Ecological Applications, 22(5), 2012, pp. 1665–1674

Experimental investigation of false positive errors in auditory species occurrence surveys

David A. W. Miller,[1,5] Linda A. Weir,[1] Brett T. McClintock,[2] Evan H. Campbell Grant,[1] Larissa L. Bailey,[3] and Theodore R. Simons[4]

[1]*United States Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708 USA*
[2]*National Marine Mammal Laboratory, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, Washington 98115 USA*
[3]*Colorado State University, Department of Fish, Wildlife and Conservation Biology, Fort Collins, Colorado 80523 USA*
[4]*United States Geological Survey, North Carolina Cooperative Fish and Wildlife Research Unit, Department of Biology, North Carolina State University, Raleigh, North Carolina 27695 USA*

Abstract. False positive errors are a significant component of many ecological data sets, which in combination with false negative errors, can lead to severe biases in conclusions about ecological systems. We present results of a field experiment where observers recorded observations for known combinations of electronically broadcast calling anurans under conditions mimicking field surveys to determine species occurrence. Our objectives were to characterize false positive error probabilities for auditory methods based on a large number of observers, to determine if targeted instruction could be used to reduce false positive error rates, and to establish useful predictors of among-observer and among-species differences in error rates. We recruited 31 observers, ranging in abilities from novice to expert, who recorded detections for 12 species during 180 calling trials (66 960 total observations). All observers made multiple false positive errors, and on average 8.1% of recorded detections in the experiment were false positive errors. Additional instruction had only minor effects on error rates. After instruction, false positive error probabilities decreased by 16% for treatment individuals compared to controls with broad confidence interval overlap of 0 (95% CI: −46 to 30%). This coincided with an increase in false negative errors due to the treatment (26%; −3 to 61%). Differences among observers in false positive and in false negative error rates were best predicted by scores from an online test and a self-assessment of observer ability completed prior to the field experiment. In contrast, years of experience conducting call surveys was a weak predictor of error rates. False positive errors were also more common for species that were played more frequently but were not related to the dominant spectral frequency of the call. Our results corroborate other work that demonstrates false positives are a significant component of species occurrence data collected by auditory methods. Instructing observers to only report detections they are completely certain are correct is not sufficient to eliminate errors. As a result, analytical methods that account for false positive errors will be needed, and independent testing of observer ability is a useful predictor for among-observer variation in observation error rates.

Key words: false negative; false positive; misclassification; occupancy; occurrence; species richness; survey.

Introduction

A fundamental challenge when making inferences about plant and animal populations is that their status and dynamics are nearly always imperfectly observed (Yoccoz et al. 2001). Ecological data reflect not only underlying ecological processes of interest, but also the observation process by which data were collected (Royle and Dorazio 2008). Accurate statistical inference necessitates reducing the potential for observation error by carefully designing studies to minimize such error and by properly accounting for these errors when using model-based estimation methods. This process requires a well-developed understanding of the types and causes of observation error in ecological studies (Simons et al. 2009).

One of the most commonly used methods for surveys of species occurrence is to establish apparent occupancy based on auditory, or a combination of auditory and visual, cues. Auditory surveys are used for many taxa, including as a primary sampling method for birds (Sauer et al. 2003, Simons et al. 2007), amphibians (Royle 2004, Weir and Mossman 2005), and bats (O'Farrell et al. 1999). As a result, many large-scale, long-term biodiversity surveys, such as the North American Breeding Bird Survey (Sauer et al. 2003), North American

Manuscript received 28 November 2011; accepted 27 February 2012; final version received 20 March 2012. Corresponding Editor: D. Brunton.
[5] E-mail: davidmiller@usgs.gov

Amphibian Monitoring Program (Weir and Mossman 2005), British Breeding Bird Survey (Newson et al. 2005), and Swiss Breeding Bird Survey (Kéry and Royle 2009), rely on auditory detections. Many of these surveys are volunteer-based, representing an extremely valuable but potentially error-prone data source that is increasingly employed for ecological monitoring (Cohn 2008, Silvertown 2009).

When investigating patterns of species occurrence and abundance both false negative and false positive errors can occur. False negative errors occur because it is generally impossible to detect every individual within a sampled area. Ecologists have long recognized this problem (e.g., Petersen 1896, Lincoln 1930, Jackson 1933) and have developed an extensive set of statistical approaches to address non-detection (e.g., Williams et al. 2002, Buckland et al. 2004, MacKenzie et al. 2006). False positive detections occur when species or individuals that are absent are erroneously detected. False positive errors are often assumed unimportant and have largely been ignored. However, a growing body of evidence shows that they are a frequent occurrence for many ecological data sets (e.g., O'Farrell et al. 1999, Acevedo et al. 2009, Knapp et al. 2009, Wright et al. 2009, Yoshizaki et al. 2009, Boessenkool et al. 2010, McClintock et al. 2010*b*, *c*, Shea et al. 2011, Farmer et al. 2012, Molinari-Jobin et al. 2012).

In the case of species occurrence data, the importance of false positive errors was demonstrated by Royle and Link (2006), who proposed a general estimator of occupancy when both false negative and false positive errors occur. Results from simulated and real data sets show that false positive detection probabilities even as low as 1–3% can lead to substantial overestimation of occupancy (Royle and Link 2006, Miller et al. 2011) and bias in estimators of colonization and extinction (McClintock et al. 2010*a*). For example, under a broad set of simulated conditions, Miller et al. (2011) found that when false positive detections occurred only 1% of the time, the standard occupancy estimator (MacKenzie et al. 2006) was positively biased by an average of 0.06 (and up to 0.29). This bias becomes even more severe with higher false positive detection probabilities. Using field data from pickerel frog (*Lithobates palustris*) call surveys, Miller et al. (2011) found that occupancy estimates were 2.15 times greater when false positive detections were not accounted for. Biases that result from false positive errors are greatest when species are rare (Miller et al. 2011). The costs associated with inaccurate inference owing to these errors can be substantial (Dalton 2009, Roberts et al. 2010)

Many important insights about auditory surveys have come from recent studies using an automatic playback system that enables field conditions to be simulated, originally developed by Simons et al. (2007). These studies have extended our understanding of the detection process and the validity of statistical estimators that attempt to account for imperfect detection (Simons et al. 2007, 2009, Alldredge et al. 2008, McClintock et al. 2010*a*; among others). A key finding of these studies, as well as others (e.g., Bart 1985, Genet and Sargent 2003, Lotz and Allen 2007, Campbell and Francis 2011, Farmer et al. 2012), is that false positive errors are more common than generally acknowledged. McClintock et al. (2010*a*, *b*) was the first to examine false positive detections in detail. They found that false positive error probabilities varied greatly among observers and among species played in the study, false positive errors rates were affected by distance from a call and ambient noise, and that error rates in the study could produce substantial bias in standard occupancy estimators.

Most studies rely on the belief that observer experience and training is sufficient to reduce or eliminate false positive errors. Even if efforts to reduce false positive errors increase false negative errors, established statistical methods can more efficiently deal with additional non-detections (but no false positives) than cases where both error types occur simultaneously (Royle and Link 2006, Miller et al. 2011). Thus, if observers are able to recognize and eliminate questionable observations, establishing study protocols that involved instructing observers to do so could greatly improve data quality.

If it is not possible to eliminate false positive errors through study design and training, then accurate inference will need to rely on model-based approaches that account for false positive errors (Royle and Link 2006, Miller et al. 2011). One method to improve occupancy estimates is to incorporate covariates that predict variation in detection error rates. For large-scale surveys, which use many observers and monitor large numbers of species, predicting among-observer and among-species variation will be especially useful. In the case of variation among observers, predictors of ability that are easily collected include measures of experience (e.g., years of experience conducting surveys), self-assessment of ability, and independent testing of skills. For among species variation, calls that are outside of the optimal hearing range (i.e., high or low spectral frequency) may be harder to hear and easier to confuse, and the number of calling individuals may also relate to error rates. Expectations may also affect error rates. Observers may tune out commonly occurring species thus increasing false negatives. Alternatively observers may over report commonly heard species thus increasing false positives.

A common protocol for anuran surveys is to record an index of abundance in addition to whether a species was observed or not (Royle 2004, Weir and Mossman 2005). Observers may be less likely to make false positive errors when a calling species is abundant. If it is possible to assign greater certainty to observations where the species was recorded as abundant, that could be exploited when estimating occupancy (Miller et al. 2011, Molinari-Jobin et al. 2012).

We present results from a study of false positive (and false negative) errors for methods typically used to determine patterns of species occurrence from auditory cues (e.g., Royle 2004). We focus on the following objectives: (1) quantifying false positive detection probabilities under conditions representative of standard field data collection protocols with a diverse variety of species and species combinations; (2) determining whether observer training and instruction is sufficient to eliminate or significantly reduce false positive errors; (3) examining potential predictors of among-observer (e.g., measures of ability and proficiency) and among-species (e.g., call characteristics, intensity, and frequency) differences in false positive error probabilities that may be used to improve study designs and to predict observation errors for model based approaches; and (4) determining if there is a relationship between false positive errors and the abundance of calling individuals.

Methods

Study design

We simulated conditions encountered during call surveys of pond-breeding amphibians using an automated playback system (Simons et al. 2007). The system uses a remote computer to control a series of amplified mp3 players that can be placed at in different configurations and at varying distances from observers (see Plate 1). This allowed us to design and reproduce precise sequences of frog calls simultaneously at multiple locations.

Our goal was to mimic conditions typical of many amphibian surveys where observers stand at or near a wetland or other habitat patch and calls come from one direction. To simulate the heterogeneity in calling volume and intensity that is encountered during surveys, we varied distances to calling individuals and the abundance of calling individuals at a site. We placed 12 speakers in a short grass field, with four players each at 20, 40, and 60 m in a single direction from observers. For each species we developed sound file using professionally produced recordings (used with permission of Lang Elliot, NatureSound Studio, Ithaca, New York, USA). Each file containing only a single species at one of two calling abundances: a single individual calling and multiple individuals with overlapping calls. Volume was equalized for all recordings to ~80 dB when measured 1 m from the speakers. This value is near sound pressure typical of many of the amphibians (Gerhardt 1975, Pough et al. 1992) and volumes used by McClintock et al. (2011*b*). True volumes in the field will depend on conditions, species, and number of calling individuals and our primary goal was to create a gradient of easy to hard-to-hear calls based on the combination of broadcast volume and distance to speaker. Dominant spectral frequencies for each species were measured using the frequency analysis tool in Adobe Audition (1024 Blackman-Harris fast Fourier transform, 0dBFS).

The experiment was conducted at the Patuxent Research Refuge in Maryland, USA. We recruited 31 volunteer observers that were familiar with the calls used in the study. Observers varied in ability and previous experience, ranging from professional ecologists with amphibian taxonomic expertise and years of experience monitoring amphibians, volunteers who regularly participated in monitoring surveys (e.g., North American Amphibian Monitoring Program), and observers that had only participated in one or two amphibian surveys. Observers were divided among 5 groups, with each group participating during a morning or an afternoon on 12, 13, or 14 November 2010. We chose the time of year to avoid any calling from wild amphibians.

Each observer participated in two sessions that each took ~50 minutes. During each session, they recorded detections for 90 separate calling trials. Each trial was meant to mimic a single call survey for a given wetland or listening station (i.e., a site). During each trial, calls of 0–4 species that came from a pool of 12 potential species were played (Table 1). We defined an observation as the data recorded for a single species by an observer during a given trial. The total number of observations during the experiment was (number of potential species) × (number of calling trials) × (number of observers). For each trial, we played calls of the selected species for 22 seconds, followed by 10 seconds of silence before the start of the next trial. Trials were grouped in sets of 10, after which a longer ~1-minute break was taken. Between the two sessions, a ~15-minute break occurred, during which an experimental treatment was administered. In addition to recording species that were heard, observers also recorded calling intensity following a protocol commonly used for amphibian surveys (1, individuals can be counted, there is space between the calls; 2, calls of individuals can be distinguished but there is some overlapping of calls; and 3, full chorus, calls are constant, continuous, and overlapping [Weir and Mossman 2005]).

Observers were told 12 species, all of which occur in central Maryland, could be heard during the experiment (Table 1). Two species, American bullfrog and green treefrog, were never played. The other species varied in the proportion of trials in which they occurred and with which species they co-occurred (Table 1). The playlist for each 90-trial session was randomized separately using the same rules for assembly. We included a diverse set of species and species combinations, limiting combinations to species likely to call at similar times of year. The basic design included three pairs of focal species that were played in different combinations of calling intensity and distances from observers, with each pair used for 30% of trials (Table 1). Trials for each of the focal pairs could also include calls for zero to two additional background species that came from a pool of four potential species assigned to each focal pair. The remaining 10% of trials included occasions where no call was played or a single background species was played.

TABLE 1. List of potential species given to observers and their role in the design of the experiment.

Common name	Focal/background species	Dominant frequency (Hz)	Proportion of trials
American toad (*Anaxyrus americanus*)	background species (1, 2, and 3)	1550	0.12
Fowler's toad (*A. fowleri*)	focal species (3)	1770	0.20
Northern cricket frog (*Acris crepitans*)	focal species (3)	3960	0.20
Gray treefrog (*Hyla versicolor*)	background species (3)	3000	0.07
Green treefrog (*H. cinerea*)	not played	1980	0
Spring peeper (*Pseudacris crucifer*)	focal species (2), background species (1 and 3)	2750	0.31
Upland chorus frog (*P. feriarum*)	focal species (2), background species (1)	3040	0.26
American bullfrog (*Lithobates catesbeianus*)	not played	300	0
green frog (*L. clamitans*)	background species (3)	340	0.07
Pickerel frog (*L. palustris*)	focal species (1), background species (2)	1300	0.31
Southern leopard frog (*L. sphenocephalus*)	focal species (1), background species (2)	1250	0.26
Wood frog (*L. sylvaticus*)	background species (1 and 2)	1460	0.12

Notes: Dominant frequency is the peak spectral frequency estimated from clips of single individuals used in the experiment. Proportion of trials is the proportion of all trials in which the species was played. Species played during a calling occasion all came from one of three species combinations that included two focal and four background species (numbers in parentheses designate the species pools in which the species was included).

The Appendix includes a detailed description of how playlists were assembled as well as the actual playlists.

To test whether training could be used to reduce or eliminate false positive errors, we conducted a before-after experiment by assigning observers to treatment or control groups. The treatment, which consisted of additional instruction aimed at reducing false positive errors, was applied between the first and second session. Results of a power analysis prior to the experiment using false positive probabilities in the range observed by McClintock et al. (2010*b*) demonstrated that having at least 20 observers in a treatment group would be sufficient to detect what we considered large effects (>50% reduction). We calculated that for 20 observers: the probability the confidence interval on the estimated effect would include no effect was <2% if false positive probabilities were halved, and <0.3% if false positive detections were reduced by 75%. Using this as a guideline, we randomly assigned 21 observers to the treatment group. To control for potential differences between the first and second session, the remaining 10 observers did not receive the treatment. Treatments and controls were distributed evenly among the five groups.

At the start of the first session, all observers were instructed to record observations of species presence and calling intensity as if they were collecting data for a standard auditory survey. After the first session ended, we told participants we wanted to solicit their impressions of the first session. This feedback was solicited from individuals in the control group, while those assigned to the treatment group were given a new set of instructions for the second session. They were explicitly told that during the second session they should only record species that they were absolutely certain had been played and for which they were certain about the species identification. They were told not to be concerned that this would likely lead to more false negative errors. In addition, we told participants that (1) in all past experiments using the electronic broadcast system all observers recorded false positive detections; (2) false positive errors were harder to correct for when estimating species occurrence probabilities so it was very important to eliminate these errors even if false negative errors increased; and (3) false positive errors could result both from either misidentifying a call or by thinking a call occurred when none had.

Prior to participating, observers were asked to complete an online frog call test similar to ones used by the North American Amphibian Monitoring Program (*available online*).[6] The online test selects randomly among hundreds of sound files, making each test session different. Each sound file is approximately 30 seconds in duration and includes calls of one or more anuran species. For this study, observers completed a test with 15 sound files, which included only the 12 species used in the field experiment and where each species occurred at least once. Observers were also asked to fill out a short questionnaire meant to gauge their experience and abilities. They were asked to self-assess their ability to recognize calls typical of the area on a five-point scale (1, low proficiency, to 5, expert) and to report years of experience conducting amphibian and bird call surveys.

All data were independently entered into a database twice and records were cross-checked for mismatches, which were corrected using the original data sheets. The error rate for data entry was 0.0016, meaning the probability a record would be entered incorrectly twice was ~ 1 in 360 000. As a result, false positive errors were unlikely to result from data entry errors.

Analysis

For data analyses, we divided observations (i.e., a record for a single species by an observer during a trial) into cases where the species call was played and cases where the species call was not played. These groups correspond to the two primary response variables that were used in analyses: the false negative probability,

[6] http://www.pwrc.usgs.gov/frogquiz/

TABLE 2. Descriptions of statistical models used to predict false positive probabilities (FP), false negative probability (FN), and the false positive probability conditional on detection (FPCD).

Model	Response	Effect of interest	Fixed effects	Random effects
1	FP, FN	among-observer and among-species variability	none	**OBS, SPC**
2	FP, FN	treatment effects	TRT, SSN, **TRT × SSN**	OBS, SPC, SSN[OBS]
3	FP, FN	self-assessed ability (SAA)	**SAA**, TRT, SSN, TRT × SSN	OBS, SPC, SAA[SPC]
4	FP, FN	total test score (TOT)	**TOT**, TRT, SSN, TRT × SSN	OBS, SPC, TOT[SPC]
5	FP, FN	test correct (QC)	**QC**, TRT, SSN, TRT × SSN	OBS, SPC, QC[SPC]
6	FN	test incorrect (QI)	**QI**, TRT, SSN, TRT × SSN	OBS, SPC, QI[SPC]
7	FP, FN	years anuran call surveys (YF)	**YF**, TRT, SSN, TRT × SSN	OBS, SPC, YF[SPC]
8	FP, FN	years all call surveys (YC)	**YC**, TRT, SSN, TRT × SSN	OBS, SPC, YC[SPC]
9	FN	dominant spectral frequency (HZ), proportion of trials (PT), calling abundance (ABUND)	**HZ, PT, ABUND**, TRT, SSN, TRT × SSN	OBS, SPC
10	FN	HZ^2, PT, ABUND	**HZ^2, PT, ABUND**, TRT, SSN, TRT × SSN	OBS, SPC
11	FP	HZ, PT	**HZ, PT, ABUND**, TRT, SSN, TRT × SSN	OBS, SPC
12	FP	HZ^2, PT	**HZ^2, PT, ABUND**, TRT, SSN, TRT × SSN	OBS, SPC
13	FPCD	recorded calling intensity (RCI)	**RCI**	OBS, SPC, RCI[SPC]

Notes: All models include both fixed and random effects. Observer (OBS) and species (SPC) were included as a random effect in all models. Random effects for differences among observers or species in slopes of a covariate are denoted using brackets to specify the grouping variable. For the experimental treatment, we included effects of the treatment (TRT) and session (SSN). For each model, the effects of primary interest are shown in boldface type.

which is the probability of not detecting a species for a trial, given the species was played and the false positive probability, which is the probability of recording a species during a trial, given the species was not played. In a couple cases we report results for a third class of observations, where our interest was in the probability that an observation is a false positive error given that a positive detection was recorded.

Data were analyzed using generalized linear mixed models (Bolker et al. 2009) fit in R (v. 2.13; R Development Core Team 2011) using the lme4 package (version 0.999375-39; *available online*).[7] All models were fit with a binomial error distribution and a logit-link function. Response variables in models included the false negative probability, false positive probability, or false positive probability conditional on detection. We included observer and species as random effects in all models ($n = 31$ and 12 levels, respectively). We also included random slope effects among observers and species when appropriate. For example, in examining how a factor affected among observer error probabilities we allowed the effect to vary among species. In all cases we allowed the random intercept and slope components to be correlated among levels of the random effects. We relied on estimates of effect sizes and confidence intervals to make inference about the importance of different parameters used in models.

Table 2 gives a detailed description of the statistical models. We first fit a general model to estimate the total variation among observers and species in error rates (model 1). Next we assessed the effect of the training treatment (model 2). We included effects for before and after (session) and between treatment and control (treatment) with the effect of the training measured by the interaction between treatment and session.

We considered six measures of observer ability as predictors of among-observer variation in error probabilities (models 3–8). These were: self-assessed ability (ranked from 1 to 5); total score on the online test (0–100; 100 × [number correct − number incorrect]/number possible); proportion correct on online test (proportion of species calls played that were detected; used for modeling false negative probability only); proportion incorrect on online test (number of false positive detections divided by total calls played; used for modeling false positive probability only); years experience conducting amphibian call surveys; and summed years of experience conducting bird and amphibian call surveys. Our interest was in determining the relative value of each measure as a predictor of error rates; therefore, we examined each predictor individually. We standardized predictors to have a mean of 0 and standard deviation (SD) of 1, so that effect sizes were estimated for a 1 SD change in the predictor.

We also considered three predictors for among species differences in error rates: the dominant frequency (Hz) of the species' call, the proportion of trials where the species call was played, and for false negative detections, the calling intensity (i.e., single individual calling or chorus of calling individuals). All effects were fit simultaneously (models 9–12). For the dominant frequency we considered two alternative relationships with detection error rates: a linear relationship or a quadratic relationship (i.e., detection errors were either rare or common at intermediate values). To simplify the quadratic model we only included the second-order term, standardizing the Hz for all trials to have a mean of 0 and SD of 1 and squaring this value. Thus, the minimum or maximum was forced to occur at the mean frequency played. We selected between the two alterna-

[7] http://lme4.r-forge.r-project.org/

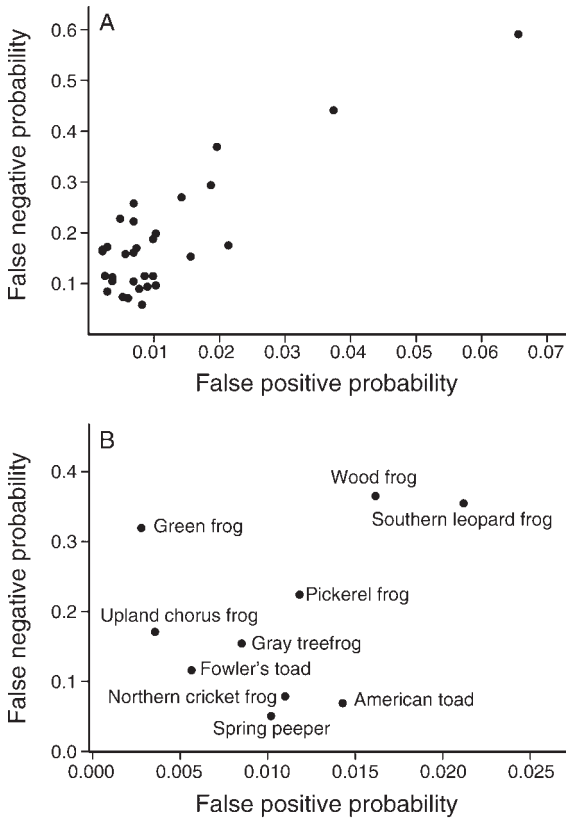

FIG. 1. Estimated false negative and false positive error rates for (A) each observer and (B) each species played. Estimates are from models with single random effects for observer and for species, respectively.

tive parameterizations based on the smaller deviance (number of parameters was the same for both models) and present results for the better fitting alternative.

Finally, we assessed whether false positive errors were less likely when observers recorded a high calling intensity. In this case, we estimated the probability a detection was a false positive error, conditional on the species being detected. We fit the categorical variable for the recorded calling intensity (1, 2, or 3) as a fixed effect predictor (model 13). We did not include data for false positives from American bullfrogs and green treefrogs because they were not played during the experiment.

RESULTS

The 31 observers recorded detections for 12 potential species during 180 trials, generating a total of 66 960 observations. There were 10 562 observations when species were played and 56 398 observations for species that were not played. Observers made 2065 false negative non-detections errors and 8497 true positive detections for a mean false negative error probability of 0.196. Observers made 746 false positive detection errors and 55 652 true negative non-detections for a false positive error probability of 0.0132. This resulted in 8.1% of all positive detections being false positive errors and an average of 0.134 false positive errors per observer per trial. Thus the average observer made a false positive error once every 7.5 trials. All observers had both false negative and false positive errors (SD for among observer effects: 0.116 and 0.0125, respectively). Error rates of each type were strongly correlated among observers ($r^2 = 0.712$; Fig. 1A). There was also substantial variation in error rates among species (SD = 0.120 for false negatives and SD = 0.057 for false positives), although correlation between the two error types was weak ($r^2 = 0.093$; Fig. 1B).

Additional training had a limited effect on false positive errors. Consistent with the expected direction of effects, training yielded a 16% decrease in false positive probabilities (CI = –46%, 30%) and a 26% increase in false negative probabilities (–3%, 60%), but confidence intervals included zero. Effect sizes of the treatment on the logit-scale were –0.180 (–0.627, 0.267) and 0.270 (–0.038, 0.578), respectively. Observers varied greatly in how error rates changed between the first session (pre-treatment) and the second session (post-treatment; Fig. 2). False positive errors decreased for 15 of the 21 treatment observers compared to decreases for 6 of the 10 controls.

One observer did not complete the online test and therefore was not included in analyses of observer ability. The mean proportion of correctly identifying calls on the online test was 0.88 (false negative probability = 0.12). Only three observers recorded no false positives during the online test and overall false positive errors were more common for the online test than the field experiment. The average probability of recording a species that was not played was 0.026, which resulted in 9.2% of all recorded detections being false positive errors. The mean self-assessed ability of observers was 3.6 (on a scale from 1–5) and observers averaged 3.9 years of experience conducting amphibian call surveys (range 0–20 years) and 11.3 combined years of bird and amphibian call survey experience (range 0–50).

All measures of observer ability were related to error rates in the predicted direction (Fig. 3). However, self-assessed ability and scores on the online test were much better predictors of error rates than length of experience conducting call surveys. Effect sizes are standardized to represent the effect (on the logit scale) of a 1 SD change in the predictor variable and thus are directly comparable among predictors. Self-assessed ability was a strong predictor of both false negative and false positive probabilities ($\beta = -0.52$, CI = [–0.77, –0.28] and –0.56 [–0.90, –0.22], respectively). Scores from the online test were an even better predictor, with overall score (–0.57 [–0.79, –0.36] and –0.70 [–0.98, –0.42], respectively) fitting better than the proportion correct for false negatives (–0.54 [–0.77, 0.32]) and the proportion incorrect for false positives (0.64 [0.35, 0.92]). Years of experience conducting both bird and amphibian call surveys was a better predictor of false negatives and false

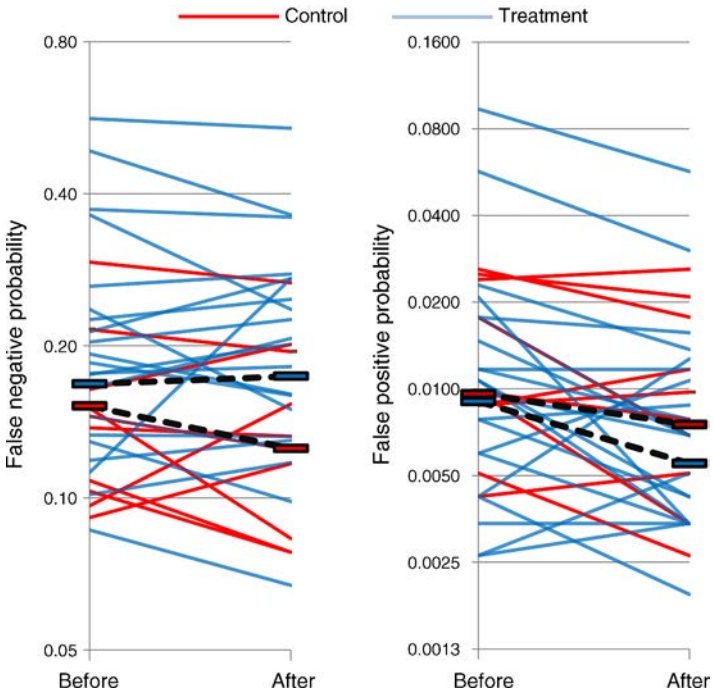

FIG. 2. Estimated error rates before and after the treatment was applied for control observers (red lines) and treatment observers (blue lines). The mean value for each group is shown by the boxes and dashed black lines. Values are plotted on a log scale so that a similar change from before to after for any two observers represents a similar proportional change in errors.

positives (–0.39 [–0.63, –0.15] and –0.21 [–0.58, 0.16], respectively), than only years of experience conducting amphibian call surveys (–0.12 [–0.40, 0.16] and –0.08 [–0.45, 0.28], respectively).

There was a negative linear relationship between false negative detection and the call frequency measured in Hz ($\beta = -0.60$, CI = [–1.04, –0.16]) and false negatives decreased when a chorus rather than single individual was played (–0.16 [–0.29, –0.03]). The proportion of trials a species' call was played had little effect on false negative probability (0.13, [–0.26, 0.52]). False positive errors were positively related to how often a species was played (0.43 [0.00, 0.86]). There was no support for a linear relationship between call frequency (Hz) and false positive errors (0.01 [–0.42, 0.45]).

Consistent with predictions, the probability that a recorded detection was a false positive decreased as the recorded calling intensity increased. False positive errors represented 8.3% of all detections recorded as calling intensity 1, 7.0% of all 2's, 3.9% of all 3's. However, confidence intervals for effects included 0 (difference in false positive probability between recorded 1's and 2's on a logit scale was –0.52 [–1.32, 0.28] and between 1's and 3's was –0.88 [–2.06, 0.30]).

DISCUSSION

Our results provide further evidence that false positive detections are an important source of observation error for many ecological studies. During field trials, 8% of all recorded detections were false positive errors, and false positive detections were recorded by all observers. These results corroborate previous findings using the experimental system employed here (McClintock et al. 2010*a*, *b*) and the growing evidence across a range of presence–absence data sources that false positive errors are a relevant source of observer error (Miller et al. 2011). Given the consequences of even low levels of false positive errors when estimating species occurrence and dynamics (Royle and Link 2006, McClintock et al. 2010*a*, *c*, Miller et al. 2011), addressing false positive error should be a priority in both the design and analysis of occupancy studies. Although observer ability was negatively associated with false positive error probabilities, even the most highly qualified observers made errors with non-trivial frequency. Importantly, we did not find evidence that significant reductions in false positive errors can be achieved by simply emphasizing the importance of eliminating this error type during observer training.

In general, it is easier to account for statistical bias from false negative errors than bias from false positive errors (Royle and Link 2006, McClintock et al. 2010*a*, Miller et al. 2011). Intuitively, it would seem that instructing observers to minimize false positive errors would be a straight-forward strategy for improving the quality of field data. The ability to reduce false positive errors in this manner assumes that observers are willing and able to forego recording uncertain detections. However, we found that the average observer only marginally reduces false positive errors when further instruction is applied, and the effect is inconsistent among observers. Thus, instruction and training may be insufficient to eliminate false positive errors.

Our measures of observer ability were all negatively related to false negative and false positive error rates. However, length of experience was a much poorer

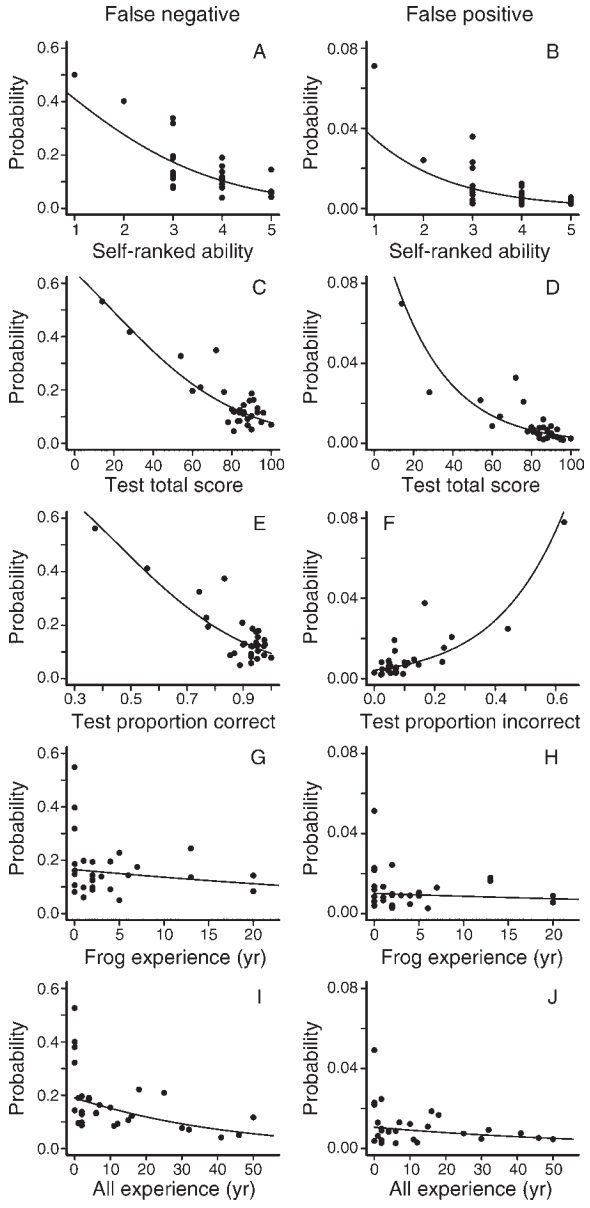

FIG. 3. Estimated relationship of observer ability to (A, C, E, G, I) false negative and (B, D, F, H, J) false positive error rates. Observer ability was measured based on (A, B) self-assessed ability, (C, D) total score on the online test, (E) proportion correct on the online test, (F) proportion incorrect on the online test, (G, H) years of experience doing amphibian call surveys, and (I, J) sum of years of experience doing both amphibian and bird call surveys.

predictor than either the self-assessed ability of observers or the online test we applied before the experiment. We found a similar effect for observer abilities for both false negatives and false positives, which resulted in strong among-observer correlation in error rates. It was somewhat surprising that self-assessment was nearly as predictive as a direct test of abilities. However, we put no pressure on participants to prove they were experts in calls surveys, which may have resulted in a more honest answer than situations where observers feel the need to prove a high ability. Whether self-assessment would hold for other observer pools is unclear.

A promising method to improve estimates will be to incorporate observer ability as a predictor of error rates when using standard statistical estimators, although results of a similar study by Farmer et al. (2012) suggest the relationship to ability may differ between common and rare species. Most approaches to population abundance and species occurrence estimation allow detection to be modeled as a function of covariates (Williams et al. 2002). Predictors of false positive error types can be incorporated in occupancy estimators by specifying these parameters to be a linear function of a predictor (Royle and Link 2006, Miller et al. 2011). This should also improve discrimination of detection types (false vs. true positive) when among-observer heterogeneity impedes accurate estimation (e.g., Fitzpatrick et al. 2009). Regular testing, as is used for the North American Amphibian Monitoring Program (see footnote 6), is not only useful for screening unqualified observers, but should be used when possible to help account for detection variation in data analyses.

False positive errors were more frequent for commonly occurring species, indicating that observer expectations can affect error rates. Thus, frequencies of occurrence in a survey could also be a useful predictor of false positive error rates. Finally, although false positives errors were rarer when observers recorded that many rather than a few individuals were calling, error rates were still significant for all recorded calling intensities. This suggests the assumption made by Miller et al. (2011) that false positive errors for high call indices were negligible was probably incorrect. Instead it is more appropriate to assign some degree of uncertainty to records even when high abundance was recorded when using model-based approaches.

Others have suggested automatic recording systems, which record calls for later processing, as an alternative to reduce observation error and increase sampling ability for acoustic surveys (Acevedo and Villanueva-Rivera 2006). However, detection errors from recordings, including false positives, can still be substantial (O'Farrell et al. 1999, Acevedo et al. 2009, Waddle et al. 2009). These occur both in cases where identification is done by observers and when processing is done using computing software. The large number of sampling occasions generated by these data sources may further exacerbate the analytical problems that result from false positive errors. When false positive errors occur, positive bias in estimators of species occurrence will increase as the amount of sampling increases (Miller et al. 2011). Potentially, error rates may be less variable among sites. This reduced heterogeneity should make it easier to account for detection errors using model-based approaches. In addition, it may be possible in some cases to independently estimate false positive error rates when these devices are used (e.g., Acevedo et al. 2009). If so,

PLATE 1. We examined false positive detection errors using an experimental setup that mimicked conditions experienced during an amphibian call survey. In the background are speakers broadcasting frog and toad calls, in the middle ground are participants, and in the foreground is the computer which controls calling timing and sequence. Photo credit: T. R. Simons.

one could integrate independent estimates of false positive error rates with estimators that account for false positives (e.g., Royle and Link 2006, Miller et al. 2011).

ACKNOWLEDGMENTS

We especially thank the 31 observers who volunteered their time to participate in the study. Lang Elliott provided permission for the use of his professional sound recordings for the experiment. Eli Rose assisted with estimating spectral frequencies from sound files, and Shannon Bielew assisted with data entry. Jim Nichols provided useful advice on study design. This is contribution 401 of the US Geological Survey's Amphibian Research and Monitoring Initiative. Any use of trade names does not imply endorsement by the U.S. government.

LITERATURE CITED

Acevedo, M. A., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecological Informatics 4:206–214.

Acevedo, M. A., and L. J. Villanueva-Rivera. 2006. Using digital recording systems as effective tools for the monitoring of birds and amphibians. Wildlife Society Bulletin 34:211–214.

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.

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

Boessenkool, S., B. Star, R. P. Scofiled, P. J. Seddon, and J. M. Walters. 2010. Lost in translation or deliberate falsification? Genetic analyses reveal erroneous museum data for historic penguin specimens. Proceedings of the Royal Society B 277:1057–1064.

Bolker, B., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24:127–135.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2004. Advance distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford, UK.

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.

Cohn, J. P. 2008. Citizen Science: can volunteers do real research? BioScience 58:192–197.

Dalton, R. 2009. Still looking for that woodpecker. Nature 463:718–719.

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

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.

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.

Gerhardt, H. C. 1975. Sound pressure levels and radiation patterns of the vocalizations of some North American frogs and toads. Journal of Comparative Physiology 102:1–12.

Jackson, C. H. N. 1933. On the true density of tsetse flies. Journal of Animal Ecology 2:204–209.

Kéry, M., and J. A. Royle. 2009. Inference about species richness and community structure using species-specific occupancy models in the National Swiss Breeding Bird Survey MHB. Pages 639–656 *in* D. L. Thomson, E. G. Cooch and M. J. Conroy, editors. Modeling demographic processes in marked populations. Springer, New York, New York, USA.

Knapp, S. M., B. A. Craig, and L. P. Waits. 2009. Incorporating genotyping error into non-invasive DNA-based mark–recapture population estimates. Journal of Wildlife Management 73:598–604.

Lincoln, F. C. 1930. Calculating waterfowl abundance on the basis of banding returns. United States Department of Agriculture Circular 118. USDA, Washington, D.C., USA.

Lotz, A., and C. R. Allen. 2007. Observer bias in anuran call surveys. Journal of Wildlife Management 72:675–679.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, J. E. Hines, and L. L. Bailey. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier, San Diego, California, USA.

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

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

McClintock, B. T., J. D. Nichols, L. L. Bailey, D. I. MacKenzie, W. L. Kendall, and A. B. Franklin. 2010*c*. Seeking a second opinion: uncertainty in disease ecology. Ecology Letters 13:659–674.

Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. 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.

Molinari-Jobin, A., M. Kéry, E. Marboutin, P. Molinari, I. Koren, C. Fuxjäger, C. Breitenmoser-Würsten, S. Wölfl, M. Fasel, I. Kos, M. Wölfl, and U. Breitenmoser. 2012. Monitoring in the presence of species misidentification: the case of the Eurasian lynx in the Alps. Animal Conservation http://dx.doi.org/10.1111/j.1469-1795.2011.00511.x

Newson, S. E., R. Woodburn, D. G. Noble, and S. R. Baillie. 2005. Evaluating the Breeding Bird Survey for producing national population size and density estimates. Bird Study 52:42–54.

O'Farrell, M. J., B. W. Miller, and W. L. Gannon. 1999. Identification of free-flying bats using the Anabat detector. Journal of Mammology 80:11–23.

Petersen, C. G. J. 1896. The yearly immigration of young plaice into the Limfjord from the German Sea. Report of the Danish Biological Station 6:5–84.

Pough, F. H., W. E. Magnusson, M. J. Ryan, K. D. Wells, and T. L. Taigen. 1992. Behavioral energetics. Pages 395–436 *in* M. E. Feder and W. W. Burggren, editors. Environmental physiology of the amphibians. University of Chicago Press, Chicago, Illinois, USA.

R Development Core Team. 2011. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, D. L., C. S. Elphick, and J. M. Reed. 2010. Identifying anomalous reports of putatively extinct species and why it matters. Conservation Biology 24:189–196.

Royle, J. A. 2004. Modeling abundance index data from anuran calling surveys. Conservation Biology 18:1378–1385.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations and communities. Academic Press, San Diego, California, USA.

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., J. E. Fallon, and R. Johnson. 2003. Use of North American Breeding Bird Survey data to estimate population change for bird conservation regions. Journal of Wildlife Management 67:372–389.

Shea, C. P., J. T. Peterson, J. M. Wisniewski, and N. A. Johnson. 2011. Misidentification of freshwater mussel species (Bivalvia: Unionidae): contributing factors, management implications, and potential solutions. Journal of the North American Benthological Society 30:446–458.

Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology and 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.

Simons, T. R., K. H. Pollock, J. M. Wettroth, M. W. Alldredge, K. Pacifici, and J. Brewster. 2009. Sources of measurement error, misclassification error, and bias in auditory avian point count data. Pages 237–254 *in* D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. Modeling demographic processes in marked populations. Environmental and Ecological Statistics 3. Springer Science and Business Media, New York, New York, USA.

Waddle, H., T. F. Thigpen, and B. M. Glorioso. 2009. Efficacy of automatic vocalization recognition software for anuran monitoring. Herpetological Conservation and Biology 4:384–388.

Weir, L. A., and M. J. Mossman. 2005. North American Amphibian Monitoring Program (NAAMP). Pages 307–313 *in* M. Lannoo, editor. Amphibian declines: the conservation status of United States species. University of California Press, Berkeley, California, USA.

Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations: modeling, estimation, and decision making. Academic Press, San Diego, California, USA.

Wright, J. A., R. J. Barker, M. R. Schofield, A. C. Frantz, A. E. Byrom, and D. M. Gleeson. 2009. Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65:833–840.

Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. Trends in Ecology and Evolution 16:446–453.

Yoshizaki, J., K. H. Pollock, C. Brownie, and R. A. Webster. 2009. Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks. Ecology 90:3–9.

Supplemental Material

Appendix

Description of species playlists used in the experiment (*Ecological Archives* A022-088-A1).