Observer Effects and Avian-Call-Count Survey Quality: Rare-Species Biases and Overconfidence

Author(s): Robert G. Farmer, Marty L. Leonard and Andrew G. Horn
Source: The Auk, 129(1):76-86.
Published By: The American Ornithologists' Union
URL: http://www.bioone.org/doi/full/10.1525/auk.2012.11129

The Auk 129(1):76–86, 2012
© The American Ornithologists' Union, 2012.
Printed in USA.

Observer Effects and Avian-Call-Count Survey Quality: Rare-Species Biases and Overconfidence

Robert G. Farmer,[1] Marty L. Leonard, and Andrew G. Horn

Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax, Nova Scotia B3H 4J1, Canada

Abstract.—Wildlife monitoring surveys are prone to nondetection errors and false positives. To determine factors that affect the incidence of these errors, we built an Internet-based survey that simulated avian point counts, and measured error rates among volunteer observers. Using similar-sounding vocalizations from paired rare and common bird species, we measured the effects of species rarity and observer skill, and the influence of a reward system that explicitly encouraged the detection of rare species. Higher self-reported skill levels and common species independently predicted fewer nondetections (probability range: 0.11 [experts, common species] to 0.54 [moderates, rare species]). Overall proportions of detections that were false positives increased significantly as skill level declined (range: 0.06 [experts, common species] to 0.22 [moderates, rare species]). Moderately skilled observers were significantly more likely to report false-positive records of common species than of rare species, whereas experts were significantly more likely to report false-positives of rare species than of common species. The reward for correctly detecting rare species did not significantly affect these patterns. Because false positives can also result from observers overestimating their own abilities ("overconfidence"), we lastly tested whether observers' beliefs that they had recorded error-free data ("confidence") tended to be incorrect ("overconfident"), and whether this pattern varied with skill. Observer confidence increased significantly with observer skill, whereas overconfidence was uniformly high (overall mean proportion = 0.73). Our results emphasize the value of controlling for observer skill in data collection and modeling and do not support the use of opinion-based (i.e., subjective) indications of observer confidence. *Received 13 June 2011, accepted 14 December 2011.*

Key words: acoustic survey, bias, birds, call count surveys, citizen science, detection, nondetection, observer effects.

Effets des observateurs et qualité des inventaires par le dénombrement des chants : biais sur les espèces rares et excès de confiance

Résumé.—Les inventaires fauniques sont sujets aux erreurs de non-détection et de faux positifs. Afin de déterminer les facteurs qui influent sur l'incidence de ces erreurs, nous avons effectué une enquête sur internet par simulation de points d'écoute d'oiseaux et mesuré les taux d'erreur parmi les observateurs volontaires. À l'aide de vocalisations similaires d'espèces d'oiseaux rares et communes, nous avons mesuré les effets de la rareté des espèces et des compétences des observateurs, ainsi que l'influence d'un système de récompenses qui encourage explicitement la détection des espèces rares. Les niveaux auto-déclarés de compétences plus élevés et les espèces communes ont indépendamment prédit moins de non-détections (étendue des probabilités : 0,11 [experts, espèces communes] à 0,54 [observateurs moyens, espèces rares]). Les proportions globales de détections qui étaient de faux positifs ont augmenté significativement alors que les niveaux de compétences diminuaient (étendue : 0,06 [experts, espèces communes] à 0,22 [observateurs moyens, espèces rares]). Les observateurs moyennement compétents étaient significativement moins susceptibles de rapporter de faux enregistrements positifs d'espèces rares que d'espèces communes, alors que les experts étaient significativement plus susceptibles de rapporter des faux positifs d'espèces rares que d'espèces communes. La récompense pour détecter correctement des espèces rares n'a pas influé significativement sur ces patrons. Parce que les faux positifs peuvent aussi résulter du fait que les observateurs surestiment leurs propres compétences (« excès de confiance »), nous avons finalement testé si les certitudes des observateurs d'avoir enregistré des données sans erreurs (« confiance ») tendaient à être incorrectes (« excès de confiance »), et si ce patron variait avec les compétences. La confiance des observateurs augmentait significativement avec les compétences de l'observateur, alors que l'excès de confiance était uniformément élevé (proportion moyenne globale = 0,73). Nos résultats soulignent l'importance de tester les compétences des observateurs lors de la collecte de données et de la modélisation et n'encouragent pas l'utilisation des indications de la confiance de l'observateur basées sur l'opinion (c'est-à-dire subjectives).

[1] E-mail: farmerb@dal.ca

The Auk, Vol. 129, Number 1, pages 76–86. ISSN 0004-8038, electronic ISSN 1938-4254. © 2012 by The American Ornithologists' Union. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press's Rights and Permissions website, http://www.ucpressjournals.com/reprintInfo.asp. DOI: 10.1525/auk.2012.11129

This article contains sound files that may be accessed by reading the full-text version of this article online at dx.doi.org/10.1525/auk.2012.11129.

Broad-scale and long-term ecological data sets collected by volunteers form an increasingly important component of contemporary wildlife management (Silvertown 2009). Among their many uses, these data sets monitor populations of birds (Link and Sauer 1998, Julliard et al. 2006, Kery and Schmid 2006, Hewson et al. 2007), anurans (de Solla et al. 2005, Lotz and Allen 2007, North American Amphibian Monitoring Program [see Acknowledgments]), invertebrates (Kremen et al. 2011, Maritimes Butterfly Atlas 2011 [see Acknowledgments]), and many marine organisms (e.g., Goffredo et al. 2010, Ward-Paige et al. 2010). Each survey typically records point-count or detection–nondetection data, or both, from a given location over a known time interval, providing broad spatial and temporal data coverage at a minimal cost.

In surveys of birds and anurans, a substantial proportion of detections are made by ear, without visual confirmation of a species' identity (Dawson and Efford 2009). Unfortunately, accurate auditory identifications can be difficult because many species sound alike (e.g., Robbins and Stallcup 1981, McClintock et al. 2010a). In field settings, different habitats and background noises also affect detection probability (Pacifici et al. 2008). Consequently, data collected by auditory surveys generally incorporate some amount of unavoidable observation error.

In spite of such error, volunteer surveys can be scientifically valuable if analyzed appropriately (i.e., if uncontrolled variability in detectability can be reduced to less than that of population variability; Johnson 2008). The need among managers for good-quality, broad-scale, long-term ecological data is increasing because of recent and ongoing challenges to global ecosystem stability (e.g., U.S. North American Bird Conservation Initiative Committee 2010). Hence, developing methods to extract such information from these surveys is a highly topical and active research concern (Elphick 2008). Reducing the influence of observer error is an important component of this research.

Observer-level errors on detection–nondetection surveys can be divided into two main types: nondetections and false positives (Royle and Link 2006). Nondetections occur when a species is present but not recorded, whereas false positives occur when a species is absent but is nonetheless recorded. False positives are more serious errors because they usually result from the misidentification of a species that is actually present; thus, they are often accompanied by concurrent nondetections (Bart 1985). Under most wildlife survey designs (including our own), the absence of a species is not explicitly recorded; hence, we refer to "nondetections" instead of "false negatives," because the latter term implies a declaration-of-absence.

The problem of incomplete detection (i.e., nondetection) in animal surveys has been studied for decades, particularly in the avian literature (e.g., Bart 1985, Marsden 1999), and direct estimation of corresponding probabilities is becoming routine (e.g., Diefenbach et al. 2003, Pellet and Schmidt 2005, Etterson et al. 2009; but see Rosenstock et al. 2002, Johnson 2008). False-positive probabilities, on the other hand, are typically assumed to be negligible (e.g., MacKenzie et al. 2009) and, therefore, have received less attention. There is growing evidence from studies of anurans and birds, however, that the frequency of false positives in auditory surveys can be appreciable. For instance, a controlled experiment measuring frog and toad call recognition errors found that 5% of all detection records were incorrect (McClintock et al. 2010a). Similarly, a study that modeled the occupancy of five bird species using repeated field visits along a North American Breeding Bird Survey (BBS) route estimated false-positive probabilities per detection event of up to 0.165 (Royle and Link 2006). Four other sets of controlled birdsong simulations showed that false positives comprised 1–14% of the total number of detections (mean = 5.8%; Bart 1985, Simons et al. 2007, Alldredge et al. 2008, Campbell and Francis 2011). Mathematical simulations have shown that failing to account for false positives of these magnitudes can lead to substantially biased estimates of species occupancy parameters (Royle and Link 2006, McClintock et al. 2010b, Miller et al. 2011).

At present, there are limited practical opportunities to correct for both false positives and nondetections simultaneously. Current published approaches that make such corrections ("misclassification models") require data from replicated surveys (e.g., multiple visits made during the same season; Royle and Link 2006, McClintock et al. 2010b, Miller et al. 2011). Unfortunately, without some indication of the reliability of the observation (Miller et al. 2011), misclassification modeling may yield biased occupancy estimators in the presence of varying levels of observer skill (Fitzpatrick et al. 2009) and when error rates are not consistent among sites (McClintock et al. 2010b). By design, they are also not suitable for surveys that lack replicated data (e.g., most BBS routes, which are surveyed once annually). Collectively, most current study designs and modeling approaches therefore have a limited ability to address important detection errors.

One approach to reducing the influence of detection errors is to address factors that contribute to their occurrence (Raitt 1981, Johnson 2008, McClintock et al. 2010a). We propose that an observer's "state of mind," which we define here as being the sum of conscious and unconscious biases that can affect decision-making behavior (e.g., Croskerry 2002, Lane et al. 2007), might constitute such a factor. Although previous authors have speculated that an observer's "attitude" (Faanes and Bystrak 1981), "carelessness" (Robbins and Stallcup 1981), and preferences and expectations (Balph and Balph 1983) might lead to identification errors on call count surveys, to our knowledge, there has been little quantitative research addressing this overall theme in ecology.

Nonetheless, such sources of error could be quite important. For instance, birdwatchers—and possibly surveyors of other taxa—are often motivated to detect and report the presence of rare species (Sullivan et al. 2009), and we hypothesize that such a preference might bias an observer to both detect more rare species under ambiguous circumstances ("observer expectancy effects"; Miller and Turnbull 1986, Lane et al. 2007) and, similarly, to be more attentive to the sounds of rare species ("search-image" detection biases; Callahan et al. 2003). These biases might lead to correspondingly fewer nondetections and more false positives for rare species than for common species. On the other hand, rarer species could instead be prone to more nondetections than common species if an observer arbitrarily rules out the possibility of a given rare species being present at all, on the basis of its rarity (the "playing the odds" bias; Croskerry 2002).

Exploring this theme, Bart (1985) reanalyzed an experimental call-count survey data set, in part to determine whether observers tend to detect particular species more often than others. He indeed found that detection error rates varied among species; however, his focus was not on the detection of rare versus common species specifically. Two recent studies have shown that detection error rates vary among rare and common species: species that call less often on field recordings of bird choruses tend to be associated with greater numbers of detection errors than frequently calling species (Rempel et al. 2005, Campbell and Francis 2011). However, those studies did not test for mechanisms underlying this pattern, for instance whether these errors tend to arise from a lack of observer knowledge or confusion with common species. Further research is thus needed that specifically controls for the effects of observer skill and the potential for rare and common species to sound alike.

Along with biases for or against the detection of rare species, unfounded observer confidence ("overconfidence"; Moore and Healy 2008) in a particular species identification might also be an important source of detection errors. An overconfident observer tends to overestimate his or her performance on a given task and, thus, is more prone to making detection errors than less overconfident observers, all else being equal. Given that overconfidence tends to occur more commonly among self-assessed experts than among novices (Larrick et al. 2007), and that many call count surveys involve expert volunteers (e.g., Sauer et al. 1994, Genet and Sargent 2003), overconfidence might explain a number of false-positive errors in survey data sets. However, we are not aware of research that has quantified its prevalence in this ecological context.

We used an Internet-based survey that mimicked an avian field point count to address these knowledge gaps. We determined rates of nondetections, false positives, and overconfidence among a set of volunteer observers to determine (1) whether observers of varying skill levels are more or less prone to detect rare species more or less often than similar-sounding common species; (2) whether an explicit incentive to correctly detect rare species affects error rates; and lastly, (3) whether overconfidence is common among observers of different skill levels.

METHODS

We created an Internet-based survey designed to mimic what observers might hear during an avian point count. The survey was composed of 16 simulated bird choruses ("scenarios") of known species, each lasting 30 s. We recruited volunteer observers to participate in the survey using e-mails sent to rare-bird and natural-history e-mail listservers in the Maritimes provinces, Canada ($n = 3$ listservers), and the northeastern United States (hereafter "New England"; $n = 3$ listservers; see Acknowledgments), and by word-of-mouth. Upon visiting the survey website, observers were first presented with an introductory page asking that they have a basic familiarity with the vocalizations of 38 candidate bird species, which we indicated might be presented in the survey. Only 12 species were actually used. We provided hyperlinks to examples of each candidate species' vocalizations. Observers were told that the featured choruses typified birds found in mixed or predominantly coniferous forest habitats (including wet brush) of eastern North America, but they were given no further information about the structure or contents of the testing scenarios.

Following this initial screening, observers were asked to declare their skill level from a list of five options (No Experience, Beginner, Moderate, Advanced, and Expert) that were not defined further. They were then asked to listen to each of the 16 scenarios once, manually beginning playback of each new scenario when ready, and then to indicate which birds were heard in each scenario using only the checklist of 38 candidate species. Replaying the scenario was possible but explicitly discouraged. Observers were not asked to count the number of individuals calling. Finally, to gauge their confidence and test for overconfidence, the survey asked observers to indicate at the end of each scenario whether they believed that they had correctly identified all species that were present.

We created all scenarios using audio samples of vocalizations (i.e., calls and songs) collected with permission from the Macaulay Library of the Cornell Laboratory of Ornithology and modified to remove background noises and normalize volume levels using the free audio manipulation software AUDACITY (see Acknowledgments). Each scenario featured the vocalizations of 6 species sampled with replacement from a pool of 12 species (consisting of 6 similar-sounding species pairs of opposing rarities; Table 1). With the exception of the

TABLE 1. Species used in the survey scenario recordings, grouped by species pairs (A–F) and rarity classes (common or rare), assigned according to the number of Maritimes Breeding Bird Atlas squares in which each species was present (percentage of 1,499 possible squares, in parentheses).

Species pair	Common name	Scientific name	Rarity
A	Alder Flycatcher	*Empidonax alnorum*	Common (65.7)
A	Olive-sided Flycatcher	*Contopus cooperi*	Rare (30.2)
B	American Robin	*Turdus migratorius*	Common (84.5)
B	Rose-breasted Grosbeak	*Pheucticus ludovicianus*	Rare (26.6)
C	Black-capped Chickadee	*Poecile atricapillus*	Common (77.6)
C	Boreal Chickadee	*P. hudsonicus*	Rare (38.8)
D	Dark-eyed Junco	*Junco hyemalis*	Common (70.2)
D	Palm Warbler	*Setophaga palmarum*	Rare (37.6)
E	Swainson's Thrush	*Catharus ustulatus*	Common (60.4)
E	Veery	*C. fuscescens*	Rare (37.0; M only)[a]
F	Song Sparrow	*Melospiza melodia*	Common (73.2)
F	Lincoln's Sparrow	*M. lincolnii*	Rare (27.0)

[a] "M only" indicates species that are relatively rare in the Maritimes but common in New England and, thus, were scored as "Common" for New England survey results.

Black-capped and Boreal chickadees (for which we used *chick-a-dee*–type calls), all vocalizations used in the scenarios were songs. Vocalizations ranged in length from 0.8 s (Alder Flycatcher) to 2.5 s (Song Sparrow) and were repeated three times per species, arranged arbitrarily within the scenarios.

We overlapped the transitions between ~90% of successive vocalizations to make scenarios comparable to a natural field situation. The maximum length of time between the remaining nonoverlapping vocalizations was ~1 s. To add standardized natural background noise to each scenario, we also superimposed a sequence of ambient cricket noises taken from a Macaulay audio sample on the sequence of bird vocalizations (maximum cricket amplitude [dB] was <1% of peak birdsong amplitude). The loudness of each vocalization was consistent among all species (sound files 1 and 2; listen to audio files by reading the full-text version of this article online at dx.doi. org/10.1525/auk.2012.11129).

One member of each species pair was randomly assigned to half of the scenarios; the second half of the scenarios featured the other member. In this way, no two members of a species pair appeared together simultaneously. Hence, false positives involving the species pairs could largely be interpreted as mistakes for the rarer or for the common variant. All scenarios had six distinct vocalizations (representing one member of each of the six species pairs), repeated three times each (Table 2).

We duplicated each scenario and alternated the duplicates randomly alongside the originals, for a total of 16 scenarios presented to each observer (Table 2). We informed observers that every second scenario would be "scored" and that correctly detecting rarer species was worth more points than correctly detecting common species. We then posted and regularly updated the top five high scores alongside user ID codes on the survey website. Our intent was to create and measure the effect of an explicit incentive to detect rare species on detection error rates. Observers were not told that the scenarios were duplicated, and we assumed that the scenarios were too similar-sounding and complex to be recognized as such. We did not expect this randomized, alternating design to show any important learning-effects biases; nonetheless, we controlled for any such systematic differences between earlier and later scenarios (see below).

We traded off the statistical need to present a large number of scenario replicates to our observers against the need to present realistic (longer) survey lengths. Our survey length of 30 s was substantially shorter than that of typical roadside point counts, which tend to last for 3–5 min, but roadside anuran survey research has shown that most species detections occur within the first 60 s (Shirose et al. 1997). Also, the species richness we presented was small per scenario (n = 6) and, thus, arguably manageable under these constraints. Hence, we assume that the challenge posed to our volunteer observers was appreciable, but not unreasonable.

Modeling details.—We defined a correct detection as occurring when a species that was present in a scenario was reported as such, and false-positive detection as occurring when a species that was not present in a scenario was similarly reported as being present. The probability of making a correct detection for a given species is equivalent to 1 minus the probability of making a nondetection error; here, we modeled correct detections in place of nondetections because the conceptual interpretation is more intuitive. Rates of correct (and non-) detections vary independently of false positives.

To determine the effect of species rarity on the incidence of false positives, we first recognized that many "phantom" species (sensu Bart and Schoultz 1984, McClintock et al. 2010b) that did not appear in any scenario were nonetheless identified repeatedly from the survey's list of 38 candidate species (Table 3). As with the "playback" (i.e., non-phantom) species pairs, we defined each phantom species as being either rare or common so that their data records could be modeled. Here, we determined the relative rarity values for each phantom species again using the Maritimes Breeding Bird Atlas detection records (2006–2010), but using the percentage of atlas squares occupied by the most abundant of the "rare" playback species (38.8%) as the threshold value distinguishing "rare" phantom species from "common" phantom species (Table 3). The maximum percentage of atlas squares occupied by "rare" phantom species was 18.3%; the minimum percentage of atlas squares occupied by a "common" phantom species was 61.4% (Table 3).

One phantom species (Eastern Phoebe) was common in the northeastern United States compared with most Maritimes regions (Weeks and Harmon 2011). Hence, we scored its detection records as "rare" if observations came from Maritimes survey participants and "common" if they came from New England survey participants. To simplify statistical analyses, we also arbitrarily discarded detection records for phantom species detected <7 times out of 4,025 total detection records (Table 3). We also discarded records from the single "Beginner" because there was no replication of this skill level.

We used generalized linear mixed models (GLMMs) to determine expected probabilities of correct detections and expected proportions of all detections that were false positives. Generalized linear mixed models incorporate random effects structures that recognize group-level deviations from overall patterns (Venables and Ripley 2002). We modeled correct

TABLE 2. Summary of experimental design of our Internet-based survey. "Scenarios" are the separate audio tracks played sequentially to the observers.

Item	n
Scenarios	16 (8 unique × 2 for incentive treatment)
Total species	12 (6 vocalization group pairs × 2)
Number of scenarios in which a given species is present	8 (4 × 2 for incentive treatment)
Species vocalizing per scenario	6
Discrete vocalizations per species, per scenario	3
Discrete vocalizations per scenario (all species)	18

Table 3. Species that were candidates for detection but not included in the scenario recordings (“phantom” species). Rarity classes (Common or Rare) were assigned to those species that were reported at least 7 times among 4,025 species records (n = 19 records or <0.5% of the total). Rarity classes were assigned according to the number of Maritimes Breeding Bird Atlas (MBBA) squares in which each species was present (percentage of 1,499 possible squares, in parentheses); “rare” species were found in <38.8% of atlas squares.

Common name	Scientific name	Rarity[a]
American Woodcock	*Scolopax minor*	
Barred Owl	*Strix varia*	
Belted Kingfisher	*Ceryle alcyon*	
Black-and-white Warbler	*Mniotilta varia*	Common (61.4)
Black-throated Green Warbler	*Setophaga virens*	
Common Grackle	*Quiscalus quiscula*	
Common Nighthawk	*Chordeiles minor*	
Common Yellowthroat	*Geothlypis trichas*	
Eastern Phoebe	*Sayornis phoebe*	Rare (18.3; M only)[b]
Eastern Towhee	*Pipilo erythrophthalmus*	
European Starling	*Sturnus vulgaris*	
Fox Sparrow	*Passerella iliaca*	Rare (11.1)
Great Horned Owl	*Bubo virginianus*	
Hairy Woodpecker	*Picoides villosus*	
Hermit Thrush	*Catharus guttatus*	Common (72.4)
Ovenbird	*Seiurus aurocapilla*	
Pine Warbler	*S. pinus*	Rare (4.8)
Red-eyed Vireo	*Vireo olivaceus*	Common (74.6)
Red-tailed Hawk	*Buteo jamaicensis*	
Rock Pigeon	*Columba livia*	
Scarlet Tanager	*Piranga olivacea*	Rare (6.2)
Eastern Whip-poor-will	*Caprimulgus vociferus*	
Willow Flycatcher	*Empidonax traillii*	Rare (1.9)
Wilson’s Warbler	*Cardellina pusilla*	Rare (14.3)
Yellow Warbler	*S. petechia*	
Yellow-rumped Warbler	*S. coronata*	Common (73.4)

[a] Rarity and percent MBBA square values were calculated only for those phantom species detected 7 times or more among all observers and scenarios, because only data from these phantom species were included in the predictive models.

[b] “M only” indicates species that are relatively rare in the Maritimes but common in New England and, thus, scored as “Common” for New England survey results.

detections and false-positive proportions as binomial responses and incorporated random effects structures that accounted for differences in error rates among observers (both models) and species pairs (correct detection model only). Our choice to model false positives as proportions of all detections is consistent with previous studies (e.g., Simons et al. 2007, Alldredge et al. 2008, McClintock et al. 2010a).

Each of the models of correct detections and false positives allowed us to estimate error rates while measuring the influence of several predictors. We used the GLMMs to model (1) how the rates of each type of error varied among rare and common species; (2) the effect of the incentive treatment rewarding the correct detection of rare species over common ones; (3) how observer skill was related to error rates; and (4) any skill- and incentive-dependent differences (interactions) in the detection of species of each rarity class. To correct for skill-dependent changes in observer ability over the course of the survey (e.g., learning, changes in interest level), we also included (5) the chronological scenario number and its interaction with observer skill as additional covariates.

To predict the probability of making a correct detection for a given species rarity class and scenario, we built a data set consisting of a record for each correct detection (1 = the species was present and detected) and each nondetection (0 = the species was present but not detected) and excluding all false positives. We used a total of 4,416 records of correct detections (n = 2,864) and nondetections (n = 1,552).

We modeled the expected probability of making a correct detection for a given species on a given scenario as a mixed-effects Bernoulli process using the package lme4 in R, version 2.13.0 (Bates and Maechler 2010, R Development Core Team 2011). In this model, in addition to recognizing differences in correct detection probability among observers as random intercepts, we also recognized variation in mean correct detection probability between species pair-groups, given that some species’ calls are more easily detected than others (Alldredge et al. 2007):

$$\begin{aligned}\text{logit}[P(Y_{ijkl} = 1)] &= \beta_0 + \beta_1 \times \text{rarity}_i + \beta_2 \times \text{skill}_j + \beta_3 \times \text{scenario}_k \\ &+ \beta_4 \times \text{skill}_j : \text{scenario}_k + \beta_5 \times \text{skill}_j : \text{rarity}_i + \beta_6 \times \text{incentive}_k \\ &+ \beta_7 \times \text{rarity}_i : \text{incentive}_k + b_{1j} + b_{2l} \qquad (1)\end{aligned}$$

where $Y_{ijkl} = 1$ when a species is correctly scored as being present; i = 1 of 2 rarity classes; $j = 1,\ldots, 52$ observers; $k = 1,\ldots, n_j$ scenarios completed per observer; and $l = 1,\ldots, 6$ species pairs. Random effects b_{1j} and b_{2l} are independently and normally distributed intercepts for observers and for species pairs, respectively, with means zero and with standard deviations estimated from the data.

We then constructed a second data set to model the expected proportion of all detections for each observer-within-scenario that were false positives for each of the rare and common species groups. For instance, if observer A, listening to scenario 1, incorrectly reported the presence of two rare species and one common species and correctly reported four common species and three rare species, his false-positive proportion would be 0.2 and 0.1 for rare and common species, respectively. In total, we modeled 1,429 false positive proportions. We estimated the proportion of false positives per rarity class per scenario as a mixed-effects binomial process with the same predictors as equation 1, but with a simpler random-effects structure, as follows:

$$\begin{aligned}\text{logit}[P(\text{Y}_{ijk} = 1)] &= \beta_0 + \beta_1 \times \text{rarity}_i + \beta_2 \times \text{skill}_j + \beta_3 \times \text{scenario}_k \\ &+ \beta_4 \times \text{skill}_j : \text{scenario}_k + \beta_5 \times \text{skill}_j : \text{rarity}_i + \beta_6 \times \text{incentive}_k \\ &+ \beta_7 \times \text{rarity}_i : \text{incentive}_k + b_{1j} \qquad (2)\end{aligned}$$

where Y_{ijk} is the proportion of all detections that were incorrect (false-positive) for a given observer, scenario, and species rarity class; i = 1 of 2 rarity classes; $j = 1,\ldots, 52$ observers; $k = 1,\ldots, n_j$ scenarios completed per observer; and b_{1j} is a normally distributed random intercept for observers with mean zero and standard deviation estimated from the data.

To measure and model confidence levels among survey participants, we first asked observers at the end of each scenario whether they believed that they had correctly accounted for all species present. If they answered “yes,” we considered that scenario and its responses to be “confident.” We then calculated the proportion of scenarios completed by each observer that were confident.

We also calculated the proportion of overconfident scenarios. We defined an overconfident scenario as one in which an observer made at least one detection error while also declaring confidence. This measure thus indicated the probability that a given observer's declaration of confidence was incorrect.

Using generalized linear models (GLMs), we modeled both the incidence of declared confidence and the incidence of overconfidence as functions of observer skill. Our confidence data were collected at a different resolution than our detection data; here, each observer contributed one confidence record per scenario. Accordingly, we built the following models:

$$\text{logit}[P(Y_{1ij} = 1)] = \beta_0 + \beta_1 \times \text{skill}_i \qquad (3)$$

$$\text{logit}[P(Y_{2ik} = 1)] = \beta_0 + \beta_1 \times \text{skill}_i \qquad (4)$$

where $Y_{1ij} = 1$ occurs when a participant declares that a particular survey scenario was scored entirely correctly ("declared confidence") and $Y_{2ik} = 1$ occurs when a declaration of confidence is incorrect ("overconfidence"); $i = 1,\ldots, 52$ observers; $j = 1,\ldots, n_i$ scenarios completed per observer; and $k = 1,\ldots, m_i$ confident scenarios per observer.

All models were checked for fit quality by examining conventional or binned residual plots (Gelman and Hill 2007), and results were compared visually with plotted raw data to check for consistency. Unless otherwise specified, results are presented as means ± SD.

Results

We modeled data from observers representing three self-reported skill levels: "Moderate" ($n = 17$), "Advanced" ($n = 26$), and "Expert" ($n = 9$), from the Canadian provinces of New Brunswick, Nova Scotia, and Prince Edward Island and the New England states of Maine, New Hampshire, and Vermont. Most observers (80.8%) completed all 16 scenarios (mean number of scenarios completed = 14.15 ± 3.31). We suspect that those who failed to complete all 16 scenarios largely did so in error, rather than out of fatigue or disinterest. This is because the survey was composed of four webpages containing four scenarios each, and most of the missed scenarios were in groups of four sequential scenarios located on the same webpage.

Observers with higher skill levels were significantly more likely to correctly detect any given species than observers of lower skill levels (Fig. 1A and Table 4). Across all skill levels, all observers were also equally and significantly less likely to correctly detect rare species than common ones (Fig. 1A and Table 4). Neither the incentive nor the scenario number was significantly related to correct detection rates among and within observers and skill levels (Table 4).

The expected proportion of species correctly detected per scenario for each skill level ranged from 0.61 (95% CI: 0.40–0.79; Moderate) to 0.89 (95% CI: 0.77–0.95; Expert) for common species, and from 0.46 (95% CI: 0.27–0.67; Moderate) to 0.81 (95% CI: 0.63–0.91; Expert) for rare species (Fig. 1A). Subtracting these values from 1.0 gives a set of nondetection probabilities that range from 0.11 (Expert skill, common species) to 0.54 (Moderate skill, rare species).

Summed across both species rarity groups, the proportion of false positives declined significantly with increasing skill level (Table 5). However, skill level also interacted significantly with species rarity. Here, moderately skilled observers falsely detected common species more often than rare species, whereas experts falsely detected rare species more often than common ones (Fig. 1B and Table 5). Again, neither the incentive nor the scenario number was significantly related to the occurrence of false positives across or within skill levels (Table 5).

The expected proportion of false positives per scenario for each skill level ranged from 0.061 (95% CI: 0.043–0.085; Expert) to 0.218 (95% CI: 0.170–0.280; Moderate) for common species, and from 0.119 (95% CI: 0.087–0.164; Expert) to 0.120 (95% CI: 0.091–0.157; Moderate) for rare species (Fig. 1B). A tabular summary of the correct detection and false-positive frequencies, indexed by species and observer skill level, is given in the Appendix.

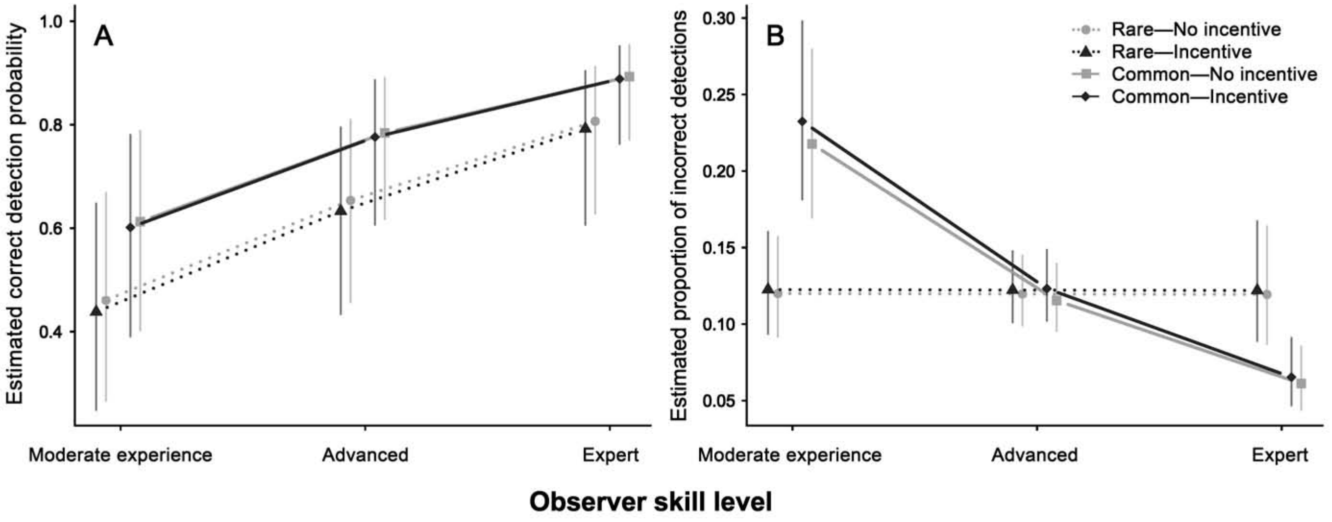

Fig. 1. Summary of (A) the predicted probability of correctly detecting a species and (B) the predicted proportion of false positives per scenario, grouped by both species rarity and whether the incentive to detect rare species was in effect, ordered by self-reported skill level. Error bars are 95% confidence intervals.

Table 4. Factors that affected the probability of correctly detecting a species on a given scenario (Equation 1; n = 4,416 correct and nondetections distributed among 52 observers). In this binomial model, σ_{b1}, the standard deviation about the observer random effects, is 0.72; and σ_{b2}, the standard deviation about the species-pair random effects, is 1.02. All values are on the logit scale.

Factor	Estimate	SE	z	P
(Intercept)	−0.516	0.547	−0.943	0.345
Rarity	−0.576	0.228	−2.531	0.011
Skill	0.820	0.198	4.143	<0.001
Scenario	0.017	0.023	0.727	0.467
Incentive	−0.046	0.102	−0.454	0.650
Skill: scenario	0.001	0.012	0.115	0.908
Rarity: skill	−0.039	0.115	−0.337	0.736
Rarity: incentive	−0.045	0.147	−0.307	0.759

Table 5. Factors that affected the number of false positives for a given scenario and species rarity group (Equation 2, n = 1,429 counts of false positives distributed among 52 observers). In this binomial model, σ_{b1}, the standard deviation about the observer random effects, is 0.41. All values are on the logit scale.

Factor	Estimate	SE	t	P
(Intercept)	−0.893	0.287	−3.114	0.002
Rarity	−1.228	0.224	−5.473	<0.001
Skill	−0.598	0.149	−3.997	<0.001
Scenario	0.0003	0.022	0.016	0.987
Incentive	0.065	0.101	0.651	0.515
Skill: scenario	−0.004	0.012	−0.370	0.711
Rarity: skill	0.632	0.112	5.649	<0.001
Rarity: incentive	−0.044	0.147	−0.302	0.762

The proportion of scenarios for which an observer declared confidence increased significantly with self-assessed observer skill (β_1 = 1.376 ± 0.148, P < 0.001; Equation 3 and Fig. 2A), with model-estimated values increasing from 0.079 (Moderate; 95% CI: 0.06–0.11) to 0.575 (Expert; 95% CI: 0.50–0.65). Among those surveyors who declared confidence on at least one survey scenario (n = 28), there was no significant difference in the amount of overconfidence among skill levels (β_1 = −0.366 ± 0.288, P = 0.204; Equation 4 and Fig. 2B). Model-estimated proportions of overconfident scenarios (overall mean = 0.73) ranged from 0.80 (Moderate; 95% CI: 0.64–0.90) to 0.66 (Expert; 95% CI: 0.54–0.76); this difference was not statistically significant.

Discussion

We found significant relationships between detection error rates and each of observer skill and species rarity. In our models, the

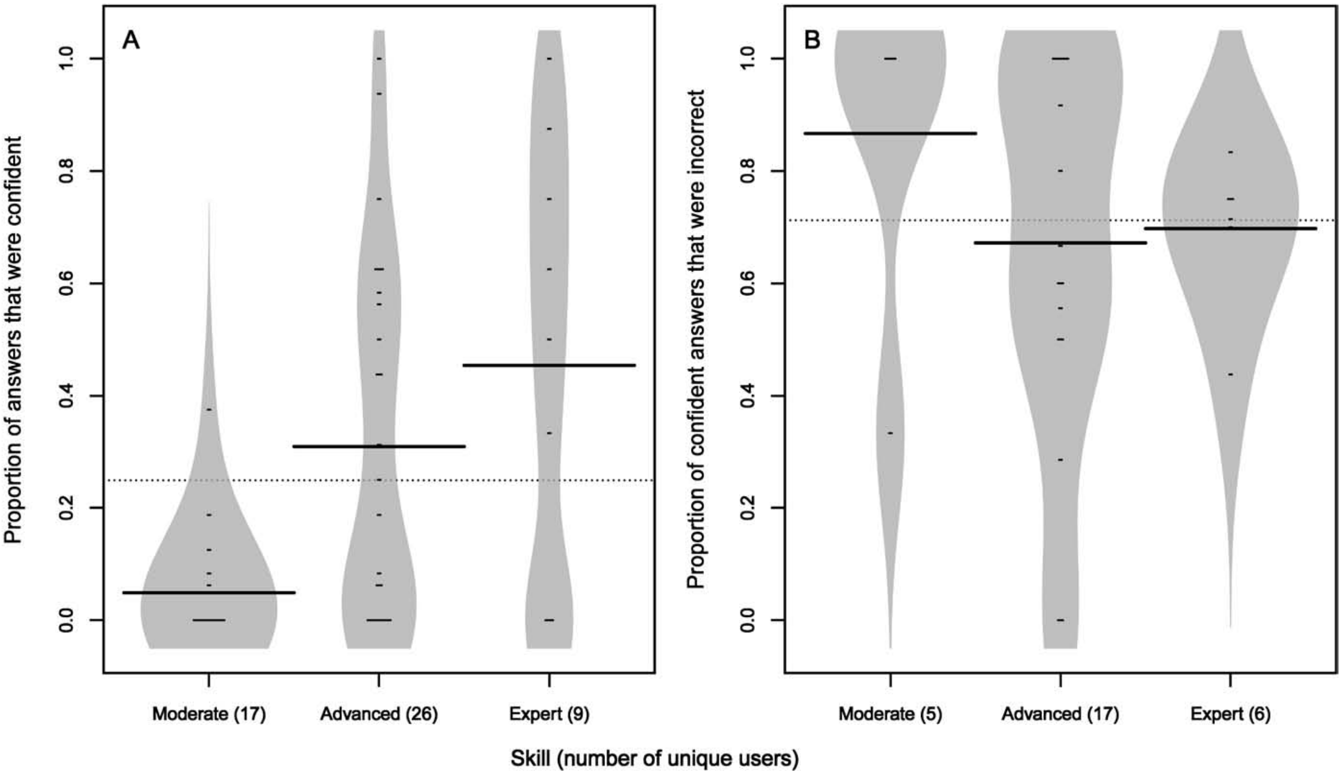

Fig. 2. Beanplot summary of the proportion of scenarios that were (A) "confident" or (B) "overconfident" (i.e., proportion of confident scenarios with at least one error), grouped by self-reported skill level. "Confident" scenarios were those in which an observer declared that he or she had correctly detected all species present. In each "bean," small ticks correspond to individual values for a given skill level and are scaled by length according to their frequency. The longer solid lines are the mean values for each bean, and the dotted line is the overall mean. Note that these are raw data values; modeled predictions differ only slightly.

probability of making a nondetection error decreased with observer skill and among common species, as did the proportion of responses that were false positives. A significant interaction between skill and species rarity for false positives also indicated that among moderately skilled observers, the majority of false positives were of common species, and among experts, the majority of false positives were of rare species. We also found no evidence that an incentive to detect rare species affected error rates. Finally, observers of all skill levels were overconfident, with 73% of scenarios completed by confident observers of any skill level having at least one error. Below, we address each of these findings in turn.

The range of observed nondetection error probabilities (0.11–0.54) is consistent with the results from similar experiments. For instance, Alldredge et al. (2007) calculated values ranging from 0.17 to 0.59, depending on the species and singing rate. Similarly, Simons et al. (2007) found probabilities ranging from 0.26 to 0.68 overall, and Bart (1985) reported a probability of 0.30 on average. Using unretouched field recordings, Campbell and Francis (2011) also reported a value of 0.23. Contributing to these errors were slower singing rates (Alldredge et al. 2007), louder background noise (Simons et al. 2007), and increased local species rarity (Campbell and Francis 2011). Our research shows that, controlling for similarity in vocalizations, increasing rarity at the population scale and decreasing observer skill are also important predictors of species detection.

The observed frequency of false positives (0.06–0.22) was also consistent with past research. For instance, Lotz and Allen (2007) found similar proportions of scenarios that had at least one incorrectly detected anuran (in the absence of similar-sounding equivalents; 0.19 and 0.238, two regions), and Campbell and Francis (2011) found that ~14% of bird detection records could not be confirmed from simultaneous field recordings, which suggests that they were false positives. Our results were, however, higher than some previously published rates—from Simons et al. 2007 (0.01–0.04), Alldredge et al. 2008 (0–0.01), and McClintock et al. 2010a (0.05)—possibly because we had ambiguous candidate species broadcast over a relatively short period (i.e., higher difficulty), and likely a lower average observer skill level. Although Royle and Link (2006) also found the probabilities of detecting a bird species, given its absence ("p_{10}"), to range from 0.007 to 0.165, this statistic differs from what has been calculated from most studies, including ours (i.e., proportion of all detections that are incorrect), and so is not directly comparable. Nonetheless, both our results and this related observation suggest that false-positive rates in avian field detection data are nontrivial.

Several previous studies have shown significant differences among individual observers in their ability to detect and identify animal sounds (e.g., Shirose et al. 1997, Link and Sauer 1998, McLaren and Cadman 1999, Alldredge et al. 2007). However, few have found relationships specifically tied to observer skill, probably because most used a homogeneous group of expert participants, who are all highly competent in spite of differences in their amateur or professional status, or in their high absolute levels of experience (e.g., Genet and Sargent 2003, Lotz and Allen 2007, McClintock et al. 2010a). Conversely, our more heterogeneous group demonstrated expected decreases in detection errors with increasing observer skill. Our use of self-assessment of observer skill therefore appeared to successfully capture real differences in ability; this suggests that self-assessment can be an efficient alternative to quizzes or other more elaborate testing approaches (e.g., Genet and Sargent 2003, McClintock et al. 2010a).

Not surprisingly, we found that rare species were correctly identified less often than their common variants among all skill levels (Fig. 1A and Table 4). Interestingly, more skilled observers tended to submit false-positive records of rare species more often than common ones, whereas the reverse was the case for moderately skilled observers, who incorrectly detected common species more often than rare species (Fig. 1B and Table 5). One explanation for this interaction might be that more experienced observers have a greater familiarity with rarer species than novices and, therefore, may be aware of a greater number of alternatives for a given vocalization (e.g., Faanes and Bystrak 1981).

These results further suggest that naively modeled data collected mostly from experts may overestimate the occupancy or abundance of rare species. Where similar-sounding rare and common species are not present simultaneously, the nondetection errors associated with these false positives would also underestimate occupancy or abundance of common species. Conversely, surveys using less-skilled volunteers would overestimate common species occupancy and underestimate occupancy for similar-sounding rare species. Hence, our data support existing evidence that heterogeneous mixtures of surveyor skill levels can lead to biased detection and occupancy estimates (e.g., Fitzpatrick et al. 2009).

In light of these detection biases, survey designers must control for skill level among participants (e.g., Kepler and Scott 1981, Genet and Sargent 2003), incorporating rare-species interaction effects as appropriate. This is important when working with both single-visit and single-observer designs (present study) and repeated-sampling designs (the preferred approach; e.g., Fitzpatrick et al. 2009). Independent of any skill effects, the nontrivial nondetection and false-positive rates we observed also emphasize that neither form of error can be ignored.

Contrary to our expectations, we found no evidence that an incentive to correctly detect rare species contributed to differences in detection error rates across or within skill levels (Tables 4 and 5). We therefore have no evidence that the intrinsically competitive designs of surveys such as eBird (Sullivan et al. 2009)—which publishes observers' names alongside their detection records and encourages the detection of rarities—could introduce bias and affect error rates. However, our survey design offered only a weak incentive—in particular, no guarantee of publicity among one's peers. Hence, we suggest that these results be regarded as preliminary.

Finally, we found that observers of higher self-assessed skill levels tended to be more confident about the correctness of their identifications than less-skilled observers (Fig. 2A). These more-skilled observers also tended to have fewer false-positive and more correct responses (Fig. 1). Thus, the higher confidence of experts was justified in principle. However, our specific measurement of observer confidence was whether observers believed that they had made zero detection errors on a given survey scenario, and this specific outcome was actually quite rare. We found a consistent overconfidence among observers of all skill levels (Fig. 2B). Thus, an apparent increase in the level of observer confidence with increasing self-assessed skill seems to have outpaced the proportionately smaller increase in actual ability, causing the level of overconfidence to remain consistent across observers of different skill levels.

A promising model-based approach to account for the nontrivial instances of both nondetection errors and false positives in detection survey data requires that observers provide a measure of the reliability of each species detection (Miller et al. 2011). Because we found widespread levels of overconfidence in our data set, we believe that a subjective declaration of certainty (e.g., a rating of observer confidence from 1 to 10; Larrick et al. 2007) for use as such a reliability measure may not be appropriate, and more objective measures such as the anuran chorus-intensity values used by Miller et al. (2011) are preferable. For bird surveys, observers could also note the call type (e.g., the *chick-burr* call for a Scarlet Tanager, a highly confident identification, vs. its less distinctive, Robin-like song, a less confident identification) or, more generally, the type of detection method used (e.g., heard vs. seen; Miller et al. 2011). Recording such detailed detection evidence is not an impractical option, because it has already been successfully implemented on broad scales in several Canadian breeding bird atlases (e.g., Maritimes Breeding Bird Atlas 2006–2010), which require observers to classify detections using a range of breeding evidence codes. Another important complementary strategy is to emphasize to volunteers the value of being conservative with one's species identifications, for instance recording no observations when in doubt (sensu McClintock et al. 2010a), which can reduce the incidence of false positives arising from overconfidence.

In sum, our results show that an observer's state of mind has important implications for detection errors. Rates of nondetections and false positives vary with species rarity and with observer skill, indicating skill-dependent biases in the detection of rare species. Furthermore, overconfidence may be an important factor contributing to these errors. Therefore, approaches to managing these differences that focus on controlling for differences in observer skill and encouraging observer objectivity should improve survey data quality. We hope that this research leads to increasingly fruitful use of the valuable, ongoing contributions of thousands of volunteers.

Acknowledgments

We thank the volunteer observers who participated in the survey, including members of the Nova Scotia Bird Society and Bird Studies Canada, as well as readers of the e-mail listservers NatureNS, NatureNB, PEIBirds, Maine Birding, Vermont Birds, and NH-Birds. We also thank G. Breed and J. Flemming for helpful discussions. Research was conducted under Dalhousie University's Social Sciences and Humanities Human Research Ethics Board approval no. 2009-2065, and with financial support from the Natural Sciences and Engineering Research Council of Canada to R.G.F. (postgraduate scholarship) and M.L.L. (Discovery grant). AUDACITY software is available at audacity.sourceforge.net. The Maritimes Breeding Bird Atlas is at www.mba-aom.ca/. The Maritimes Butterfly Atlas is at www.accdc.com/butterflyatlas.html. The North American Amphibian Monitoring Program is at www.pwrc.usgs.gov/naamp/.

Literature Cited

Alldredge, M. W., K. Pacifici, T. R. Simons, and K. H. Pollock. 2008. A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities. Journal of Applied Ecology 45:1349–1356.

Alldredge, M. W., T. R. Simons, and K. H. Pollock. 2007. Factors affecting aural detections of songbirds. Ecological Applications 17:948–955.

Balph, D. F., and M. H. Balph. 1983. On the psychology of watching birds: The problem of observer-expectancy bias. Auk 100:755–757.

Bart, J. 1985. Causes of recording errors in singing bird surveys. Wilson Bulletin 97:161–172.

Bart, J., and J. D. Schoultz. 1984. Reliability of singing bird surveys: Changes in observer efficiency with avian density. Auk 101:307–318.

Bates, D., and M. Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. [Online.] Available at cran.r-project.org/web/packages/lme4/index.html.

Bevier, L. R., A. F. Poole, and W. Moskoff. 2005. Veery (*Catharus fuscescens*). *In* The Birds of North America Online (A. Poole, Ed.). Cornell Lab of Ornithology, Ithaca, New York. [Online.] Available at bna.birds.cornell.edu/bna/species/142.

Callahan, J. S., A. L. Brownlee, M. D. Brtek, and H. L. Tosi. 2003. Examining the unique effects of multiple motivational sources on task performance. Journal of Applied Social Psychology 33:2515–2535.

Campbell, M., and C. M. Francis. 2011. Using stereo-microphones to evaluate observer variation in North American Breeding Bird Survey point counts. Auk 128:303–312.

Croskerry, P. 2002. Achieving quality in clinical decision making: Cognitive strategies and detection of bias. Academic Emergency Medicine 9:1184–1204.

Dawson, D. K., and M. G. Efford. 2009. Bird population density estimated from acoustic signals. Journal of Applied Ecology 46:1201–1209.

De Solla, S. R., L. J. Shirose, K. J. Fernie, G. C. Barrett, C. S. Brousseau, and C. A. Bishop. 2005. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. Biological Conservation 121:585–594.

Diefenbach, D. R., D. W. Brauning, and J. A. Mattice. 2003. Variability in grassland bird counts related to observer differences and species detection rates. Auk 120:1168–1179.

Elphick, C. S. 2008. How you count counts: The importance of methods research in applied ecology. Journal of Applied Ecology 45:1313–1320.

Etterson, M. A., G. J. Niemi, and N. P. Danz. 2009. Estimating the effects of detection heterogeneity and overdispersion on trends estimated from avian point counts. Ecological Applications 19:2049–2066.

Faanes, C. A., and D. Bystrak. 1981. The role of observer bias in the North American Breeding Bird Survey. Pages 353–359 *in* Estimating Numbers of Terrestrial Birds (C. J. Ralph and J. M. Scott, Eds.). Studies in Avian Biology, no. 6.

Fitzpatrick, M. C., E. L. Preisser, A. M. Ellison, and J. S. Elkinton. 2009. Observer bias and the detection of low-density populations. Ecological Applications 19:1673–1679.

Gelman, A., and J. Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, United Kingdom.

Genet, K. S., and L. G. Sargent. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildlife Society Bulletin 31:703–714.

Goffredo, S., F. Pensa, P. Neri, A. Orlandi, M. Scola Gagliardi, A. Velardi, C. Piccinetti, and F. Zaccanti. 2010. Unite research with what citizens do for fun: "Recreational monitoring" of marine biodiversity. Ecological Applications 20:2170–2187.

Hewson, C. M., A. Amar, J. A. Lindsell, R. M. Thewlis, S. Butler, K. Smith, and R. J. Fuller. 2007. Recent changes in bird populations in British broadleaved woodland. Ibis 149:14–28.

Johnson, D. H. 2008. In defense of indices: The case of bird surveys. Journal of Wildlife Management 72:857–868.

Julliard, R., J. Clavel, V. Devictor, F. Jiguet, and D. Couvet. 2006. Spatial segregation of specialists and generalists in bird communities. Ecology Letters 9:1237–1244.

Kepler, C. B., and J. M. Scott. 1981. Reducing bird count variability by training observers. Pages 366–371 *in* Estimating Numbers of Terrestrial Birds (C. J. Ralph and J. M. Scott, Eds.). Studies in Avian Biology, no. 6.

Kery, M., and H. Schmid. 2006. Estimating species richness: Calibrating a large avian monitoring programme. Journal of Applied Ecology 43:101–110.

Kremen, C., K. S. Ullman, and R. W. Thorp. 2011. Evaluating the quality of citizen-scientist data on pollinator communities. Conservation Biology 25:607–617.

Lane, K. A., J. Kang, and M. R. Banaji. 2007. Implicit social cognition and law. Annual Review of Law and Social Science 3:427–451.

Larrick, R. P., K. A. Burson, and J. B. Soll. 2007. Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). Organizational Behavior and Human Decision Processes 102:76–94.

Link, W. A., and J. R. Sauer. 1998. Estimating population change from count data: Application to the North American Breeding Bird Survey. Ecological Applications 8:258–268.

Lotz, A., and C. R. Allen. 2007. Observer bias in anuran call surveys. Journal of Wildlife Management 71:675–679.

MacKenzie, D. I., J. D. Nichols, M. E. Seamans, and R. J. Gutiérrez. 2009. Modeling species occurrence dynamics with multiple states and imperfect detection. Ecology 90:823–835.

Marsden, S. J. 1999. Estimation of parrot and hornbill densities using a point count distance sampling method. Ibis 141:377–390.

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

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010b. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. Ecology 91:2446–2454.

McLaren, A. A., and M. D. Cadman. 1999. Can novice volunteers provide credible data for bird surveys requiring song identification? Journal of Field Ornithology 70:481–490.

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

Miller, D. T., and W. Turnbull. 1986. Expectancies and interpersonal processes. Annual Review of Psychology 37:233–256.

Moore, D. A., and P. J. Healy. 2008. The trouble with overconfidence. Psychological Review 115:502–517.

Pacifici, K., T. R. Simons, and K. H. Pollock. 2008. Effects of vegetation and background noise on the detection process in auditory avian point-count surveys. Auk 125:600–607.

Pellet, J., and B. R. Schmidt. 2005. Monitoring distributions using call surveys: Estimating site occupancy, detection probabilities and inferring absence. Biological Conservation 123:27–35.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Raitt, R. J. 1981. Chairman's introductory remarks: Observer variability. Page 326 *in* Estimating Numbers of Terrestrial Birds (C. J. Ralph and J. M. Scott, Eds.). Studies in Avian Biology, no. 6.

Rempel, R. S., K. A. Hobson, G. Holborn, S. L. Van Wilgenburg, and J. Elliott. 2005. Bioacoustic monitoring of forest songbirds: Interpreter variability and effects of configuration and digital processing methods in the laboratory. Journal of Field Ornithology 76:1–11.

Robbins, C. S., and R. W. Stallcup. 1981. Problems in separating species with similar habits and vocalizations. Pages 360–365 *in* Estimating Numbers of Terrestrial Birds (C. J. Ralph and J. M. Scott, Eds.). Studies in Avian Biology, no. 6.

Rosenstock, S. S., D. R. Anderson, K. M. Giesen, T. Leukering, and M. F. Carter. 2002. Landbird counting techniques: Current practices and an alternative. Auk 119:46–53.

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835–841.

Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North American Breeding Bird Survey. Auk 111:50–62.

Shirose, L. J., C. A. Bishop, D. M. Green, C. J. MacDonald, R. J. Brooks, and N. J. Helferty. 1997. Validation tests of an amphibian call count survey technique in Ontario, Canada. Herpetologica 53:312–320.

Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology & Evolution 24:467–471.

Simons, T. R., M. W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk 124:986–999.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

U.S. North American Bird Conservation Initiative Committee. 2010. The State of the Birds 2010. [Online.] Available at www.stateofthebirds.org/2010/pdf_files/State%20of%20the%20Birds_FINAL.pdf.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S, 4th ed. Springer, New York.

Ward-Paige, C. A., C. Mora, H. K. Lotze, C. Pattengill-Semmens, L. McClenachan, E. Arias-Castro, and R. A. Myers. 2010. Large-scale absence of sharks on reefs in the greater-Caribbean: A footprint of human pressures. PLoS ONE 5: e11968.

Weeks, H. P., Jr., and P. Harmon. 2011. Eastern Phoebe (*Sayornis phoebe*) *In* The Birds of North America Online (A. Poole, Ed.). Cornell Lab of Ornithology, Ithaca, New York. [Online.] Available at bna.birds.cornell.edu/bna/species/094.

Associate Editor: D. H. Johnson

Appendix. Summary of correct-detection and false-positive data, grouped by species and observer skill level. "Correct" count data are the total number of correct detections for a given species among all observers and scenarios ("observer-scenarios"). The proportion correct (in parentheses) is the number of correct detections divided by the total number of times that species was played among all observer-scenarios. "False-positive" count data are the total number of false positives for a given species among all observer-scenarios. The proportion of false positives per scenario (in parentheses) is the number of false positives divided by the total number of observer-scenarios. This value is different from the modeled proportion of false positives per scenario (results here are summarized across multiple observers and scenarios).

			Moderate (248 observer-scenarios)		Advanced (372 observer-scenarios)		Expert (116 observer-scenarios)	
Group	Species[a]	Rarity	Correct (proportion correct)	False positive (per scenario)	Correct (proportion correct)	False positive (per scenario)	Correct (proportion correct)	False positive (per scenario)
A	ALFL	C	89 (0.7)	17 (0.07)	167 (0.86)	3 (0.01)	53 (0.95)	0 (0)
A	OSFL	R	96 (0.8)	8 (0.03)	171 (0.97)	7 (0.02)	60 (1)	4 (0.03)
B	AMRO	C	87 (0.69)	48 (0.19)	134 (0.71)	55 (0.15)	57 (0.92)	12 (0.1)
B	RBGR	R	55 (0.45)	8 (0.03)	88 (0.48)	9 (0.02)	41 (0.76)	2 (0.02)
C	BCCH	C	116 (0.92)	34 (0.14)	179 (0.95)	35 (0.09)	56 (0.9)	1 (0.01)
C	BOCH	R	52 (0.43)	6 (0.02)	100 (0.55)	2 (0.01)	39 (0.72)	3 (0.03)
D	DEJU	C	32 (0.25)	28 (0.11)	70 (0.37)	39 (0.1)	26 (0.42)	21 (0.18)
D	PAWA	R	25 (0.2)	14 (0.06)	62 (0.34)	25 (0.07)	21 (0.39)	13 (0.11)
E	SWTH	C	63 (0.5)	8 (0.03)	146 (0.76)	20 (0.05)	54 (0.95)	0 (0)
E	VEER	R[b]	89 (0.73)	19 (0.08)	153 (0.85)	16 (0.04)	59 (1)	1 (0.01)
F	SOSP	C	53 (0.44)	5 (0.02)	118 (0.65)	6 (0.02)	48 (0.92)	3 (0.03)
F	LISP	R	50 (0.39)	16 (0.06)	101 (0.53)	36 (0.1)	54 (0.84)	15 (0.13)
Phantom	BAWW	C		2 (0.01)		1 (0)		0
Phantom	EAPH	R[b]		22 (0.09)		5 (0.01)		0
Phantom	FOSP	R		3 (0.01)		32 (0.09)		1 (0.01)
Phantom	HETH	C		34 (0.14)		12 (0.03)		0
Phantom	PIWA	R		23 (0.09)		34 (0.09)		10 (0.09)
Phantom	REVI	C		4 (0.02)		14 (0.04)		0
Phantom	SCTA	R		14 (0.06)		26 (0.07)		7 (0.06)
Phantom	WIFL	R		11 (0.04)		14 (0.04)		1 (0.01)
Phantom	WIWA	R		5 (0.02)		23 (0.06)		12 (0.1)
Phantom	YRWA	C		7 (0.03)		13 (0.03)		5 (0.04)

[a] Abbreviations: ALFL = Alder Flycatcher (*Empidonax alnorum*); OSFL = Olive-sided Flycatcher (*Contopus cooperi*); AMRO = American Robin (*Turdus migratorius*); RBGR = Rose-breasted Grosbeak (*Pheucticus ludovicianus*); BCCH = Black-capped Chickadee (*Poecile atricapillus*); BOCH = Boreal Chickadee (*P. hudsonicus*); DEJU = Dark-eyed Junco (*Junco hyemalis*); PAWA = Palm Warbler (*Setophaga palmarum*); SWTH = Swainson's Thrush (*Catharus ustulatus*); VEER = Veery (*C. fuscescens*); SOSP = Song Sparrow (*Melospiza melodia*); and LISP = Lincoln's Sparrow (*M. lincolnii*); BAWW = Black-and-white Warbler (*Mniotilta varia*); EAPH = Eastern Phoebe (*Sayornia phoebe*); FOSP = Fox Sparrow (*Passerella iliaca*); HETH = Hermit Thrush (*C. guttatus*); PIWA = Pine Warbler (*S. pinus*); REVI = Red-eyed Vireo (*Vireo olivaceus*); SCTA = Scarlet Tanager (*Piranga olivacea*); WIFL = Willow Flycatcher (*E. traillii*); YRWA = Yellow-rumped Warbler (*S. coronata*).

[b] Rare only in the Maritimes provinces