Subspecies status and methods explain strength of response to local versus foreign song by oscine birds in meta-analysis

Timothy H. Parker[1,2,6]
Emma I. Greig[3]
Shinichi Nakagawa[4]
Marcelina Parra[1]
Anthony C. Dalisio[5]

1. Department of Biology, Whitman College, Walla Walla, WA, USA
2. Department of Biological Sciences, Macquarie University, North Ryde, Australia
3. Cornell Lab of Ornithology, Ithaca, NY, USA
4. Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Australia
5. Department of Biological Sciences, Emporia State University, KS, USA.
6. Corresponding author: parkerth@whitman.edu; +1.509.526.4777, Department of Biology, Whitman College, 345 Boyer Ave, Walla Walla, WA 99362, USA

key words: deceptive mimicry, meta-analysis, playback, pseudoreplication, selective reporting, signalling

word count: 10227 excluding references, tables, figures

This is a post-print of a published paper:
Parker, T.H., E. I. Greig, S. Nakagawa, M. Parra, A. C. Dalisio. 2018. Subspecies status and methods explain strength of response to local versus foreign song by oscine birds in meta-analysis. Animal Behaviour. 142: 1-17. doi: 10.1016/j.anbehav.2018.05.023

Our data are available on the Open Science Framework:
Parker TH, Greig E, Nakagawa S, Parra M, Dalisio A (2018) Data and code from “Sub-species status and methodological choices explain strength of territorial response to local versus foreign song by oscine birds in meta-analysis.” Open Science Framework. doi:10.17605/OSF.IO/W2MVP

To understand the implications of geographic variation in vocal culture in songbirds, researchers have often compared territorial responses to playback of local songs versus responses to playback of songs from ‘foreign’ conspecifics. This body of work has the potential to help us move towards a general understanding of factors driving divergence in signal recognition. We conducted a systematic review and meta-analysis of 57 playback studies to explain variation in strength of response to local versus foreign songs. Studies with incomplete reporting of results had elevated effects due to selective reporting. Studies that used small numbers of stimuli as exemplars (pseudoreplication) had more variable effects than studies without severe pseudoreplication. Whether or not we controlled for pseudoreplication, we found greater response to playback of local song than to foreign song. In investigating potential biological drivers of the variation in strength of experimental effects, we found that the difference in territorial response to local versus foreign song was stronger if the foreign song was recorded from another subspecies than if the foreign song was recorded from the same subspecies as the focal individuals. Indices of risk of accidental response to heterospecific song did not coherently explain response to foreign conspecific songs, nor did factors expected to influence individual experience with foreign conspecific songs. Thus, although oscine songbirds clearly react more aggressively to local song than to foreign song and variation in the strength of this effect is influenced by methodological choices and sub-species status, considerable variation in the strength of response to local versus foreign song playback remains to be explained.

Understanding the factors influencing receiver response to conspecific signals has long been a major part of behavioural research (Wiley 1983). Response to conspecific signals is especially interesting when these signals differ geographically (e.g., Danner et al. 2011; Searcy and Andersson 1986). Geographic variation in vocal signals is common and is often particularly striking in the oscine songbirds thanks to learning from local conspecifics (Beecher and Brenowitz 2005). In songbirds, receiver responses to geographically variable vocal signals have been studied for decades with song playback experiments in which focal territorial individuals are played either local or foreign song. In such playback experiments, it seems that the most common result is for the territorial individual to respond more strongly to playback of the local song than to playback of the foreign song (e.g., Bradley et al. 2013; Lemon 1967; McGregor 1983; Podos 2007), though this has never been quantified. However, these playback experiments (Appendix 1) reveal substantial variability in response to foreign and local song as measured in hundreds of statistical tests from dozens of species and populations studied around the world (Parker et al. 2018). Thus, these experiments are an excellent resource for seeking to understand variability in response to geographically divergent signals. Explanations for this variability might plausibly come from a combination of evolutionary and developmental processes and from methodological differences among studies.

One possible explanation for variable response to foreign playback is rooted in signalling theory, which predicts that receivers should evolve to respond to signallers with whom they interact and whose signals convey relevant information while simultaneously avoiding response to irrelevant signallers with whom they do not interact (Wiley 1994). Relevant interactions often involve mate attraction or competition with conspecifics, so an important axis of discrimination will be between conspecifics and heterospecifics (Amézquita et al. 2011). Any receiver can be imagined to have a degree of permissiveness for what it will consider relevant in any given context, and with respect to recognizing mates and competitors we can think of this permissiveness as a “window of recognition” (Hudson and Price 2014). The boundaries of the window of recognition for conspecifics should depend in part on the acoustic environment created by the singing of heterospecifics. Most songbirds probably face selection to avoid responding to heterospecific songs without reducing response to conspecifics (e.g., Shizuka 2014) as appears to be the case in other taxa (Amézquita et al. 2011; Symes 2014). This selection may be stronger in the presence of more species or of more closely-related species (Hamao 2016). Thus, selection to discriminate between conspecific and heterospecific song might plausibly narrow the window of conspecific recognition and thus as a by-product reduce response to foreign conspecific song.

Although this variation in the signalling environment may drive the evolution of divergent patterns of discrimination against foreign conspecific song, variation in song discrimination could also emerge as a plastic developmental response to individual experience. For instance, interaction with individuals from other conspecific vocal cultures could serve to broaden what is recognized as a conspecific song (Wright and Dorin 2001). Such interactions could occur in any number of scenarios, but might be more likely in migratory populations or in populations in which individuals often disperse across cultural boundaries or gradients (Colbeck et al. 2010).

Another possibility is that evolutionary divergence of populations leads to divergent signals and divergent signal-recognition. This hypothesis seems likely to be true especially given that even naïve juveniles may preferentially learn songs of their own subspecies over conspecific song from other subspecies (Nelson 2000). Stronger response to song from the local subspecies has received some support (Alström and Olsson 1999; Petrinovich 1981; Turcokova et al. 2011), but is sometimes contradicted (Tietze et al. 2012). However, divergence of response to song has itself been used as evidence of evolutionary divergence (Freeman and Montgomery 2017; Randler et al. 2012).

Although biological hypotheses may explain variability in response to foreign and local songs, it is also likely that various methodological differences among playback studies have influenced the distribution of published effects. Of particular interest in playback studies is the effect of pseudoreplication of song stimuli, in which a relatively small number of distinct stimuli are used in a larger number of trials. When the number of stimuli is smaller than that number of otherwise independent trials, trials with the same stimulus are not true independent replicates (Kroodsma 1989). This form of pseudoreplication has become less common since it was first identified in the song playback literature (Kroodsma et al. 2001), but it characterizes nearly all of the early song playback literature and some more recent papers as well (Fig S1). One way to think about the problem with this sort of pseudoreplication is that as the number of stimuli declines, the probability that the chosen stimuli reliably represent the distribution of stimuli in the population from which they are drawn also declines. In the extreme case, it is easy to see that a single recording of a single individual might tend to induce a weaker or a stronger response in playback trials than the average response from a series of stimuli recorded from multiple individuals. In other words, pseudoreplicated studies should produce more variable and thus less reliable results than those from studies in which different stimuli were used for each trial. However, this hypothesis has never been tested empirically.

We used meta-analysis of published studies (Appendix 1) of response to playback by oscine birds to assess several hypotheses about the determinants of signal recognition. Before testing our primary biological hypotheses, we tested several hypotheses that might explain variation in effect size as a function of the methods of the original study, including whether or not the original study suffered from pseudoreplication. We explored three potential biological explanations for variation in effect size. Our first such hypotheses was that the risk of accidentally responding to heterospecifics drives increased discrimination against foreign conspecifics. If this mechanism were operating, then we expected greater difference in response (i.e. reduced response to foreign song relative to local song) at sites with congeners or confamilials present, or at sites expected to have higher songbird species diversity overall. Our second biological hypothesis was that individual experience drives discrimination, either due to direct experience with particular songs or experience with a diversity of conspecific song types. If this mechanism were operating, then we expected birds that move longer distances during their lifetime to be more likely to respond to foreign songs than more sedentary birds, birds being played foreign songs from nearby dialects to respond more strongly to those songs than those hearing songs recorded at much greater distances, and birds that are physically isolated from the foreign song by geographic barriers to respond less strongly to foreign songs than birds separated from foreign song by occupied habitat. Our final hypothesis was that evolutionary divergence drives song discrimination. In this scenario we expected stronger differences in response between foreign song from different sub-species and local song than between foreign song from within the same subspecies and local song.

METHODS

Locating and screening primary studies

We conducted a systematic review of studies in which the response of wild territorial oscine birds to playback of recorded song on their territory was assessed using the playback of local songs and the playback of foreign songs. To locate studies, we searched the complete Web of Science database (year 1900 to present) with the following search terms [play*back* (song* or sing*)] on 20 September 2016 which produced 1521 records, and on 27 September 2016 [dialect recogn* bird*] which produced 78 records, [geograph* recogn* vocal*] which produced 127 records, [dialect* (song* or sing*) foreign*] which produced 26 records, and [dialect* (song* or sing*) local] which produced 145 records for a total of 1748 unique records (Fig. S2). We examined each title and rejected all papers that were obviously not applicable, for instance because study subject was not an oscine songbird or because the stated topic differed dramatically from playback of song in the wild. At this first stage, we read the abstracts of all papers not rejected based on their title and again filtered out papers that were obviously unsuitable. This left us with 128 studies. We then examined the full text versions of papers themselves to determine suitability for our analysis. Finally, for each of the 44 papers that were judged suitable for our analysis, we read its literature cited and identified any potentially relevant papers that had not been identified in our Web of Science search. This turned up an additional 27 papers, 19 of which were suitable for inclusion in our analysis. From the 63 eligible studies, we identified 499 relevant statistical effects from 148 experiments. However, some studies that reported results from experiments that met our criteria unfortunately did not report results in sufficient detail from one or more statistical tests to allow for inclusion of those tests in our analyses. In these cases, we sought to contact the author via email to request the missing information. Some authors we could not locate, but we contacted 15 authors and received additional information from 7 of these authors. In total, we had to exclude 120 incompletely reported statistical effects from 31 experiments reported in 14 papers because we could not obtain sufficient information from the authors. Our final sample size was 379 usable statistical effects from 130 experiments and 57 published studies of 44 species (Fig S2, a PRISMA [Moher et al. 2009] style diagram; see also Appendix 1).

We only included studies in which the focal birds receiving the playback were defending a territory. In part because most of the studies we located were from North Temperate species, in most of the studied species the males are the primary singers and the males respond most aggressively to playback. However, we also included studies of species in which both sexes sing and in which both sexes join in territory defence.

Amongst studies that evaluated avian responses to playback of foreign and local song, we further constrained inclusion in our analysis in an attempt to create a relatively homogeneous body of work from which to draw conclusions. First, although we initially gathered some papers from non-oscine birds, prior to completing data gathering we decided to limit the study to oscine songbirds because most song playback research has been conducted on members of this large group and because we expected that song learning and song function are likely most similar within this monophyletic group. We also limited our analysis to studies of responses to conspecific songs, though we included studies in which the foreign song recordings were either within or across subspecies. We did not consider studies of response to heterospecifics because if selection fairly consistently acts against response to heterospecifics, such studies would complicate our efforts to determine the factors that limit song recognition within conspecifics. Further, we would still face the decision of deciding which heterspecific song playback was acceptable (within genus, within family, any oscine?), and so we opted for the clear and justified choice of limiting our analyses to within-species songs. We used the most up-to-date taxonomic classification, so if an original study reported a foreign song as coming from a different subspecies, but the taxa had since been classified as separate species, we excluded the study. However, some studies used playback response as evidence to determine whether a taxon should be considered a separate species, or just a separate subspecies. In this scenario the strength of response in a playback trial could determine whether that trial was included in our analysis because a weak response could lead the author to classify the two populations as separate species, thus excluding the result from our analysis, whereas a strong response could lead to their classification as conspecific, therefore leading to inclusion in our analysis. This would bias our sample against studies that found weak response to foreign playback recordings when those recordings were from other subspecies. To minimize this problem, we only used papers seeking to determine if taxa were separate species (or separate sub-species) if those papers based their conclusion regarding taxonomic status primarily on non-song evidence. Several studies used hybrid recordings which combined components of local and foreign songs in an effort to identify the components of the song that most effectively promoted differential response. We excluded these experiments because of the impossibility of determining, a priori, whether such a song could have been perceived as local or foreign. We also excluded the two cases where individuals of the studied species were likely not defending territories (as individuals, pairs, or cooperative groups). Further, we excluded a study in which playback trials involved taxidermic mounts that differed in plumage such that song differences were confounded by plumage patterns. If a study used identical taxidermic mounts in both local and foreign playback, we retained the study for our analyses.

Not all studies comparing response to foreign versus local song used local songs that met our criteria. We excluded studies in which the ‘local’ song was recorded at a spatial scale dramatically larger than an individual bird was likely to regularly encounter while territorial during the breeding season. In practice, studies either recorded their ‘local’ stimulus songs from within or close to their study site, or from a much broader region, often extending up to 100’s of km from the site of the playback, sometimes without providing evidence for dialect constancy at that scale. In many species, geographic variation in song is substantial over much smaller spatial scales, and so we wanted to reduce the chance of including trials in which both treatments were perceived as foreign by the focal individual. We recognize that some species exhibit dialects that appear to be relatively uniform across 100's of km, and it may be biologically plausible to consider such distant songs to be 'local' in some species. However, we excluded all cases in which some or all 'local' songs were recorded >100 km from the playback site (or where we could not exclude this possibility) in the interests of adopting a consistent decision rule and avoiding judgments for which we often lacked sufficient empirical basis (for details of excluded studies, see Parker et al. 2018). We also excluded studies in which the 'local' song recording was taken from an adjacent neighbour, as there is a well-demonstrated tendency for reduced aggression towards playback of known-neighbour song, presumably due to the 'dear-enemy' effect (Brooks and Falls 1975; Godard 1993; Hyman 2005; Wei et al. 2011; Wilson and Vehrencamp 2001).

We relied on a variety of reported statistical information to derive standard effect sizes from primary studies for use in our meta-analysis. If treatment means and associated standard deviations (or standard errors) were reported, we used this information to calculate Hedge's d. However, because many studies did not include means, we chose to convert all effects, including Hedge's D, to Fisher's Z transformation of the correlation coefficient. In the absence of means or a reported correlation coefficient (r), we used test statistics (F, t, χ^2) and standard conversion formulae (Rosenberg et al. 2013). If these were unavailable, we used p-values by first converting p to the standard normal deviate (Z score, using the function qnorm in R), then to the correlation coefficient (r). Some papers included means and standard deviation or standard error only in figures, and in these cases, we used the program WebPlotDigitizer (3.5) to extract values from figures (Rohatgi 2017). In several other papers, the only usable information was in the form of counts of responses to different treatments. In these cases, we conducted Fisher's Exact tests in R (using the function fisher.test in R) to generate p-values for conversion to Fisher's Z. We provide details of the derivation of each effect size in our archived data (Parker et al. 2018).

Analyses

We analysed data with R version 3.4.2 (R_Core_Team 2017). All R code is publicly archived (Parker et al. 2018). We conducted meta-analysis with the metafor (version 1.9.9) package (Viechtbauer 2010). We conducted our test for publication bias with the MCMCglmm (version 2.24) package (Hadfield 2010; Hadfield and Nakagawa 2010) to implement Bayesian generalized linear mixed models using Markov chain Monte Carlo methods (e.g. Horváthová et al. 2012).

In meta-analysis, individual effects are weighted according to the sample size from which they were derived. Calculations are designed so that studies based on larger samples more strongly influence the meta-analytic average because as sample sizes increase, effects should converge on the true population value (Gurevitch and Hedges 1999). When standardizing effects as

Fisher's transformation of the correlation coefficient (Zr) as we did, effects are weighted by an inverse function of the corresponding variance (Vz) which is a function of sample size (($Vz = 1/(n-3)$, where n = the sample size of the study). Determining sample size is typically straightforward unless authors have failed to report sufficient information. However, in some cases non-independence in the designs of primary studies can create ambiguity (Noble et al. 2017). In most playback studies we located, researchers used the number of trials as the sample size, or depending on study design, the number of focal individuals receiving playback. However, many of the studies in our data set used recordings of relatively few individuals as the stimuli across multiple trials. In the extreme cases, which are common in our data set, a single recording of a single individual was used to represent all foreign songs, and a single recording of a single individual was used to represent all local songs. As we describe in the introduction, if the number of stimuli is smaller than that number of otherwise independent trials, the trials sharing an identical stimulus are not independent replicates, thus creating a form of pseudoreplication (Kroodsma 1989; Kroodsma et al. 2001).

We wished to understand the effects of this pseudoreplication on the conclusions of published studies and on our ability to draw conclusions from meta-analysis. If psuedoreplication acts as theory predicts by overestimating the degree to which the set of stimuli represents the population of available stimuli, all else being equal, we should observe greater variation among effect sizes from pseudoreplicated studies than among effect sizes from studies in which different stimuli were used for each trial. To test this prediction, we first classified each effect as being derived from an experiment that was (1) not pseudoreplicated (typically a different stimulus for each trial), or (2) moderately pseudoreplicated (number of stimuli >25% of the number of otherwise independent trials), or (3) severely pseudoreplication (number of stimuli no more than 25% of the number of otherwise independent trials). We then compared the degree of variability in effect size among these three categories using an F-test test to compare variances between two groups (using the function var.test in R). Finally, to further understand the effects of pseudoreplication on the reliability of studies and on our ability to draw conclusions from meta-analytic synthesis of these studies, we performed a sensitivity analysis (Noble et al. 2017) by conducting our meta-analyses in three different ways. In the first, we used the number of independent trials (ignoring non-independence of stimuli) as sample size, in the second we used the number of independent stimuli as sample size, and in the third, we used the number of trials but we excluded effects from severely pseudoreplicated studies. In the second case, multiple studies used <4 independent stimuli. Since $Vz = 1/n-3$, n must be >3 to calculate Vz. To generate a Vz value for these cases where $n < 4$, we assigned an $n = 4$. This produces the maximum possible Vz value (1). If pseudoreplication introduces substantial variance into playback experiments, we expected that a meta-analysis using sample sizes based on number of trials would show less convincing effects of moderator variables than our other two forms of analyses which account for pseudoreplication.

We conducted all of the remaining analyses three times; once in each of the three data sets (trials = n, stimuli = n, no severe pseudoreplication). We first estimated the overall average strength of the difference between responses to local versus foreign song. In all tests at this stage and the later stages of analyses, we included study identity and experiment identity nested within study as random effects to account for non-independence of effects within experiments and studies. These random effects do not account for non-independence arising from different measurements taken in the same experiment (e.g. response song rate and approach distance; see Noble et al. 2017). Thus, we modelled this correlated structure as a covariance matrix which also included sampling error variance (i.e., Vz); we set the correlation among response measurements as 0.5 (justification for this method presented in Noble et al. 2017). We also accounted for phylogenetic non-independence as a random effect in all analyses regardless of whether we detected phylogenetic structure in effect size. We used an ultrametric phylogenetic tree obtained from Jetz et al. (2012; http://birdtree.org/; Fig S3) using an Ericson backbone. We converted this ultrametric phylogenetic tree into a correlation matrix which was fitted to the rma.mv function in metafor.

With this basic model, we also estimated the overall heterogeneity (I^2), which is typically considered the proportion of total heterogeneity attributable to variance among studies (Rosenberg 2013). However, because the ratio assumes constant variance within studies which is not reasonable when sample sizes vary, I^2 might be better interpreted as an indicator of inconsistency in effect sizes among studies (Borenstein et al. 2009). Besides an overall estimate of I^2 for each of our three data sets, we also partitioned I^2 among the random effects to determine the extent to which inconsistencies among effect sizes could be attributable to particular sources of variance. For these analyses, we included not only the two random effects described in the previous paragraph, but also a third random effect that we selected for inclusion in tests of our main biological hypotheses (see following two paragraphs, also results). This additional random effect was response variable type (Table 1).

Next, we sought to identify any factors or covariates that might explain variation in effect size but that do not directly relate to our three primary biological hypotheses. Any of these secondary covariates or factors that explained a likely biologically meaningful portion of the variance in effect we retained to include in all subsequent models testing our primary hypotheses. The secondary variables we examined included (1) response variable type (Table 1), (2) whether all effects from a study were reported in sufficient detail for inclusion in meta-analysis, (3) the degree to which foreign song appeared distinct from local song to human observers, and (4) the repertoire size category of individuals within the population.

We tested for an effect of response variable type because different studies quantified the aggressive response of territorial birds to playback of conspecific song in different ways. Most of these response measurements could be readily classified into one of several categories (Table 1). The most common measurements dealt either with the approach to the speaker or with singing behaviour. Approach measurements assumed that closer or more rapid approaches to the speaker meant greater aggression, and the song measurements were typically based on the assumption that more song meant greater aggression. The number of flights and the number of non-song ('call') vocalizations were two other categories of response. Some studies simply noted whether the observer concluded the bird responded, and a sizable number of studies sought to summarize multiple response variables across the above categories by generating some sort of composite score, often with the use of principal component analysis. We had a strong a priori expectation that, at least in some systems, not all categories of response convey the same information. For instance, in some species individuals reduce song rate in response to playback and instead increase flights and approach the speaker (Parra et al. 2017). Thus it may be that song rate is a less reliable measure of aggressive response to playback than approach to speaker.

For our purposes, full reporting of results meant that both the strength and direction of every reported statistical effect could be derived from reported statistical information The relevance of full reporting of all effects lies in the observation that when a portion of results is not fully reported, those under-reported results are often a biased subset and tend to be from weaker, non-significant tests (Parker 2013; Parker et al. 2016). Thus the reported effects in these studies with incomplete reporting should be an upward biased subset, and controlling for this effect should help us detect any other patterns among these effect sizes. We describe our assessment of other forms of publication bias later in this section.

Although most studies in this data set examined systems in which foreign songs were readily distinguishable from local songs by human observers, this was not true of all studies. We classified each pair of studied populations as having a degree of distinctiveness between songs that was (a) weak or apparently absent (n= 14 effects), (b) moderate (n = 89 effects), or (c) strong (n = 276 effects). In most cases, we based our classifications on statements or statistical tests from the published study in question. Systems we classified as being strongly distinctive were those in which a human observer could reliably classify songs according to population of origin (for instance, if a paper reported ‘distinct dialects’), or a statistical method, such as discriminant function analysis, classified the vast majority (> 95%) of songs correctly. We classified distinctiveness as ‘moderate’ if there appeared to be substantial overlap in the distributions of measured song traits between populations or if statistical classifications made mistakes > 5% of the time. When authors reported no evident distinction between populations, or reported that differences were difficult to detect, we classified this as a case of “weak or no” distinctiveness. This classification was subjective in some cases, but it allowed us to test for, and then if needed control for, a relationship between effect size and distinctiveness of local from foreign song. We report the basis for each classification in our archived data (Parker et al. 2018).

Individual repertoire size varies substantially among species, and differences in repertoire size could plausibly influence how individuals respond to different versions of conspecific song. We do not have a strong a priori hypothesis linking average individual repertoire size to variation in response to foreign song, but we tested for an effect here to reduce the chance that we were overlooking an important source of variation that should be controlled in later analyses. For this exploratory analysis, we classified individual repertoires into three categories, ‘one song’ in which most individuals produce a single song type (n = 279 effects), ‘small repertoires’ in which most individuals sing two or three song types (n = 47 effects), and ‘large repertoires’ in which most individuals sing four or more song types (n = 53 effects).

We made decisions regarding the inclusion of the four secondary variables described above prior to conducting our later stages of analyses. For each secondary variable, we considered the strength of our a priori expectation of an effect, as well as the observed effect sizes and the p-values across all three of our approaches to accounting for pseudoreplication. We then decided upon a single set of secondary variables that we would include in tests of all three biological hypotheses with all three approaches to account for pseudoreplication.

Before going on to discuss the tests of our biological hypotheses, we note here that besides testing for effects of incomplete reporting on effect sizes as described above, we also assessed the extent to which publication bias was likely to have influenced the results of our meta-analysis in two other ways. We used Nakagawa and Santos's (2012) modification of 'Egger's regression', which detects asymmetry in funnel plots (Egger et al. 1997). In this test, evidence consistent with publication bias would be an intercept significantly different from zero in the relationship of standardized 'meta-analytic residuals' (representing a combination of within-study effects and their corresponding sampling-error effects) and the inverse of the square root of the meta-analytic variance (Nakagawa and Santos 2012). Egger's regression on meta-analytic residuals is most appropriate because the non-independence is controlled for in residuals and the residuals from the full models are less likely to be influenced by heterogeneity in data. We conducted this test with the two complete data sets (the data set not accounting for pseudoreplication and the data set accounting for pseudoreplication by using number of stimuli as sample size). Further, we conducted trim-and-fill tests, which could detect funnel asymmetry and adjust such asymmetry on the same datasets. Trim and fill is based on the assumption that observed effects derived from samples should be distributed symmetrically around the mean 'true' effect, but with decreasing variability as sample size increases. Trim and fill estimates asymmetry and the number of apparently 'missing' studies that would need to be added to eliminate asymmetry (Duval and Tweedie 2000a; Duval and Tweedie 2000b).

Our later stages of analyses were devoted to testing our three primary biological hypotheses. Our first set of models assessed the hypothesis that selection to minimize response to heterospecific song leads populations to narrow their template for conspecific song recognition and that this narrow template also limits response to foreign conspecific song. To test this hypothesis, we assumed that a greater number of sympatric heterospecifics from similar species would mean greater selection to avoid responding to heterospecific song. We used three different indices of the number of sympatric heterspecifics. In model 1, we used the number of sympatric congeners present at the playback study location during the breeding season as determined by current taxonomy as used by ebird and range maps of those congeners at the study location generated on ebird (eBird 2012). In model 2, we used a local relatives score, in which we assigned a site a score of '2' if at least one congener was present locally during the breeding season, a '1' if no congeners, but at least one confamilial was present locally during breeding, and a '0' if no congeners or confamilials were present locally during the breeding season. In model 3, we wished to capture variability in overall bird species richness among sites. To do this, we used both latitude and whether or not the site was on a remote (non-continental-shelf) oceanic island. Avian species richness is generally much greater at lower latitudes, and is much lower on isolated oceanic islands, but the slope of the richness relationship does not differ between islands and mainlands (Kalmar and Currie 2007), and so we did not fit an interaction term between island and latitude. Finally, model 4 included all of the above variables together.

Our second primary hypothesis was that experience interacting with conspecifics singing foreign song increases the tendency for individual to respond to foreign conspecific song. Thus when exposure to foreign songs is common, aggressive responses to foreign song should more closely match responses to local songs and effect sizes in our meta-analysis should be smaller for populations that have more experience with the foreign song in question or more experience with foreign songs in general. In our first test of this hypothesis (model 5), we included two predictor variables, the geographic distance between the study site and the location where the foreign songs were recorded, and whether the foreign song came from an isolated population (disjunct from the focal population). Because the effect of distance should be strongest when two populations are connected and therefore have a possible conduit for cultural transmission, we also included an interaction term between isolation and distance. This analysis used geographic distance as a proxy for the probability of interaction across cultures. We used the base 10 log of distance because we expect that the strongest effects of distance would be over small and medium distances, and that among studies that considered foreign song recorded a long distance away, small differences in distances among studies would show little effect. In model 6 we included a factor that we termed ‘movement’. For this variable, we classified all populations as (1) defending territories year-round (n = 104 effects), (2) dissolving or partially dissolving territories during the non-breeding season but not migrating out the local area (n = 165 effects), (3) short distance migrants (typically <1000 km and within a continent; n = 54 effects), or (4) long distance migrants (typically >1000 km and to different continent; n = 56 effects). In model 7 we assessed movement, but excluded cases where the foreign population was disjunct, and thus movement would be unlikely to facilitate cross-population contact. In this model, we again included log 10 of distance between focal and foreign song, and we included an interaction term between movement and distance because the effects of distance should differ based on the type of movement common in the population. Finally, model 8, with the full data set, included isolation, distance, movement, and both an interaction between isolation and distance and an interaction between movement and distance.

Our third major hypothesis was that as populations diverge evolutionarily, they become less likely to recognize each other’s signals and thus discrimination is a function of evolutionary divergence. To test this hypothesis, in model 9 we compared effect sizes between studies in which the foreign song was recorded from within the same sub-species to effect sizes from studies in which the foreign song was recorded from a different sub-species. We predicted weaker responses to foreign songs (and thus stronger effect sizes) when the song came from different subspecies.

Finally, we tested a global model (model 10) that included all of the predictors from the prior nine models testing our three primary biological hypotheses.

RESULTS

The overall effect of local versus foreign playback was moderate and highly statistically significant regardless of the how we accounted, or did not account, for pseudoreplication (Table 2). Territorial oscines directed more aggression towards playback of local songs than towards playback of foreign conspecific songs (Fig 1). Our estimates of total heterogeneity (I^2) varied dramatically depending on how we dealt with pseudoreplication (Table 3). When we ignored pseudoreplication and used trial number as our sample size, our analysis indicated high heterogeneity, but when we instead used the number of stimuli as our sample size, the analysis indicated relatively low heterogeneity. We found intermediate levels of heterogeneity when we excluded severely pseudoreplicated studies and set sample size equal to trial number. When we portioned heterogeneity among the random effects in our models, we found somewhat different patterns for the three different methods of accounting for pseudoreplication. When using trial number as sample size, almost half of the observed heterogeneity appeared to stem from unexplained variability among studies, but this dropped to 5% when we used the number of stimuli as sample size, and to basically 0% when we excluded cases of severe pseudoreplication. In contrast, when including all effect sizes in the analyses, no heterogeneity was associated with phylogeny, but 21% of observed heterogeneity was attributable to phylogeny when we eliminated severely pseudoreplicated studies from the analyses. Similarly, almost none of the observed heterogeneity could be attributed to variation among response variable categories when we included all effect sizes in our analyses, but 8% of heterogeneity was attributable to response variable category when we excluded effects from severely pseudoreplicated trials (Table 3).

Our results show that pseudoreplication of stimuli, especially when it is severe, elevates sampling variance. Our F-test to compare variance found significantly more variance among severely pseudoreplicated studies than among studies lacking pseudoreplication or among studies with moderate pseudoreplication (Table 4). Moderately pseudoreplicated studies did not have elevated variance relative to non-pseudoreplicated studies (Table 4). Further, as described below, the clearest and strongest methodological and biological effects appeared stronger when accounting for pseudoreplication than when ignoring it. Finally, we see that the funnel plot based on trial number as sample size, ignoring pseudoreplication, does not resemble a funnel because of elevated variation among effect size estimates associated with large numbers of (mostly pseudoreplicated) trials (Fig 2).

Our assessment of moderator variables was similar among the three groups of analyses (Tables 5, S1, S2). As we expected, those studies reporting a full set of results (either in the original paper or in response to our request for additional information), had a lower average effect size than those from which results were missing (Tables 5, S1, S2; Fig 1). This effect was moderately strong and thus justified the inclusion of this variable as a fixed effect in later tests of biological hypotheses. Once controlling for this effect of underreporting, we found no evidence of additional publication bias using the Egger's regression (funnel asymmetry; trial number as n: $t = -0.419$, $df = 377$, $p = 0.675$; stimuli number as n: $t = -1.496$, $df = 377$, p = 0.136). This result was corroborated by the result from the trim-and-fill test (estimated missing studies; trial number as n: 0 ± 1.41, p = 0.5; stimuli number as n: 0 ± 1.41, p = 0.5; for the test for no missing studies on the left side of the funnel for both meta-analyses; Figs 2-3).

The average magnitude of statistical effects differed dramatically among the response variable categories. However, there was substantial variance within categories, and so even the large differences among categories were not statistically significant. Because the differences in magnitude of effect were so large, we decided to include response variable category as a random effect in later tests of biological hypotheses. Also, these large differences in magnitude of effect led us to explore further the role of response variable in influencing effect size. Because we had hypothesized a priori that vocal responses to playback might be less consistent than whole-body movement responses, we decided to conduct a more powerful post hoc test for this effect. We only conducted this test using number of stimuli as sample size. For this analysis, we lumped the response variable categories into 'vocal', which included both calling and song, and 'movement', which included any measurement that was at least partially a function of movement or approach. This meant 'movement' included all response variable categories besides calling and song. This post hoc analysis significantly supported the hypothesis that vocalizations are a moderately less reliable way to measure response to playback than indices of whole-body movement, such as approach to the speaker and number of flights (Table 6; Fig 1).

The evident distinctiveness between local and foreign song positively predicted the strength of the reported effect size. However, this effect only exceeded 0.1 and the p-value only dropped below 0.05 in the analysis that did not control for pseudoreplication (Tables 5, S1, S2). Although the reality and importance of this effect must remain uncertain, we decided to include this variable as a covariate in our later analyses because the effect was relatively consistent and was in the predicted direction. It is worth noting, however, that in the subsequent analyses, the effect of the distinctiveness between local and foreign song was typically weaker than in the initial tests.

As repertoire size increased, effect sizes tended to become larger, but the effect was weak and not significant (Tables 5, S1, S2). Further, this effect was not strongly predicted a priori. Because of the poor explanatory value of this variable, we did not include it in any later tests of biological hypotheses.

We found mixed support for the hypothesis that selection against response to heterospecific song constrains the breadth of the song recognition template, and therefore limits response to foreign conspecific song (models 1-4, Table 7, S3, S4). When not accounting for pseudoreplication, all effect sizes were < 0.11 and typically much smaller, and no p-values approached significance (Table S3). When controlling for pseudoreplication either by using number of stimuli as the sample size or by excluding effects from severely pseudoreplicated experiments, effect sizes were similar, but there was a consistent effect (Z = 0.096-0.140) of local relatives score that was modestly significant (p = 0.015 to 0.030, models 2, 4). Populations that co-occurred with congeners or confamilials had a weak tendency to respond less strongly to foreign conspecific song than did populations with no confamilials present (Fig 1).

Interpretation of this effect requires caution, especially in light of the absence of an effect of the number of locally sympatric congeners.

We found no support for the hypothesis that exposure to foreign conspecifics increases the strength of response to foreign conspecific song. Nearly all of our selected indices of exposure to foreign songs showed weak and non-significant effects (models 5-8, Table 8, S5, S6). The only effect associated with p-value < 0.05 was of the opposite sign to the prediction. In this one model (in which the analysis excluded both severely pseudoreplicated effects and studies in which the population singing the foreign song was disjunct from the local population) there was a modest trend ($Z = 0.109$, $p = 0.035$, model 8, Table S6) but it was in the direction opposite to that predicted, with decreasing strength of response to foreign songs associated with greater population movement and thus greater expected exposure to foreign song.

Our analyses revealed strong support for the hypothesis that evolutionary divergence drives divergence in response to song (model 9, Table 9, S7). When a foreign song was recorded from a different subspecies, it elicited substantially weaker responses than foreign song recorded from within the same subspecies (Fig 1). This effect was most evident when accounting for pseudoreplication though it did not differ significantly in strength among our analyses.

Analyses of global models, in which all variables from the tests of our three biological hypotheses were combined, did not reveal any strong patterns not already evident in the simpler models (model 10, Table 10, S8).

DISCUSSION

We found an unambiguous pattern across a wide range of songbird species for territorial individuals to respond more strongly to playback of local conspecific song than to playback of foreign conspecific song. However, there was heterogeneity in this effect among studies, and we were able to identify some sources of this variability, both biological and methodological. We found strong evidence that classification into distinct subspecies is associated with reduced response to conspecific song from across the subspecific divide; aggressive response to foreign conspecific song was lower on average if the foreign song was from a different subspecies than if it was from the same subspecies. We also found suggestive evidence that the presence of closely related species may drive increased song discrimination, and thus reduced response to foreign conspecific song, though the lack of concordance among related variables weakens this conclusion. We found no support for the hypothesis that the strength of response to foreign conspecific song is a function of experience with foreign song. In contrast, we found compelling evidence of three important methodological impacts on inferences from playback studies. First, we found that when authors reported full results from only a subset of statistical effects, these reported effects tended to be larger than the average effects from studies with a complete set of results. Second, we found more variable, and thus less reliable, results from experiments in which a small number of stimuli were used repeatedly in a large number of trials. Further, if we did not account for this pseudoreplication, the effects of the local relatives score and of evolutionary divergence were less convincing. Finally, we found that when response to playback was measured at least in part as movement of the focal individual, response differences between local and foreign song were stronger on average than response differences measured only as vocalization of the focal individual.

Biological implications

The moderately consistent and highly significant tendency for territorial songbirds to respond more strongly to playback of local song than to playback of foreign song raises important questions about why birds fail to respond strongly to foreign songs. For instance, is this a failure in recognition due entirely to insufficient experience with the stimuli (Colbeck et al. 2010), or has selection acted directly to reduce response to foreign song because individuals singing foreign song pose limited territorial threats (Nicholls 2008; Tomback et al. 1983; Turcokova et al. 2011)? Of course, as weakly suggested by our meta-analysis, indirect selection to avoid response to heterospecifics could also play a role in limiting response to foreign song (Amézquita et al. 2011; Hamao 2016; Symes 2014). This is a rich area for future empirical work.

The tendency for experiments to often reveal stronger response to local than to foreign song has long been recognized. Our quantification of this effect confirms prior qualitative understanding and provides a useful benchmark against which to compare future findings. It also allows for the robust rejection of any lingering hypotheses that predict other patterns, such as the deceptive mimicry hypothesis that posited that immigrant individuals match local songs to appear local and thus reduce aggression directed at them from established local individuals (Payne 1981; Payne 1983; Rohwer 1982).

The relatively large reduction in the strength of response to foreign song from different subspecies is interesting for several reasons. First, response to song is sometimes used as evidence when assessing evolutionary divergence (Alström et al. 2008; Matessi et al. 2001; Randler et al. 2012). Before we conducted our analyses, it was not clear to us that we would find an effect of subspecies status because many species show much weaker response to foreign song than to local song even if the foreign song is from within the same subspecies, response to other subspecies' song is sometimes strong (Parker et al. 2018), and oscine songbirds appear slow to evolve the ability to show differential response to songs of closely related heterospecific lineages (Freeman et al. 2017). Despite the convincing difference in average effects we found between playback from within subspecies and playback across subspecies, there is enough variation in effect size even with large samples (Fig 3) to suggest that strength of response to foreign song should not be used as primary evidence when assessing sub-species status. Recent work provides some insights into both why we might expect generally weaker response to song from different subspecies, and also why we might expect substantial variability within subspecies. Songs diverge in acoustic space as populations diverge (Freeman and Montgomery 2017). The extent to which this divergence is due to drift (Goodfellow and Slater 1986; Lachlan and Slater 2003; Williams and Slater 1990) or to selection from divergent acoustic environments (Baker 2006; Kirschel et al. 2009; Slabbekoorn and Smith 2002) remains an open question. There is good evidence, however, that song recognition templates eventually develop or evolve to track these changes in song (Freeman et al. 2017), but this only happens reliably when the acoustic differences become large (Freeman and Montgomery 2017). Presumably, if our meta-analysis had included studies of responses to heterospecific congener song, we would have observed even weaker responses to those songs than to songs from conspecifics belonging to different subspecies.

If subspecies classification typically represents evolutionary divergence, then we expect songs to have diverged more consistently between subspecies than among populations within a subspecies. In fact, we judged local and foreign song to be ‘strongly’ distinguishable from each other in 15 of 17 (88%) of cases when the foreign song was from a different subspecies, but we judged only 38 of 56 (68%) to be ‘strongly’ distinguishable when the foreign song was from the same subspecies. At first glance, some of our other results appear to provide a counter example. We observed only a weak relationship (at best) of song distinctiveness and differential response to foreign versus local song in playback studies. However, this turns out to be an unconvincing counter example because our index of song distinctiveness does not account for absolute degree of acoustic divergence, but rather consistency of divergence. In other words, if songs from two populations were very similar acoustically, but the existing acoustic divergence were consistent between populations such that individuals could be reliably classified by population based on song, we would consider that strongly distinctive. In contrast average acoustic divergence could be large between populations, but if variance were high such that some songs from different populations appeared similar, this could be classified as only moderately distinctive. It seems likely that in our data set degree of acoustic divergence between population pairs is highly variable both within and across subspecies. Thus it remains possible that the variance in effects that we observed may be due to variance in acoustic divergence (Freeman and Montgomery 2017). Future work quantifying acoustic divergence, in terms of both degree and variability, between the populations in our meta-analysis would be informative, but gaining access to the playbacks used by all researchers or acquiring new recordings from all focal and treatment populations was well beyond the scope of this paper. We also could benefit from molecular data regarding the degree of evolutionary divergence between local populations and the populations from which foreign song was recorded. Subspecies status may not consistently represent the degree of evolutionary divergence, and so our conclusions about the effects of evolutionary divergence on response to foreign song should be considered tentative. Unfortunately the necessary molecular data are not available for most of our population pairs, and so this analysis must also be left for the future. We can, however, rule out the possibility that our subspecies effect was due to greater distance between the playback site and the recording location of the foreign song for populations in different subspecies since we found no relationship between this distance and effect size in our meta-analysis.

Our lack of strong support for a role of either selection to avoid response to heterospecifics or experience with foreign song to explain variability in response to foreign conspecific song is not, unfortunately, strong evidence against these mechanisms. In testing the hypothesis that the risk of responding to heterospecific song selects for a narrow template of conspecific song recognition, and thus reduced response to foreign conspecific song, we worked from an assumption that the degree of local coexistence with other bird species was a suitable index of the strength of selection to avoid responding to heterospecific song. All else being equal, more other species present should mean more occupied acoustic space and a greater chance of similarity with other species. However, this might not be true if, for instance, sound transmission obstacles in some environments constrain the acoustic space (Kirschel et al. 2009; Slabbekoorn and Smith 2002) more dramatically than in other environments or other factors drive non-random occupancy of acoustic space (Wilkins et al. 2013). Ideally, we would have multivariate measures of acoustic similarity between the focal species and the other species in the local community at each study site. However, at present that is a monumental task and so must remain a challenge for the future. Interestingly, we found some evidence in support of the heterospecific selection hypothesis when we used our local relatives score (2 = congenerics present, 1 = confamilials [but no congenerics] present, 0 = no confamilials present). To the extent that the degree of acoustic similarity among songs is a decreasing function of species relatedness, this may be our best index. However, there was no difference in response between birds at sites with congenerics present and those at sites with only confamilials present, and if anything, the effect was stronger at sites with only confamilials (Fig 1), which is counter to prediction. Further, we found no relationship between effect size and the number of congeners present, which is surprising since presumably as the number of congeners present increased, the odds of the presence of a similar song would also increase. Our other analyses seeking to capture effects of selection to minimize response to heterospecific song assumed that latitude and island isolation are sufficient indices of local species richness. Latitude is not a perfect index for richness, in part because latitude itself is an index of potential causal factors such as plant productivity that are imperfect predictors of avian richness. However, avian richness within locations differs dramatically across major differences in latitude, and so variation in latitude captures a sizable portion of variation in avian richness (Hillebrand 2004), and we suspect that this is not a major obstacle to effectiveness of our analyses. Likewise, species richness is much lower on isolated islands (MacArthur and Wilson 1967), and so we feel confident in our use of this variable as an index of richness.

Our indices of potential exposure to foreign song are also imperfect, and tests of the exposure hypothesis would benefit from detailed natural history knowledge about each studied population (Parra et al. 2017). Ideally, we would know the spatial scale over which vocal culture changes, the likelihood of individual movements at those scales, and the correspondence between those movements and seasonal variation in singing. For instance we assume that movement outside of the breeding season corresponds to exposure to foreign song. This is certainly true in some species (DeWolfe and Baptista 1995; Parra et al. 2017), but in some others singing may be too limited outside of the breeding season to provide migrants with meaningful exposure to diverse conspecific signals. As with the acoustic characterization of the communities in which all these experiments were conducted, obtaining the extensive knowledge needed across a large number of populations is a major undertaking that we must leave to future researchers.

All meta-analyses are constrained by available data, and in our case, available data were skewed towards studies of North temperate species, especially in Europe and North America. This skew means that males were the primary territorial respondents to playback in many of our studies, though female contributions to territorial defence are common and substantial in other parts of the world. This means that we should be cautious when generalizing beyond the species examined here. That said, we did have data from species elsewhere in the world, and studies that included territorial responses by females also supported the basic findings of our meta-analysis of greater aggression towards local song. We likely further reduced the equity in our geographic representation with our decision to exclude non-oscines as the playback studies of these species have often been performed in the Neotropics. However, the potentially dramatic differences in song learning between oscines and suboscines (Touchton et al. 2014) seemed likely to create more heterogeneity in effects than we expected from differences in geographic location within oscines, and this concern has been supported by work published since we completed our analyses (Freeman et al. 2017). Fortunately, recent studies using playback of bird song (Freeman and Montgomery 2017; Freeman et al. 2017) have dramatically increased availability of data from outside Europe and North America.

Methodological implications

Of the three methodological effects we observed, the inflated effect sizes from the studies with incomplete reporting has the broadest implication. There is strong evidence that incomplete reporting of results is widespread across sub disciplines in behaviour, ecology, and evolutionary biology (Parker et al. 2016), and there is good reason to expect that incompletely reported results are often a biased subset (Parker 2013). The results from our meta-analysis of song playback studies show this effect quite clearly and so demonstrate the need for transparency in reporting of results to facilitate clear interpretation of results and effective meta-analysis (Gerstner et al. 2017).

One of the more striking results of our study is the higher variability in observed effect size from experiments with severe pseudoreplication of song stimuli. When researchers used the same few stimulus songs repeatedly in much larger numbers of trials that were otherwise independent, they often found either an unusually strong or an unusually weak difference in response to foreign versus local song relative to studies in which large numbers of different stimuli were used. This effect was also evident in the estimates of heterogeneity among studies. When we did not account for pseudoreplication, we observed higher heterogeneity in effect size among studies. This was in part because when we excluded severely pseudoreplicated studies, we excluded studies with substantial sampling variance, and thus eliminated considerable unexplained variance. Second, when we used stimulus number instead of trial number as the sample size, this reduced the per-study sample sizes of many studies (often dramatically) and since within-study variance in meta-analysis is inversely proportional to sample size, this increased within-study variances. Because within study variance increased, among-study variance became a smaller portion of overall variance. Our partitioning of heterogeneity among random effects also provides hints of the effect of pseudoreplication. Heterogeneity associated with study identity was higher when not accounting for pseudoreplication than in other analyses, presumably because pseudoreplication was a characteristic of individual studies.

When testing biological hypotheses with our meta-analyses, if we ignored pseudoreplication, we found weaker evidence for some of the patterns we observed. For instance, the pattern of reduced response to foreign songs from different subspecies relative to foreign songs from the same subspecies was less convincing when ignoring pseudoreplication, as was our tentative conclusion that populations with closer relatives present respond less strongly to foreign conspecific songs. Ignoring pseudoreplication also made it more difficult to detect the effect of incomplete reporting on the size of reported effects. In other words, the elevated sampling variance associated with pseudoreplication of song stimuli masked our ability to detect meaningful variation among experimental effects from song playback experiments. These results indicate that a small number of 'representative' songs used as stimuli within a treatment group (e.g., one or a few 'foreign' or 'local' songs) fail to actually represent the variability in stimuli that characterize the population from which the stimulus was drawn. Thus results from these strongly pseudoreplicated studies poorly represent the true range of biological effects. This is clear empirical support for the recommendations to avoid pseudoreplication of stimuli in playback experiments made several decades ago based on statistical principles (Kroodsma 1989; McGregor et al. 1992). Although pseudoreplication of stimuli in playback experiments declined in response to the initial calls for improved methods (Kroodsma et al. 2001), researchers continue to conduct playback studies with severe pseudoreplication of stimuli (see, Parker et al. 2018), and so we hope that our result demonstrating the poor reliability of such studies will help finally to end this practice.

Our analyses suggest that, across oscine species, the most consistently useful methods of measuring response to song playback are those tracking body movements. The differential response to local versus foreign song was considerably weaker when measured by calling or singing rate than by variables that capture body movement such as approach distance or flight rate. We expect that this is not because song is consistently a poor indicator of aggression, but rather that in some species elevated singing rate is not a primary response to detection of a territorial intrusion, and that in these species, direct approach, possibly with the intent to chase the rival from the territory, is the primary response (Parra et al. 2017). Thus we suggest that future playback studies, especially those working in unfamiliar systems, use measurements of approach to the speaker as the index of response.

Conclusions

The song playback trials conducted in oscine songbirds provide overwhelming evidence that territorial responses tend to be stronger to local conspecific songs than to foreign conspecific songs. Although we found evidence that evolutionary divergence within species, as indicated by subspecific status, is associated with reduced response to conspecific song, we were unable to detect any other strong patterns that might explain how the window of song recognition develops or evolves. Further work with considerably more detailed study of song structures and individual experience will be needed to make progress with this question. Although substantial work remains before we can answer our primary biological question, our meta-analysis provided robust insights into several methodological issues. We showed the importance of thorough reporting of all statistical results, of using large numbers of independent stimuli in playback trials, and in measuring behavioural response to playback by observing movement of the focal individual.

ACKNOWLEDGMENTS

We would like to thank the following authors for sharing information relevant to our analysis: C. Branch, H. Lampe, K. Mortega, D. Nelson, T. Osiejuk, M. Packert, J. Podos, L. Ratcliffe, P. Slater, D. Tietze. Our manuscript was improved by suggestions from anonymous reviewers and T. Freeberg. This work was completed while T. Parker was on a sabbatical supported by Whitman College. S. Griffith hosted T. Parker at Macquarie University for a portion of this sabbatical. The authors declare no conflicts of interest.

T. Parker conceived of the study, developed hypotheses, checked or conducted all data extraction from papers used in meta-analysis, wrote some R code, conducted most data analyses, wrote much of the manuscript, and edited the manuscript. E. Greig developed hypotheses, obtained much of the natural history information relevant to study locations and species, wrote portions of the manuscript, and edited the manuscript. S. Nakagawa designed or helped design all major analyses, wrote most R code, wrote portions of the manuscript, and edited the manuscript. M. Parra and A. Dalisio identified and obtained relevant papers and extracted data from papers during preliminary work on this project and were essential to initiation of the project.

REFERENCES

Alström P, Olsson U (1999) The golden-spectacled warbler: a complex of sibling species, including a previously undescribed species. Ibis 141:545-568 doi:10.1111/j.1474-919X.1999.tb07363.x

Alström PER, Rasmussen PC, Olsson U, Sundberg PER (2008) Species delimitation based on multiple criteria: the spotted bush warbler *Bradypterus thoracicus* complex (Aves: Megaluridae). Zoological Journal of the Linnean Society 154:291-307 doi:10.1111/j.1096-3642.2008.00418.x

Amézquita A, Flechas SV, Lima AP, Gasser H, Hödl W (2011) Acoustic interference and recognition space within a complex assemblage of dendrobatid frogs. Proc Natl Acad Sci USA 108:17058-17063

Baker MC (1982) Vocal dialect recognition and population genetic consequences. Am Zool 22:561-569

Baker MC (1994) Loss of function in territorial song: comparison of island and mainland populations of the singing honeyeater (*Meliphaga virescens*). Auk 111:178-184

Baker MC (2006) Differentiation of mating vocalizations in birds: acoustic features in mainland and island populations and evidence of habitat-dependent selection on songs. Ethol 112:757-771 doi:10.1111/j.1439-0310.2006.01212.x

Baker MC, Thompson DB, Sherman GL, Cunningham MA (1981) The role of male vs male interactions in maintaining population dialect structure. Behav Ecol Sociobiol 8:65-69

Baker MC, Tomback DF, Thompson DB, Theimer TC, Bradley DC (1984) Behavioral consequences of song learning: discrimination of song types by male white-crowned sparrows. Learning and Motivation 15:428-440 doi:http://dx.doi.org/10.1016/0023-9690(84)90007-9

Balaban E (1988) Cultural and genetic variation in swamp sparrows (*Melospiza georgiana*) 2. Behavioral salience of geographic song variants. Behaviour 105:292-322 doi:10.1163/156853988x00052

Becker PH (1977) Geographische variation des gesanges von winter- und sommergoldhāhnchen (*Regulus regulus*, *R. ignicapillus*). Vogelwarte 29

Beecher MD, Brenowitz EA (2005) Functional aspects of song learning in songbirds. Trends Ecol Evol 20:143-149

Bjerke TK (1984) The response of male redwings *Turdus iliacus* to playback of different dialects. Fauna Norvegica Ser C, Cinclus 7:24-27

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Introduction to meta-analysis. John Wiley & Sons, Chichester, UK

Bradley DW, Molles LE, Waas JR (2013) Local–foreign dialect discrimination and responses to mixed-dialect duets in the North Island kōkako. Behav Ecol 24:570-578 doi:10.1093/beheco/ars201

Branch CL, Pravosudov VV (2016) Do male mountain chickadees discriminate between local and non-local elevation intruders? Ethol 122:376-388 doi:10.1111/eth.12487

Briefer E, Rybak F, Aubin T (2010) Are unfamiliar neighbours considered to be dear-enemies? PLoS ONE 5:e12428 doi:10.1371/journal.pone.0012428

Brooks RJ, Falls JB (1975) Individual recognition by song in white-throated sparrows. I. Discrimination of songs of neighbors and strangers. Can J Zool 53:879-888 doi:10.1139/z75-101

Colbeck GJ, Sillett TS, Webster MS (2010) Asymmetric discrimination of geographical variation in song in a migratory passerine. Anim Behav 80:311-318 doi:10.1016/j.anbehav.2010.05.013

Danner JE, Danner RM, Bonier F, Martin PR, Small TW, Moore IT (2011) Female, but not male, tropical sparrows respond more strongly to the local song dialect: implications for population divergence. Am Nat 178:53-63 doi:10.1086/660283

Derryberry EP (2011) Male response to historical and geographical variation in bird song. Biol Lett 7:57-59 doi:10.1098/rsbl.2010.0519

DeWolfe BB, Baptista LF (1995) Singing behavior, song types on their wintering grounds and the question of leap-frog migration in Puget Sound white-crowned sparrows. The Condor 97:376-389 doi:10.2307/1369024

Dingle C, Poelstra JW, Halfwerk W, Brinkhuizen DM, Slabbekoorn H (2010) Asymmetric response patterns to subspecies-specific song differences in allopatry and parapatry in the gray-breasted wood-wren. Evolution 64:3537-3548 doi:10.1111/j.1558-5646.2010.01089.x

Duval S, Tweedie R (2000a) A nonparametric ‘‘trim and fill’’ method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association 95:89-98

Duval S, Tweedie R (2000b) Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56:455-463

eBird (2012) eBird: An online database of bird distribution and abundance [web application]. eBird. http://www.ebird.org. Accessed March, 2017 2017

Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. Br Med J 315:629-634

Freeman BG, Montgomery GA (2017) Using song playback experiments to measure species recognition between geographically isolated populations: a comparison with acoustic trait analyses. Auk:857-870 doi:10.1642/AUK-17-63.1

Freeman BG, Montgomery GA, Schluter D (2017) Evolution and plasticity: divergence of song discrimination is faster in birds with innate song than in song learners in Neotropical passerine birds. Evolution 71:2230-2242 doi:10.1111/evo.13311

Gerstner K, Moreno-Mateos D, Gurevitch J, Beckmann M, Kambach S, Jones HP, Seppelt R (2017) Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. Methods in Ecology and Evolution 8:777-784 doi:10.1111/2041-210X.12758

Godard R (1993) Tit for tat among neighboring hooded warblers. Behav Ecol Sociobiol 33:45-50 doi:10.1007/BF00164345

Goodfellow DJ, Slater PJB (1986) A model of bird song dialects. Anim Behav 34:1579-1580

Grant BR, Grant PR (2002a) Lack of premating isolation at the base of a phylogenetic tree. Am Nat 160:1-19 doi:10.1086/339987

Grant BR, Grant PR (2002b) Simulating secondary contact in allopatric speciation: an empirical test of premating isolation. Biol J Linnean Soc 76:545-556 doi:10.1046/j.1095-8312.2002.00076.x

Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. Ecology 80:1142-1149

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R Package. J Stat Softw 33:1-22

Hadfield JD, Nakagawa S (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology 23:494-508 doi:doi:10.1111/j.1420-9101.2009.01915.x

Hamao S (2016) Asymmetric response to song dialects among bird populations: the effect of sympatric related species. Anim Behav 119:143-150 doi:http://dx.doi.org/10.1016/j.anbehav.2016.06.009

Hansen P (1984) Neighbor-stranger song discrimination in territorial yellowhammer *Emberiza citrinella* males, and a comparison with responses to own and alien song dialects. Ornis Scandinavica 15:240-247 doi:10.2307/3675932

Hansson MC, Bensch S, Brannstrom O (2000) Range expansion and the possibility of an emerging contact zone between two subspecies of Chiffchaff *Phylloscopus collybita ssp*. J Avian Biol 31:548-558 doi:10.1034/j.1600-048X.2000.1310414.x

Harris MA, Lemon RE (1974) Songs of song sparrows: reactions of males to songs of different localities. Condor 76:33-44 doi:10.2307/1365982

Hillebrand H (2004) On the generality of the latitudinal diversity gradient. Am Nat 163:192-211 doi:10.1086/381004

Horváthová T, Nakagawa S, Uller T (2012) Strategic female reproductive investment in response to male attractiveness in birds. Proc R Soc Lond Ser B-Biol Sci 279:163-170

Hudson EJ, Price TD (2014) Pervasive reinforcement and the role of sexual selection in biological speciation. Journal of Heredity 105:821-833 doi:10.1093/jhered/esu041

Hyman J (2005) Seasonal variation in response to neighbors and strangers by a territorial songbird. Ethol 111:951-961 doi:10.1111/j.1439-0310.2005.01104.x

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. Nature 491:444-448 doi:http://www.nature.com/nature/journal/v491/n7424/abs/nature11631.html#supplementary-information

Kalmar A, Currie DJ (2007) A unified model of avian species richness on islands and continents. Ecology 88:1309-1321 doi:10.1890/06-1368

Kirschel ANG, Blumstein DT, Cohen RE, Buermann W, Smith TB, Slabbekoorn H (2009) Birdsong tuned to the environment: green hylia song varies with elevation, tree cover, and noise. Behav Ecol 20:1089-1095 doi:10.1093/beheco/arp101

Koetz-Trowse AH, Westcott DA, Congdon BC (2012) Discrimination of song dialects in relation to song similarity and geographical distance in a rainforest passerine. Emu 112:189-198 doi:10.1071/mu11056

Kreutzer M (1974) Stereotypie et variations dans les chants de proclamation territoriale chezle troglodyte (*Troglodytes troglodytes*). Revue du Comportement Animal 8:270-286

Kreutzer M (1987) Reactions of cirl buntings (*Emberiza cirlus*) to playback of an atypical natural song: the use of own and neighbors' repertoires for song recognition. Journal of Comparative Psychology 101:382-386

Kroodsma DE (1989) Suggested experimental designs for song playbacks. Anim Behav 37:600-609 doi:10.1016/0003-3472(89)90039-0

Kroodsma DE, Byers BE, Goodale E, Johnson S, Liu W-C (2001) Pseudoreplication in playback experiments, revisited a decade later. Anim Behav 61:1029-1033 doi:http://dx.doi.org/10.1006/anbe.2000.1676

Kroodsma DE, Meservey WR, Whitlock AL, Vanderhaegen WM (1984) Blue-winged warblers (*Vermivora pinus*) recognize dialects in type II but not type I songs. Behav Ecol Sociobiol 15:127-131 doi:10.1007/bf00299380

Lachlan RF, Slater PJB (2003) Song learning by chaffinches: how accurate, and from where? Anim Behav 65:957-969

Lampe HM, Baker MC (1994) Behavioral response to song playback by male and female white-crowned sparrows of two subspecies. Bioacoustics 5:171-185 doi:10.1080/09524622.1994.9753243

Latruffe C, McGregor PK, Tavares JP, Mota PG (2000) Microgeographic variation in corn bunting (*Miliaria calandra*) song: quantitative and discrimination aspects. Behaviour 137:1241-1255

Leader N, Wright J, Yom-Tov Y (2002) Dialect discrimination by male orange-tufted sunbirds (Nectarinia osea): Reactions to own vs. neighbor dialects. Ethol 108:367-376 doi:10.1046/j.1439-0310.2002.00771.x

Leader N, Wright J, Yom-Tov Y (2005) Acoustic properties of two urban song dialects in the orange-tufted sunbird (*Nectarinia osea*). Auk 122:231-245

Lemon RE (1967) The response of cardinals to songs of different dialects. Anim Behav 15:538-545 doi:10.1016/0003-3472(67)90056-5

Lemon RE, Harris M (1974) The question of dialects in the songs of white-throated sparrows. Can J Zool 52:83-98 doi:10.1139/z74-013

Liu IA, Lohr B, Olsen B, Greenberg R (2008) Macrogeographic vocal variation in subspecies of swamp sparrow. Condor 110:102-109 doi:10.1525/cond.2008.110.1.102

MacArthur R, Wilson EO (1967) The Theory of Island Biogeography. Princeton University Press, Princeton, NJ, USA

Martens J, Meincke C (1989) Territorial song of the siberian chiffchaff (*Phylloscopus collybita tristis*) and playback experiments within a central European population (*Ph. c. collybita*). Journal für Ornithologie 130:455-473 doi:10.1007/bf01918465

Matessi G, Dabelsteen T, Pilastro A (2001) Subspecies song discrimination in a Mediterranean population of the reed bunting Emberiza schoeniclus intermedia. Italian Journal of Zoology 68:311-314

Matthysen E (1997) Geographic variation in the occurrence of song types in Nuthatch *Sitta europaea* populations. Ibis 139:102-106 doi:10.1111/j.1474-919X.1997.tb04508.x

McGregor PK (1983) The response of corn buntings to playback of dialects. Z Tierpsychol 62:256-260 doi:10.1111/j.1439-0310.1983.tb02155.x

McGregor PK et al. (1992) Design of playback experiments: the Thornbridge Hall NATO ARW Consensus. In: McGregor PK (ed) Playback and Studies of Animal Communication. Plenum Press, New York, pp 1-9

Milligan MM, Verner J (1971) Inter-populational song dialect discrimination in the white-crowned sparrow. The Condor 73:208-213 doi:10.2307/1365840

Moher D, Liberati A, Tetzlaff J, Altman DG, The PG (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. PLOS Medicine 6:e1000097 doi:10.1371/journal.pmed.1000097

Mortega KG, Flinks H, Helm B (2014) Behavioural response of a migratory songbird to geographic variation in song and morphology. Frontiers in Zoology 11 doi:10.1186/s12983-014-0085-6

Morton ES, Howlett J, Kopysh NC, Chiver I (2006) Song ranging by incubating male blue-headed vireos: the importance of song representation in repertoires and implications for song delivery patterns and local/foreign dialect discrimination. J Field Ornithol 77:291-301 doi:10.1111/j.1557-9263.2006.00055.x

Nakagawa S, Santos EA (2012) Methodological issues and advances in biological meta-analysis. Evol Ecol 26:1253-1274 doi:10.1007/s10682-012-9555-5

Nelson DA (1998) Geographic variation in song of Gambel's white-crowned sparrow. Behaviour 135:321-342

Nelson DA (2000) A preference for own-subspecies' song guides vocal learning in a song bird. Proc Natl Acad Sci USA 97:13348-13353 doi:10.1073/pnas.240457797

Nelson DA, Soha JA (2004) Perception of geographical variation in song by male Puget Sound white-crowned sparrows, *Zonotrichia leucophrys pugetensis*. Anim Behav 68:395-405

Nicholls JA (2008) Site specificity in advertisement calls and responses to playbacks of local and foreign call variants in satin bowerbirds, *Ptilonorhynchus violaceus*. Behav Ecol Sociobiol 62:831-841 doi:10.1007/s00265-007-0509-9

Noble DWA, Lagisz M, O'Dea RE, Nakagawa S (2017) Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. Molecular Ecology 26:2410-2425 doi:10.1111/mec.14031

Osiejuk TS, Bielecka A, Skierczynski M (2012) What exactly is 'local song' in a population of ortolan buntings with a common dialect? J Ethol 30:133-142 doi:10.1007/s10164-011-0306-9

Osiejuk TS, Ratynska K, Dale S (2007) What makes a 'local song' in a population of ortolan buntings without a common dialect? Anim Behav 74:121-130

Päckert M, Martens J, Hofmeister T (2001) Vocalisations of firecrests from the islands of Madeira and Mallorca (*Regulus ignicapillus madeirensis, R. i. balearicus*). Journal für Ornithologie 142:16-29 doi:10.1007/bf01651735

Parker KA, Hauber ME, Brunton DH (2010) Contemporary cultural evolution of a conspecific recognition signal following serial translocations. Evolution 64:2431-2441 doi:10.1111/j.1558-5646.2010.01013.x

Parker TH (2013) What do we really know about the signalling role of plumage colour in blue tits? A case study of impediments to progress in evolutionary biology. Biological Reviews 88:511-536

Parker TH et al. (2016) Transparency in ecology and evolution: real problems, real solutions. Trends in Ecology & Evolution 31:711-719 doi:10.1016/j.tree.2016.07.002

Parker TH, Greig E, Nakagawa S, Parra M, Dalisio A (2018) Data and code from “Sub-species status and methodological choices explain strength of territorial response to local versus foreign song by oscine birds in meta-analysis.” Open Science Framework. doi:10.17605/OSF.IO/W2MVP

Parra M, Dalisio AC, Jensen WE, Parker TH (2017) Male territorial aggression does not drive conformity to local vocal culture in a passerine bird. Ethol 123:800-810 doi:10.1111/eth.12648

Payne RB (1981) Population structure and social behavior: models for testing the ecological significance of song dialects in birds. In: Alexander RD, Tinkle DW (eds) Natural selection and social behavior: recent research and new theory. Chiron Press, New York,

Payne RB (1983) The social context of song mimicry - song matching dialects in indigo buntings (*Passerina cyanea*). Anim Behav 31:788-805

Pellerin M (1983) Variability of response of the corn bunting, *Emberiza calandra*, to songs of different dialects. Behav Processes 8:157-163 doi:10.1016/0376-6357(83)90004-9

Petrinovich L (1981) The responses of white-crowned sparrows to songs of different dialects and subspecies. Z Tierpsychol 57:1-14

Podos J (2007) Discrimination of geographical song variants by Darwin's finches. Anim Behav 73:833-844 doi:10.1016/j.anbehav.2006.11.001

Podos J (2010) Acoustic discrimination of sympatric morphs in Darwin's finches: a behavioural mechanism for assortative mating? Philosophical Transactions of the Royal Society B: Biological Sciences 365:1031-1039 doi:10.1098/rstb.2009.0289

Podos J, Dybboe R, Jensen MO (2013) Ecological speciation in Darwin's finches: parsing the effects of magic traits. Current Zoology 59:8-19 doi:10.1093/czoolo/59.1.8

R_Core_Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Randler C, Foerschler MI, Gonzalez J, Aliabadian M, Bairlein F, Wink M (2012) Phylogeography, pre-zygotic isolation and taxonomic status in the endemic Cyprus Wheatear *Oenanthe cypriaca*. J Ornith 153:303-312 doi:10.1007/s10336-011-0744-8

Ratcliffe LM, Grant PR (1985) Species recognition in Darwin finches (*Geospiza*, Gould) 3. Male responses to playback of different song types, dialects and heterospecific songs. Anim Behav 33:290-307 doi:10.1016/s0003-3472(85)80143-3

Reichard DG (2014) Male dark-eyed juncos (*Junco hyemalis*) respond differentially to playback of local and foreign song. Wilson Journal of Ornithology 126:605-611

Ripmeester EAP, Mulder M, Slabbekoorn H (2010) Habitat-dependent acoustic divergence affects playback response in urban and forest populations of the European blackbird. Behav Ecol 21:876-883 doi:10.1093/beheco/arq075

Rohatgi A (2017) WebPlotDigitizer, 3.5 edn., Austin, TX, USA

Rohwer S (1982) The evolution of reliable and unreliable badges of fighting ability. Am Zool 22:531-546

Rosenberg MS (2013) Moment and least-squares based approaches to metaanalytic inference. . In: Koricheva J, Gurevitch J, Mengersen K (eds) Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, USA, pp 108-124

Rosenberg MS, Rothstein HR, Gurevitch J (2013) Effect sizes: conventional choices and calculations. In: Koricheva J, Gurevitch J, Mengersen K (eds) Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, USA, pp 61-71

Searcy WA, Andersson M (1986) Sexual selection and the evolution of song. Annual Review of Ecology and Systematics 17:507-533 doi:10.2307/2097007

Searcy WA, Nowicki S, Hughes M (1997) The response of male and female song sparrows to geographic variation in song. The Condor 99:651-657

Searcy WA, Nowicki S, Hughes M, Peters S (2002) Geographic song discrimination in relation to dispersal distances in song sparrows. Am Nat 159:221-230 doi:10.1086/338509

Searcy WA, Nowicki S, Peters S (2003) Phonology and geographic song discrimination in song sparrows. Ethol 109:23-35 doi:10.1046/j.1439-0310.2003.00835.x

Shizuka D (2014) Early song discrimination by nestling sparrows in the wild. Anim Behav 92:19-24 doi:http://dx.doi.org/10.1016/j.anbehav.2014.03.021

Slabbekoorn H, Smith TB (2002) Habitat-dependent song divergence in the little greenbul: an analysis of environmental selection pressures on acoustic signals. Evolution 56:1849-1858

Slater PJB, Catchpole CK (1990) Responses of the two chaffinch species on Tenerife *(Fringilla teydea* and *F. coelebs tintillon*) to playback of the song of their own and the other species. Behaviour 115:143-152

Soha JA, Poesel A, Nelson DA, Lohr B (2016) Non-salient geographic variation in birdsong in a species that learns by improvisation. Ethol 122:343-353 doi:10.1111/eth.12474

Stock M, Bergmann HH (1988) The song of the robin (*Erithacus rubecula superbus*) from Tenerife (Canary Islands): structure and reactions to playback. Zoologische Jahrbücher Abteilung für allgemeine Zoologie und Physiologie der Tiere 92:197-212

Symes LB (2014) Community composition affects the shape of mate response functions. Evolution 68:2005-2013 doi:10.1111/evo.12415

Thompson AD, Baker MC (1993) Song dialect recognition by male white-crowned sparrows: effects of manipulated song components. The Condor 95:414-421 doi:10.2307/1369364

Tietze DT, Wassmann C, Martens J (2012) Territorial song does not isolate yellowhammers (*Emberiza citrinella*) from pine buntings (*E. leucocephalos*). Vertebrate Zoology 62:113-122

Tomback DF, Thompson DB, Baker MC (1983) Dialect discrimination by white-crowned sparrows: reactions to near and distant dialects. Auk 100:452-460 doi:10.2307/4086539

Touchton JM, Seddon N, Tobias JA (2014) Captive rearing experiments confirm song development without learning in a tracheophone suboscine bird. Plos One 9:9 doi:10.1371/journal.pone.0095746

Turcokova L, Pavel V, Chutny B, Petrusek A, Petruskova T (2011) Differential response of males of a subarctic population of Bluethroat *Luscinina svecica svecica* to playbacks of their own and foreign subspecies. J Ornith 152:975-982 doi:10.1007/s10336-011-0680-7

Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Softw 36:1-48 doi:10.18637/jss.v036.i03

Wei M, Lloyd H, Zhang Y (2011) Neighbour–stranger discrimination by yellow-bellied tit *Parus venustulus*: evidence for the “dear-enemy” effect. J Ornith 152:431-438 doi:10.1007/s10336-010-0609-6

Westcott DA, Kroon FJ (2002) Geographic song variation and its consequences in the golden bowerbird. Condor 104:750-760 doi:10.1650/0010-5422(2002)104[0750:gsvaic]2.0.co;2

Wiley RH (1983) The evolution of communication: information and manipulation. In: Halliday TR, Slater PJB (eds) Animal Behaviour, vol 2, Communication. Blackwell, Oxford, UK, pp 156-189

Wiley RH (1994) Errors, exaggeration, and deception in animal communication. In: Real L (ed) Behavioral Mechanisms in Ecology. University of Chicago Press, Chicago, pp 157-189

Wilkins MR, Seddon N, Safran RJ (2013) Evolutionary divergence in acoustic signals: causes and consequences. Trends Ecol Evol 28:156-166 doi:http://dx.doi.org/10.1016/j.tree.2012.10.002

Williams JM, Slater PJB (1990) Modelling bird song dialects - the influence of repertoire size and numbers of neighbors. J Theor Biol 145:487-496

Wilson PL, Vehrencamp SL (2001) A test of the deceptive mimicry hypothesis in song-sharing song sparrows. Anim Behav 62:1197-1205

Wright TF, Dorin M (2001) Pair duets in the yellow-naped Amazon (Psittaciformes: *Amazona auropalliata*): responses to playbacks of different dialects. Ethol 107:111-124 doi:10.1046/j.1439-0310.2001.00632.x

Table 1. Our categorization of response variables used in studies of wild bird response to playback of local versus foreign song.

Category of response variable	Description of category	number of effects in category
any	observations of “any response” - for instance whether the focal bird either sang or approached the speaker	11
approach	measures of approach distance or time to approach the speaker	142
calling	measures of non-song vocalization	16
composite	scores combining different categories of information, for instance summing across song and approach categories. Also, principal components (PCs) combining multiple response measurements from multiple categories (PCs combining responses within a categories remain in that category; e.g., a PC summarizing different measures of approach is classified as ‘approach’)	50
flight	measures of flight rate or latency to fly	61
singing	measures of song rate, number of songs, or latency to sing	89
misc.	anything not classified above	10

Table 2. Average effect size with associated statistics as drawn from three different random effects meta-analytic models assessing the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song. Larger values of effect size indicate stronger response to local song versus foreign song.

model	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
All effects, sample size = trial n (379 effects)	0.323	0.039	8.25	< 0.0001	0.246	0.400
All effects, sample size = stimuli n (379 effects)	0.309	0.034	8.84	< 0.0001	0.236	0.371
Extreme pseudoreplication excluded, sample size = trial n (180 effects)	0.297	0.037	8.04	< 0.0001	0.225	0.370

Table 3. Heterogeneity from three different random effects meta-analytic models assessing the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song. Total heterogeneity is partitioned among the three random effects in our mixed models, response variable (measured response to playback), phylogeny, and the identity of the individual studies.

model	total I^2	response variable I^2	phylogeny I^2	study I^2
All effects, sample size = trial n (379 effects)	0.76	0.00	0.00	0.42
All effects, sample size = stimuli n (379 effects)	0.23	0.01	0.00	0.05
Extreme pseudoreplication excluded, sample size = trial n (180 effects)	0.63	0.08	0.21	0.00

Table 4. Results of F-tests to compare variances among effect sizes between severely pseudoreplicated studies and non-pseudoreplicated studies, severely pseudoreplicated studies and moderately pseudoreplicated studies, and moderately pseudoreplicated studies and non-pseudoreplicated studies. Effect sizes were all derived from experiments comparing responses of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song

comparison	F	95% CI	Num df	Den df	p
severe vs. no pseudo.	2.49	1.77 – 3.44	198	111	< 0.0001
severe vs. moderate	2.44	1.61 – 3.54	198	67	< 0.0001
moderate vs. no pseudo.	1.02	0.67 – 1.60	67	111	0.903

Table 5. Analyses of the effect of potential moderator variables on the strength of effect from experiments testing the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song. In this set of analyses, we used the number of independent stimuli as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. N = 379 effects.

variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
intercept	0.345	0.305	1.13	0.258	-0.253	0.943
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.268	0.121	-2.21	0.027	-0.506	-0.030
response variable category						
any	.	.	.	.	.	.
approach	-0.102	0.213	-0.48	0.632	-0.519	0.315
calling	-0.383	0.231	-1.66	0.097	-0.834	0.070
composite	-0.158	0.219	-0.72	0.470	-0.586	0.271
flight	-0.146	0.217	-0.67	0.501	-0.572	0.280
misc.	-0.119	0.224	-0.53	0.596	-0.559	0.321
singing	-0.233	0.216	-1.08	0.281	-0.655	0.190
distinctiveness between local and foreign song	0.094	0.053	1.78	0.075	-0.010	0.198
repertoire size	0.057	0.044	1.30	0.194	-0.029	0.144

Table 6. Post hoc analyses of the effect of broadly defined response variable category (vocal vs. whole body movement) on the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song. In this set of analyses, we used the number of independent stimuli as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. N = 379 effects.

variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
intercept	0.388	0.182	2.13	0.033	0.031	0.744
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.260	0.124	-2.10	0.036	-0.504	-0.017
distinctiveness between local and foreign song	0.063	0.049	1.28	0.202	-0.034	0.159
response variable category						
movement	.	.	.	.	.	.
vocal	-0.130	0.044	-2.99	0.003	-0.215	-0.045

Table 7. Models testing predictions of the hypotheses that selection to avoid responding to songs of heterospecifics constrains response to foreign conspecific song in territorial oscine songbirds. In this set of analyses, we used the number of independent stimuli as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. N = 379 effects.

model and variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
1.						
intercept	0.356	0.189	1.88	0.060	-0.015	0.727
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.258	0.128	-2.02	0.043	-0.508	-0.008
distinctiveness between local and foreign song	0.059	0.050	1.17	0.242	-0.040	0.157
local sympatric congeners	-0.005	0.015	-0.32	0.747	-0.035	0.025
2.						
intercept	0.137	0.204	0.67	0.502	-0.263	0.537
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.234	0.116	-2.01	0.045	-0.462	-0.006
distinctiveness between local and foreign song	0.093	0.052	1.80	0.072	-0.008	0.194
local relatives score	0.096	0.044	2.18	0.029	0.010	0.183
3.						
intercept	0.269	0.204	1.32	0.187	-0.131	0.670
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.257	0.124	-2.07	0.039	-0.500	-0.013
distinctiveness between local and foreign song	0.062	0.049	1.26	0.209	-0.035	0.159
latitude	0.002	0.002	0.87	0.386	-0.002	0.006
island status						
mainland	.	.	.	.	.	.
isolated oceanic island	0.040	0.090	0.44	0.660	-0.137	0.217
4.						
intercept	0.065	0.219	0.30	0.767	-0.365	0.495
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.230	0.117	-1.96	0.050	-0.459	0.000
distinctiveness between local and	0.090	0.052	1.74	0.082	-0.011	0.191

foreign song						
local sympatric congeners	-0.018	0.016	-1.09	0.276	-0.049	0.014
local relatives score	0.116	0.050	2.33	0.020	0.019	0.214
latitude	0.002	0.002	0.79	0.429	-0.003	0.006
oceanic island						
no	.	.	.	.	.	.
yes	0.055	0.087	0.64	0.525	-0.115	0.226

Table 8. Models testing predictions of the hypotheses that individual history of exposure to foreign conspecific song determines individual response to foreign conspecific song in territorial oscine songbirds. In this set of analyses, we used the number of independent stimuli as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. N = 379 effects. Model 8 only includes effects where the foreign population was separated by distance (N = 239) and excludes effects where the foreign population was disjunct from the local populations.

model and variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
5.						
intercept	0.374	0.179	2.09	0.037	0.023	0.725
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.246	0.121	-2.03	0.042	-0.484	-0.009
distinctiveness between local and foreign song	0.032	0.051	0.63	0.528	-0.068	0.132
average distance to foreign song (log)	0.000	0.000	0.97	0.330	0.000	0.000
isolation vs. distance						
distance	.	.	.	.	.	.
isolation	0.058	0.082	0.71	0.480	-0.103	0.219
distance (log) * isolation vs. distance	0.000	0.000	0.01	0.994	0.000	0.000
6.						
intercept	0.350	0.214	1.64	0.102	-0.069	0.769
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.249	0.114	-2.18	0.029	-0.472	-0.025
distinctiveness between local and foreign song	0.045	0.050	0.90	0.367	-0.053	0.143
movement	0.010	0.041	0.24	0.809	-0.070	0.089
average distance to foreign song (log)	0.000	0.000	0.32	0.753	0.000	0.000
movement * distance (log)	0.000	0.000	0.21	0.837	0.000	0.000
7.						
intercept	0.324	0.223	1.45	0.147	-0.114	0.762
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.234	0.119	-1.97	0.049	-0.468	-0.001
distinctiveness between local and foreign song	0.036	0.052	0.70	0.485	-0.066	0.139
average distance to foreign song (log)	0.000	0.000	0.31	0.753	0.000	0.000
isolation vs. distance						

distance	.	.	.	.	.	.
isolation	0.053	0.085	0.63	0.528	-0.112	0.219
movement	0.018	0.044	0.42	0.674	-0.067	0.104
distance (log) * isolation vs. distance	0.000	0.000	0.00	1.000	0.000	0.000
movement * distance (log)	0.000	0.000	0.11	0.911	0.000	0.000
8.						
intercept	0.127	0.310	0.41	0.683	-0.481	0.735
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.350	0.195	-1.80	0.072	-0.731	0.032
distinctiveness between local and foreign song	0.103	0.057	1.81	0.071	-0.009	0.214
movement	0.060	0.059	1.01	0.312	-0.056	0.176
distance (log)	0.000	0.000	0.97	0.332	0.000	0.000
movement * distance (log)	0.000	0.000	-0.80	0.424	0.000	0.000

Table 9. Model testing predictions of the hypotheses that evolutionary divergence of populations reduces the strength of individual response to foreign conspecific song by territorial oscine songbirds with the number of independent stimuli as the sample size (N = 379).

model and variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
9. Number of stimuli as per-study n						
intercept	0.579	0.183	3.17	0.002	0.221	0.938
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.218	0.112	-1.94	0.053	-0.438	0.002
distinctiveness between local and foreign song	0.031	0.047	0.66	0.511	-0.061	0.123
Subspecies status of foreign song						
different subspecies	.	.	.	.	.	.
same subspecies	-0.210	0.075	-2.79	0.005	-0.358	-0.063

Table 10. Global models including all the predictor variables from all our tests of our three primary biological hypotheses explaining variability in response to playback of local versus foreign conspecific song in oscine birds with the number of independent stimuli as the sample size (N = 379).

model and variable	mean effect	SE	Z	p	95% CI (lower)	95% CI (upper)
10. Number of stimuli as per-study n						
intercept	0.392	0.271	1.45	0.148	-0.139	0.924
availability of results						
some results unavailable	.	.	.	.	.	.
all results are available	-0.200	0.129	-1.55	0.120	-0.453	0.052
distinctiveness between local and foreign song	0.038	0.055	0.69	0.492	-0.070	0.146
local sympatric congeners	-0.007	0.018	-0.40	0.690	-0.042	0.028
local relatives score	0.122	0.059	2.06	0.039	0.006	0.238
latitude	0.001	0.003	0.29	0.771	-0.005	0.007
oceanic island						
no	.	.	.	.	.	.
yes	0.025	0.103	0.25	0.805	-0.176	0.227
average distance to foreign song (log)	0.000	0.000	-0.19	0.850	0.000	0.000
isolation vs. distance						
distance	.	.	.	.	.	.
isolation	0.104	0.098	1.05	0.292	-0.089	0.296
movement	-0.028	0.059	-0.48	0.633	-0.144	0.088
distance (log) * isolation vs. distance	0.000	0.000	0.24	0.808	0.000	0.000
movement * distance (log)	0.000	0.000	0.05	0.957	0.000	0.000
Subspecies status of foreign song						
different subspecies	.	.	.	.	.	.
same subspecies	-0.222	0.110	-2.03	0.042	-0.437	-0.008

Appendix 1. Studies, listed by species, that reported results from experiments comparing responses of oscine songbirds to playback of local versus foreign song. Studies identified as unusable are those that reported results from relevant experiments, but which did so with insufficient detail for inclusion of any effects in meta-analysis. For details of data extracted from each study, see archived data (Parker et al. 2018).

species	study
Alauda arvensis	(Briefer et al. 2010)
Ammodramus savannarum	(Soha et al. 2016)
Callaeas wilsoni	(Bradley et al. 2013)
Cardinalis cardinalis	(Lemon 1967)
Certhidea fusca	(Grant and Grant 2002a)
Certhidea olivacea	(Grant and Grant 2002a)
Emberiza calandra	(Latruffe et al. 2000)
	(McGregor 1983)
	(Pellerin 1983)
Emberiza cirlus	(Kreutzer 1987)
Emberiza citrinella	(Hansen 1984)
	(Tietze et al. 2012)
Emberiza hortulana	(Osiejuk et al. 2007)
	(Osiejuk et al. 2012)
Erithacus rubecula	(Stock and Bergmann 1988)
Fringilla coelebs	(Slater and Catchpole 1990)
Geospiza conirostris	(Ratcliffe and Grant 1985)
Geospiza difficilis	(Grant and Grant 2002b)
	(Ratcliffe and Grant 1985)
Geospiza fortis	(Podos 2007)
	(Podos et al. 2013)
	(Podos 2010)
	(Ratcliffe and Grant 1985)
Geospiza fuliginosa	(Ratcliffe and Grant 1985)
Geospiza magnirostris	(Ratcliffe and Grant 1985)
Geospiza scandens	(Ratcliffe and Grant 1985)
Henicorhina leucophrys	(Dingle et al. 2010)
Junco hyemalis	(Reichard 2014)
Luscinia svecica	(Turcokova et al. 2011)
Meliphaga virescens	(Baker 1994)
Melospiza georgiana	(Balaban 1988)
	(Liu et al. 2008)
Melospiza melodia	(Harris and Lemon 1974) [unusable]
	(Searcy et al. 2003)
	(Searcy et al. 2002)
	(Searcy et al. 1997)

Nectarinia osea	(Leader et al. 2005)
	(Leader et al. 2002)
Orthonyx spaldingii	(Koetz-Trowse et al. 2012)
Parus minor	(Hamao 2016)
Philesturnus rufusater	(Parker et al. 2010)
Phylloscopus collybita	(Hansson et al. 2000)
	(Martens and Meincke 1989)
Poecile gambeli	(Branch and Pravosudov 2016)
Poecile varius	(Hamao 2016)
Prionodura newtonia	(Westcott and Kroon 2002)
Ptilonorhynchus violaceous	(Nicholls 2008) [unusable]
Regulus ignicapillus	(Becker 1977)
	(Päckert et al. 2001) [unusable]
Regulus regulus	(Becker 1977)
Saxicola torquata	(Mortega et al. 2014)
Setophaga caerulescens	(Colbeck et al. 2010)
Sitta europaea	(Matthysen 1997)
Troglodytes troglodytes	(Kreutzer 1974)
Turdus iliacus	(Bjerke 1984)
Turdus merula	(Ripmeester et al. 2010)
Vermivora cyanoptera	(Kroodsma et al. 1984)
Vireo solitarius	(Morton et al. 2006)
Zonotrichia albicollis	(Lemon and Harris 1974) [unusable]
Zonotrichia capensis	(Danner et al. 2011)
Zonotrichia leucophrys	(Baker 1982) [unusable]
	(Baker et al. 1981)
	(Baker et al. 1984)
	(Derryberry 2011)
	(Lampe and Baker 1994)
	(Milligan and Verner 1971) [unusable]
	(Nelson and Soha 2004)
	(Nelson 1998)
	(Petrinovich 1981)
	(Thompson and Baker 1993)
	(Tomback et al. 1983)

Table S1

Analyses of the effect of potential moderator variables on the strength of effect from experiments testing the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song

Variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
Intercept	0.238	0.248	0.96	0.338	-0.249	0.725
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.186	0.107	-1.73	0.084	-0.396	0.025
Response variable category						
Any	—	—	—	—	—	—
Approach	-0.171	0.099	-1.74	0.083	-0.364	0.022
Calling	-0.471	0.128	-3.67	0.0002	-0.722	-0.220
Composite	-0.158	0.112	-1.41	0.158	-0.378	0.062
Flight	-0.206	0.105	-1.97	0.049	-0.410	-0.001
Misc.	-0.153	0.128	-1.20	0.230	-0.404	0.097
Singing	-0.216	0.103	-2.10	0.036	-0.417	-0.014
Distinctiveness between local and foreign song	0.124	0.058	2.13	0.033	0.010	0.239
Repertoire size	0.061	0.048	1.28	0.202	-0.033	0.154

In this set of analyses, we used the number of independent trials (ignoring pseudoreplication in number of stimuli) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 379 effects.

Table S2

Analyses of the effect of potential moderator variables on the strength of effect from experiments testing the response of territorial oscine songbirds to the playback of local conspecific song versus foreign conspecific song

Variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
Intercept	0.103	0.220	0.47	0.639	-0.327	0.534
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.204	0.123	-1.67	0.096	-0.445	0.036
Response variable category						
Singing	—	—	—	—	—	—
Approach	0.151	0.054	2.79	0.005	0.045	0.258
Calling	-0.206	0.092	-2.25	0.025	-0.386	-0.026
Composite	0.066	0.064	1.04	0.298	-0.059	0.192
Flight	0.104	0.061	1.71	0.087	-0.015	0.224
Misc.	0.153	0.095	1.61	0.107	-0.033	0.338
Distinctiveness between local and foreign song	0.086	0.052	1.65	0.100	-0.017	0.189
Repertoire size	0.047	0.044	1.08	0.281	-0.039	0.134

In this set of analyses, we excluded effects with severe pseudoreplication and used the number of independent trials (ignoring moderate pseudoreplication of stimuli) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 180 effects.

Table S3

Models testing predictions of the hypotheses that selection to avoid responding to songs of heterospecifics constrains response to foreign conspecific song

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
1.						
Intercept	0.234	0.201	1.16	0.244	-0.160	0.627
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.155	0.110	-1.41	0.160	-0.372	0.061
Distinctiveness between local and foreign song	0.090	0.057	1.57	0.115	-0.022	0.202
Local sympatric congeners	-0.017	0.018	-0.94	0.345	-0.051	0.018
2.						
Intercept	0.033	0.234	0.14	0.887	-0.425	0.492
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.145	0.112	-1.30	0.193	-0.364	0.073
Distinctiveness between local and foreign song	0.110	0.058	1.90	0.057	-0.003	0.223
Local relatives score	0.069	0.057	1.22	0.223	-0.042	0.180
3.						
Intercept	0.063	0.224	0.28	0.778	-0.376	0.502
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.163	0.110	-1.49	0.138	-0.379	0.052
Distinctiveness between local and foreign song	0.104	0.057	1.82	0.069	-0.008	0.216
Latitude	0.003	0.003	1.15	0.251	-0.002	0.008
Island status						
Mainland	—	—	—	—	—	—
Isolated oceanic island	0.049	0.106	0.46	0.643	-0.159	0.257
4.						
Intercept	-0.093	0.265	-0.35	0.726	-0.613	0.427
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.137	0.115	-1.19	0.232	-0.362	0.088
Distinctiveness between local and foreign song	0.108	0.059	1.82	0.068	-0.008	0.224
Local sympatric congeners	-0.028	0.020	-1.44	0.150	-0.066	0.010
Local relatives score	0.106	0.062	1.69	0.090	-0.017	0.228
Latitude	0.003	0.003	1.05	0.295	-0.002	0.008
Oceanic island						
No	—	—	—	—	—	—
Yes	0.063	0.109	0.58	0.565	-0.151	0.276

In this set of analyses, we used the number of independent trials (ignoring pseudoreplication in number of stimuli) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 379 effects.

Table S4

Models testing predictions of the hypotheses that selection to avoid responding to songs of heterospecifics constrains response to foreign conspecific song

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
1.						
Intercept	0.304	0.196	1.55	0.120	-0.080	0.688
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.184	0.131	-1.41	0.158	-0.440	0.072
Distinctiveness between local and foreign song	0.056	0.050	1.11	0.269	-0.043	0.154
Local sympatric congeners	-0.004	0.011	-0.37	0.712	-0.026	0.018
2.						
Intercept	0.060	0.214	0.28	0.779	-0.359	0.479
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.162	0.125	-1.29	0.197	-0.408	0.084
Distinctiveness between local and foreign song	0.093	0.051	1.84	0.066	-0.006	0.193
Local relatives score	0.114	0.050	2.27	0.023	0.016	0.213
3.						
Intercept	0.173	0.217	0.80	0.424	-0.251	0.597
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.172	0.127	-1.35	0.176	-0.421	0.077
Distinctiveness between local and foreign song	0.062	0.049	1.25	0.210	-0.035	0.159
Latitude	0.003	0.002	1.15	0.250	-0.002	0.007
Island status						
Mainland	—	—	—	—	—	—
Isolated oceanic island	0.070	0.091	0.77	0.439	-0.108	0.248
4.						
Intercept	-0.018	0.239	-0.08	0.939	-0.486	0.450
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.151	0.132	-1.15	0.250	-0.409	0.106
Distinctiveness between local and foreign song	0.087	0.052	1.68	0.094	-0.015	0.190
Local sympatric congeners	-0.012	0.012	-1.01	0.312	-0.035	0.011
Local relatives score	0.140	0.057	2.44	0.015	0.028	0.253
Latitude	0.001	0.002	0.58	0.561	-0.003	0.006
Oceanic island						
No	—	—	—	—	—	—
Yes	0.100	0.095	1.06	0.290	-0.085	0.285

In this set of analyses, we excluded effects from severely pseudoreplicated experiments and used the number of independent trials (ignoring moderate pseudoreplication) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 180 effects.

Table S5

Models testing predictions of the hypotheses that individual history of exposure to foreign conspecific song determines individual response to foreign conspecific song

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
5.						
Intercept	0.208	0.197	1.06	0.291	-0.178	0.593
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.175	0.113	-1.55	0.121	-0.395	0.046
Distinctiveness between local and foreign song	0.085	0.059	1.44	0.151	-0.031	0.202
Average distance to foreign song (log)	0.000	0.000	1.29	0.198	0.000	0.000
isolation vs. distance						
Distance	—	—	—	—	—	—
Isolation	0.009	0.088	0.11	0.916	-0.162	0.181
Distance (log) * isolation vs. distance	0.000	0.000	-0.78	0.436	0.000	0.000
6.						
Intercept	0.235	0.251	0.93	0.350	-0.258	0.727
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.174	0.111	-1.57	0.117	-0.392	0.043
Distinctiveness between local and foreign song	0.083	0.062	1.35	0.176	-0.037	0.204
Movement	-0.006	0.048	-0.12	0.904	-0.101	0.089
Average distance to foreign song (log)	0.000	0.000	-0.30	0.762	0.000	0.000
Movement * distance (log)	0.000	0.000	0.65	0.515	0.000	0.000
7.						
Intercept	0.250	0.255	0.98	0.328	-0.251	0.750
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.179	0.114	-1.57	0.117	-0.402	0.045
Distinctiveness between local and foreign song	0.080	0.064	1.25	0.210	-0.045	0.204
Average distance to foreign song (log)	0.000	0.000	0.12	0.905	0.000	0.000
isolation vs. distance						
Distance	—	—	—	—	—	—
Isolation	0.006	0.089	0.07	0.945	-0.168	0.180
Movement	-0.012	0.050	-0.24	0.811	-0.110	0.086
Distance (log) * isolation vs. distance	0.000	0.000	-0.72	0.475	0.000	0.000
Movement * distance (log)	0.000	0.000	0.62	0.538	0.000	0.000
8.						
Intercept	0.093	0.340	0.27	0.784	-0.573	0.759
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.231	0.178	-1.30	0.195	-0.580	0.118
Distinctiveness between local and foreign song	0.134	0.078	1.72	0.085	-0.018	0.287
Movement	0.026	0.066	0.40	0.689	-0.103	0.156
Distance (log)	0.000	0.000	0.37	0.714	0.000	0.000
Movement * distance (log)	0.000	0.000	0.05	0.960	0.000	0.000

In this set of analyses, we used the number of independent trials (ignoring pseudoreplication in number of stimuli) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 379 effects. Model 8 only includes effects where the foreign population was separated by distance (*N* = 239) and excludes effects where the foreign population was disjunct from the local populations.

Table S6

Models testing predictions of the hypotheses that individual history of exposure to foreign conspecific song determines individual response to foreign conspecific song

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
5.						
Intercept	0.348	0.187	1.86	0.063	-0.019	0.714
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.191	0.123	-1.55	0.120	-0.431	0.050
Distinctiveness between local and foreign song	0.021	0.051	0.40	0.688	-0.080	0.121
Average distance to foreign song (log) isolation vs. distance	0.000	0.000	0.96	0.336	0.000	0.000
Distance	—	—	—	—	—	—
Isolation	0.079	0.088	0.90	0.368	-0.093	0.250
Distance (log) * isolation vs. distance	0.000	0.000	0.25	0.800	0.000	0.000
6.						
Intercept	0.195	0.226	0.86	0.389	-0.248	0.638
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.178	0.120	-1.48	0.139	-0.414	0.058
Distinctiveness between local and foreign song	0.053	0.050	1.07	0.285	-0.044	0.150
Movement	0.042	0.044	0.97	0.330	-0.043	0.128
Average distance to foreign song (log)	0.000	0.000	1.10	0.271	0.000	0.000
Movement * distance (log)	0.000	0.000	-0.61	0.543	0.000	0.000
7.						
Intercept	0.159	0.230	0.69	0.489	-0.292	0.610
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.162	0.121	-1.34	0.181	-0.400	0.076
Distinctiveness between local and foreign song	0.030	0.052	0.58	0.562	-0.071	0.131
Average distance to foreign song (log) isolation vs. distance	0.000	0.000	1.17	0.242	0.000	0.000
Distance	—	—	—	—	—	—
Isolation	0.100	0.087	1.15	0.250	-0.070	0.270
Movement	0.064	0.046	1.39	0.165	-0.026	0.154
Distance (log) * isolation vs. distance	0.000	0.000	0.46	0.647	0.000	0.000
Movement * distance (log)	0.000	0.000	-0.94	0.350	0.000	0.000
8.						
Intercept	-0.145	0.276	-0.53	0.599	-0.687	0.397
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.128	0.148	-0.87	0.384	-0.418	0.161
Distinctiveness between local and foreign song	0.092	0.054	1.72	0.085	-0.013	0.198
Movement	0.109	0.052	2.11	0.035	0.008	0.210
Distance (log)	0.000	0.000	1.52	0.128	0.000	0.000
Movement * distance (log)	0.000	0.000	-1.20	0.231	0.000	0.000

In this set of analyses, we excluded effects from severely pseudoreplicated experiments and used the number of independent trials (ignoring moderate pseudoreplication) as the sample size. Larger values of effect size indicate stronger response to local song versus foreign song. *N* = 180 effects. Model 8 only includes effects where the foreign population was separated by distance (*N* = 124) and excludes effects where the foreign population was disjunct from the local populations.

Table S7

Models testing predictions of the hypotheses that evolutionary divergence of populations reduces the strength of individual response to foreign conspecific song

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
9. Number of trials as per-study n						
Intercept	0.392	0.211	1.86	0.063	-0.021	0.804
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.167	0.108	-1.55	0.121	-0.378	0.044
Distinctiveness between local and foreign song	0.077	0.056	1.36	0.175	-0.034	0.187
Subspecies status of foreign song						
Different subspecies	—	—	—	—	—	—
Same subspecies	-0.176	0.076	-2.32	0.021	-0.325	-0.027
9. Severe pseudorep. excluded						
Intercept	0.546	0.178	3.07	0.002	0.198	0.895
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.157	0.110	-1.43	0.153	-0.373	0.059
Distinctiveness between local and foreign song	0.020	0.046	0.45	0.656	-0.070	0.110
Subspecies status of foreign song						
Different subspecies	—	—	—	—	—	—
Same subspecies	-0.226	0.068	-3.33	0.001	-0.360	-0.093

The two models are the same except that the first uses the number of independent trials (ignoring pseudoreplication) as the sample size (*N* = 379) and the second excludes severely pseudoreplicated studies and uses the number of independent trials (ignoring moderate pseudoreplication) as sample size (*N* = 180).

Table S8

Global models including all the predictor variables from all our tests of our three primary biological hypotheses explaining variability in response to playback of local versus foreign conspecific song in oscine birds

Model and variable	Mean effect	SE	*Z*	*P*	95% CI (lower)	95% CI (upper)
10. Number of trials as per-study *N*						
Intercept	0.163	0.316	0.51	0.607	-0.457	0.782
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.143	0.120	-1.20	0.231	-0.378	0.091
Distinctiveness between local and foreign song	0.075	0.065	1.16	0.247	-0.052	0.203
Local sympatric congeners	-0.026	0.022	-1.20	0.232	-0.068	0.016
Local relatives score	0.107	0.065	1.65	0.099	-0.020	0.235
Latitude	0.002	0.003	0.66	0.509	-0.004	0.008
Oceanic island						
No	—	—	—	—	—	—
Yes	0.047	0.123	0.38	0.705	-0.195	0.288
Average distance to foreign song (log)	0.000	0.000	-0.28	0.782	0.000	0.000
isolation vs. distance						
Distance	—	—	—	—	—	—
Isolation	0.016	0.117	0.13	0.895	-0.214	0.245
Movement	-0.020	0.058	-0.34	0.733	-0.134	0.094
Distance (log) * isolation vs. distance	0.000	0.000	-0.07	0.946	0.000	0.000
Movement * distance (log)	0.000	0.000	0.48	0.631	0.000	0.000
Subspecies status of foreign song						
Different subspecies	—	—	—	—	—	—
Same subspecies	-0.136	0.104	-1.31	0.191	-0.340	0.068
10. Severe pseudorep. excluded						
Intercept	0.215	0.301	0.71	0.475	-0.375	0.805
Availability of results						
Some results unavailable	—	—	—	—	—	—
All results are available	-0.140	0.139	-1.00	0.315	-0.412	0.133
Distinctiveness between local and foreign song	0.028	0.058	0.48	0.635	-0.086	0.142
Local sympatric congeners	-0.007	0.015	-0.49	0.621	-0.037	0.022
Local relatives score	0.148	0.069	2.17	0.030	0.014	0.283
Latitude	0.000	0.004	0.00	0.998	-0.007	0.007
Oceanic island						
No	—	—	—	—	—	—
Yes	0.077	0.106	0.73	0.467	-0.131	0.286
Average distance to foreign song (log)	0.000	0.000	0.39	0.698	0.000	0.000
isolation vs. distance						
Distance	—	—	—	—	—	—
Isolation	0.123	0.108	1.14	0.255	-0.089	0.335
Movement	0.022	0.071	0.30	0.761	-0.118	0.161
Distance (log) * isolation vs. distance	0.000	0.000	0.06	0.948	0.000	0.000
Movement * distance (log)	0.000	0.000	-0.40	0.689	0.000	0.000
Subspecies status of foreign song						
Different subspecies	—	—	—	—	—	—
Same subspecies	-0.196	0.106	-1.85	0.065	-0.403	0.012

The two models are the same except that the first uses the number of independent trials (ignoring pseudoreplication) as the sample size (*N* = 379) and the second excludes severely pseudoreplicated studies and uses the number of independent trials (ignoring moderate pseudoreplication) as sample size (*N* = 180).

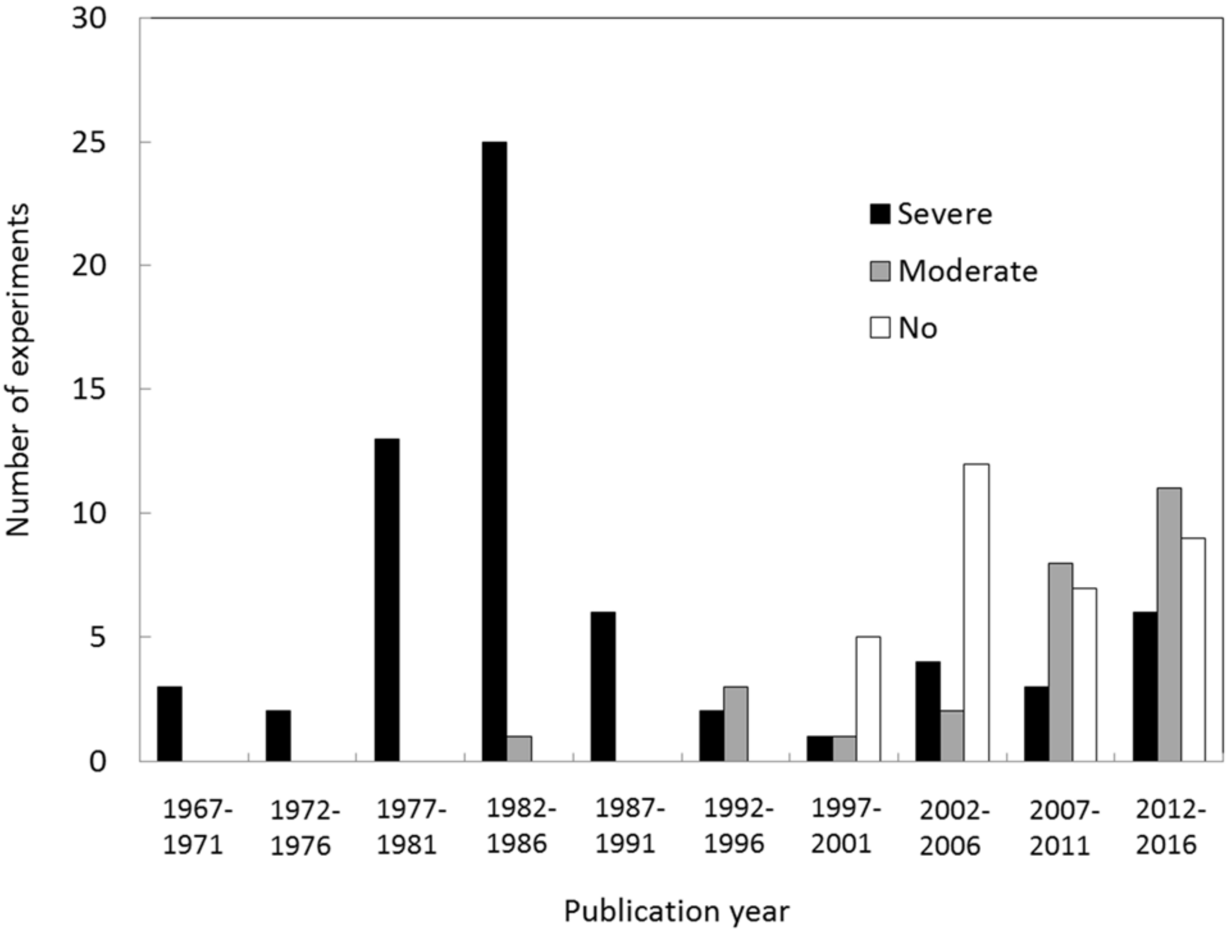

Figure S1. Pseudoreplication severity in 5-year publication date intervals among experiments comparing territorial songbird responses to playback of conspecific local versus conspecific foreign song. This histogram shows the number of experiments with severe (black bars: number of stimuli no more than 25% of the number of otherwise independent trials), moderate (grey bars: number of stimuli >25% of the number of otherwise independent trials), or no (white bars: typically a different stimulus for each trial) pseudoreplication of stimulus songs. Early studies were all severely pseudoreplicated. Although experiments with no pseudoreplication become common in the late 1990s, experiments with moderate and severe pseudoreplication have continued to be published.

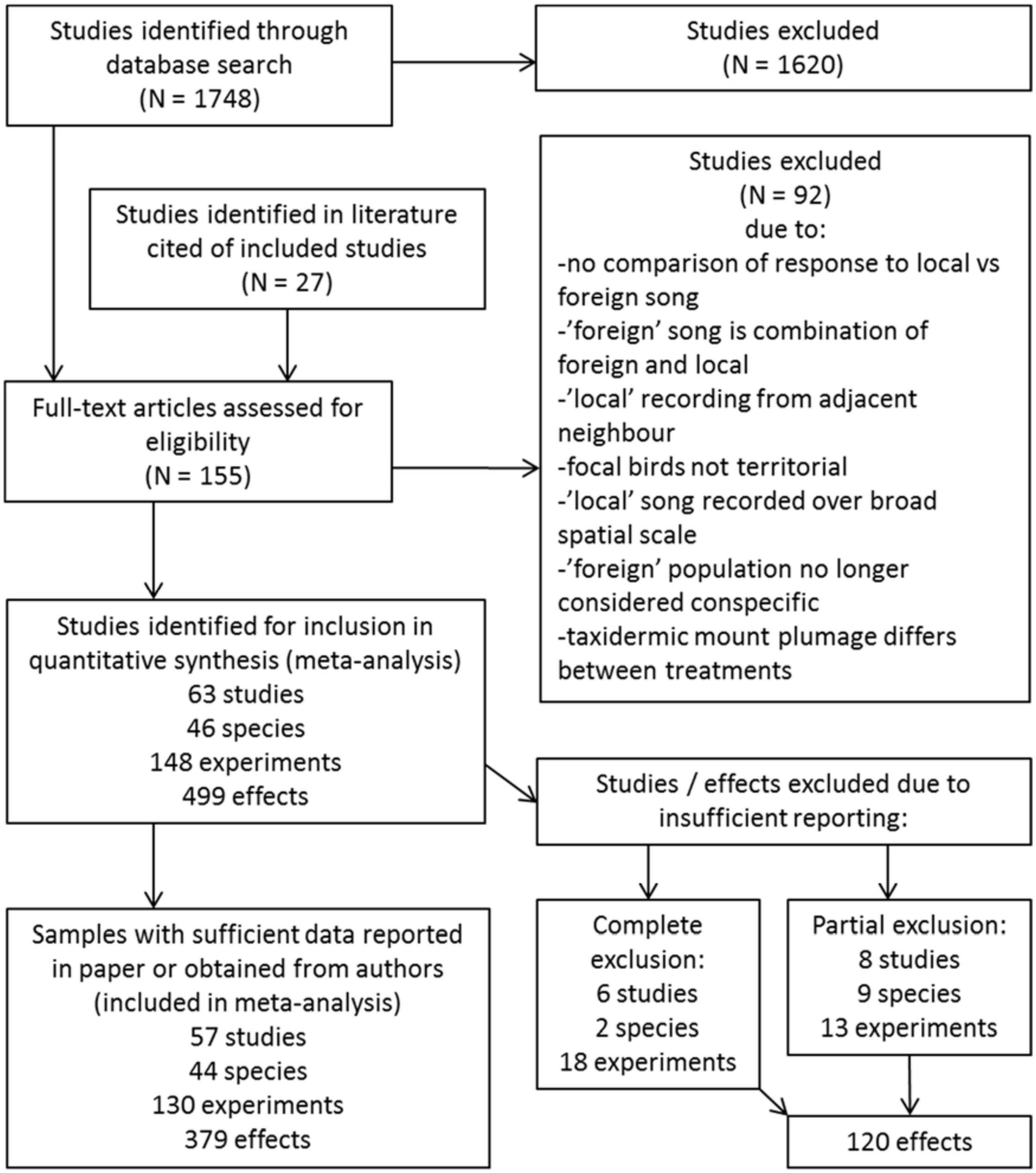

Figure S2. PRISMA-style diagram outlining the number of studies located and screened for inclusion in our systematic review of experiments comparing territorial songbird responses to playback of conspecific local versus conspecific foreign song.

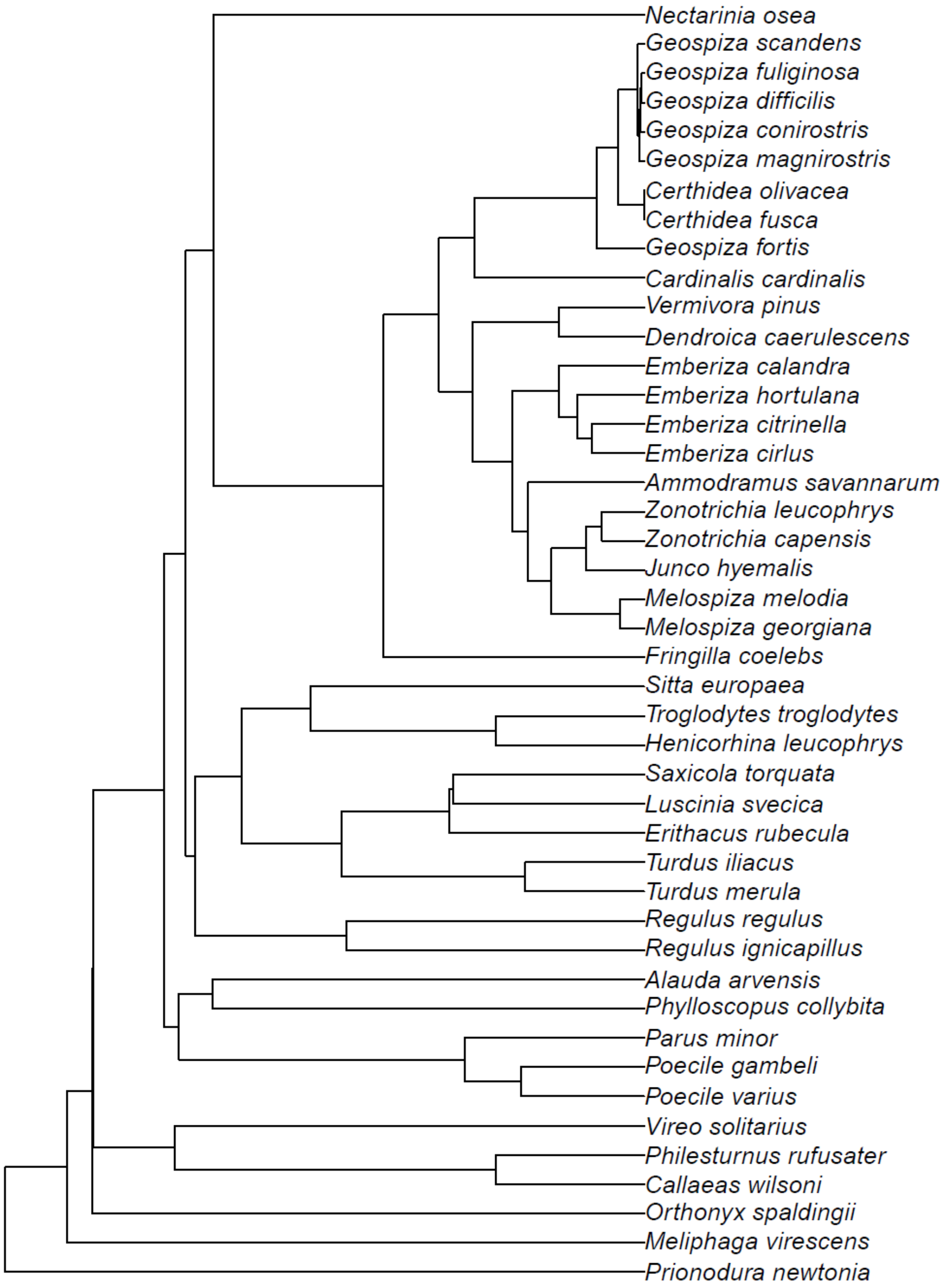

Figure S3. Ultrametric phylogenetic tree with an Ericson backbone obtained from Jetz et al. (2012; http://birdtree.org/) and converted into correlation matrix for our meta-analyses by fitting it to the rma.mv function in metaphor.

Reference

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature, 491*, 444–448. doi:http://www.nature.com/nature/journal/v491/n7424/abs/nature11631.html#supplementary-information.