ARTICLE

Is hearing believing? Patterns of bird voice misidentification in an online quiz

Bento Collares Gonçalves[1,3,4] and Gonçalo Ferraz[2]

[1] Instituto de Biociências, Universidade Federal do Rio Grande do Sul, CEP 91540-000, Porto Alegre, RS, Brazil.
[2] Departamento de Ecologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, CEP 91501-970, Porto Alegre, RS, Brazil.
[3] Current address: Department of Ecology and Evolution, Stony Brook University, 650 Life Sciences Building, 11794-5245, Stony Brook, NY, USA.
[4] Corresponding author: bentocollares@gmail.com

Received on 12 March 2015. Accepted on 12 August 2016.

ABSTRACT: This study aims to uncover patterns of species identification error in bioacoustic surveys of central Amazon birds. To quantify errors, we developed an on-line quiz based on vocalizations of an undisclosed set of 41 antbird (Thamnophilidae) and woodcreeper (Dendrocolaptinae) species. We invited experts to answer the quiz and obtained 820 answers from 20 participants. The answers were compared to the results of a binomial experiment with a success probability of 0.5; *i.e.* we examined whether participants identified species correctly more often than expected by the toss of a coin with a 50% chance of producing the right identification. We also examined whether species were correctly identified more often than expected under a similar coin toss experiment. Quiz answers were compiled in a triangular matrix showing species ranked by taxonomic order on both axes. From the triangular matrix we can ask whether closely-related species were mistaken for each other, *i.e.* confused, more often than distantly-related species. We tested this hypothesis with a null model approach that compared the mean taxonomic distance between confused species in the observed matrix to the distribution of mean taxonomic distances between confused species in 10,000 randomized matrices. Finally, we drew a dendrogram to represent the similarity between species with regard to the distribution of identification errors. The 20 participants who took the quiz showed substantial variation in their ability to identify species correctly. Fourteen species were correctly identified more often than expected at random, while only one was misidentified more often than expected at random. The observed mean distance between confused species was smaller than all of the mean distances from the randomized, null-model matrices, indicating that confusions are more frequent between closely related species than between distant ones.

KEY-WORDS: bioacoustic sampling, bird vocalizations, central Amazon, Dendrocolaptinae, false positives, misidentification, Thamnophilidae.

INTRODUCTION

Bioacoustic monitoring of wildlife based on autonomous recording units has seen remarkable technical progress in the last decade, both in aquatic (Sousa-Lima *et al.* 2013) and in terrestrial (Fristrup and Mennitt 2012) environments. Such progress has made it easier to obtain presence/absence data for species with conspicuous vocalizations like some insects, anurans, and a large variety of birds. Autonomous recording combines four features that make it a particularly cost-efficient sampling technique: the possibility of sampling in all directions from one observation point; relatively high detection probability when visibility is low; the possibility of simultaneously sampling many sites with moderate to low effort; and, last but not least, a permanent record of animal signals that can be easily reviewed to correct doubtful identifications. In spite of its convenient features, autonomous recording is still liable to errors, like all field-sampling techniques. Acoustic recordings may easily miss species that are present at a site (false negatives), or they may lead to identification errors, which can result in the mistaken record of a species that is actually not present at a site (false positives).

There is a large body of literature offering modeling solutions for estimating biological parameters based on data with false-negative errors (MacKenzie *et al.* 2002). False positive errors, on the other hand, have received relatively less attention, and the analytical solutions to deal with them are in an earlier stage of development (Miller *et al.* 2013). Nonetheless, the relevance of false positives is evident, especially in site-occupancy surveys, where they can lead to measurable errors in occupancy estimates even when they represent as little as 1% of detections (McClintock *et al.* 2010). Occupancy models that take false positive errors into account, in cases where some amount of identification error is expected, produce substantially different, less biased estimates of occupancy than those models that ignore false positives (Miller *et al.* 2011).

Perhaps the greatest source of false positive sampling errors in bioacoustic data is the great similarity between many species' sounds (McClintock *et al.* 2010). This similarity, compounded by other factors such as lack of visual information and various types of background noise, results in some unavoidable amount of identification error (Farmer *et al.* 2012). There are many possible causes for false-positive errors, but even when causes are unknown, it is possible to improve site-occupancy and species distribution inferences if we have some notion of which species are easier or harder to identify (Miller *et al.* 2011). Knowledge of how easy it is to mistake one species for another can also be useful in novel probabilistic methods of taxonomic classification (Somervuo *et al.* 2016).

In the present study, we asked a group of experts to identify vocalizations of 41 Amazon Forest bird species, and used their answers to quantify which species were most likely to be mistakenly identified. In doing this, we addressed three specific questions: a) To what extent are experts capable of correctly identifying bird sounds? b) Which species are more difficult to identify? and finally, c) Is the taxonomic distance between two species related to the probability of mistaking one of those species for the other?

We based our survey of expert identifications in an on-line quiz which presented users with recordings of bird species occurring in the Biological Dynamics of Forest Fragments Project (BDFFP) study area, on the southwest end of the Guiana shield region, 80 km north of the city of Manaus, Brazil (2.4°S; 59.9°W). This area is particularly fit for a study about false positives in bioacoustic sampling because it has a rich avifauna (Cohn-Haft *et al.* 1997); it has a very good reference collection of bird vocalizations (Naka *et al.* 2008); and its bird fauna is relatively well known, compared to other regions of the Amazon. The combination of these three factors facilitates the emergence of a fairly large community of experts who can identify regional birds from their vocalizations. Although the study area has about 400 bird species, this study will focus on a small subset of species to construct a bird identification quiz that represents a meaningful part of the avifauna but is short enough to engage a reasonable number of collaborating experts.

METHODS

Construction of a vocalization library

The first step of this study was to assemble a vocalization library with species from the family Thamnophilidae (antbirds) and subfamily Dendrocolaptinae (woodcreepers, family Dendrocolaptidae) occurring in the BDFFP area. We chose these two groups for four reasons: i) they are almost entirely represented by understory birds, and thus easier to hear and record; ii) most of the species in these groups are common in Amazonia, very vocal, and well known; iii) their songs are rather simple and stereotypical when compared to oscine passerines and iv) they have relatively well-resolved phylogenies (Irestedt *et al.* 2004, Moyle *et al.* 2009). The latter attribute allows us to ask if taxonomic distance between two given species bears any relationship with the probability of mistaking one of those species for the other.

The recordings used in this study were obtained from 1) the Ferraz Lab autonomous recordings database, 2) the Xeno-Canto Foundation on-line database, and 3) from the commercially available CD "voices of the Brazilian Amazon" (Naka *et al.* 2008). To minimize sound quality differences within the quiz, we individually edited recordings using software Adobe Audition 5.5 to standardize duration, background noise and signal amplitude. By doing this, we aimed to ensure that variations in identification success were determined mostly by variation in characteristics of the vocalizations. Nevertheless, in order to present quiz users with an aural experience that was somewhat faithful to that experienced in the field, we did not attempt to completely eliminate background noise and other imperfections. In the end, our quiz library contained 82 vocalizations from 41 species (13 woodcreepers and 28 antbirds), with two different vocalizations for each species. One recording of *Thamnophilus punctatus* was removed from the study after the quiz application because six observers raised doubts about the possibility of correctly identifying the vocalization.

On-line quiz

In order to quantify identification errors, we designed an on-line quiz using the software Wondershare Quiz Creator. The quiz consisted of 41 questions, selected at random from a pool of 82. Each question presented an audio recording and a sonogram, which illustrate the vocalization of one focal species. To answer each question, experts had to listen to the recording and fill a blank space with the name of the species that they believed to be featured in the recording. Since the 41 questions in every test are picked at random, the number of questions per species and the number of species heard in a single test are subject to some variation. However, random selection of questions was done without replacement, thus preventing any species from being heard more than twice in one test. To ensure that the quiz was done in one take, each user had a time limit of 30 minutes to finish answering all the questions. Since the quiz software does not check for typing mistakes, we developed a script in the R package (R Development Core Team 2013) that compares expert answers to a list of species from the study area and corrects typing mistakes. Corrections were applied only in cases where the given answer was five or less characters away from one species on the list. Answers that were more than five characters away from every species in the list were flagged for manual verification.

Quiz participants

The search for quiz participants followed an e-mail thread that started with a collaboration request, instructions to complete the on-line quiz, and a brief description of the study goals. The request was sent to a list of thirty experts, defined here as individuals with professional or graduate-level experience in identifying Amazon Forest birds by their vocalizations. Everyone on this list was personally known to us as a competent field researcher or recommended to us by ornithologists with more than 25 years of experience identifying Amazon bird vocalizations. We had a total of 20 quiz takers, which inevitably had variable skills in identifying the study species: two were professional field guides, nine were graduate students, and nine were professional ornithologists. Some participants had more experience in visual than aural identification while others knew Amazon bird vocalizations well but not necessarily the vocalizations from the study area. These sources of variability in observation skill are unavoidable and contribute to the misidentification that we want to study.

Binomial analyses of identification data

We measured the performance of each expert in identifying vocalizations by the proportion of quiz questions that he or she answered correctly. To sort performances between exceptionally good, average, or exceptionally bad, we performed a binomial test. The test is based on the null hypothesis of equal probability of getting answers right or wrong. The null scenario is equivalent to assuming that, in each question, the participant tosses an unbiased coin that has a right answer on one side and a wrong answer on the other. The binomial test quantifies the probability P of such participant obtaining a result just as extreme, or more extreme than the one obtained in the quiz. “More extreme” means “with a greater number of correct answers”, or “with a greater number of wrong answers”, depending on which end of the distribution the participant falls. We obtain P from an implementation of the Binomial distribution formula in the R core Package (R Development Core Team 2013), and apply a two-tailed approach to testing the null hypothesis. When the probability of getting a number of correct answers greater than or equal to the observed was ≤ 0.025 (*i.e.* performance lies in the upper tail of our distribution), we considered that performance exceptionally good. On the other hand, when the probability of getting a number of correct answers smaller than or equal to the observed was ≤ 0.025 (*i.e.* performance lies in the lower tail of our distribution), we considered the performance exceptionally bad and excluded the answers of the observer from subsequent steps in the analysis. Our decision to exclude responses from experts with exceptionally bad performance is an attempt to direct the subsequent part of our analysis to identification mistakes that stem from the similarity between vocalizations and not so much from the observer's lack of previous contact with the species. In all cases where P > 0.025, we considered that the participant had a standard performance.

As a second step in our study, we compared difficulty of identification across species (using the answers from participants with standard or exceptionally good performance). This comparison followed the same approach as the comparison between participants, with the difference that here, the number of coin tosses in the binomial distribution is the total number of times, *N*, that the quiz presented any expert with a vocalization of the focal species. Since quiz questions are randomly sampled, the value of *N* was slightly different among species (mean = 15.82, SD = 4.49). In the comparison among species, the two-tailed test based on the binomial distribution allowed us to identify which species are particularly difficult or particularly easy to identify. A value of P ≤ 0.025 means it is highly unlikely that a species would present a result as extreme as, or more extreme than observed, under the null hypothesis that the probability of a correct identification equals 0.5.

Multinomial analysis

The binomial analyses described above looked only at whether quiz answers were right or wrong. In the multinomial part of our methods, however, we take advantage of the fact that, even though there is only one way to be right, there are many different ways of being wrong. At the most superficial level, we considered three kinds of wrong answers: blank answers, where users did not write anything or declared that they could not answer; off-site answers, where users named a species that does not occur in the study area; and plain-wrong answers, where users named a species which does occur in the study area but does not appear in the recording. From here on, in evaluating the frequency of confusions between species of the BDFFP area, we restrict our analysis to right answers and plain-wrong answers alone. Furthermore, our quantification of confusions is symmetric, *i.e.* an answer where the expert writes the name of species *b* while listening to the voice of species *a*, counts as a confusion between *a* and *b* in the same way as an answer where the expert writes the name of species *a* while listening to the voice of species *b*. The number of confusions between species *a* and *b* is the sum of confusions in both directions.

Correct and plain-wrong answers by all experts with standard and exceptionally good performance were compiled in a triangular matrix with the same list of species in rows and columns. Cells along the diagonal of this triangular matrix show the number of times each species was correctly identified; cells in the sub-diagonal show the number of confusions between the respective row and column species. We sorted species along columns and rows according to taxonomic relatedness, following the classification by Remsen *et al.* (2014). Two species in consecutive positions on the matrix are separated by one unit of taxonomic distance and are taxonomically closer than two species separated by one or more positions in the list. To investigate whether it was easier to confuse taxonomically close than taxonomically distant species, we used a null model approach (Gotelli & Graves 1996) where we compared the average distance between confused species in the observed confusion matrix (measured in positions in the ranking) to the distribution of average distances between confused species in a set of 10,000 randomized, or "null", confusion matrices.

The null model approach tests the null hypothesis that relatedness between two species has no effect on the probability of confusion, *i.e.* the observed distance does not significantly depart from the distribution of random distances. The lower the observed distance relative to the distribution of "null" distances, the easier it is to reject the null hypotheses and the stronger the support for the idea that relatedness does influence confusion. The randomization algorithm that generates the null matrices has two key restrictions: 1) the number of wrong answers per species is kept constant across random matrices; and 2) the probability that each species is picked as a wrong answer is also kept constant across randomizations. The first restriction ensures that randomizations do not change the basic difficulty of correctly identifying each species. The second restriction is a conservative choice to ensure that if observers have some species bias when offering wrong answers, that bias won't be lost in the null matrices. We experimented with other, less restrictive, algorithms and obtained qualitatively similar results.

To get a quantification and graphic presentation of the possibilities of confusion between species we generated a dendrogram based on the observed identification errors. To transform the number of confusions between two given species into a similarity measurement, we converted our confusion matrix into a matrix of Canberra distances between species (Lance and Williams 1967). The Canberra distance between species vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are given by:

$$d^{CAN}(\boldsymbol{x},\boldsymbol{y}) = \frac{n}{NZ} * \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ are species-specific identification vectors with length equal to the total number of species, *n*, and elements representing the number of times that the vocalizations of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, were identified as vocalization of species *i* = 1, ..., *n*.

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$$
$$\boldsymbol{y} = (y_1, y_2, \dots, y_n)$$

The denominator NZ in the distance formula is the number of coordinate pairs $(x_{i,} y_i)$ that are different from (0,0); within the sum, terms that are divided by zero are treated as zero. We used Canberra distances as implemented in the R stats package (R Development Core Team 2013), where multiplication by the *n*/NZ factor treats cases where both x_i and y_i are zero as missing data. This factoring is useful for ensuring that two species will not be deemed more similar only because they were never confused with a third species. With Canberra distances in hand, we represented the confusions among species in the form of a dendrogram, where our study species are positioned according to information in the confusion matrix of Figure 1. We drew the dendrogram using a Lance-Williams clustering analysis (Lance and Williams 1966) with the complete-linkage clustering method (farthest neighbors clustering). In the process of drawing our dendrogram, we tested different combinations of inter-specific distance metrics and clustering algorithms. None of the distance metrics commonly used to construct phylogenies was designed for the type of data in our confusion matrix, which has a large number of values that are equal or close to zero. In the end, we settled on the Canberra distance with a Lance-Williams clustering algorithm because this option gave us the simplest results, which could be easily related to the distribution of confusions observed in Figure 1. Our use of Canberra distances is also justified by the frequent use of this metric as a dissimilarity index on ranked lists and other strictly positive, discrete variables in computer science (Jurman *et al.* 2009).

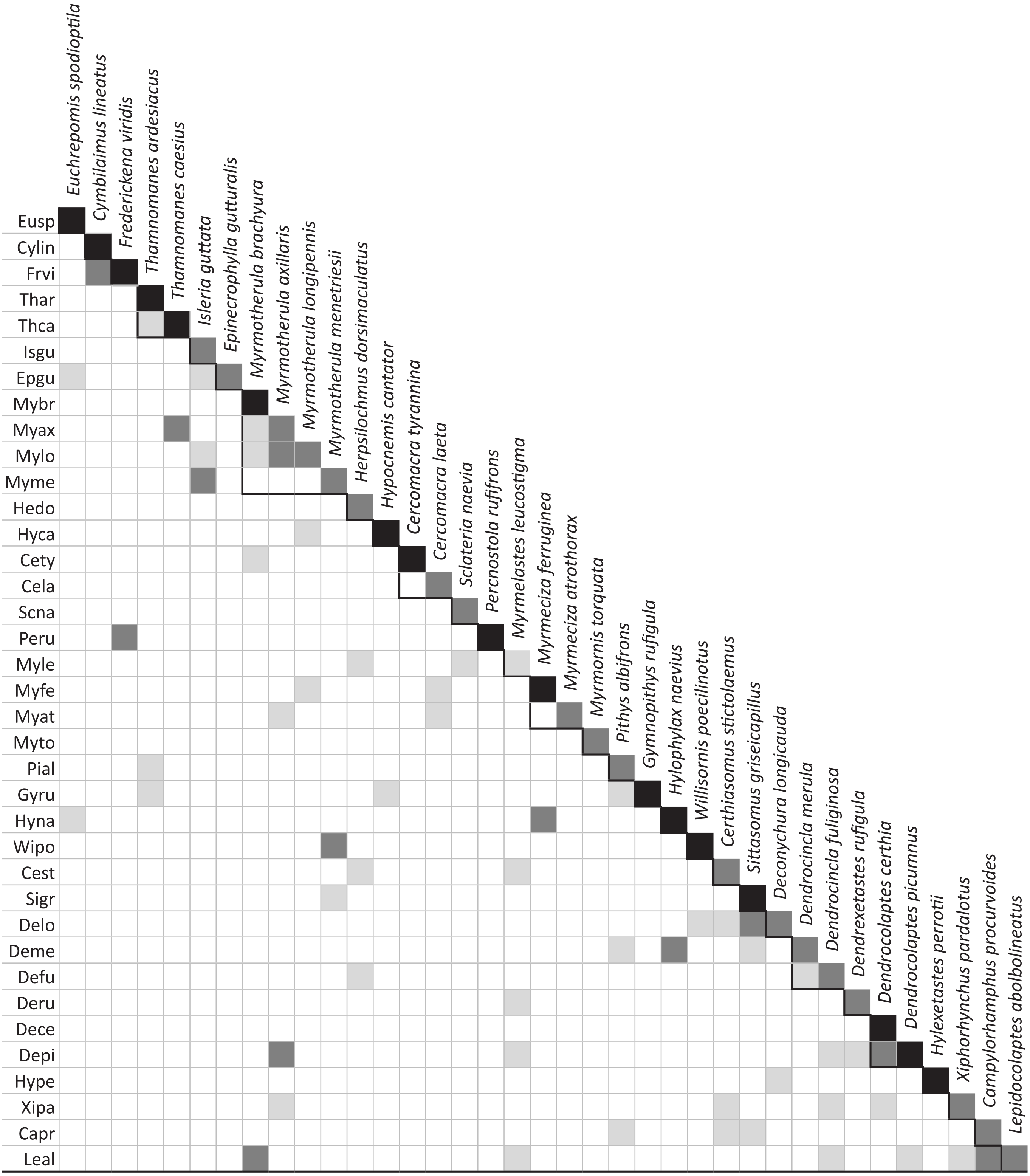

FIGURE 1. Triangular confusion matrix summarizing the quiz results, with correct answers on the diagonal and wrong answers to the left of the diagonal. Species are sorted in taxonomic order across rows and columns; codes on the left are abbreviations of the species names on the right. The color of each cell corresponds to the number of times the column species was identified as the corresponding row species: white stands for 0, light grey for 1, dark grey for values ≥ 2 and ≤ 10, and black for values > 10. Confusions between species of the same genus are outlined by a thin black line.

RESULTS

We obtained the collaboration of 20 experts, each of whom took the bird identification quiz once. The joint results from the 20 tests returned 820 answers. Out of those, 469 (57%) were correct identifications, 179 (22 %) were left blank, 128 (16%) were mistaken by species that occur in the study area and the remaining 44 answers

(5%) were mistaken by other species that do not occur in the study area. Expert performances in the test varied substantially: seven (35%) had an exceptionally good performance, and four (20%) had an exceptionally poor performance, *i.e.* they answered correctly less often than would be expected in a binomial experiment with 0.5 probability of guessing a question right. The remaining 10 participants performed at the standard level (Table 1).

Upon excluding answers from participants with an exceptionally poor performance, we summarized the results by species, as shown in Table 2. Evidently, some species are easier to identify than others: out of 41 species tested, 14 (34%) were identified correctly much more often than if the answer were determined by the toss of a fair coin (including *Cymbilaimus lineatus*, *Thamnophilus murinus* and *Thamnophilus punctatus*, which were always correctly identified by all observers). Only one species, *Myrmelastes leucostigma*, was so hard to identify that the experts got the species right less often than expected by the toss of a fair coin. For the 26 remaining species (64%) we found no evidence of difference between the outcome of the test and the results of a Binomial experiment with probability of success equal to 0.5. That is, the majority of species were neither extremely easy nor extremely hard to identify.

The null model analysis of the confusion matrix (represented in Figure 1) shows that confusions were more frequent between taxonomically closer species than between relatively distant ones. The observed mean distance between confused species of 5.8 taxonomic units was lower than every single one of the 10,000 simulated mean distances (Figure 2). The probability of obtaining a distance as low as the observed one is thus lower than 0.0001; we reject the null hypotheses with P < 0.0001. The two species that were most frequently confused were the antbirds *Willisornis poecilonotus* and *Myrmotherula menetriesii* with six confusions out of 38 times in which either species was heard (16%). The dendrogram generated from the Canberra distance matrix is consistent with the confusions found in the triangular matrix (Figure 3). Thirteen out of 18 (72%) branches on the dendrogram correspond to confusion points on the triangular matrix. Note how the antbirds *Isleria guttata* and *Myrmotherula menetriesii* stand out for being the pair of species separated by the shortest Canberra distance.

TABLE 1. Bird-voice identification results for the 20 experts involved in this study, showing the number of blank answers ("Blank"), answers with a species that does not occur in the study area ("Off-site"), and answers with a wrong species from the study area ("Plain wrong"). The column "Correct" shows the number of correct answers. "P" indicates the binomial probability of obtaining a number of correct answers as extreme or more extreme than the observed, given the total number of trials and a probability of success equal to 0.5. Rows F, J, K, L, O, R, S, and T add to 40, and not to 41 trials, because they included the *T. punctatus* recording that was removed from the analyses.

Observer	Blank	Off-site	Plain wrong	Correct	P
Observer A*	23	0	8	10	0.0007
Observer B	5	3	9	24	0.1744
Observer C**	0	0	1	40	<0.0001
Observer D	11	0	4	26	0.0586
Observer E*	3	22	7	9	0.0002
Observer F	14	4	6	16	0.1340
Observer G*	25	0	9	7	<0.0001
Observer H	16	1	5	19	0.4372
Observer I*	22	1	8	10	0.0007
Observer J**	4	3	4	29	0.0032
Observer K**	4	1	2	33	<0.0001
Observer L**	0	1	1	38	<0.0001
Observer M**	5	0	7	29	0.0057
Observer N	8	0	10	23	0.2663
Observer O**	1	1	3	35	<0.0001
Observer P	11	1	5	24	0.1744
Observer Q	11	3	4	23	0.2663
Observer R	14	0	4	22	0.3179
Observer S**	0	0	13	27	0.0192
Observer T	1	4	11	24	0.1340

* Right answer probability significantly lower than 0.5 in a two-tailed test with P = 0.05.

** Right answer probability significantly higher than 0.5 in a two-tailed test with P = 0.05.

TABLE 2. Summary of species-specific quiz results, showing the number of times each species was left in blank ("Blank"), mistaken for a species outside the study area ("Off-site"), or mistaken for a species from the study area ("Plain wrong"). Columns "Correct" and "n" show the number of correct answers and the number of times the species was heard by participants, respectively. "P" is the binomial probability of obtaining a number of correct answers as extreme, or more extreme than observed, given n attempts and a probability of success equal to 0.5.

Species	Blank	Off-site	Plain wrong	Correct	n	P
Euchrepomis spodioptila	3	1	1	12	17	0.0717
Cymbilaimus lineatus**	1	0	0	13	14	<0.0001
Frederickena viridis	1	0	4	14	19	0.0318
Thamnophilus murinus**	0	0	0	18	18	<0.0001
Thamnophilus punctatus**,***	0	0	0	8	8	0.0039
Thamnomanes ardesiacus	6	0	0	16	22	0.0262
Thamnomanes caesius	7	0	3	12	22	0.4159
Isleria guttata	3	1	5	2	11	0.0327
Epinecrophylla gutturalis	4	2	1	10	17	0.3145
Myrmotherula brachyura	7	0	2	13	22	0.2617
Myrmotherula axillaris	0	0	6	8	14	0.3953
Myrmotherula longipennis	0	1	5	6	12	0.6128
Myrmotherula menetriesii	4	0	4	8	16	0.5982
Herpsilochmus dorsimaculatus	4	0	2	7	13	0.5000
Hypocnemis cantator	6	0	0	12	18	0.1189
Cercomacra cinerascens**	2	0	0	18	20	0.0002
Cercomacra tyrannina**	1	2	0	18	21	0.0007
Cercomacra laeta	1	0	2	8	11	0.1133
Sclateria naevia**	0	0	1	8	9	0.0195
Percnostola rufifrons**	0	0	2	15	17	0.0012
_Myrmelastes leucostigma_*	3	0	5	1	9	0.0195
Myrmeciza ferruginea**	6	0	0	18	24	0.0113
Myrmeciza atrothorax	6	0	1	8	15	0.5000
Myrmornis torquata**	0	2	1	10	13	0.0461
Pithys albifrons	6	0	4	6	16	0.2272
Gymnopithys rufigula**	2	1	2	17	22	0.0084
Hylophylax naevius	2	0	3	13	18	0.0481
Willisornis poecilinotus	5	1	5	11	22	0.5841
Certhiasomus stictolaemus	2	4	2	4	12	0.1208
Sittasomus griseicapillus**	0	0	3	18	21	0.0007
Deconychura longicauda	4	0	3	7	14	0.6047
Dendrocincla merula	2	0	5	7	14	0.6047
Dendrocincla fuliginosa	3	1	3	8	15	0.5000
Glyphorynchus spirurus**	1	0	0	14	15	<0.0001
Dendrexetastes rufigula	6	2	0	10	18	0.4072
Dendrocolaptes certhia**	1	1	2	13	17	0.0245
Dendrocolaptes picumnus	3	1	3	13	20	0.1316
Hylexetastes perrotii**	0	0	1	11	12	0.0032
Xiphorhynchus pardalotus	2	0	3	10	15	0.1508
Campylorhamphus procurvoides	0	0	1	2	3	0.5000
Lepidocolaptes albolineatus	2	1	5	5	13	0.2905

* Difficult species, with a number of correct answers lower than expected in a two-tailed test with significance level P = 0.05.

** Easy species, with a number of correct answers higher than expected in a two-tailed test with significance level P = 0.05.

*** One of the _T. punctatus_ vocalizations used in the study had its identification questioned by experts and was removed from results.

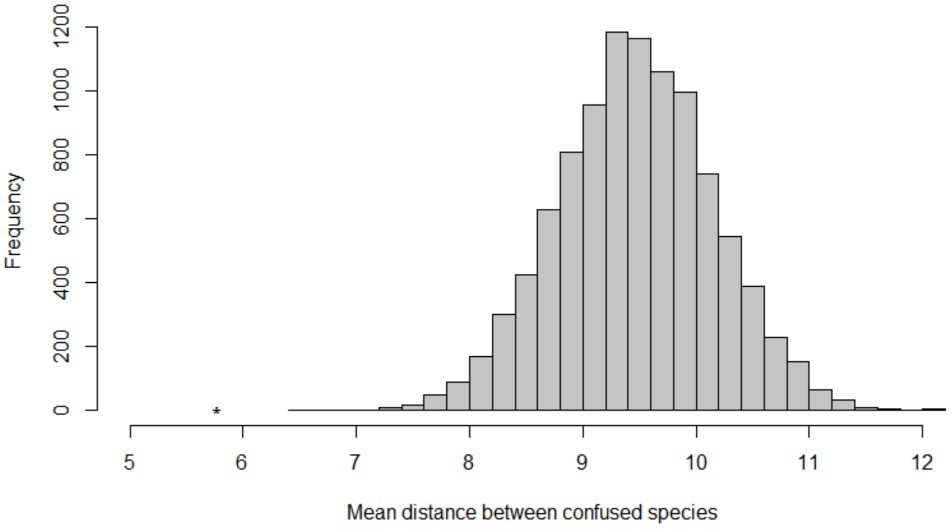

FIGURE 2. Observed mean taxonomic distance between confused species (*) and histogram of the simulated mean distances between confused species in 10,000 randomized matrices. Values on the *y-axis* indicate the number of random matrices with a mean distance between confused species equal to the corresponding value in the *x* axis.

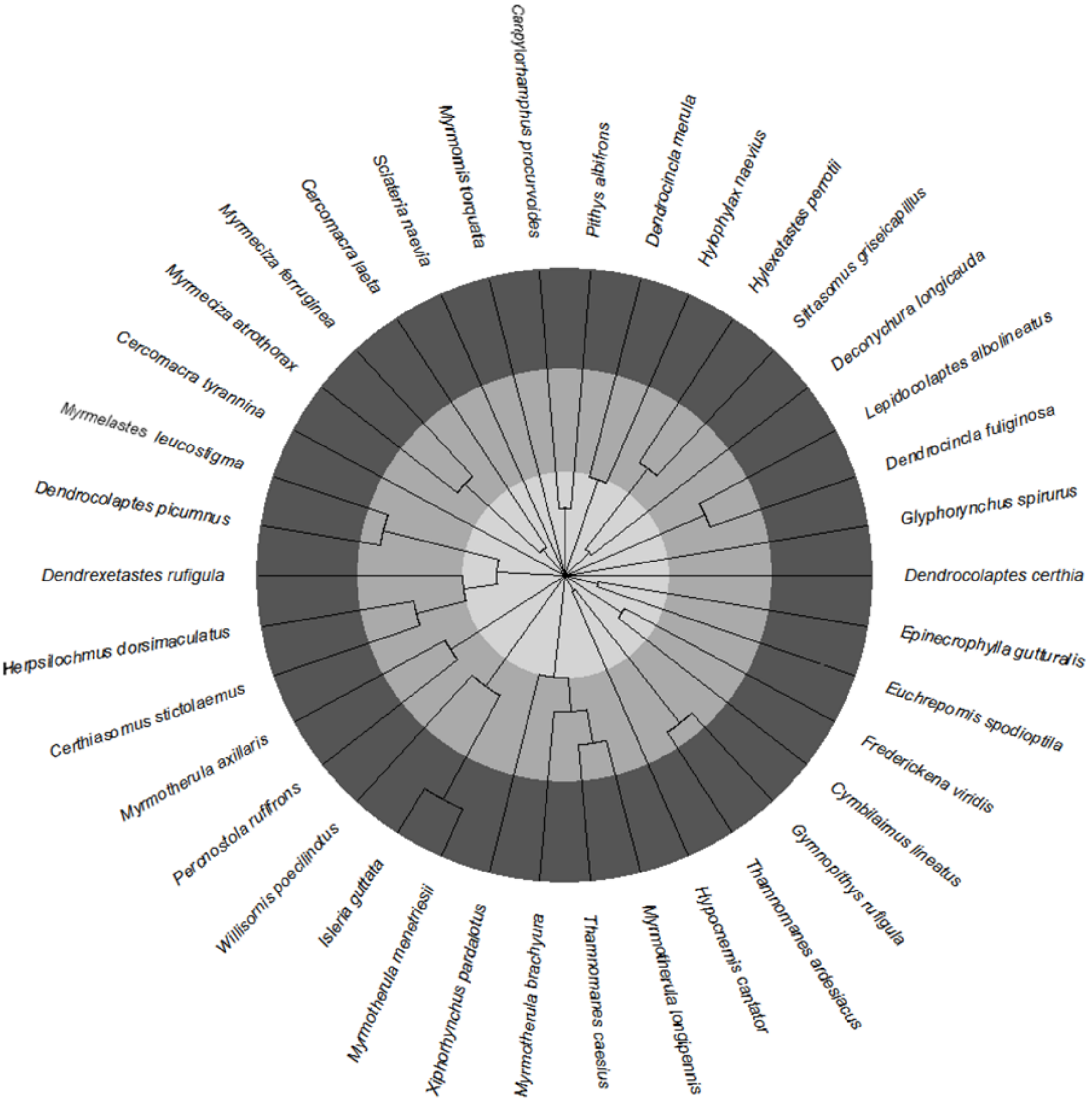

FIGURE 3. Confusion dendrogram based on the Canberra distance between species and Lance-Williams clustering algorithm. The distance from a branching point and the outer edge of the graphic is proportional to how easily observers could tell the two branches apart. Branching points in the dark gray area separate species that were frequently confused, while branching points in the light gray area separate easily distinguishable species or groups of species. For simplicity, this figure omits species that were never mistaken by other species.

DISCUSSION

Our results document the pervasiveness of errors in the identification of bird vocalizations, suggesting that such errors are inescapable and widespread in ornithological surveys (see also Lees *et al.* 2014). Even among identifications made by experts taking an on-line quiz, with the opportunity of listening to a fairly good recording while observing the respective sonogram, we found that more than 25% of identifications were wrong. Understanding how these errors happen is a key step towards lowering their frequency and improving our ability to obtain unbiased estimates of wildlife population parameters from bioacoustic data. We found considerable variation among experts in their ability to identify vocalizations, as well as substantial variation among species, in the frequency with which they were correctly identified. Although there are many possible errors, the probability of confusion between closely related species is higher than between relatively more distant ones, even when focusing on a phylogenetically restricted set of species. We acknowledge that our online quiz may have presented difficulties that are atypical of real-world processing of bioacoustics data, such as the relatively short time limit for answering questions, the lack of precise geographical information on where the recording was done, and the absence of environmental cues such as microhabitat and time of the day; nonetheless, these results are a motivation to improve ornithological training, to use sampling techniques that keep a permanent record of observations, and, most importantly, to incorporate the very real possibility of identification error in analyses of bioacoustics data.

Knowing that different experts have different backgrounds, it should come as no surprise that they performed very differently from each other in the identification quiz. Backgrounds varied in more than one way: while some experts learned the vocalizations in the field and probably relied mostly on sound for their quiz answers, others learned mostly in the lab, while processing audio recordings, and were more likely to take clues from the sonogram. There was also geographical variation in the backgrounds, with some experts having direct experience of listening to bird vocalizations from our study area and others having learnt mostly from experience in other parts of the Amazon. Experts from the latter group will be more likely to err by giving names of species that were not part of the study – especially when they are not informed about the geographic origin of the recordings. While it is unavoidable that different people will recall auditive memories differently, this problem could be minimized through the use of spaced-repetition learning (Donovan & Radosevich 1999) supported by digital tools (*e.g.* Cerqueira *et al.* 2013). Field practice will help observers memorize the voices of animals that they encounter most frequently; spaced-repetition learning, on the other hand, offers a means for adjusting the time studying each species, not according to the opportunity of encounter, but to how well the observer recalls one particular sound.

The observed variation among species with regard to ease of identification helps to sort out which species can be reliably studied based on bioacoustic data and which certainly require caution. Among the species in our study, *Cymbilaimus lineatus*, *Thamnophilus murinus* and *Thamnophilus punctatus* stand out for never having been mistaken by other species. Why would it be so? *T. murinus* and *C. lineatus* are respectively the fourth and seventh most frequently detected species among the antbirds and woodcreepers in our autonomous recordings database. The *T. punctatus'* song ends with a very peculiar rhythmic pattern, which could be the reason why it is particularly hard to confuse with other songs. These three species summarize what we believe to be two main factors facilitating correct identifications: commonness, already reported to play a role in species detection by Farmer *et al.* (2012), and peculiarity of the vocalization. On the opposite end of the difficulty spectrum, *Myrmelastes leucostigma*, stood out for being the only species with evidence for a correct identification probability lower than 0.5. *M. leucostigma*, along with the recurrently confused *Willisornis poecilonotus* and *Myrmotherula menetriesii*, may hold clues for understanding what makes a vocalization difficult to identify. Clearly, some species will be confused with each other because they sound alike—such as *W. poecilinotus* and *M. menetriesii*. However, the vocalization of *M. leucostigma* was confused with half a dozen species that don't particularly sound like each other. We don't know what caused these errors but wonder if there are acoustic traits that make a vocalization particularly difficult to memorize, regardless of its resemblance with other vocalizations. Besides the inherent difficulty of a sound and the obvious pairwise resemblance between species, it is also interesting to ask whether there are broader patterns that help one predict what are the most likely confusions. Both the dendrogram and the null model results support the reasonable idea that increasing phylogenetic relatedness increases the probability of confusion between species vocalizations. Our metric of relatedness is crude, but the final result is a contribution to understanding what types of misidentifications to expect as well as a motivation to take a detailed look at those exceptional situations where frequent confusion arises between unrelated species. This should be an incentive for keeping permanent records of bioacoustic surveys so that inevitable errors can be corrected and understood.

We see the work reported here as a first step towards understanding what are the most frequent misidentifications between species in the bioacoustic surveys of central Amazon birds. This work could be usefully expanded to a larger set of species and an online quiz where participants are informed *a priori* about the geographical context of the questions. We did not anticipate this to be a problem, but in hindsight, we believe we might be able to learn more about the possibility of misidentification if experts had a basis for excluding species that do not belong in our sample. A complementary work that could throw further light on the causes for confusion would be to quantify distance between vocalizations based not on expert answers to the quiz but on quantitative measures of the frequency and tempo of vocalizations. It would be particularly interesting to confront results from the two approaches and find out in what circumstances two vocalizations that have similar measures may be easily distinguished by the observers as well as when observers fail to discriminate sounds that are measurably different.

Knowledge of which animal sounds are most difficult to identify will contribute towards decreasing false positive errors and improving the quality of bioacoustic data. It is important to keep in mind, however, that as much as one values data quality and observer training, identification errors will never go away permanently. Whether the observer is a human being or a machine, there will be a non-negligible possibility of error. Future work should aim not only at reducing errors, but also at incorporating the possibility of errors in the analysis of bioacoustics data. Consideration of identification errors is particularly important when estimating population parameters from surveys of animal sounds. A reduction in parameter estimation bias can go a long way in advancing scientific knowledge and supporting management decisions. We hope that our results help improve the quantification of uncertainty about Amazon bird identification, and ultimately advance knowledge of their distribution and population dynamics.

ACKNOWLEDGEMENTS

We are grateful to Aída Rodrigues, Humberto Mohr, Luiza Figueira, Gilberto Fernandez, Francisco Carvalho Diniz, Marconi Cerqueira, and Gabriel McCrate, for their participation on a pilot quiz that was crucial for solving operational problems and fine-tuning quiz parameters. Quiz questions included recordings kindly shared by Josh Engel, Andrew Spencer, Jeremy Minns, Gabriel Leite, Nick Athanas, and the Xeno-Canto Foundation. Sandra Müller, Renato Silvano, and two anonymous reviewers read, commented and improved an earlier version of this manuscript. This work was only possible thanks to the help of ornithologists Marconi Cerqueira, Mario Cohn-haft, Thiago Costa, Sidnei Dantas, Thiago Orsi Laranjeiras, Curtis Marantz, Bret Whitney and 13 other experts who chose to remain anonymous. We are thankful for their feedback and generosity with their time.

REFERENCES

Cerqueira, M. C.; Cohn-Haft, M.; Vargas, C. F.; Nader, C. E.; Andretti, C. B.; Costa, T. V. V.; Sberze, M.; Hines, J. E.; Ferraz, G. & Burgman, M. 2013. Rare or elusive? A test of expert knowledge about rarity of Amazon Forest birds. *Diversity and Distributions*, 19: 710–721.

Cohn-Haft, M.; Whittaker, A. & Stouffer, P. C. 1997. A new look at the "species-poor" central Amazon: the avifauna north of Manaus, Brazil. *Ornithological Monographs*, 48: 205–235.

Donovan, J. J. & Radosevich, D. J. 1999. A meta-analytic review of the distribution of practice effect: now you see it, now you don't. *Journal of Applied Psychology*, 84: 795–805.

Farmer, R. G.; Leonard, M. L. & Horn, A. G. 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence. *Auk*, 129: 76–86.

Fristrup, K. M. & Mennitt, D. 2012. Bioacustical monitoring in terristrial environments. *Acoustics Today*, 8: 15–24.

Gotelli, N. J. & Graves, G. R. 1996. *Null models in ecology*. Washington: Smithsonian Institution Press.

Irestedt, M.; Fjeldså, J.; Nylander, J. A. & Ericson, P. G. 2004. Phylogenetic relationships of typical antbirds (Thamnophilidae) and test of incongruence based on Bayes factors. *BMC Evolutionary Biology*, 4: 23.

Jurman, G.; Riccadonna, S.; Visintainer, R. & Furlanello, C. 2009. Canberra distance on ranked lists. *Proceedings, Advances in Ranking* – NIPS 09 Workshop 22–27.

Lance, G. N. & Williams, W. T. 1966. Computer programs for hierarchical polythetic classification ("similarity analyses"). *Computer Journal*, 9: 60–64.

Lance, G. N. & Williams, W. T. 1967. Mixed-data classificatory programs I - agglomerative systems. *Australian Computer Journal*, 1: 15–20.

Lees, A. C.; Naka, L. N.; Aleixo, A.; Cohn-Haft, M.; Piacentini, V. Q.; Santos, M. P. D.; & Silveira, L. F. 2014. Conducting rigorous avian inventories: Amazonian case studies and a roadmap for improvement. *Revista Brasileira de Ornitologia*, 22: 107–120.

MacKenzie, D. I.; Nichols, J. D.; Lachman, G. B.; Droege, S.; Royle, J. A. & Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilites are less than one. *Ecology*, 83: 2248–2255.

McClintock, B. T.; Bailey, L. L.; Pollock, K. H. & Simons, T. R. 2010. Experimental investigation of observation error in anuran call surveys. *Journal of Wildlife Management*, 74: 1882–1893.

Miller, D. A.; Nichols, J. D.; Gude, J. A.; Rich, L. N.; Podruzny, K. M.; Hines, J. E. & Mitchell, M. S. 2013. Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. *PLoS One*, 8: e65808.

Miller, D. A.; Nichols, J. D.; McClintock, B. T.; Grant, E. H. C.; Bailey, L. L. & Weir, L. A. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, 92: 1422–1428.

Moyle, R. G.; Chesser, R. T.; Brumfield, R. T.; Tello, J. G.; Marchese, D. J. & Cracraft, J. 2009. Phylogeny and phylogenetic classification of the antbirds, ovenbirds, woodcreepers, and allies (Aves: Passeriformes: infraorder Furnariides). *Cladistics*, 25: 386–405.

Naka, L. N.; Stouffer, P. C.; Cohn-Haft, M.; Marantz, C. A.; Whittaker, A. & Bierregaard, R. O. J. 2008. *Voices of the Brazilian Amazon, v. 1. Birds of the terra firme forests north of Manaus: Guianan area of endemism.* Manaus: Editora INPA.

R Development Core Team. 2013. *A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Version 3.0.2. http://www.R-project.org/*

Remsen, J. V. J.; Cadena, C. D.; Jaramillo, A.; Nores, M.; Pacheco, J. F.; Pérez-Emán, J.; Robbins, M. B.; Stiles, F. G.; Stotz, D. F. & Zimmer, K. J. 2014. *A classification of the bird species of South America. American Ornithologists' Union. Version [20/05/2014] http://museum.lsu.edu/~Remsen/SACCBaseline.html.*

Somervuo, P.; Koskela, S.; Pennanen, J.; Nilsson, R. H. & Ovaskainen, O. 2016. Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, 32: 2920–2927.

Sousa-Lima, R. S.; Norris, T. F.; Oswald, J. N. & Fernandes, D. P. 2013. A review and inventory of fixed autonomous recorders for passive acoustic monitoring of marine mammals. *Aquatic Mammals*, 39: 23–53.

Associate Editor: Alexander Lees.