Contents

31 March 3016

Dear editor of Animal Behavior:

Thank you very much for the effort that you and the four reviewers have provided for my manuscript. I am deeply grateful. I welcome any challenge to anything I am thinking or writing, in either substance or style. I have skimmed through the reviews now, and will go through each comment shortly.

Before I do that, however, I would like to state here that I am providing some background for this manuscript in another file, entitled "Kroodsma Correspondence on Performance Studies."

There is so much behind my manuscript that neither editors nor reviewers have seen, and for "legal reasons" I have kept track of every communication that has occurred on this topic. It is my understanding that this document will become part of the Animal Behavior archives. Some time in the future, animal behavior historians will look back on this period and ask "What were they thinking about birdsong and sexual selection? Why were they thinking *that*??" I would like my document to be an aid to their understanding at that point, some time in the future, whenever these kinds of documents become freely accessible to anyone who asks.

Response to editor

From the editor:

If there are any attachments uploaded by the reviewers, you may view these by logging in as an author and checking the folder named "Submissions Needing Revision" and in the Action column select "View Reviewer Attachments".

PLEASE NOTE: The journal would like to enrich online articles by visualising and providing geographical details described in Animal Behaviour articles. For this purpose, corresponding KML (GoogleMaps) files can be uploaded in our online submission system. Submitted KML files will be published with your online article on ScienceDirect. Elsevier will generate maps from the KML files and include them in the online article.

Animal Behaviour features the Interactive Map Viewer, http://www.elsevier.com/googlemaps. Interactive Maps visualize geospatial data provided by the author in a GoogleMap. To include one with your article, please submit a .kml or .kmz file and test it online at http://elsevier-apps.sciverse.com/GoogleMaps/verification before uploading it with your submission.

> **KROODSMA**: I do not believe that I have any materials to enrich the manuscript in the above fashion.

Date Revision Due Apr 28, 2016

Dear Don,

The reviews on your manuscript, "Birdsong Performance Studies: A Contrary View" are now in and are appended below. Although the reviewers and I found the paper interesting, in my judgment the paper is not ready for publication in its present form and I must consequently reject it. However, I strongly encourage you to consider a resubmission of this manuscript if you feel you can satisfactorily address the reviewers' concerns.

> **KROODSMA:** I will do my best to revise. I am going to take the liberty of speaking frankly in this document, not mincing words as I would have to if these words were going to be public.

There are two major issues. The first concerns the issue of confirmation bias - what it takes to prove it, and whether you really want to attempt that. I think it is subsidiary to and distracting from the major issues, which are the methodological, empirical and interpretative flaws in studies claiming to support the hypothesis. These problems may come to exist because of confirmation bias (or for some other reason), but it is these problems, not confirmation bias, that should be discussed in a research critique in Animal Behaviour (as opposed to a more general analysis submitted, say, to a journal that focuses on issues in philosophy of science).

KROODSMA: Ok; more below.

The second issue concerns the tone of the manuscript. I know you have tried to 'tone it down', but the residues of the much stronger original ms remain, and they defeat your fundamental goal, which is for researchers to look at studies on the performance hypothesis with a more critical eye and to consider plausible competing hypotheses.

KROODSMA: Ok**;** more below

Concerning confirmation bias, Rev. 1 notes that your paper's aim of demonstrating "how confirmation bias taints the literature on birdsong and sexual selection … needs more explication if it's to remain in the ms, which only asserts confirmation bias, and doesn't demonstrate it". I believe all the reviewers concur on this point.

I would go further and remove reference to confirmation bias, which imputes a particular psychological process to the researchers, which you may believe but which you can't really prove, and focus instead on building your case that the evidence for the performance hypothesis is incomplete, that the studies purporting to demonstrate it are flawed, and that there is an alternative hypothesis (which they ignore) that is as at least as convincing as their favored hypothesis, if not more so. It is true that Type 1 errors (false positives) are likely to occur in a research program that is plagued by confirmation bias, and this may have happened here, but it is these incorrect conclusions that are of interest to scientists, not the putative psychological process that may have produced them. (As I suggested above, discussions of confirmation bias are better suited for sources focusing on philosophy science.) As Rev. 1 implies, one can only assert confirmation bias, whereas one can actually demonstrate flaws in research design, misreadings of the data and failure to consider competing hypotheses or attempt strong inference tests. Furthermore, most of us realize that all scientists fall victim to confirmation bias, to one degree or another. Let he who is without conformation bias cast the first stone. Don, perhaps you are totally clean in this regard, but I would not make that claim for myself, nor for most of my scientific colleagues in science. I strive to avoid it, and expect my colleagues to do the same, but as many psychologists have shown (e.g., Wason), confirmation bias is a natural human tendency, and although scientists are supposed to be Feynman-like and above it, the reward structure of science (granting agencies, journals, tenure review committees, etc.) inevitably encourages it. What keeps the science honest, in the long run, is the natural dynamic where OTHER scientists, who are not enamored of YOUR favored hypothesis, try to disprove it.

This is what you are trying to do here, but you hurt your cause to the extent that you focus on proving the working of this putative psychological process rather than on simply trying to

disprove the actual hypothesis, which you do by pointing out the flaws in the 'supporting' research and by establishing the plausibility of and evidence for a competing hypothesis.

> **KROODSMA:** Regarding "tainted by confirmation bias." I am guilty of many things, one of which is accepting the words of a trusted sage in the field of birdsong, one who knows all of the players and motivations and consequences. When he suggested that I put in the words "tainted by confirmation bias," I queried him, saying those were pretty strong words from him, but he was comfortable with those words. I thought they were pretty strong, but I added them to the manuscript. Now I being asked to remove them. I can easily do that. I will adjust the wording there, with the fear that what I write in substitution may be even more offensive or counterproductive.
>
> I think that the editor has inadvertently added a new phrase to the literature, or at least one that I am unfamiliar with: "conformation bias." I like that. I think there are really two issues involved in my essay. One of them is the need to conform to the already published material on a topic, especially if major players in the field are promoting that idea, because that almost certainly guarantees publication. It is the need to conform that creates the kind of literature that I am critiquing.

All four reviewers identify the second issue, the tone of the manuscript, and point out that it undermines your fundamental goal of encouraging researchers to examine supposed demonstrations of the performance hypothesis more critically, and to turn their attention to plausible alternative hypotheses.

Here are the reviewers' most salient comments on tone:

Rev 1: "My last general comment has to do with the tone of the manuscript. There are several places in it where a point was amplified to an extent that struck me as hyperbolic. The manuscript would be stronger if it were more understated at those points, which I've also detailed [in many specific suggestions] below".

Rev 2 [from the comments to the editor]: "My only real criticism of the commentary is stylistic. The core arguments are strong, but the tone in some places is too harsh, to the point that it might detract from the message. Instead of coming off in a curmudgeon-sounding diatribe, the work is a constructive (and instructive) critique and can be written more as such. Kroodsma can take the high road here both in terms of the argument, and how he presents the argument, and by doing so I think will make the pitch stronger to a wider audience".

Rev 3: "The paper makes a number of other useful points that may stimulate future research on vocal signals. However its hypercritical nature (suggesting that the criticized authors are doing bad science) may be unproductive".

KROODSMA: This reviewer understates my point. Some of these authors are not doing science at all, in my opinion, but are instead doing what Feynman calls "science that isn't science. But I had better stop there; details are in the supplemental file.

Rev 5*: "I suspect that a number of things brought up in this paper are correct. I also suspect that a dispassionate and well-designed research program would show that the motor constraint hypothesis is at best a weak predictor of sexual selection in birds (despite its popularity). A number of ideas have floated around in the behavior literature over the years that have not been particularly useful. A vetting of these ideas is a terrific endeavor. The motor constraints hypothesis may (or may not) be on this list. However, a proper vetting of these ideas requires a level-headed and robust discussion of why the ideas are inadequate. From my perspective, this paper falls short on both fronts. It seemed quite odd to me and not an auspicious beginning in the paper to be told that 'I offer apologies to those who feel my approach too frank, or too blunt, or overkill'. After reading the paper, I agree that the apology is warranted. Rhetoric of this level really does not belong in Animal Behaviour, in part because it correctly implies that the paper will be anything but a robust, level-headed discussion of the facts".

*Note: there was no Reviewer 4.

Reviewer 5 is the most blunt about this criticism, but all the reviewers are making essentially the same point: that the tone is counterproductive, that the critique needs to be a strictly sober, measured account of the facts, not a seemingly personal attack on the motives of these researchers (Podos in particular). Otherwise your critique will be viewed skeptically, as a personal attack. Stick to just the facts, do not shoot your argument in the foot.

KROODSMA: I am happy to consider any comment about any offensive word that I have used in the manuscript. That said, I want it absolutely clear here (in this response to the reviewers; I will not make that claim in the manuscript) that I believe there is strong scientific and ethical misconduct in what has transpired in this literature. And, I again have to stop there, but all the details are in the file available to the editors.

When you revise your paper, please prepare a detailed explanation of how you have dealt with all of the reviewers' and Editor's comments. Refer to the Guide for Authors (on the main menu of the Elsevier Editorial System at http://ees.elsevier.com/anbeh) for details of our house style and for a list of file types that are acceptable for revised papers. Log in to the Elsevier Editorial System as an Author to submit your response to the comments and your revised paper. Changes in the revised paper should be highlighted; you may either use the Track Changes or Highlight Tools in MS Word, or underline your changes. Please submit both the highlighted version and the non-highlighted version of the revised paper.

KROODSMA: I have used track changes.

As you revise your manuscript, please note that the journal's guidelines require that you address any animal welfare issues arising from your study either in the main text of the Methods section or in a separate subsection of the Methods headed Ethical Note. Even if your study involves only invertebrates, please address all ethical implications of the experimental design and procedures, including any procedures taken to minimize adverse impacts on the welfare of subjects or to enhance their welfare. For a specific list of potential topics see: http://www.elsevier.com/framework_products/promis_misc/ethyanbe.doc. You must also include IACUC (USA) or UAREB (Canada) protocol numbers, approving institution, and any other relevant details about the approval for the work. For further information on what ethical information to include, please consult the "Animal Welfare" and "Methods" sections of the journal's "Guide for Authors" and "A Guide to Ethical Information Required for Animal Behaviour Papers" (http://www.elsevier.com/framework_products/promis_misc/ethyanbe.doc).

> **KROODSMA:** I don't think that any of these animal welfare issues are relevant for what I have done, which is to review the literature, and to record some birds, leaving no trace, disturbing no birds, etc.

We would like to receive the revised paper within 30 days. If you think you will be unable to revise your manuscript in that time, please let the Journal Manager know (yanbe@elsevier.com).

Thank you for sending this interesting work to Animal Behaviour. I hope you find the enclosed reviews helpful and look forward to seeing a revision.

All the best,

Editor of Animal Behavior

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~

Response to Reviewer #1

Reviewer #1: This manuscript sets out to do two things. First, it tries to show that there's no evidence for the performance hypothesis — in particular, that some of the main evidence purported to support it is better explained by song learning aimed at acquiring a "normal" song. The manuscript achieves this aim to some degree, although the case could be made more convincing on several points, which I've detailed below.

Still on the first aim, I also wonder whether this ms is enough of an advance beyond what's already been accomplished by Cardoso's nice work and Akçay and Beecher's critique. I guess it would be more visible than those previous papers, and it takes aim at more studies on the performance hypothesis, but its (repeated) core arguments are few (and arguable; see below), so I'm not sure.

KROODSMA: I agree that Cardoso has done nice work, but as I point out, he seems to be disregarded. Zollinger et al. dismissed his work in a lengthy list of how one does good science, implying that Cardoso does not do good science. Cardoso et al. also worked on juncos, which do not imitate songs like the favored subjects of this research do (chipping sparrows, swamp sparrows), so that claims of Cardoso are a little bit of a stretch (now addressed in the manuscript under Cardoso's work).

Akçay and Beecher's miss the main fatal flaws of Goodwin and Podos (2014), as I detail in my manuscript.

At any rate, I think the flaws of the performance hypothesis need to be spelled out in excruciating detail, with clarity in the figures, so that it is plainly obvious to anyone who considers this topic. My goal is to make it so obvious that no one in future sexual selection studies would make the same mistakes.

Also, although I agree strongly with the conclusions of Cardoso et al., it's not 100% how they got there. According to a 1962 study (Marler, P., M. Kreith, and M. Tamura. 1962. Song development in hand-raised Oregon juncos. Auk. 79:12-30.), juncos don't imitate each other very closely, which results in a lot of song variation within a population, making it perhaps difficult to assign songs of a given male to a "song type." Yet Cardoso in 2007 merely say "We assigned syllable types by visual inspection of spectrograms, based on the shape of elements within syllables" (p. 1052). I can find no illustrations of the kind of variation that Cardoso assigned to song types. Without any offering to readers of a figure showing how they generated these song type groupings, there remain doubts about their data. In contrast, I can offer figures for the chipping and swamps to show how there are definite song types, and the males learn them in the laboratory, etc.

The ms's second aim is grander: "to demonstrate how confirmation bias taints the literature on birdsong and sexual selection" (l. 54-55). I think this aim needs more explication if it's to remain in the ms, which only asserts confirmation bias, and doesn't demonstrate it. True, later papers may not have mentioned or questioned the arguable points of previous papers (including those that the present ms argues were dead wrong), but is there a causal connection between this apparent insouciance and the (arguably) poor experimental designs, biased observations, etc of those later workers? I think readers need stronger evidence for that claim; the ms's criticisms of previous papers do not quite establish that (specific examples given below).

KROODSMA: To me, "confirmation bias" simply means that, consciously or not, the work is influenced by failure to consider alternatives; the words do not imply intention, at least to me. The fact that no alternatives are considered to explain this kind of research is, for me, by very definition, confirmation bias. If one didn't already believe the previous studies, then one would have a more open mind and not make the same mistakes. So, in my mind, by definition, all the work that I critique is "confirmation bias." I believe that

was also the opinion of the sage whom I consulted for the original wording, but there's no harm in removing the words if they are a stumbling block to understanding the issues.

OK, I omitted the entire last paragraph of the Introduction and substituted another.

My last general comment has to do with the tone of the manuscript. There's several places in it where a point was amplified to an extent that struck me as hyperbolic. The manuscript would be stronger if it were more understated at those points, which I've also detailed below.

KROODSMA: "Several" is less than the million that were there before. I'm making progress. I'm happy to considering changing any offensive wording that remains.

On to the specific comments that I've summarized above:

l. 48 "simply biologically implausible, if not impossible" seems a bit hyperbolic. Surely the hypothesis could work in theory (and thus is plausible). Rather, your point is that it flies in the face of the facts.

KROODSMA: I thought I'd look up definitions. Hyperbole: "unnecessary, more than necessary exaggeration." Is my statement really hyperbole? I don't get it. It's a statement of fact. It's highly implausible given how performance scores of males are distributed, next to impossible given that distribution. Maybe this comment comes early in the reviewer's reading of the ms, and that later this comment might not have been made, given the nature of the data that I reveal. I have added some key words here that perhaps spell out why I think it implausible or impossible: "furthermore, given how song performance measures are distributed among song types and among males, the hypothesis becomes biologically implausible, if not impossible"

Full Definition of *hyperbole*

: extravagant exaggeration (as “mile-high ice-cream cones”)

Simple Definition of *extravagant*

: more than is usual, necessary, or proper

l. 64 What sort of social bond? I ask partly for general clarity, but also because of the point I raise below re whether youngsters actively prefer particular tutors (l. 98-99).

KROODSMA: I can change this to "social and aggressive interactions."

l. 65 The word "unequivocal" conveys the very kind of certitude that's being challenged here, so I'd replace it with "strong" or, perhaps better, just leave it out and make the factual case (which is plenty strong already).

KROODSMA: I believe that anyone who looks at the facts will accept that the data are unequivocal. It is facts that I think are so important, and when one goes out and simply observes what marked birds do, one comes up with facts that have no explanation other than that youngsters learn the songs of the adults next to them. This is about as unequivocal as it gets. It's not a hypothesis, not a guess, not something made up, but absolutely unequivocal. It is this kind of fact that is glibly ignored in Goodwin and Podos (2014), for example. I'm going to leave the word in, as I think "unequivocal" is rock solid. I welcome anyone to read the paper by Liu and come to another conclusion. . . . So I changed the word to "solid."

l. 94-95 Rather than asserting your belief that the methods match previous studies, it would be more convincing to state (or show) that they do (well, at least adequately for the purpose at hand).

KROODSMA: That is exactly what I have done in the previous sentences, I think. I have laid out exactly how I have measured the songs, and I believe that those methods match previous studies. I think that's all that is important here.

l. 98-99 Why couldn't they do both? E.g., a young male might choose to learn from the tutor who sings the highest quality song that the youngster can manage. Is there any such bias in tutor choice (you seem to have the data to test it)?

KROODSMA: In lines 100-102, I address this issue: "and there is no evidence for song learning in any songbird species or especially in chipping sparrows (Liu and Kroodsma, 1999, 2006) that a male is in any way limited in what naturally occurring trill rate he can learn." Yes, it would be very nice if a young male knew at the outside how "good" he was, then found a male with the appropriate trill rate, somewhere between 7 and 30, that matched his own prowess, but there's no evidence anywhere that chipping sparrows or any other species are limited in learning any naturally occurring trill rate. I've changed the text here, taken the thought out of parentheses:

NEW TEXT: 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 male whose song he can master, thus choosing an appropriate tutor male with a trill rate somewhere between seven and 25 (as in Figure 2), but there is no evidence for song learning in any songbird species or especially in chipping sparrows (Liu and Kroodsma, 1999, 2006) that a male is in any way limited in what naturally occurring trill rate he can learn.

Also, at this point, some readers might counter that trill rate might nevertheless signal male quality, if the birds took song type into account in their assessment. This possibility is discussed (and rejected) farther along in the paper, but it might be good to reference that discussion here. I myself found that "yeah, but …" in my head to be distracting — until I was finally relieved to find it covered later on.

KROODSMA: see above.

l. 140, 174 I would remove "ornithological" from both "basic ornithological data" and "basic ornithological facts". Basic ornithological data are things like weight, clutch size, etc, but here we're into more specialized data, so the phrase sounds a bit hyperbolic, at least on first reading.

KROODSMA: I sense a palpable disdain for *basic ornithological facts* that get in the way of favored explanations. Which males sang which songs are, in my opinion, basic ornithological facts in this business. When the focus in lab meetings (professor + graduate students) is on finding the hook that will capture the most interest among readers, regardless of whether the interpretation fits "basic ornithological facts," I think there's a severe problem. I like the word "ornithological" because there is a science of "ornithology" that is seemingly eschewed, belittled, as a second-rate endeavor. And, perhaps I should point out that I am emeritus in the same department where Podos has a position. I know how these things work. Check to see how many publications Podos has in "basic ornithological journals." Just one, in Emu, where Podos is a distant author and someone else wrote the paper. So, for better or worse, I'm going to leave the words in, as they make a point that I would like to make, though the point will be lost on most readers.

l. 156-157 Here's the mention of social bonds again. Sorry if I've missed it, but it'd be good to spell out what's meant by that a bit more, perhaps around l. 64 where the phrase first occurs.

KROODSMA: Yes, I can spell that out. It's "social interactions," with the young male and the adult engaged in ferocious back-and-forth aggressive interactions, apparently fighting over territorial space.

Figure 5. Conveys the point very nicely, especially following so closely on the heels of the plot in Figure 4.

KROODSMA: Thank you. I think the figures are worth more than 1000 words. Those figures could have easily been drawn with the data various researchers had in hand. All these kinds of data were available for swamp sparrows, for example, but the authors chose not to publish this kind of "basic, descriptive, ornithological" information. Mere description is a second rate endeavor. Also, for banded wrens, the authors had all of these kinds of data in hand, could have graphed them in a few minutes time, but instead said in the manuscript that it was not yet known whether individual males were consistent in their song performance.

l. 175-177 See comment re l. 98-99, above.

KROODSMA: I think I have addressed it there.

l. 177-180 Why? The pre-dawn gatherings are lek-like, and there's many features of leks that might well be cooperative (albeit not egalitarian). Also, Goodwin and Podos didn't "simply assume" that the gatherings are cooperative; they inferred it from what they observed (however prematurely and, probably, wrongly).

> **KROODSMA:** I can reword this so that I think it will be clearer. See revised text.

l. 197-199 I would remove this. It ends one of the most compelling paragraphs in the manuscript with a pretty obvious statement about science in general. Since we should all know and ascribe to that (albeit with varying degrees of earnestness, no doubt), less sympathetic readers will take it as patronizing — which won't help the paper's cause.

> **KROODSMA:** Here is something I wrote back in 2004 to one of these authors: ***In my view, science is the search for truth regardless of how good the story is, whereas "marketing or advertising" is the search for a good story regardless of the truth, or regardless of how good the data are***. . . . Can't spell out the details here, as they're rated "confidential," to the editors only . . . . But, ok, out of respect for this reviewer, I'll take out the part about "truth" and just leave the last part of this statement. Done.

l. 204 ff. True, but this is the catch-22 of most playback studies using manipulated stimuli. One should always worry about it, but often the best way to make any progress is to try it, warts and all, and then do follow-up experiments to test whether the results are artifactual, whether other variables are more important, etc. All to say, do you have another way to do these manipulations in general, or are you leading up to your main point about Goodwin and Podos in particular on l. 217-218? If it's the latter, I would emphasize the fact that the manipulations were not counterbalanced (see next comment); if the former, it sounds like a counsel of despair, unless you've got another way (play back a range of natural songs with differing trill rates, perhaps? but then, how do we know trill rate is the key stimulus?).

> **KROODSMA:** "Counterbalanced" is a good word; I could try to work that in. All I'm saying here is that if the authors are going to claim that pigs can fly, they had better be honest about what information they have. And if they make a statement about assuring that the trill rates are within the population range and therefore the song stimuli are normal, they are concealing key information that could offer an entirely different, uninteresting interpretation of their study. It is not my task to design their experiment. It is their task to be honest about what they have done, and what they have not done. Not disclosing the severe artificiality of their stimuli is, in my opinion, highly deceptive and dishonest.

l. 217-218 But if they counterbalanced stretched vs sped-up songs, then faster isn't necessarily more or less normal. Your alternative explanation might still hold, but I think it's more that Goodwin and Podos, as far as I can tell, do not tell us whether stretched and sped-up songs are equally represented in the data (or included as a factor in any analyses).

KROODSMA: I don't know exactly how to characterize the problem, and the way I have characterized it above is probably offensive (authors are dishonest). Given all that I have seen in this research, however, I am convinced there is less regard for what actually happens in the real world and more regard for how good a story can be generated. If there were more regard for uncovering truths about the world . . . (again, I need to stop here).

l. 1. 253-257 This point is only slightly relevant, since the performance hypothesis need not apply one standard across the whole species (whether or not that's been implied by previous work); different populations might calibrate to different standards (especially if there's geographical differences in morphology).

KROODSMA: Agreed that morphological differences in birds could make a difference in their songs, if trill rate were an accurate measure of male ability/quality. The median trill rate in Massachusetts is about 12% faster than in Michigan; I am going to leave it to the reader to imagine what kind of morphological differences might be required in a bird to achieve that kind of difference. But then look at the point I make in Figure 8, showing that birds 20 km apart have different performance values. Such differences over 20 km can't be attributed to morphology. I will leave my statement as is, and if someone wants to make a point of it, I will let them. . . . Well, actually, now I've come back here and can say that I've added some text to clarify.

l. 276-278 Again, variation within song types could still be telling (see second comment under l. 98-99, above); the "yeah, but …" in my head got louder again at this point.

KROODSMA: 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?

l. 279-280 No need to say it twice.

KROODSMA: Sorry that this reviewer thinks I'm saying it twice. I am using different words to drive the point home. I think it is crucial that the point be driven home in as explicit a wording as possible.

l. 295 ff. For readers to be convinced that Podos et al 2004 is better explained as a process by which sparrows try to produce normal song, you need to more explicitly describe what Podos et al did, and why that wasn't good enough (bearing in mind my catch-22 comment, l. 204 above; if manipulated songs are abnormal, the design should be counterbalanced — something that Podos

et al 2004, and its predecessor Podos et al 1999, were more careful about, I think, than Goodwin and Podos).

> **KROODSMA:** It's not what they did experimentally, but it's how they mine their data for tidbits that are consistent with their favored explanation, on calibration; as a result, I don't think that telling anything about the experimental design is important. Everything I want the reader to know is here.
>
> I ask for honesty, of the kind of "utter honesty" that Feynman advocates. If a good experiment can't be done, one can't settle for a poor, incomplete experiment and then claim one has done the defining work on the topic. Honesty is simply hard to come by in this literature, and that is my primary point, though I simply can't be explicit about that in the way that it would be healthy to do so.

l. 297-303 Could you summarize the results, and why the authors took them to be consistent with the calibration hypothesis? I appreciate the difficulties in doing so (it's a tough paper to comprehend, at least for me), but all that's said in the present ms is that the authors found something that they said was consistent with calibration.

> **KROODSMA:** In a manuscript approaching 11,000 words, and when the first editor of Animal Behavior wanted something no more than 3000 words, I am reluctant to add to the length of this ms with details that I think are not all that relevant. At some point, I simply have to make a statement, and if someone wants to go see the details, I have to refer them to the manuscript. The details of exactly what the authors chose to report are unimportant, I believe; instead, it's the concealing of an alternative explanation by selective reporting of what they found.

l. 304-307 Sorry, I don't get the impression of certitude from these claims; on the contrary, "consistent with" conveys uncertainty. Moreover, "consistent with" is less a red flag signaling suppressed alternatives, or inconsistent data, so much as a failure to provide convincing evidence. I would revise this paragraph accordingly.

> **KROODSMA:** I copy two quotes from "Thinking, Fast and Slow." p. 87. “It is the consistency of the information that matters for a good story, not its completeness.” p. 81. "A deliberate search for confirming evidence . . . seeks data likely to be compatible with the beliefs they currently hold." . . . No, it's not certitude that the authors proclaim, but by reporting only evidence consistent with their preconceived beliefs, they develop a good story, but not a complete story. They are, in my opinion, dishonest about what their study has shown if they do not also tell what other explanations their data are consistent with.

l. 320-322 Same comment as for l. 197-199.

KROODSMA: My paper is about how to do science, and how not to do science. If I can't make a simple, candid statement like this, and drive home the point of this discussion, I fear that my point will be lost on all too many of the people who would perpetuate this kind of work, perhaps not on performance studies, but on the next craze of sexual selection studies to emerge. Something blunt has to be said that encapsulates all of the ills of these kinds of studies.

l. 332 ff. Why not just say that the work of Ballentine et al. was propagated in later papers, especially since the phrasing ("… in support of their own work …") casts aspersions on the later authors' scientific integrity (as opposed to their uncritical acceptance of previous findings — which is also not good, but could be caused by inattention, haste, etc — rather than self-interest).

KROODSMA: OK. I will delete a couple of words: "support of." I realize that "in their own work" then becomes somewhat redundant, but it pushes the point home a little closer to where I think it needs to be.

l. 362-370 Neither the ease of choosing experimental stimuli nor the number of times a word or its variants are cited in a paper is compelling evidence for biased observers. I would remove this paragraph; it's a critique that could be leveled at any study, unless you can muster better evidence that observer bias was particularly rife in this study.

KROODSMA: I will respectfully disagree. I have now returned to the Discussion what I at one time wrote, but because of space considerations, removed it. The topic is "word choice," and how to describe using neutral rather than functional terms.

l. 376-378 Could you put some numbers on "considerably more reverberation", or at least provide some evidence that the difference would make a difference? The frequencies at which the two sizes of parabola are no longer directional aren't all that different, and, at about 1000 Hz and 750 Hz (ignoring roll-off), are well below the sounds of interest here.

KROODSMA: It's not the roll-off frequency that is important. The larger parabola has twice the area of the smaller one, providing a far more effective physical amplifier. To achieve signals of equal amplitude with the smaller parabola, one has to boost the gain electronically, which significantly boosts the reverberation. I have used the PBR330 and found it practically useless as a sound collecting device. . . . For over a year I have known I needed to go out and test microphones with different song types. I have now done that, and the results are rather remarkable, calling into question every bandwidth measurement ever made for these kinds of studies. See Figure 9 of revised manuscript.

l. 379-392 This is really the response part of the criticism in the previous paragraph; i.e., the end result of any difference in reverberation between the recordings. I would combine the two paragraphs (or at least the two reasons, as one). Also, for the criticism to apply, there'd need to

be a bias in the conditions under which high- vs low-performance songs were recorded. That makes sense for DuBois et al 2009, as you point out l. 435-437, but how do Ballentine et al 2004's recording methods invite such bias? Again, without a bit more here, this is a criticism that could be guessed at for any study.

> **KROODSMA:** I agree. I've adjusted so that all of the reverb is in one paragraph, and then I add another paragraph on another alternative explanation that came to me last night at 3a.m. as I rolled over in bed and couldn't get this stuff out of my mind.

l. 404 Was it really a randomly selected song type? If so, I would rephrase, as that doesn't necessarily follow from the preceding sentence. If not, then I would omit the phrase.

> **KROODSMA:** OK, he "seems to choose a random . . ."

l. 405 I would omit "highly", partly to convey a more measured tone, but partly also because I'm not sure it's wildly inconsistent; not going all-out is a characteristic feature of animal contests, especially those between territorial songbirds.

> **KROODSMA:** easily omitted

l. 425-426 Signals can vary in relation to lots of disparate (yet meaningful) factors, so I don't think the title is misleading, so much as the variation in "aggressive" song use is well within the range of "neutral sessions" (which is something that should be defined, or replaced with a more explicit phrase/term — station singing, or advertising, perhaps?).

> **KROODSMA:** I respectfully disagree. When one conclusion is presented and not balanced with another true statement that songs are also modulated in neutral or nonaggressive situations, that title for a paper is highly misleading. It is the conclusion that readers are going to take from this paper. Presenting that title is, again in my opinion, simply dishonest about the findings of the study. It is not a complete statement, and by its incompleteness it is misleading and dishonest.

l. 441-449 You could turn your reasons on l. 444-449 right around to say that, if all these considerations are up in the air, the best shot one has at judging a male's best performance is that male's "best" song. I suspect that's Ballentine's rationale, and it's not so bad, given that animals often don't go all-out in contests (see comment re l. 405). Is a country forever at war but without nuclear capability more weaponized than a country at peace that has the bomb?

> **KROODSMA:** Yes, I suppose one could. If the authors want to make that defense, I will let them do so. Then I wonder under what conditions the male actually uses that super-song that tells who he is. So far, we have no such information. We continue to stretch and stretch to try to figure out how what has been published might in some way show something conclusive about performance.

l. 460-463 Why assume that a repetitious author is trying to make the statement true? More likely, they're just really convinced they're right. Again, I would rephrase or omit this sentence; it impugns the authors unnecessarily, and thus doesn't help the present manuscript's case.

> **KROODSMA:** It is the effect on the reader that is important here. See my comments above for lines 304-307. I am trying to stress that the author has a responsibility to be honest about the existence of alternative explanations. This is an honesty issue. Without that honesty, published works become marketing devices for favored ideas and for authors themselves, and that's not how one is supposed to do science, current trends to the contrary. I don't know how I can say that authors have this responsibility without "impugning." At some point, one must simply say that what we are doing is wrong.

l. 467 "Shunned" implies active avoidance. Your point is that the key analysis hasn't been done; I would leave the authors' motivations for readers to decide.

> **KROODSMA:** words adjusted. I believe the analyses have been done and not reported, of course. When Illes et al. have all of the data in hand and state that they don't yet know if males are consistent in their performance values, I find it hard to believe that the analysis wasn't actually done and found wanting. But, then, I'm a skeptical, cynical kind of guy. Not a nice person, as it has been claimed. A curmudgeon.

l. 492-497 See my comments re l. 376-378.

> **KROODSMA:** Yes. I've also addressed this issue back there.

l. 498 This is always true, not just "when investigators are deeply committed to an hypothesis". Moreover, you don't know how deeply committed they are; they might well be, but they can deny it so easily that this implication will weaken your case. I would remove it.

> **KROODSMA:** Weakens my case? I suppose so. Yes, "it" is always true. But if the simplest of alternative explanations are not mentioned, in favor of one more complex explanation, there is little doubt that the evidence says the authors are committed to confirming their pet idea. Nevertheless, I will adjust the wording here.

l. 510-519 But the study asked whether receivers discriminated between- or within-male variance in performance (or both), taking the correlation of those with the sender's quality as a starting point (based on previous studies). If males can vary their songs, that suggests the song is a conditional signal rather than an index, but it can only be a signal if receivers are sensitive to that variation, which this paper falsifies. In contrast, this paper confirms that receivers are sensitive to

the individual variation that a male is stuck with, supporting the idea (albeit weakly, in my opinion) that the variation is an indexical signal. If receivers didn't respond to either source of variation, that would falsify the idea that performance honestly conveys a male's quality. All to say, what the authors say is not inconsistent with falsifiability.

> **KROODSMA:** I have struggled with this paragraph, and I appreciate the reviewer reminding me of it. There's something that is infallible about the reasoning of the authors. There will be "honesty in signaling" no matter what the outcome. Maybe I just omit the paragraph . . . or think a little more about how to change it . . . or just admit my exhaustion and delete it, which is what I did, at no loss.

l. 520 This experiment does the very sort of counterbalancing I was calling for above. So, if you're correct about swamp sparrows aiming to produce normal songs, then why don't the sparrows lower the trill rates of the abnormally fast songs that they hear? (Perhaps because the fast songs that were presented were still within the normal range?)

> **KROODSMA:** It's a very nice experiment, and Lahti is the only person whom I critique who is semi-open about what he has done. But he reaches a point where he says he cannot say anything about the dynamics of how this study was reported, as apparently his relationship with his postdoctoral advisor had deteriorated, and he could not afford to speak out , . . . I can't answer the particular question posed here, and overall I think answering it is secondary to other issues. If the authors in the Forum want to raise this issue, it can be dealt with there, I think.

l. 555-558 I think this reading of the passage gives the authors too much credit. I read the phrase as reminding readers that the playback songs were within the normal range of variation, as one might for any stimulus that ones subjects responded positively to, not that the fast songs were responded to strictly because they were normal.

> **KROODSMA:** I'm puzzled here. Does the reviewer agree with me? Is the reviewer saying I am not hard enough on the authors? This must be a typographical error on the part of the reviewer. I'm going to leave the passage as is.
>
> I would to say here, for the record, that this is one of the most deceptively written articles I have read. The introduction sets the reader up to accept the authors' chosen conclusion, in much the same way that I objected to Podos' 2004 paper when I reviewed it in-house.
>
> The only reason I don't come down harder on Lahti et al. is because I believe Lahti was forced into this interpretation by Podos, and I believe Lahti is, by himself, an excellent science with a good future. So I cut Lahti some slack. Maybe I'm getting soft.

l. 559 ff. In contrast to my quibbles with the critiques of the other papers, everything in this critique seems pretty reasonable to me.

KROODSMA: I'm grateful that the reviewer is referring to our differences of opinion as "quibbles." Perhaps I'm wearing him/her down . . . no, I look ahead and see that is not the explanation.

l. 654 ff. I don't think any of these criticisms are different enough from the criticisms of the chipping and song sparrow papers to merit inclusion of the paper in the present ms. #4 is a rhetorical potshot that just sounds angry and doesn't help the present paper's cause.

KROODSMA: My task, as I see it, is to review convincingly as much of the evidence for the performance hypothesis as reasonable. I have omitted many papers, and just refer to them and say that I couldn't find any evidence there either. If the problems here are the same as elsewhere, that helps make my case that there's "conformation bias" running throughout this literature.

As to #4, I will say that, after multiple readings, I remained incredulous. I will nevertheless read the paper again and adjust the wording here accordingly.

l. 664 Unless I misunderstand this passage or Illes et al, this paragraph is quite unfair. Illes et al. present several statistical tests, and there's nothing in their paper that would lead me to think they failed to disclose other statistical tests. I would either remove all this or spell out the evidence more.

KROODSMA: It's possible I'm unfair. I'll restudy . . . Yes, upon closer study, I was all too generous to these authors. I have altered and expanded my discussion significantly to reveal a number of ills.

l. 677-692 This critique, too, misses the mark. Granted, the paper interprets everything in light of the performance hypothesis, which the authors could have been more circumspect about, especially given the problems raised in the present ms. But Vehrencamp et al are examining relative differences in "performance" within and between males, not absolute levels of performance. If the performance hypothesis holds, it should hold wherever birds fall on the performance graph, not just among the best birds in the species' range. A spear doesn't have to be a missile to be more effective than another spear.

KROODSMA: I will look again at this paper . . . but I don't like it from the very first sentence, as I state in my one paragraph. It makes no sense, according to all that I have revealed about chipping and swamp sparrows, and Vehrencamp et al. have the same kind of data, but have not realized the importance of it, or have chosen not to reveal it. I have tempered my one-paragraph statement to acknowledge the "excellent descriptive statistics" that the authors have collected, but continue to fault them for perpetuating the performance assumptions. It is this kind of hype that keeps the falsehoods coming.

l. 693-740 Yes, I quite agree about Cardoso's nice work, which covers much of the same ground as the present ms.

> **KROODSMA:** Yes, but see my earlier critique of Cardoso et al. on what a song type is.

l. 721-732 What is the main point of this paragraph, or rather the reason for its inclusion? I don't think it's here to say that Cardoso et al's results have been discredited, because Cardoso et al have been brought in to support the present ms. Yet that's how this paragraph sounds. If, instead, the paragraph is meant to say Zollinger et al agree with the present ms's standards of good science, then it's ironic that they do so at Cardoso's expense. Some rephrasing is needed to avoid taking out Cardoso with friendly fire. Perhaps his response to Zollinger et al, which I thought was pretty good, could be brought to bear. Even then, however, it'll take some care to handle the Zollinger et al critique adequately without throwing the present ms's argument off the rails, and just when we're so close to the end.

> **KROODSMA:** I cannot say what needs to be said, and I am allowing readers to come to their own conclusion. Podos is hellishly hypocritical to sign on publicly to a statement about good science when his private record is so bad. Everything I have tried to write here has been censored by my friendly critics. I will rethink how to say this without using the word "hypocrisy" . . . and have given it my best try.

l. 754-757 I don't know what's gained by drawing in these very general opinion pieces (one a commencement address and another a letter to the editor) on herd mentality in science. It seems to relate to the second aim of the present ms, its critique of confirmation bias in song and sexual selection work overall, but the ms hasn't really developed that theme. By this point, the ms has presented a reasonable case against the performance hypothesis, but has only asserted, rather than shown, that confirmation bias in particular has driven acceptance of the hypothesis. Closing the ms with an implied (and undemonstrated) accusation of bias among those working on the performance hypothesis — and implying that any respondents to the present ms might not aspire to or even know about the principles of science expressed in Feynman 1985 or Gitzen 2007 — does not help the ms's cause.

> **KROODSMA:** I have tried so many Discussions and in the end deleted them all. None of them will be acceptable. In the end, I've added the section about bias in word choice and added a small paragraph about what we need to do going forward.
>
> But what I think is really needed is for those who publish on sexual selection to read this kind of material. And read Prum as well, though I just refer to this passage and don't quote it in the paper:
>
> > . . . the study of sexual selection has become a weak science that largely seeks to confirm the adaptive hypotheses it assumes—i.e. that natural selection on mating preferences is the determining force in intersexual

> selection. In this intellectual environment, failure to confirm an honest indication or adaptive signaling hypothesis merely means that the researchers have failed to work hard enough to do so . . . the possibility that traits are not indicating anything is rarely even entertained. Sexual selection has become a field in which the role of natural selection on mating preferences is usually assumed, rarely discussed, largely beyond testing and even redefined into the definition of sexual selection itself. (Prum 2012:2253)
>
> . . . the goal of much empirical work in intersexual selection is to confirm the origin of the signal honesty and sensory efficiency rather than to test its existence. . . . In confirmationist research, negative results are interpreted as failure to have yet looked hard enough to find the evidence of additional selection on preferences . . . Much of intersexual selection research is an extant remnant of the "adaptationist programme" (Gould and Lewontin 1979) in which the deterministic power of natural selection is assumed and alternative explanations are defined out of existence or treated as irrelevant (Prum 2010:3086)

Minor comments:

l. 10 Might be best to specify "biological significance", since we're talking about a scatterplot.

> **KROODSMA:** good

l. 29-31 Must it? Couldn't song be an acoustic flag (or guide) to reveal the singer's location, elaborated by the need for effective transmission, detection, and reception (and perhaps sensory bias or drive or a Fisherian process or whatever) and no more?

> **KROODSMA:** yes, could be. I could change wording . . . If a male has the "wrong songs" for a dialect, the wrong song for the species, etc., I think it says something about his relative quality. I realize that's a very different "quality" than is part of the performance literature, but it is quality. Something in his singing must convey something about who he is, his quality, etc. But it is unsettling when a local prairie warbler sings the song of something like a Prothonotary and is on territory and very successful for about 6 years.

l. 34 Rephrase "confirming that relationship", since your point is that they didn't directly look at it (i.e., they used surrogate measures, correlational evidence, etc).

> **KROODSMA:** ok. It is "confirmation bias" that is the context for the word "confirming," but I'll leave out confirmation bias and adjust the wording here

l. 38 and l. 40 Change commas to semicolons.

KROODSMA: changed one to a colon

l. 48 Insert "is" before "simply" (assuming you don't mean it's become [the sentence's verb] any more implausible than it ever was).

KROODSMA: good

l. 53 Omit "however" [it's already in the previous sentence, which the present sentence does not contradict]

l. 120, 125 What do you mean by lek-like? If you're just pointing out the similarity of this situation to a lek, then I'd rephrase to make that clearer, since you've already described the situation quite nicely, and yes, it is indeed lek-like. If you instead mean that there's additional features that make the situation lek-like, then I would spell them out — especially for l. 125, which implies something additional is going on.

KROODSMA: I'll adjust wording. L 125 is simply a recap of the paragraph, nothing new. I do introduce the word "competitively," which seems like a logical inference given the behaviors of the birds.

~ ~ ~ ~ ~ ~ ~ ~ ~

Response to Reviewer #2

Reviewer #2: The author makes a strong and convincing argument against the commonly-accepted 'motor constraints' hypothesis put forward by Podos and colleagues. The hypothesis argues that the closer a male songbird can get to the 'high performance' upper limit line comparing frequency bandwidth of notes with trill rate of notes, the higher a quality of male it is, which should impact female choice of males. I teach animal communication, and admit that I have presented the 'motor constraints' hypothesis as a given in the field. After reading the author's commentary, however, I am convinced by his counter-arguments. I think his work will be an important and influential contribution to the field - it should certainly generate useful debate. Even more importantly, hopefully the author's work here will motivate stronger studies in the field to address these important questions.

KROODSMA: I am glad #2 is convinced of my arguments. I wasn't sure about #1!

Other comments by line number

112-133 It was not entirely clear why this section is in the paper. I can see some reasons why, but a stronger transition into the section and then out of the section might make for a stronger story.

> **KROODSMA:** OK. I'll introduce this entire section a little more clearly when I begin the section on CHIPPING SPARROW.

Figure 4 It is brilliant and informative to include 0s on the axes to bring in the blank space to the left and below the scatter of data, to illustrate the 'species atypical' song variation regions.

> **KROODSMA:** I felt that the usual graphical representation was deceptive, intentional or not. I wanted to show there was an area of "normal behavior" surrounded by areas of "abnormal" behavior, and that one are of abnormal behavior might not be any more important than others. All lines delimiting the data are, by definition, "performance limits" of one kind or another, as males do not perform outside those boundaries.

246-257 It struck me that perhaps habitat constraints (vegetation thickness, noise, etc.) could also explain some of the variation on trill rates that is seen - has this been looked at in these studies?

> **KROODSMA:** No information.

Figure 8 Should there not be an X-axis scale with units here?

> **KROODSMA:** The "distance" to the upper line is all relative, given that the measured distance on a graph depends on the axes and a number of features of the graph. So I think that the best I can do is show the relative distance, as that is really all that is important.

292 A transition paragraph into this new 'focused critique' section would be helpful.

> **KROODSMA:** I added just a few words, but I thought that lines 293-294 were sufficient.

363-364 / 498-500 The author might wish to cite Burghardt et al. (2012, Ethology, Minimizing observer bias in behavioral studies...) to back up this important point. It is not just Ballentine et al. 2004 who are guilty of this - the vast majority of studies in the field do not have analyses conducted blind or do not report high inter-rater agreement - or at least do not report these things in the published studies.

> **KROODSMA:** Thank you. Great reference added. Burghardt, G. M., J. N. Bartmess-Levasseur, S. A. Browning, K. E. Morrison, C. L. Stec, C. E. Zachau, and T. M. Freeberg. Perspectives - Minimizing Observer Bias in Behavioral Studies: A Review and Recommendations. Ethology. 118:511-517..

371-378 The implication here is that the different recording media and equipment were used in a biased, systematic way across the data set; otherwise one might just expect more 'noise' in the

system - more variation in the data set. Does the author suspect the former - that the different equipment resulted in the differences in behavior assessed?

> **KROODSMA:** Yes, I believe that the 13" parabola would give poorer recordings. I have used this parabola, and found it relatively useless for obtaining high performance recordings. See comments on same topic for reviewer #1. I have now done tests with 5 different chipping sparrow songs and 3 different microphones. The results are sobering for these performance studies (see Figure 9 in the revised ms).

525-535 / 548-551 Please add page numbers to these quotes from Lahti et al. (2011)

592-594 / 630-633 Please add page numbers to this quote from Moseley et al. (2013)

> **KROODSMA:** good to add page numbers in above two requests.

Figure 10. It would be good to keep the parallel structure with the axes based at 0, to provide the 'species-atypical song' regions. Also, the Y-axis needs a title and units. Additionally, given these are data from a new species, perhaps a few spectrograms of banded wren songs would be good here for simple comparison's sake.

> **KROODSMA:** Yes. I have added a sentence to the figure legend, and will add the Y-axis.

662-663 This part 4) of the argument reads as too bitter/gruff, when it clearly does not have to be to make the strong argument being made here. There are other places in the manuscript where the gruffness could be toned down or eliminated, and the same argument would hold without the potential 'turn-off' language for some readers.

> **KROODSMA:** agreed. Will adjust wording

668-669 "to be comfortable WITH that result..."

> **KROODSMA:** yes

693 I would suggest the author add some transition here - perhaps even a new section to the commentary - that sets up a 'these are stronger studies' conversation. I originally read into the Cardoso et al. 2009 discussion as if it were just another in this long line of studies being critiqued - confusion lead to re-reading a couple times until it hit me that the conversation had shifted directions. A stronger transition would make this all much clearer and smoother.

> **KROODSMA:** yes. I think that a single sentence does it, before the citation from the paper: "In contrast to my remarks on all of the above papers, I applaud the conclusions of these papers by Cardoso et al."

~ ~ ~ ~ ~ ~ ~ ~ ~ ~

Response to Reviewer #3

Reviewer #3: This paper presents a pointed critique of the performance hypothesis and several studies that claim to have found evidence for it. The author argues there is no strong support for the hypothesis and that the hypothesis itself is not credible. This is based largely on a logical argument that trill rates and bandwidth cannot possibly reveal male quality because the scatter of points on the bandwidth vs trill rate graph is due to birds singing different song types, which is a result of faithful copying of neighbors' songs. The author presents his own data on two species of sparrows to illustrate this point clearly. This is a significant contribution.

The also paper makes a number of other useful points that may stimulate future research on vocal signals. However its hypercritical nature (suggesting that the criticized authors are doing bad science) may be unproductive. The reality may be that the performance in the sense used in this paper does explain some of the variation, but that a variety of other factors are important.

> **KROODSMA:** My point is that the authors are doing bad science. If this reviewer knew the background and origins of this paper, I think he/she might understand a little better what is going on. But, as I accept, no need to flog the dead horse. I have tried to be as gentle as possible and still get the points across. I'll continue to revise as reviewers point out wording that can be altered.

The author could improve the paper by clearly stating what is meant by the performance hypothesis and summarizing the history of its development. This is important because vocal performance can refer to a number of things in the literature (e.g., song rate, song complexity, song consistency).

> **KROODSMA:** I felt that I had done some of that in the Abstract and Introduction. Doing more will make the paper even longer, so I'm reluctant to add more. I have adjusted the attribution to the hypothesis here and there, to make the history clearer.

The paper is quite long and a bit rambling, repeating the same points in different places. A more concise and focused discussion would be easier to follow and more convincing.

> **KROODSMA:** Yes, I repeat the same point because different authors have repeated the same points, and I need to show the weaknesses of each paper that claims to show the biological significance of the scatter plot. I fear redundancy is inevitable.

Several statements, as noted below, attribute the hypothesis incorrectly:

Lines 11: The author states "This "motor constraints hypothesis" of Podos (1997) proposes that the closer a song plots to an upper bound on this 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."

Line 41: "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."

Line 98: The author again seems to mistakenly attribute the "performance hypothesis" to Podos (1997): the "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 sense of Podos (1997)."

Line 269: "A critical but untested feature of the Podos (1997) performance hypothesis is that songs actually provide reliable, honest signals of male quality."

Line 282: "Figure 8. Song performance measures (sensu Podos, 1997) can provide no reliable information about inherent male quality in swamp sparrows."

Yet Podos 1997 does not mention a "motor constraints hypothesis" - he proposes a "performance constraints hypothesis"

It is Podos 1996 that proposed the "motor constraints hypothesis"(this paper is cited on line 203, but is not included in the References section)

Nowhere in Podos 1997 (or in Podos 1996) can I find any statement linking performance constraints or motor constraints to quality of the singer. Podos 2001 may have first presented the measure of "vocal deviation" but he uses it to compare species and bill shape, not as a measure of male quality.

> **KROODSMA:** I have been sloppy here. I will try to attribute correctly. It is Ballentine et al. 2004 and Podos et al. 2004 who developed the performance hypothesis based on Podos 1996, 1997. For example, I now write the following:
>
> Another idea that has over the last decade gained much traction is the performance hypothesis developed by Ballentine, Hyman, and Nowicki (2004) and Podos, Peters, and Nowicki (2004), based on motor and performance constraints described by Podos (1996, 1997).

Line 16: The author states: "The scatter in the graph for songbirds is explained not by male quality but by social factors and song learning". This is the author's main criticism of the "performance hypothesis".

KROODSMA: yes. That is one main criticism, among several others that I can't state explicitly because it would not be nice.

It would be only fair to note that, although not dwelling on it, Podos 1997 and Ballentine et al. 2004 do indeed acknowledge that much of the scatter in the graph is explained by the inclusion of different song types and song learning:

Podos 1997 (p 548) states: "With respect to the present data, it is clear that factors unrelated to performance limits have shaped the expression of song. For example, much of the variation under the upper-bound regression lines in Figure 3 is undoubtedly a result of cultural evolution, which is itself dependent upon accuracy in song learning"

Ballentine et al. 2004 (p 165) state: "Some of this between-male variation is due to variation in the average deviation of different song types (i.e., some song types consistently have low deviations and others high deviations regardless of which male sang them) … and the fact that males differed in the number of low-deviation and high-deviation song types they included in their repertoires"

To me this indicates that they are not trying to explain all of the scatter (and perhaps that they don't find it very interesting).

KROODSMA: Yes, in 1997 Podos writes about how accurately the swamp sparrows learn their songs, but it seems that the accuracy in song learning drifts to how inaccurately they learn their songs when he needs support for his calibration hypothesis. I could give credit that Podos and Ballentine acknowledge that some scatter is due to different song types, but I don't think they recognize the importance of that statement, and don't find it all that interesting, and giving them credit for making the statement and not making enough of it doesn't seem useful. . . . I have now capaitalized on that statement, because I believe the observation by Ballentine et al. essentially falsifies their hypothesis. Here's what I have now written in a revised paragraph:

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" (p. 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 doesn’t reveal it, or 3) there’s no selection for low-deviation songs to convey male quality. Consequently, although it remains possible (though 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.

Line 48: states " … the hypothesis has become largely an assumption, never truly tested, and simply biologically implausible, if not impossible"

If data match the predictions of any hypothesis, then it does represent support for it, even if it isn't true. Podos 1997 provides evidence that, among sparrows that produce certain kinds of trills, performance is constrained. The evidence is the upper bound, as it suggests a tradeoff between trill rate and bandwith. It is an important point that Podos' performance constraints hypothesis does not attempt to explain the scatter of points.

> **KROODSMA:** Yes, I have reattributed the origin of the hypothesis elsewhere. And the word "support" I have addressed as well, as there is not the kind of openness needed when obvious alternative hypotheses are also supported by the data. But, I wonder, has the lower and left bound on the scatterplot been explored? I bet the birds don't sing those songs either, and if they don't sing the abnormal songs below and left of the scatterplot, what do we make of those data? Performance constraints? By definition they are performance constraints, because males don't perform songs in those blank areas of the graph. Now we quibble over what "performance" means, and we begin to see the problems with the very word.

It may be worth citing Wilson et al. 2014 (Uneven Sampling and the Analysis of Vocal Performance Constraints. Am Nat 183(2):214-228) who reanalyzed data from trill rate and bandwidth studies and found that the data for many species, including Chipping Sparrows and Swamp Sparrows, did not support a performance tradeoff; the regression slope was not statistically significant because of sampling limitations.

> **KROODSMA:** I suppose I should cite it, but the paper misses the mark in so many ways, I feel. Their purpose is not to question the primary issues that I raise, but more to detect better what is known to be true. Below, I highlight key wording in the Abstract. I do now cite Wilson et al., but not in so positive a fashion.
>
> > abstract: Studies of trilled vocalizations provide a premiere illustration of how performance constraints shape the evolution of mating displays. In trill production, vocal tract mechanics impose a tradeoff between syllable repetition rate and frequency bandwidth, with the trade-off most pronounced at higher values of both parameters. Available evidence suggests that trills that simultaneously maximize both traits are more threatening to males or more attractive to females, consistent with a history of sexual selection favoring high performance trills. Here, we identify a sampling limitation that confounds the detection and description of performance trade-offs. We reassess 70 data sets (from 26 published studies) and show that sampling limitations afflict 63 of these to some degree. Traditional upper-bound regression, which does not control for sampling limitations, detects performance trade-offs in 33 data sets; yet when

sampling limitations are controlled, performance trade-offs are detected in only 15. Sampling limitations therefore confound more than half of all performance trade-offs reported using the traditional method. An alternative method that circumvents this sampling limitation, which we explore here, is quantile regression. Our goal is not to question the presence of mechanical trade-offs on trill production but rather to reconsider how these trade-offs can be detected and characterized from acoustic data.

Line 100: The author makes an important point here: "(and there is no evidence for song learning in any songbird species or especially in chipping sparrows (Liu and Kroodsma, 1999, 2006) that a male is in any way limited in what naturally occurring trill rate he can learn)." If this is the challenge the author wants to make, then perhaps it should be made more forcefully rather than parenthetically.

> **KROODSMA:** Yes, I have removed from parentheses and stated more clearly the significance of this statement.

Line 323: The author begins criticism of Ballentine et al 2004, which in my opinion presents the strongest evidence for the hypothesis, by comparing female preferences for higher trill rates of the same a song type sung by different males.

Line 362: The author dismisses the results of Ballentine et al 2004 by suggesting they were methodologically biased. We don't actually know that there was any such bias - but the author can't believe the results so they must be wrong.

> **KROODSMA:** Based on my analyses of swamp sparrow songs, I don't think I am alone in believing that the results can't be true. I dismiss the results because they make no sense, given the biology of swamp sparrow song learning. My task is then to point out possible reasons why the study went astray. I don't dismiss because of the possible methodological biases, but must point out the possible reasons that the authors came to the conclusions that they did. Failing to collect data blindly, when there appears to be such commitment to the favored hypothesis, is perhaps sufficient. I have also expanded the analysis of this paper to take into account the "serious blow" they make to their own paper (see above commentary).

Line 747: The author states "No compelling evidence suggests that either males or females attend to trill rate, frequency bandwidth, or a combination of the two in assessing the quality of the singer, i.e., no compelling evidence supports the performance hypothesis." So, what would be a strong test, what would be compelling evidence?

> **KROODSMA:** Good question. I'd read Zollinger et al. about the characteristics of good science. I'd read Feynman. I'd ask for utter honesty and scientific integrity instead of what we get in these papers. I'd ask for blind observers, for consideration of alternative

explanations for data when they exist. I'd ask for good science, in short, but the kind of science I ask for is in precious short supply in studies of sexual selection and birdsong. See quotes by Prum somewhere above. . . . And in the end I add one paragraph in the Discussion that gives some suggestions for moving forward.

~ ~ ~ ~ ~ ~ ~ ~ ~ ~

Response to Reviewer #5

Reviewer #5: The author offers a contrary view to the notion that there is selection in birds on maximizing performance, particularly as it relates to Podo's motor constraint hypothesis. The general idea is that the papers supporting the hypothesis have been too heavily slanted towards advocacy in lieu of actually testing the hypothesis. In response to this literature, the author offers a host of reasons why we should simply reject the hypothesis.

I suspect that a number of things brought up in this paper are correct. I also suspect that a dispassionate and well-designed research program would show that the motor constraint hypothesis is at best a weak predictor of sexual selection in birds (despite its popularity).

A number of ideas have floated around in the behavior literature over the years that have not been particularly useful. A vetting of these ideas is a terrific endeavor. The motor constraints hypothesis may (or may not) be on this list. However, a proper vetting of these ideas requires a level-headed and robust discussion of why the ideas are inadequate. From my perspective, this paper falls short on both fronts.

It seemed quite odd to me and not an auspicious beginning in the paper to be told that "I offer apologies to those who feel my approach too frank, or too blunt, or overkill". After reading the paper, I agree that the apology is warranted. Rhetoric of this level really does not belong in Animal Behaviour, in part because it correctly implies that the paper will be anything but a robust, level-headed discussion of the facts.

KROODSMA: Perhaps. I have simply deleted the last paragraph of the discussion and added another. I am weary of trying to find some PC way of saying what needs to be said candidly.

As for the data: the onus is on the author of a paper such as this to exceed the standards set by the papers being criticized. Unfortunately, the first set of data offered as proof that there is no selection on maximal performance within constraints does not rise to this level. The author argues that lower variation within song types compared to variation across song types, coupled with the learning of neighbor song types by a male, means that the performance level of a song is uninformative. This is a problem.

KROODSMA: The first set of data must mean the chipping sparrows. Or perhaps the swamp sparrow data as well**.** I guess I don't see the extent of the problem. To me it is pretty obvious that the trill rate of a chipping sparrow's song cannot reveal anything about male quality, given how the male learns his song from an adult, how there's no evidence in chipping sparrows or any other species that a male is limited in what trill rate he can learn, etc. If one other reviewer had questioned the nature of my data, I'd work a little harder to try to figure out the weaknesses in my data on the two sparrow species, but lacking additional input, I don't know how to modify what I have done.

For example, American toads call at higher frequencies when the density of males is low but at lower frequencies when there are a large number of competing males. Nearly everyone knows the tungara frog example. By the author's logic, this would mean that call frequency (or chucks added to the call) is uninformative.

KROODSMA: During interactions, birds don't vary their songs in the way frogs do. If frequencies were raised or lowered in different contexts in some detectable way (as in Kentucky warblers), that would be very interesting in the context of performance. I am not dismissing what frogs do. I am addressing what birds do and the severe problems with the literature.

This is almost certainly incorrect. Similarly, escalation and de-escalation of agonistic interactions using vocal signals in birds has been demonstrated for several species. As such, the fact that a male has a diversity of calls that vary in performance level is not particularly good evidence either for or against selection on performance. Moreover, the fact that there is still a change in the structure of calls during the learning process begs the question whether there are any systematic changes in the calls, perhaps as a function of male quality. This issue is not addressed.

KROODSMA: Yes, songs change during the learning process, but they inevitably (in song-learning songbirds) match closely the tutor song, as I have demonstrated for chipping and swamp sparrows. In chipping sparrows, trill rates vary from about 7 to 25, but young birds match closely what their tutor does.

We might also like to know if there is any selective learning of song types by males. This is not addressed.

> **KROODSMA:** Yes, I believe I do address it. I discuss that there is no evidence that young birds are in any way limited in what naturally occurring trill rates they learn.

This whole discussion begs the question whether females prefer any of these calls to begin with. This is not addressed. As a result, much of the evidence offered to counter the performance hypothesis is simple conjecture, something that the author is accusing Podos and colleagues of doing.

> **KROODSMA:** I try to point out that all of the evidence is compromised by bad science. I welcome some good, objective science in support of this hypothesis, in which alternative explanations and hypotheses are honestly revealed.

The author also suggests that some of the Podos' and colleagues' results could be an artifact of their use of different sized parabolic reflectors. But is this really likely? Do the experiment. See how females respond.

> **KROODSMA:** I will elaborate on that in the text. One doesn't have to ask females if one can show that the parabolas capture sounds differently and the differences are sufficient to believe that birds would respond differently to the recordings. One of the real problems with this literature is that so few people are doing it, and no one is going to take the time to repeat or confirm someone else's work, and there's a strong need to conform and confirm so that one's own paper gets published.. . . This discussion updated to reveal the serious problems with measuring frequency bandwidths (see Figure 9 in revised ms).

In short, the author may well be correct. It is entirely likely that the motor constraint hypothesis is simply wrong. Sensory issues apparently offer another reason to say this. I understand that females may not be all that good at detecting bandwidth anyway. But there are too many holes in this manuscript and the rhetoric is too extreme for me to think that this paper is what is needed to put this issue to rest.

> **KROODSMA:** I guess that's what a Forum is all about. Let's air the issues. Let's think about how to do good science. So far, no issues have been aired, as all attempts to do so have been suppressed, even invoking university police to threaten me with criminal charges for trying to address these issues. The issues I raise need to be aired, whether right or wrong. If I'm wrong, great, but we will all have come to a better understanding of what to do next. If I'm right, ok, then let's move on as well, and do some good science to understand more about how the world works.

Instructions for resubmitting

NOTE: Upon submitting your revised manuscript, please upload the source files for your article. For additional details regarding acceptable file formats, please refer to the Guide for Authors at: http://www.elsevier.com/journals/animal-behaviour/0003-3472/guide-for-authors

When submitting your revised paper, we ask that you include the following items:

Response to Reviewers (mandatory)

This should be a separate file labeled "Response to Reviewers" that carefully addresses, point-by-point, the issues raised in the comments appended below. You should also include a suitable rebuttal to any specific request for change that you have not made. Mention the page, paragraph, and line number of any revisions that are made.

Manuscript and Figure Source Files (mandatory)

We cannot accommodate PDF manuscript files for production purposes. We also ask that when submitting your revision you follow the journal formatting guidelines. Figures and tables may be embedded within the source file for the submission as long as they are of sufficient resolution for Production. For any figure that cannot be embedded within the source file (such as *.PSD Photoshop files), the original figure needs to be uploaded separately. Refer to the Guide for Authors for additional information.

http://www.elsevier.com/journals/animal-behaviour/0003-3472/guide-for-authors

Highlights (mandatory)

Highlights consist of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point). See the following website for more information

http://www.elsevier.com/highlights