to singing any low-performance songs? And if the upper limit drawn on the graph has any relevance to song performance, and some song types are simply far from the line and therefore easy to sing, what kind of a measure of performance is it if a male sings a song slightly closer to the line when it is so easy to sing in the first place? Performance measures simply cannot be used consistently by either other

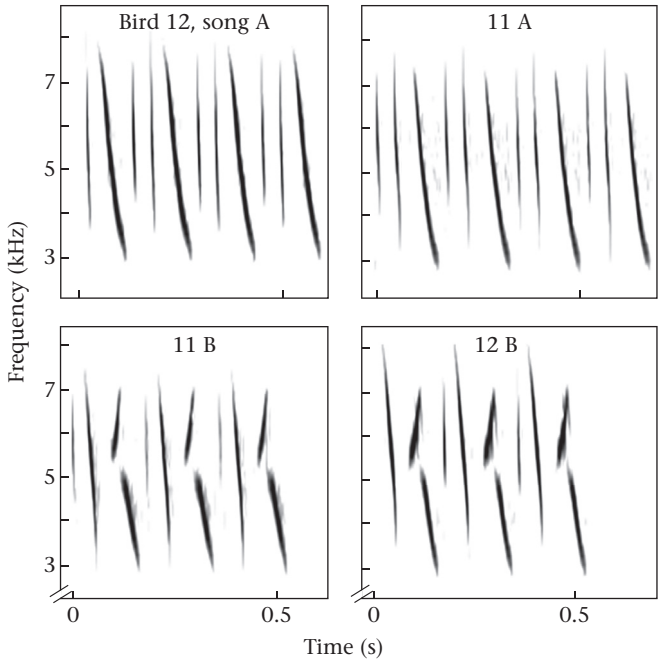

Figure 7. Swamp sparrow males within a marsh learn their songs from one another and, as a result, many songs are shared among birds in a population. From Lewiston, Michigan, U.S.A., examples are illustrated for two different song types (A and B). Songs of higher performance (lower deviation from the upper bound, as shown in upper left of Fig. 6) are in the second column. Birds 11 and 12 were immediate neighbours; bird 11 had the 'better' A song, bird 12 had the 'better' B song.

swamp sparrows or by humans to assess the relative quality of a singer. The data provide no support for the feasibility of the performance hypothesis.

A Focused Critique of Swamp Sparrows and Performance Studies

My conclusions for swamp sparrows are at odds with all of the published studies on this species. In an attempt to understand why, I next examine each of those studies in chronological order.

Podos et al. (2004): Calibration of song learning targets during vocal ontogeny in swamp sparrows, Melospiza georgiana

Podos et al. claim that, when a young swamp sparrow learns a given song type, he adjusts the trill rate or frequency bandwidth to match his own proficiency at producing that song, so as to acquire as high a performance song as he possibly can (i.e. closest to the upper bound on the graph), and they further claim repeatedly that their data are 'consistent with' or 'support' the 'calibration hypothesis'.

The impression conveyed by these claims is that, given how everything is consistent with the calibration hypothesis, it must therefore be true. 'It is the consistency of the information that matters for a good story, not its completeness' (Kahneman, 2013, page 87). The words 'consistent with' are also red flags for readers to ask what other hypotheses the data might be consistent with, or what data are not consistent with the hypothesis (i.e. what is missing in the story?).

The figures I provide on the biology of swamp sparrow song learning do not support the authors' conclusions about calibration.

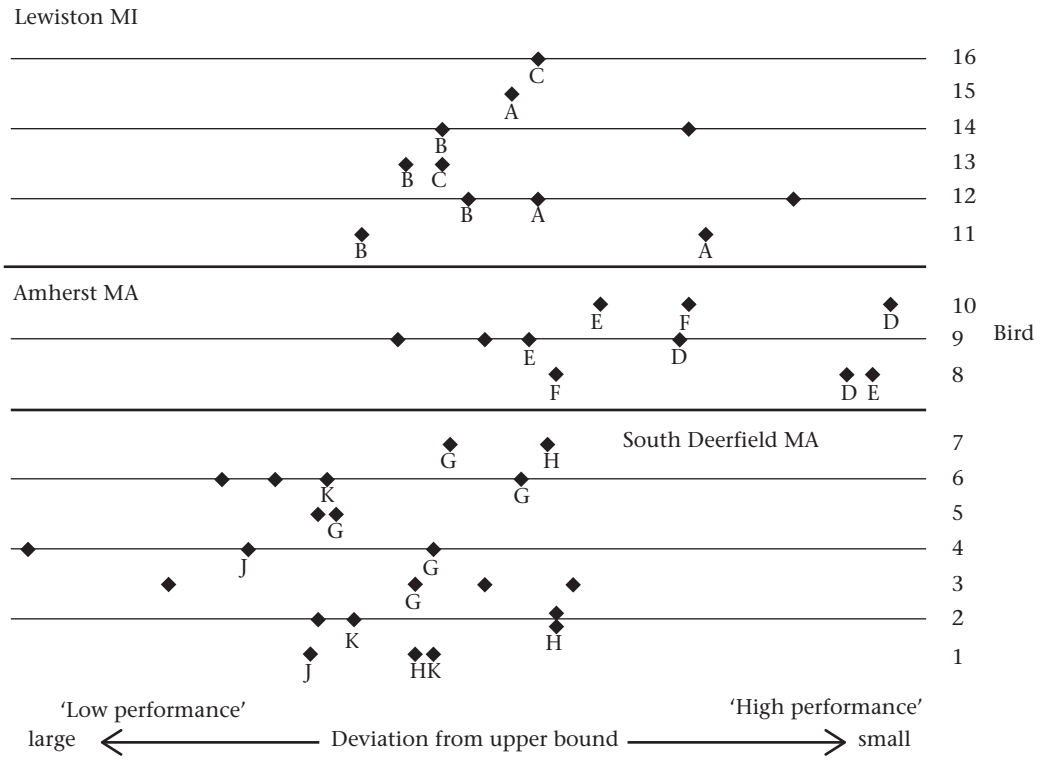

Figure 8. Song performance measures can provide no reliable information about inherent male quality in swamp sparrows. Data plotted here have been extracted from Fig. 6 by measuring the shortest distance to the upper bound for each song (all measures are relative, depending on a number of factors, so no absolute scale is provided for performance). Performance measures for different males are broadly overlapping, such that a male might have the highest performance on one of his song types but the lowest on another (e.g. compare bird 3 from South Deerfield with birds 1, 2, 5, 6 and 7; song types are lettered as in Fig. 6). Note that Amherst swamp sparrows significantly 'outperform' the South Deerfield birds only 20 km distant (the two populations were recorded within a few days of each other during late June 2015; I can think of no methodological issues that would have produced such results).

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

When learning a song, a swamp sparrow conforms to the particular song type of the local dialect; he does not adjust features of what he learns in any way consistent with an attempt to calibrate a normal, wild-type song to his own abilities. As a result, a male may have what would have to be labelled the worst 'performance' on one song type, the best on another.

Podos et al.'s data are also consistent with a very simple alternative explanation, which they failed to point out: no matter what recognizable features of a song a swamp sparrow hears, he tries to develop as normal a song as possible, making a fine-tuned effort to take whatever he hears and produce a normal, wild-type song (the only logical conclusion also for Lahti, Moseley, & Podos, 2011; see review below). Contrary to the title and the statements supporting it in this paper, there is no credible evidence that an individual male 'calibrates' normal swamp sparrow songs to his particular proficiency.

Ballentine et al. (2004): Vocal performance influences female response to male bird song: an experimental test

Female swamp sparrows are shown to display more to high-performance songs than to low-performance songs of the same song type recorded from different males. The authors conclude the following:

> *we can conclude with certainty that females are attending to subtle differences in song reflecting male motor capabilities. Thus, our results provide a crucial piece of evidence in support of the general hypothesis that female birds assess male quality on the basis of vocal performance*
>
> **(Ballentine et al., 2004, page 167)**

Those who have cited this paper in their own work are equally convinced: 'it is known that females prefer trilled songs closer to the production limit' (Illes et al., 2006, page 1907; reviewed below); 'females are able to assess a male's quality as a potential mate using vocal performance' (Dubois, Nowicki, & Searcy, 2011, page 724; reviewed below); 'females … are known to discern fine features of song in the functional contexts of mate choice' (Lachlan, Anderson, Peters, Searcy, & Nowicki, 2014, page 2).

In fact, 'As of May/June 2014, this 'highly cited paper' received enough citations to place it in the top 1% of the academic field of Plant & Animal Science based on a highly cited threshold for the field and publication year' (Web of Science, http://hcr.stateofinnovation.thomsonreuters.com/page/archives, accessed 8–25 October 2014). In a survey entitled '25 Years of *Behavioral Ecology*', a review article for the journal cites the importance of this paper:

> *the 10 articles from Behavioral Ecology which have received the most number of citations … females are more likely to solicit copulations from males capable of vocalizing at the upper boundaries of the performance limit; female swamp sparrows prefer males with the most elaborate sexual displays*
>
> **(Simmons, 2014, page 1)**

One focus of this paper is especially puzzling, in that it does not present the most relevant analysis for this kind of study. Ballentine et al. expend considerable effort demonstrating (1) that males can differ in their average performance scores (mean, CV, etc.) and (2) that males are consistent in how they produce a given song type (model II ANOVA, repeatability measures, within- versus between-male variation, their Table 1). Yet, even though extensive sampling yielded 30 different song types among 91 males (with an average repertoire of 3.1 song types/male), with all the needed data in hand, no apparent effort was made to determine whether those 3.1 performance scores for each male were consistent with each other. The performance measures as assays of male quality are useful only if the performance scores for a given male in some way reliably convey his ability. My above analyses for swamp sparrows demonstrate that these performance measures among a male's different song types are neither consistent nor reliable; a similar analysis from the data of Ballentine et al. likely would have revealed the same implausibility for the performance hypothesis.

Ballentine et al. actually deal their own performance hypothesis a serious blow when they write that 'some song types consistently have low deviations and others high deviations regardless of which male sang them, suggesting that some song types are harder to produce than others' (page 165). From that observation, one of the three following conclusions must be correct: (1) selection for low deviation (i.e. 'high performance') songs is not uniform among all song types, or (2) selection is uniform but the deviation measure does not reveal it, or (3) there is no selection for low-deviation songs to convey male quality. Consequently, although it remains possible (although largely assumed) that deviation from the upper bound could reflect the relative difficulty in producing a song, that *deviation cannot reflect male quality*, because males readily and routinely learn many song types that, according to the performance hypothesis, are easy to produce and therefore cannot reveal any intrinsic ability of the male.

All of the data thus suggest that the scatter in the plot has nothing to do with male performance, and everything to do with the need for males to conform to the song types of a given dialect, regardless of where they plot on the graph. It is an accurate production of the particular song type that seems to matter to a male, not his overall 'song proficiency or performance' as measured by the deviation from the upper bound on the scatterplot.

If deviation does not reflect male quality, why would the 10 female swamp sparrows in this study display more to high-performance songs than to low-performance songs? Briefly, I offer three possibilities for the results, which are difficult to accept at face value given all the above.

First, 'Believing is seeing', it might be said, the results stemming from nonblind observers with strong expectations for the results (i.e. observer bias; see Burghardt, Bartmess-Levasseur, Browning, Morrison, & Stec et al., 2012). From the outset, the concept of 'performance' is already a given, not a hypothesis: 'our knowledge of song production mechanisms allows us to identify *a priori* which songs are produced with *greater vocal proficiency*' (Introduction, page 163; italics mine). This conviction permeates the paper: 'males that shared the same song type also differed … in *how well* they produced these song types' (page 165). Variations of the word 'perform' with its attending assumptions and built-in biases (see Discussion) are used nearly 100 times throughout this paper. Authors who know which songs are 'best' are going to have a difficult time being objective in how they judge the birds' responses to known 'good' and 'poor' songs.

Second, alternative explanations for results are never considered in this study, and rarely, if ever, in other performance studies, but several years later Ballentine (2009) inadvertently offers another explanation for the results (see review of that article below). First year birds have lower performance scores (higher deviation) than older birds, and their songs are distinguished by higher plasticity, or lower consistency, in the trill notes. Songs were recorded by Ballentine et al. (2004) during May and June, but even during late May I have found that songs of some yearling swamp sparrows can remain highly plastic. It would be fully expected for females to respond more strongly to (higher-performance) stable songs of full adults than to (lower-performance) plastic songs of yearling males, and the females could do so without any reference

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

to the relative song deviation or performance for yearling and older males.

Third, it is assumed wrongly that measured frequency bandwidths faithfully capture the essence of a male's song, but those measured bandwidths can vary enormously depending on (1) the microphone system used, (2) the distance to the singer and (3) the particular song type being measured (Fig. 9), and undoubtedly (4) the environmental conditions under which the song was recorded. The study by Ballentine et al. is compromised by the use of two different parabolic reflectors, one with a 13-inch (33 cm) diameter (Sony PBR-330) and one with an 18-inch (45.7 cm) diameter (Saul Mineroff SME PR-1000), and there is no control for distance. In all of the performance literature, bandwidths as measured are a function of several variables and do not accurately measure bandwidth at the source (i.e. at the beak of the singing male; increasingly sophisticated analyses of these measures, as by Wilson et al. (2014), are undermined by their very inaccuracy).

There is also a large parallel literature, none of it cited in any of these studies of performance, that shows how birds vary their response to playback songs depending on how much reverberation is in the recording, prompting Morton, Gish, and Van Der Voort (1986, page 815) to write the following: 'Sufficient evidence now exists to suggest that sound degradation … should be taken into account in studies using responses to playback of bird song'. Songs recorded at greater distances from the singer would likely have more reverberation and could thus be rated 'low-performance' songs, in which case both females and males would be expected to respond less to them based on reverberation alone.

Dubois, Nowicki, and Searcy (2009): Swamp sparrows modulate vocal performance in an aggressive context

The authors' main conclusion (page 163, from the Abstract): 'we show that male swamp sparrows … increase the vocal performance of individual song types in aggressive contexts by increasing both the trill rate and frequency bandwidth'.

Male swamp sparrows were played either a control song (that of a white-crowned sparrow, *Zonotrichia leucophrys*) or an 'aggressive' song (that of a conspecific), and the authors then measured the trill rates and frequency bandwidths of the songs delivered in these two contexts. Two results stand out.

(1) The particular song types used by subjects in reply to aggressive and control contests did not differ. When it matters most, then, when a male is challenged on his territory, he seems to choose a random song from his repertoire, not a song that best conveys his overall quality. This important result is inconsistent with the performance hypothesis (although not mentioned in the abstract), yet the authors puzzlingly conclude 'we do not think this means that males are not trying to maximize their vocal performance during aggressive signaling'.

(2) The following quote, reinforcing the title and abstract, is from the Results (page 164): 'males increased both the trill rate … and the frequency bandwidth … during the aggressive trial. This results in significantly higher vocal performance … during the aggressive trial' (results are based on an overall average among $N = 23$ males, with increases of from 6.94 to

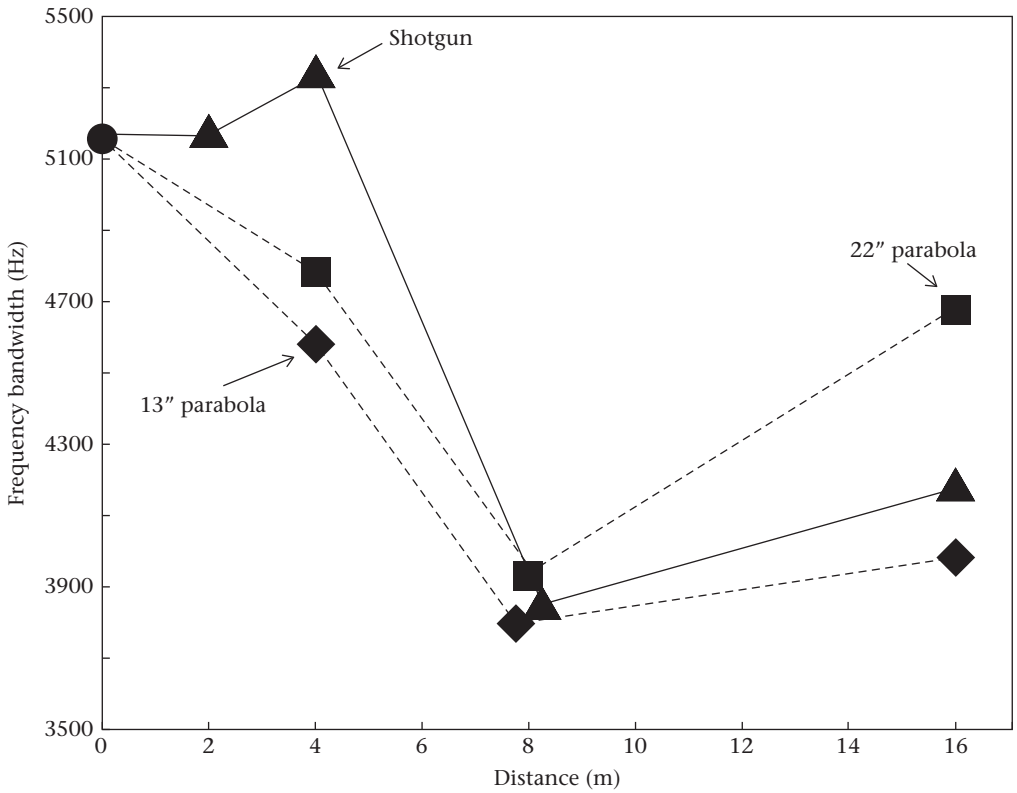

Figure 9. The frequency bandwidth of a chipping sparrow song, as measured in performance studies, varies not only with (1) the microphone used but also with (2) the distance to the singing subject. A chipping sparrow song was broadcast from a JBL Pro III speaker under calm, quiet conditions (midnight) and recorded at 2, 4, 8 and 16 m; the speaker was perched 1 m above the ground, with a direct line of sight beside several small bushes over the 16 m distance. The bandwidth value reported for 0 m was measured directly from the WAV file that was later broadcast through the speaker. Each plotted value is the median of three measurements for successive broadcasts of the same song. Microphones tested: Sennheiser ME66 shotgun, 13-inch (33 cm) Sony PBR330 parabola, 22-inch (56 cm) Telinga parabola (both parabolas with Sennheiser MKH20 microphone). Furthermore, bandwidth also varies with a third variable, song type, as four additional song types revealed strikingly different transmission patterns among them (for the ME66 shotgun, bandwidths for all five song types ranged from 83 to 104% of the original WAV file at 4 m, 74 to 100% at 8 m and 72 to 107% at 16 m).

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

7.10 syllables/s, 4870.4–4960.9 Hz). These authors would later declare that male swamp sparrows 'actively increase' and 'exaggerate' their vocal performance in aggressive situations (Dubois et al., 2011).

Yet, one must ask, how could it possibly be biologically meaningful to increase the trill rate by 2.3% or frequency bandwidth by 1.8%? In Fig. 6, for example, consider a song with a trill rate of 6.0 syllables/s and a frequency bandwidth of 4000 Hz that is 'exaggerated' to 6.1 syllables/s and a frequency bandwidth of 4072 Hz. The exaggerated data point on the scatterplot is moved a miniscule distance. If a male really wanted to increase his performance during aggressive contexts, he could switch to a more impressive song in his repertoire, but he does not do that, as if performance did not matter. Furthermore, during aggressive contexts, it is likely that males approached more closely and were recorded at shorter distances than during neutral contexts. An aggressive subject recorded at 4 m could show a 20% increase in measured bandwidth over a neutral subject at 8 m, with no actual difference in the song as delivered by the male (Fig. 9). Moreover, two years later, the authors would accept that these 'exaggerations' are biologically meaningless (Dubois et al., 2011; see below).

Also, the title of the paper may be true, but it is highly misleading, because swamp sparrows also modulate their songs in nonaggressive contexts. Using two lengthy recordings from my collection, for example, I measure that trill rates vary from 1% to 3% within a neutral session, spanning the 2% change the authors measured from neutral to aggressive contexts. Frequency bandwidth is also modulated within neutral sessions, varying by a median of 1.1% among the three measures taken from all swamp sparrows that I analysed for this study.

It should also be noted that Dubois et al. (2009) measured frequency at a resolution of 172 Hz, yet the frequency difference between neutral and aggressive contexts was reported as 91 Hz, about half the magnitude of the measurement error, thus rendering their frequency measurements inadequate.

Ballentine (2009): The ability to perform physically challenging songs predicts age and size in male swamp sparrows, Melospiza georgiana

The author 'used the highest performance song in a male's repertoire to determine each male's vocal performance', but that rationale is questionable. As discussed above, an important condition for honesty and reliability is that males consistently use songs within a relatively narrow range of performance abilities. If the performance values of males broadly overlap (see my Fig. 8), so that a male can rank highest on one of his songs and lowest on another (as also revealed in Kagawa & Soma, 2013), and a male does not even use his 'best' songs in aggressive encounters when it matters most (see Dubois et al., 2009), or does not use his 'best' song in any particular context that has yet been identified, it makes little sense to rate a male only by the one song of highest performance ability.

Also, suppose a female is to use performance, as measured in this paper, to distinguish first-year from older birds. For each male that she would want to assess, she would have to (1) listen to his entire repertoire over an extended period, (2) rate and remember each of his song types on the performance scale, (3) eventually dismiss as irrelevant all the song types of lowest performance value (but why?) and (4) focus only on the one song type that plots closest to the upper bound on the scatterplot, because that is the song type to be used to predict this male's age and quality. Before making her decisions about relative male quality, she would have to accomplish this task for a number of males, integrating over a broad range of trill rates and frequency bandwidths (see dashed rectangle in Fig. 4) among widely varying song types, simultaneously adjusting all of her evaluations for both the distance she was from the singing male and for each song type, given differences in sound transmission (Fig. 9).

Identifying a first-year bird does not require that much effort. Songs of first-year birds are typically more plastic and less repeatable than those of older birds, and this plasticity alone could readily identify a young bird in just a few songs. Merely writing repeatedly that the data 'support' the hypothesis that birds attend to performance ability, and not mentioning (less exciting) alternative explanations, does not make the hypothesis true (see also my above critique of Podos et al., 2004); authors have a responsibility to present for readers a balanced perspective on their findings.

Given that all song types were recorded from all males in this study, the author missed an opportunity to show, as I have, that song performance cannot be a reliable measure of male quality (my Figs. 6 and 8). I do not understand how this important analysis, so crucial for the performance hypothesis to be true, has not been presented in the many studies of the performance hypothesis that have had the relevant data.

Podos, Lahti, and Moseley (2009): Vocal performance and sensorimotor learning in songbirds

This review of the literature so far is typical of the unflagging support that one finds for the performance hypothesis:

> *Emerging descriptive and experimental evidence thus suggests that vocal performance varies among individuals, and suggests that singers who maximize vocal performance gain advantages in song function and ultimately in reproductive success*
>
> **(Podos et al., 2009, page 170)**

I can find no credible scientific evidence to support that conclusion, either in the literature up to 2009 or the years to follow.

Dubois et al. (2011): Discrimination of vocal performance by male swamp sparrows

Three experiments are performed. In experiment 1, males are asked to discriminate between high- and low-performance songs of the same song type as sung by different males.

> *Responses were greater toward high-performance song on all five univariate measures, and the differences were significant for three of these … This result supports our … hypothesis that males assess individual differences in vocal performance*
>
> **(Dubois et al., 2011, page 722)**

Three issues can be raised about these conclusions for experiment 1.

(1) As can be seen in my Figs. 6 and 8, song types plot in different spaces on the scatterplots, because males conform to the features of that song type when learning it. It is the conforming that is important, not any exaggeration of trill rate or frequency bandwidth to reveal a bird's prowess on a particular song type.
(2) The songs used in playbacks are the same songs that were used by Ballentine et al. (2004) and Ballentine (2009). My critique of those papers also applies here.
(3) In these kinds of playbacks, which consistently produce the expected results in tests of the performance hypothesis, credibility will be enhanced with blind observations; see Burghardt et al. (2012).

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

Experiments 2 and 3 are similar to each other, each of them asking whether males respond differently to the kind of within-male differences in vocal performance observed in Dubois et al. (2009), where trill rates and frequency bandwidths increased on average about 2% from neutral to aggressive performances. No significant differences in response were found (i.e. males responded no differently to the 'extremes' of high- and low-performance versions of a particular song type that a given male might sing).

Lahti et al. (2011): A tradeoff between performance and accuracy in bird song learning

Experimental songs are produced by adding or deleting silent intervals between song elements, yielding songs that swamp sparrows would never by themselves have produced or heard in nature. Young swamp sparrows are then tutored with these odd songs.

> *Our main finding is that birds elevated the trill rates of low-performance models, but at the expense of imitative accuracy*
>
> **(Podos et al., 2004, page 802)**
>
> *The elevation of trill rates of slowed models supports the hypothesis that birds calibrate learned vocal output to match their individual performance capabilities (Podos et al., 2004, 2009) … Prior work in swamp sparrows showed calibration*
>
> **(Podos et al., 2004, page 808)**
>
> *our data imply that selection has favored birds that …* [produce] *… trill rates that maximize birds' vocal capabilities … A bias toward increasing the performance level of songs would enable birds to indicate their performance capacities; otherwise, the quality of a tutor's song would set a ceiling on the performance level a learner could attain*
>
> **(Podos et al., 2004, page 809)**

These interpretations have problems. What is certainly true is that the young swamp sparrows removed silent intervals from odd, slowed tutor songs to produce more normal, wild-type songs. That result, however, based on abnormal, experimental songs, does not warrant any conclusion about a young swamp sparrow either in nature or in the laboratory taking a natural tutor song that it hears, foregoing 'imitative accuracy', and adjusting that song in trill rate or frequency bandwidth to match his own capabilities, all so that he can honestly broadcast his individual quality. There are no data in this paper or elsewhere demonstrating that a young swamp sparrow adjusts a normal or abnormal song to match his own individual proficiency, only data showing how young birds strive to produce normal, species-typical songs; in fact, Fig. 3a of Lahti et al. (2011) shows that the more normal the tutor song that a young bird hears, the better he will copy it, with altered models being copied less accurately.

If one wanted to test the importance of trill rate in reflecting male quality, one simple experiment would be to tutor birds with natural, slow songs and natural, fast songs. If trill rate is important, one would predict less accuracy in learning the slow songs, because males would try to 'improve' on them by speeding up the trill rates. I would predict, as I have argued throughout this paper, equally accurate copies of slow and fast models, because trill rates and frequency bandwidths, or their combination (i.e. performance), are not relevant to the birds as indicators of male quality.

One hint that the authors perceive the conflict between their data and their interpretation, however briefly, is provided in the following quote (italics mine): 'our results reveal that vocal ontogeny can be shaped … by a premium on high performance. Again, performance in this case refers to the trill rate of songs, all other features being equal, and *high performance being that of typical songs recorded from the field* as compared with our experimentally slowed versions' (page 808). Except for the 11 words in italics, throughout the paper 'performance' refers to vocal proficiencies of individuals, to an individual adjusting a tutor's song to the 'best' song he possibly can produce (i.e. relatively fast trill and broad frequency bandwidth), thus revealing his individual proficiency and quality, as in the first sentence of this quote. The brief reference to normal, wild-type songs is a puzzle.

Moseley et al. (2013): Responses to song playback vary with the vocal performance of both signal senders and receivers

The authors use previous methods (e.g. Lahti et al., 2011; Podos et al., 2004) to produce highly abnormal test stimuli: 2 s songs for playback to swamp sparrows are prepared from normal songs by either inserting or deleting silent spaces between the song elements. The manipulated songs then contain anywhere from 35% (a 'low performance' song) to 155% (a 'high performance' song) of the elements in control songs, with trill rates for those particular songs thus ranging from 35% to 155% of normal. As is evident in their Fig. 1 and my Fig. 10, three obvious features of the songs have changed from the original song: (1) the trill rate is slower or faster, (2) the quantity of stimulus is correspondingly less or more and (3) the more silence edited into or out of the song, the more abnormal it is, unlike anything a swamp sparrow has ever sung or would hear.

The two confounding variables are a serious problem. First, it is entirely reasonable to expect that a 'normal' song with three times as much stimulus as the lowest-performance song might elicit a stronger response, based on stimulus quantity alone. But the authors offer no control for such an alternative explanation for their results. Without somehow controlling for this confounding factor of stimulus quantity, one cannot attribute response strength to trill rate alone.

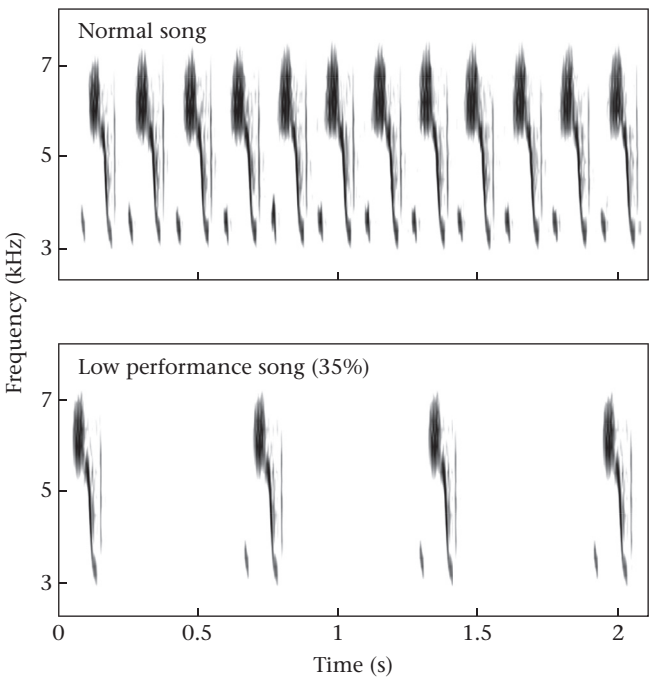

Figure 10. To produce playback stimuli, Moseley et al. (2013) altered a normal song (top) by adding or deleting silent spaces between the song elements. In this illustration (bottom), silent intervals are increased 8.3 times, producing a highly abnormal song with four syllables delivered at 35% the rate of normal.

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

Second, the stimulus songs are simply highly abnormal. To a swamp sparrow, a song slowed to 35% of normal must sound odd indeed, heard as a staccato, halting sequence of perhaps familiar song elements all out of sync, as these songs fall far outside the range of what any swamp sparrow would ever sing in nature. Two years before, in fact, these same authors (Lahti et al., 2011, page 808) had concluded that songs with trill rates below 55% or above 115% of normal were 'so unlike typical swamp sparrow songs that males do not consider them as targets for learning', that is, they are so abnormal as to not be biologically meaningful; nevertheless, Moseley et al. use songs well outside that range.

The confounding variables of stimulus quantity and abnormality do not seem to be of concern to the authors when they conclude the following: 'territorial male swamp sparrows responded significantly less strongly to low-performance than to control-performance playback stimuli, consistent with our hypothesis that receivers should attribute limited threat to low-performance songs' (page 4).

When that conclusion is rewritten to focus on one of the confounding variables, it becomes uninteresting and almost certainly unpublishable: territorial male swamp sparrows responded significantly less strongly to *abnormal* than to *normal* playback stimuli, consistent with our hypothesis that receivers should attribute limited threat to *abnormal* songs.

The authors 'predicted that subjects' tendencies to engage simulated intruders would vary positively with their own vocal performance' (page 2), that is, that more aggressive males would sing higher-performance songs, and then proved it. It is difficult to understand how that prediction arises (see Figs. 6–8), given that (1) males have several song types in their repertoire, (2) those song types vary widely in vocal performance, (3) such that song performance offers no reliable indication of male quality (Fig. 8), (4) the particular song type a male chooses to use in aggressive contexts is random with respect to the purported vocal performance capabilities of that male (Dubois et al., 2009), (5) the song he does use is not exaggerated in performance in any detectable way and (6) the authors measured the vocal performance of the responding male only by that one randomly chosen song he used during the playback responses. Any relationship between the measured song quality (especially frequency bandwidth) and the aggressive response of a male as described by the authors is likely an artefact of how close the male was to the microphone when he was recorded (see Fig. 9).

The logic is troubling throughout this paper. Two examples suffice.

(1) 'we predicted that stimuli with performance levels increased slightly would be responded to aggressively, whereas stimuli increased to the highest performance levels would be avoided, because of the higher perceived risk' (page 2) of a supernormal stimulus (from the Introduction). The highest-performance songs, by the authors' definition, can also be the most abnormal, yet the authors argue that these highly abnormal songs cause subject males to flee. The authors do not explain, however, how they can distinguish between failing to respond to a highly abnormal stimulus and fleeing an intimidating stimulus.

(2) The Discussion is a string of ad hoc explanations for why males (a) might not respond strongly to low-performance (abnormal) songs (e.g. low threat from a low-quality intruding male who is no threat in extrapair matings for the resident male), (b) might respond strongly to high-performance (perhaps relatively normal) songs (high threat for loss of paternity to intruding superior male), or (c) might not respond strongly to even higher-performance (perhaps most abnormal) songs, because then the responding male should flee, although now the apparent lack of response to the stimulus is because the test stimulus is high threat, not low threat as before.

In spite of all these issues, the authors conclude the following.

> *Taken together, our results provide a novel line of support for the hypothesis that vocal performance provides a reliable signal of aggressive threat … Most broadly, our data contribute to a general understanding of how animals respond to signals or signalers that are threatening*
>
> **(Moseley et al., 2013, page 7)**

OTHER SPECIES

Banded Wrens

I offer comments on just a few more papers, although I have reviewed many others searching for evidence supporting the performance hypothesis (e.g. Cramer, 2013; Cramer, Hall, De Kort, Lovette, & Vehrencamp, 2011; Cramer & Price, 2007; De Kort, Eldermire, Cramer, & Vehrencamp, 2009; Juola & Searcy, 2011; Kagawa & Soma, 2013; Sprau, Roth, Amrhein, & Naguib, 2013). Of all the studies reviewed in this document, I believe only one used blind observers to collect playback data (Cramer & Price, 2007), although even in that study the relative trill rates of the two playback stimuli would have been an obvious clue as to which song was which.

Illes et al. (2006): Vocal performance influences male receiver response in the banded wren

Given that each male banded wren, *Thryothorus pleurostictus*, has about 20 different songs, each learned from other males in the local dialect, the scatterplot of frequency bandwidth and trill rate contains a wealth of information (Fig. 11). Foremost, to me at least, it reveals great variation in 'performance' among different song types. Some song types are low performance, some high performance, so that like swamp sparrows, the scatter in the plot seems dictated by song type, having little if anything to do with consistent individual differences in performance. All males conform to the local dialect of songs, learning those songs whether they are deemed to be high- or low-performance on the scatterplot (again, as if performance itself did not matter).

Playback stimuli from 25 males (consisting of 12 different trill types used to construct 19 different song types) are played back to 31 males; given that it is trill types that are manipulated, the simplest way to avoid pseudoreplication would be to use 12 trill types as the unit of analysis, not 19 song types, or 25 males, and certainly not 31 independent playbacks as was done in their analyses. For each pair of playback stimuli, a slow and fast version of a song type was created by adding or deleting silent intervals between trill elements, all the while ensuring that trill rates remained within the range of natural variation for those particular trill types. A faster trill rate thus became a higher-performance song, as it was moved to the right on the scatterplot and was therefore closer to the upper bound. It is also worth noting two potentially confounding variables: (1) altering relative durations of trill notes and silent intervals could create abnormal songs in respects other than trill rate, but without additional information the degree of abnormality cannot be known; and (2) the fast and slow songs contain the same number of trill notes, so the fast songs are shorter in duration than the slow songs.

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

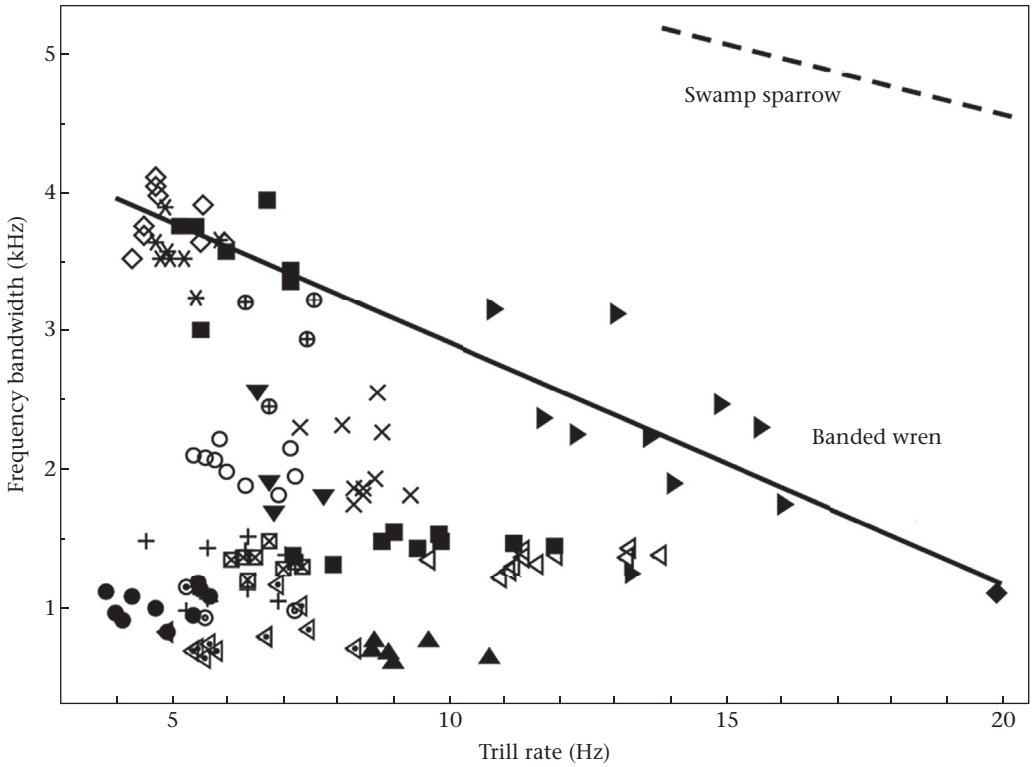

Figure 11. Reprinted with permission from Figure 1 in Illes et al. (2006). 'Graph of trill rate versus frequency bandwidth for 695 trills recorded from 13 individuals and including 16 trill types. Each symbol type represents a different trill type. The banded wren upper-bound limit is shown with a solid line and the swamp sparrow limit (Ballentine et al. 2004) with a dashed line.' Note that, as is usual for these graphs, the axes do not meet at the origin (0,0), and thus do not reveal the 'atypical space' below and to the left of the data.

Given the authors' prediction 'that males would respond more strongly to stimuli closer to the performance limit' (page 1908), the need for blind observers in the large experimental arena becomes especially important. In habitat with limited visibility that consists largely of 'tropical deciduous forest in various stages of regeneration' (Molles & Vehrencamp, 1999, page 678), with a bird up to 40 m distant, the observer must, at times, estimate 1 m movements, or estimate whether a bird is within a bounded area or not. Even though flagging is used to mark area boundaries, the task of monitoring the location of a moving bird in this habitat seems a high challenge to accomplish with much confidence.

My overall concerns are similar to those for the chipping sparrow and swamp sparrow studies. Given the distribution of data in Fig. 11, for example, birds seem far more concerned with conforming to the local dialect than with singing high-performance songs. Although abundant data are available for multiple songs types from each individual, the authors do not take advantage of that information: 'Male banded wrens differed in their performance of a given trill type, although we do not yet know whether such differences are consistent across all song types and reflect individual singing ability' (page 1911). As of 2016, that analysis seems not to have been done.

One strong summary statement requiring evaluation is the authors' conclusion that high-performance songs repel birds. The evidence for that statement seems to be as follows. (1) When the lower- and higher-performance songs are played simultaneously, birds are more likely to approach the higher-performance song first (13 of 17 birds, interpreted as an aggressive response towards the fast trill, high-performance song; $P = 0.049$). (2) For subjects that approached within 10 m of a speaker, most (18 of 25) first approached the high-performance song, again considered an aggressive response towards the high-performance song ($P = 0.043$ or 0.027). (3) For males approaching within 10 m of either stimulus, the time spent within 10 m did not differ for the lower- and higher-performance stimuli ($P = 0.182$). (4) Time spent in a larger area near the two stimuli also did not differ ($P = 0.583$). (5) In another analysis, however, the '16 males that entered the 10 m fast circle [where the fast-trill song was broadcast] at some point during the trial spent less time there the higher the performance score of their stimulus trill' ($P = 0.020$; page 1910). It is the initial strong, aggressive response towards the high-performance songs (items 1 and 2) and the subsequent reduced time spent within 10 m of higher-performance songs (item 5), in spite of no differences between low- and high-performance songs when assessed simultaneously (items 3 and 4), that leads the authors to the following conclusion: 'the subsequent decrease in aggressive response by the receiver suggests that the highest performance signals posed a threat so extreme that they effectively repelled rivals, even territory owners' (page 1911). The logic used here is challenging to accept. Why would a bird first attack the more intimidating song, then subsequently be scared by it? How do the authors choose when and why and over what time frame a song should have what effect?

Furthermore, if the 11 statistical tests reported in the Results are corrected by some multiple comparisons test (e.g. Bonferroni), and the sample size for the number of independent playbacks is reduced from 31 to 12 (or 19, or 25; see pseudoreplication comment above), none of these tests would be statistically significant, leaving nothing on which to base the overall story (see also above critique of Goodwin & Podos, 2014).

Alternative explanations are also worth considering. Perhaps in items 1 and 2 (above), the birds were actually fleeing the (a) slower, (b) longer, (c) low-performance songs with (d) low trill-note-to-

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

silent-interval ratios when they only appeared to be more aggressive towards the faster, shorter, high-performance songs with high trill-note-to-silent-interval ratios. But, then, to which feature of the songs were the birds responding? How can one be sure when a bird is fleeing a song or just not interested in it, for whatever reason? How normal are those trills in which silent intervals are added or deleted? Trill rates may be normal, but are ratios of trill note durations to silent intervals also normal?

Rather than manipulating songs and adding multiple confounding variables, it would seem that a first, worthwhile experiment would simply compare responses to naturally occurring songs that are of high and low performance. Regardless of song type, on average, if performance matters the birds should respond differently to intact, unmanipulated songs at two extremes of performance.

For all the above reasons, the results and conclusions of this paper are questionable. Its results are not questioned, however, by the community of biologists who cite it so frequently in the literature, 80 times as of April 2016.

Vehrencamp, Yantachka, Hall, and De Kort (2013): Trill performance components vary with age, season, and motivation in the banded wren

From the opening sentence of the Abstract (emphases mine): 'Acoustic displays with *difficult-to-execute* sounds are often subject to *strong sexual selection* because *performance levels* are related to the sender's *condition* or genetic *quality*' (page 409).

This study provides excellent descriptive statistics for how songs change 'with age, breeding stage, and motivation related to social context' (page 409), but the opening sentence squarely places the context and rationale for this study in the realm of performance studies and sexual selection and honest signalling, with 'difficult-to-execute' sounds revealing male quality. Everything is interpreted in this context, yet there is no obvious scientific justification for doing so and good justification for not doing so (see especially the above review of Illes et al. (2006) on the same species). According to the scatterplot of trill rate and bandwidth for banded wrens (Fig. 11), relatively few songs are difficult to execute as defined in this performance context, because most songs fall far from the upper bound on the graph. Every male 'willingly' learns many 'low-performance', easy-to-execute songs in order to have particular song types in his repertoire, *as if performance did not matter*, as if there were no selection for difficult-to-execute songs as claimed in this paper.

Dark-Eyed Juncos

*Cardoso, Atwell, Ketterson, and Price (2009): Song types, song performance, and the use of repertoires in dark-eyed juncos (*Junco hyemalis*)*

In contrast to my remarks on all of the above papers, I applaud the conclusions of this paper by Cardoso et al.

> *We found low but significant correlations of performance measures among the song types of individual males. This contrasts with highly consistent differences in performance among song types, regardless of which males sing them*
>
> **(Cardoso et al., 2009, page 901)**

> *The main conclusion from our results is that, because most of the variation in performance depends on the song type, a receiver that compares a few song types from different males is likely to obtain little information about performance differences between males*
>
> **(Cardoso et al., 2009, page 905)**

Here is the analysis for which I have been yearning, and the conclusion is much the same as the one I came to when looking at my analyses of chipping sparrows and swamp sparrows (Figs. 1–8), and the figures in Illes et al. (2006), Liu and Kroodsma (2006), and Kagawa and Soma (2013). What matters most to these singing males is to have a song type like other birds in the population, and the relative performance abilities in singing that particular song type are almost certainly irrelevant.

One cause for concern is how song types were identified: 'We assigned syllable types by visual inspection of spectrograms, based on the shape of elements within syllables' (Cardoso et al., 2007, page 1052). That is normal procedure, yet no figure illustrates the degree of similarity within and among song types, as is routinely displayed for other species (e.g. Figs. 1, 2 and 7 in this paper). For a species that is known more for improvising than imitating songs (Marler, Kreith, & Tamura, 1962), it would be reassuring to know that Cardoso et al.'s song type categories are based on categories that the birds themselves establish by imitating the details of their songs from each other. Without that reassurance, one must consider the possibility that songs with similar trill rates and bandwidths are grouped into song types, thus artificially creating song type exemplars with similar performance values.

Cardoso, Atwell, Hu, Ketterson, and Price (2012): No correlation between three selected trade-offs in birdsong performance and male quality for a species with song repertoires

Here is the same message, that performance of songs as plotted on the graph of trill rate and frequency bandwidth has little predictive value.

> *most variation in performance is found among song types rather than among males … song performance [does] not allow a good assessment of male quality in juncos, and perhaps more generally in species with song repertoires*
>
> **(Cardoso et al., 2012, page 584)** (and I would add any species without repertoires as well).

The overall work of Cardoso et al. has been criticized (Zollinger, Podos, Nemeth, Goller, & Brumm, 2012) because of how frequency bandwidths were measured (manually from sonograms). In a wide-ranging critique, Cardoso et al. are instructed on (1) proper measurements and methodology, (2) interpretation of data, (3) validity of results, (4) experimental rigour, (5) alternative explanations and hypotheses for data, (6) the ability to reject hypotheses, (7) appropriate use of scepticism, (8) problems in published papers that 'undermine the validity of the results reported and the conclusions reached' and (9) 'basic principles' of science. These five authors are concerned, more broadly, with (10) how papers failing to use 'established methodologies will have a profound adverse effect on the way the research field is viewed by the rest of the scientific community' (page e8). No one would disagree with these prescriptions for good science and the consequences of bad science (unless the 'established methodologies' are flawed; see below), but it is the overall performance literature, not Cardoso et al., that violates all of these accepted norms for good science. Ironically, the second author of this prescriptive diet is a primary contributor to and promotor of the very performance literature that I critique.

What I find surprising is that Cardoso et al. (2009, 2012) have offered a new and, in my opinion, correct interpretation of the trill rate/bandwidth scatterplot, but that contribution to science has gone unrecognized. Instead, these authors have been beaten down by a technicality, on how frequency bandwidth was measured, although the measurement advocated by Zollinger et al. (2012)

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

yields highly spurious results (Fig. 9). For the 15 citations of Cardoso et al. (2009) listed in Web of Science by authors other than Cardoso himself (April 2016), for example, no one mentions that Cardoso et al. have offered a fundamentally different interpretation for the significance of the scatterplot.

DISCUSSION

Word Choice and Inherent Bias

One root of the problem in these performance studies lies in the very words used to state the hypothesis. The word 'perform' is 'used to describe how effective or successful someone or something is' (Merriam-Webster online dictionary). Words like 'performance' and 'proficiency' are thus non-neutral, loaded terms with the implicit assumption that where a song plots on a graph tells *how well* a male sings, or how *effectively* or *successfully* or *proficiently* he sings, and therefore how *good* a male he is. With repeated use of the term 'performance', the concept is no longer a hypothesis to be tested but instead a proven fact, or an assumption so hidden that it is accepted as fact (Prum, 2010, 2012). Functional, non-neutral terms like 'performance' inevitably and unconsciously block alternative views from being entertained, as they implicitly define the universe of discourse. As a result, 'Our job as scientists … to discover truths about the world' (Simmons et al., 2011, page 1359) is severely hampered.

As Marler and Hamilton wrote a half century ago (bold emphasis mine),

> *The process of description is intimately involved with naming, and here too a degree of discipline is called for. Studies of communicatory behavior in animals have often included in their primary descriptions such terms as domination and subordination behavior, inferiority and superiority postures, intimidation, distraction, threat, and appeasement displays* [and 'performance', I might add]. *These terms are liable to prejudge the function of behavior …* ***clear separation of description from function is desirable … There should be a maximum reliance on intrinsic properties of the behavior and a minimum of interpretation***
>
> **(Marler & Hamilton, 1966, page 716)**

I have felt bound to use the same terminology in this review that is used throughout this literature, although I flinch every time I write the word 'performance', because the very attempt to address the problems is already half-defeated by the use of such a loaded word. Substituting neutral descriptive terms for functional terms can be a mind-expanding experience. Consider, for example, the terms 'low performance' and 'high performance'. The intellectual landscape is released from single-minded explanations by merely labelling these songs 'high-deviation' trills and 'low-deviation' trills (or some such descriptive terms). With the descriptive terms no longer rooted in terms that focus on only one functional interpretation, one can more comfortably acknowledge a null hypothesis and alternative hypotheses, and do one's best to falsify them in turn.

Going Forward

What information listeners extract about singers from their songs (beyond species identification) is an exciting area of research. We await good information on this topic.

Acknowledgments

I thank Pavel Linhart, Gonçalo Cardoso, David Lahti, and especially Becky Cramer for responding to an early draft of my document that was circulated to all target authors during December 2014. Others who have provided helpful advice were Çağlar Akçay, Mike Beecher, Walter Berry, Ted Miller, Gene Morton, and especially Sylvia Halkin. I also thank editor Susan Foster and the four anonymous referees for the journal who were extraordinarily conscientious and helpful.

References

Akçay, Ç., & Beecher, M. D. (2015). Team of rivals in chipping sparrows? A comment on Goodwin & Podos. *Biology Letters, 11*, 20141043.
Ballentine, B. (2009). The ability to perform physically challenging songs predicts age and size in male swamp sparrows, *Melospiza georgiana*. *Animal Behaviour, 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.
Burghardt, G. M., Bartmess-Levasseur, J. N., Browning, S. A., Morrison, K. E., Stec, C. L., Zachau, C. E., et al. (2012). Perspectives: Minimizing observer bias in behavioral studies: A review and recommendations. *Ethology, 118*, 511–517.
Byers, B. E., & Kroodsma, D. E. (2009). Female mate choice and songbird song repertoires. *Animal Behaviour, 77*, 13–22.
Cardoso, G. C., Atwell, J. W., Hu, Y., Ketterson, E. D., & Price, T. D. (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, G. C., Atwell, J. W., Ketterson, E. D., & Price, T. D. (2007). Inferring performance in the songs of dark-eyed juncos (*Junco hyemalis*). *Behavioral Ecology, 18*, 1051–1057.
Cardoso, G. C., Atwell, J. W., Ketterson, E. D., & Price, T. D. (2009). Song types, song performance, and the use of repertoires in dark-eyed juncos (*Junco hyemalis*). *Behavioral Ecology, 20*, 901–907.
Cramer, E. R. A. (2013). Physically challenging song traits, male quality, and reproductive success in house wrens. *PLoS One, 8*(3), e59208. http://dx.doi.org/10.1371/journal.pone.0059208.
Cramer, E. R. A., Hall, M. L., De Kort, S. R., Lovette, I. J., & Vehrencamp, S. L. (2011). Infrequent extra-pair paternity in the banded wren, a synchronously breeding tropical passerine. *Condor, 113*, 637–645.
Cramer, E. R. A., & Price, J. J. (2007). Red-winged blackbirds *Agelaius phoeniceus* respond differently to song types with different performance levels. *Journal of Avian Biology, 38*, 122–127.
De Kort, S. R., Eldermire, E. R. B., Cramer, E. R. A., & Vehrencamp, S. L. (2009). The deterrent effect of bird song in territory defense. *Behavioral Ecology, 20*, 200–206.
Dubois, A. L., Nowicki, S., & Searcy, W. A. (2009). Swamp sparrows modulate vocal performance in an aggressive context. *Biology Letters, 5*, 163–165.
Dubois, A. L., Nowicki, S., & Searcy, W. A. (2011). Discrimination of vocal performance by male swamp sparrows. *Behavioral Ecology and Sociobiology, 65*, 717–726.
Goodwin, S. E., & Podos, J. (2014). Team of rivals: Alliance formation in territorial songbirds is predicted by vocal signal structure. *Biology Letters, 10*, 20131083.
Goodwin, S. E., & Podos, J. (2015). Reply to Akçay & Beecher: Yes, team of rivals in chipping sparrows. *Biology Letters, 11*, 20150319.
Illes, A. E., Hall, M. L., & Vehrencamp, S. L. (2006). Vocal performance influences male receiver response in the banded wren. *Proceedings of the Royal Society B: Biological Sciences, 273*(1596), 1907–1912.
Juola, F. A., & Searcy, W. A. (2011). Vocalizations reveal body condition and are associated with visual traits in great frigatebirds (*Fregata minor*). *Behavioral Ecology and Sociobiology, 65*, 2297–2303.
Kagawa, H., & Soma, M. (2013). Song performance and elaboration as potential indicators of male quality in Java sparrows. *Behavioural Processes, 99*, 138–144.
Kahneman, D. (2013). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
Kroodsma, D. E. (2005). *The singing life of birds. The art and science of listening to birdsong*. Boston, MA: Houghton–Mifflin.
Lachlan, R. F., Anderson, R. C., Peters, S., Searcy, W. A., & Nowicki, S. (2014). Typical versions of learned swamp sparrow song types are more effective signals than are less typical versions. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20140252.
Lahti, D. C., Moseley, D. L., & Podos, J. (2011). A tradeoff between performance and accuracy in bird song learning. *Ethology, 117*, 802–811.
Liu, W.-C. (2004). The effect of neighbours and females on dawn and daytime singing behaviours by male chipping sparrows. *Animal Behaviour, 68*, 39–44.
Liu, W.-C., & Kroodsma, D. E. (1999). Song development by field sparrows (*Spizella pusilla*) and chipping sparrows (*Spizella passerina*). *Animal Behaviour, 57*, 1275–1286.
Liu, W.-C., & Kroodsma, D. E. (2006). Song learning by chipping sparrows: When, where, and from whom. *Condor, 108*, 509–517.
Marler, P., & Hamilton, W. J., III (1966). *Mechanisms of animal behavior*. New York, NY: J. Wiley.
Marler, P., Kreith, M., & Tamura, M. (1962). Song development in hand-raised Oregon juncos. *Auk, 79*, 12–30.

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015

Molles, L. E., & Vehrencamp, S. L. (1999). Repertoire size, repertoire overlap, and singing modes in the banded wren (*Thryothorus pleurostictus*). *Auk, 116*, 677–689.

Morton, E. S., Gish, S. L., & Van Der Voort, M. (1986). On the learning of degraded and undegraded songs in the Carolina wren. *Animal Behaviour, 34*, 815–820.

Moseley, D. L., Lahti, D. C., & Podos, J. (2013). Responses to song playback vary with the vocal performance of both signal senders and receivers. *Proceedings of the Royal Society B: Biological Sciences, 280*(1768), 20131401.

Nowicki, S., Hasselquist, D., Bensch, S., & Peters, S. (2000). Nestling growth and song repertoire sire in great reed warblers: Evidence for song learning as an indicator mechanism in mate choice. *Proceedings of the Royal Society B: Biological Sciences, 267*, 2419–2424.

Podos, J. (1996). Motor constraints on vocal development in a songbird. *Animal Behaviour, 51*, 1061–1070.

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

Podos, J., Lahti, D. C., & Moseley, D. L. (2009). Vocal performance and sensorimotor learning in songbirds. In M. Naguib, K. Zuberbühler, N. Clayton, & V. Janik (Eds.), *Advances in the study of behavior* (Vol. 40, pp. 159–195). San Diego, CA: Elsevier Academic Press.

Podos, J., Peters, S., & Nowicki, S. (2004). Calibration of song learning targets during vocal ontogeny in swamp sparrows, *Melospiza georgiana*. *Animal Behaviour, 68*, 929–940.

Prum, R. O. (2010). The Lande–Kirkpatrick mechanism is the null model of evolution by intersexual selection: Implications for meaning, honesty, and design in intersexual signals. *Evolution, 64*, 3085–3100.

Prum, R. O. (2012). Aesthetic evolution by mate choice: Darwin's really dangerous idea. *Philosophical Transactions of the Royal B: Biological Sciences, 367*, 2253–2265.

Simmons, L. W. (2014). 25 years of behavioral ecology. *Behavioral Ecology, 25*, 1–3.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Snijders, L., Eijk, J. V. D., Rooij, E. P. V., Goede, P. D., Oers, K. V., & Naguib, M. (2015). Song trait similarity in great tits varies with social structure. *PLoS One, 10*(2), e0116881. http://dx.doi.org/10.1371/journalpone0116881.

Sprau, P., Roth, T., Amrhein, V., & Naguib, M. (2013). The predictive value of trill performance in a large repertoire songbird, the nightingale *Luscinia megarhynchos*. *Journal of Avian Biology, 44*, 567–574.

Vehrencamp, S. L., Yantachka, J., Hall, M. L., & De Kort, S. R. (2013). Trill performance components vary with age, season, and motivation in the banded wren. *Behavioral Ecology and Sociobiology, 67*, 409–419.

Wilson, D. R., Bitton, P. P., Podos, J., & Mennill, D. J. (2014). Uneven sampling and the analysis of vocal performance constraints. *American Naturalist, 183*, 214–228.

Zollinger, S. A., Podos, J., Nemeth, E., Goller, F., & Brumm, H. (2012). On the relationship between, and measurement of, amplitude and frequency in birdsong. *Animal Behaviour, 84*(4), e1–e9.

Please cite this article in press as: Kroodsma, D., Birdsong performance studies: a contrary view, Animal Behaviour (2016), http://dx.doi.org/10.1016/j.anbehav.2016.06.015