



Tools and Technology Article

Experimental Investigation of Observation Error in Anuran Call Surveys

BRETT T. MCCLINTOCK,^{1,2} *United States Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, MD 20708, USA*

LARISSA L. BAILEY, *Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA*

KENNETH H. POLLOCK, *Biology, Biomathematics, and Statistics, Campus Box 7617, North Carolina State University, Raleigh, NC 27695, USA*

THEODORE R. SIMONS, *United States Geological Survey North Carolina Cooperative Fish and Wildlife Research Unit, Department of Biology, Campus Box 7617, North Carolina State University, Raleigh, NC 27695, USA*

ABSTRACT Occupancy models that account for imperfect detection are often used to monitor anuran and songbird species occurrence. However, presence-absence data arising from auditory detections may be more prone to observation error (e.g., false-positive detections) than are sampling approaches utilizing physical captures or sightings of individuals. We conducted realistic, replicated field experiments using a remote broadcasting system to simulate simple anuran call surveys and to investigate potential factors affecting observation error in these studies. Distance, time, ambient noise, and observer abilities were the most important factors explaining false-negative detections. Distance and observer ability were the best overall predictors of false-positive errors, but ambient noise and competing species also affected error rates for some species. False-positive errors made up 5% of all positive detections, with individual observers exhibiting false-positive rates between 0.5% and 14%. Previous research suggests false-positive errors of these magnitudes would induce substantial positive biases in standard estimators of species occurrence, and we recommend practices to mitigate for false positives when developing occupancy monitoring protocols that rely on auditory detections. These recommendations include additional observer training, limiting the number of target species, and establishing distance and ambient noise thresholds during surveys.

KEY WORDS auditory detection, aural detection, detection probability, false negative, false positive, imperfect detection, monitoring, site occupancy, species occurrence.

Since 1997, the North American Amphibian Monitoring Program (NAAMP) has conducted roadside surveys to monitor anuran populations based on male vocalizations during the breeding season (Weir and Mossman 2005). Consisting of state, federal, and nonprofit agencies, this collaborative initiative established a unified protocol for calling surveys that was adopted by >20 states and utilizes hundreds of volunteers each year (L. Weir, United States Geological Survey, personal communication). With substantial resources increasingly invested in such programs, it is important to acknowledge that imperfect detection complicates the program's ability to monitor anuran populations reliably (Mazerolle et al. 2007). Imperfect detection not only includes non-detection when a species or individual is present (i.e., false-negative error), but also detection when a species or individual is absent (i.e., false-positive error). The former received considerable attention in application of capture-recapture models (e.g., Corn et al. 2000, Bailey et al. 2004, Scherer et al. 2005) and models of species occurrence (e.g., MacKenzie et al. 2002, Pellet and Schmidt 2005, Weir et al. 2005, Brander et al. 2007), but false-positive errors receive considerably less attention (but see Genet and Sargent 2003, Royle and Link 2006, Lotz and Allen 2007).

Relative to other monitoring methods accounting for imperfect detection, occupancy methods utilizing auditory

detections can have advantages in terms of ease and quantity of data collected (Pierce and Gutzwiller 2004, MacKenzie et al. 2006, Mazerolle et al. 2007). Indeed, the United States Amphibian Research and Monitoring Initiative identified monitoring patterns of species occurrence as the most promising approach for assessing changes in amphibian population status (Hall and Langtimm 2001). However, similar to avian point count surveys (Simons et al. 2007, 2009), reliance of these surveys on auditory cues may leave them more prone to observation error than are surveys relying on detections through captures or sightings of individuals.

Despite increasing use of occupancy models to account for variation in the detection process in anuran call surveys, little is understood about the mechanisms driving this process. Different breeding behaviors and call characteristics of species influence their detectabilities, but effects of observers, ambient noise, and other environmental conditions will also influence detection. Previous studies identified temperature, precipitation, breeding behavior, competing species, time of night, moon illumination, ambient noise (e.g., wind or road traffic), and observer ability as important factors affecting anuran calling or detection (Blankenhorn 1972, Genet and Sargent 2003, Weir et al. 2005, Brander et al. 2007). However, Pellet and Schmidt (2005) found no clear patterns of detection across species of differing call intensities occupying similar habitats. Given the enormous potential for variability in detection across species and environmental conditions, there is a growing consensus that the detection process must be better understood to provide efficient study designs and make reliable inferences about

¹ E-mail: brett.mcclintock@gmail.com

² Present address: Centre for Research into Ecological and Environmental Modelling, University of St Andrews, The Observatory, Buchanan Gardens, St Andrews, Fife KY16 9LZ, UK

patterns and dynamics of species occurrence (MacKenzie et al. 2006, Mazerolle et al. 2007).

Models developed by MacKenzie et al. (2006) facilitate unbiased estimation of occupancy probability, assuming no false-positive detections and conditional on inclusion of adequate covariates to model variation in the false-negative detection process. Metrics for environmental conditions, observer abilities, or ambient noise can be identified a priori and used to explain false negatives. However, other factors related to detection that are more difficult to assess, such as variation in abundance (Royle and Nichols 2003) or calling distances (i.e., distance between calling amphibians and the observer) across sampling units, can result in detection probability heterogeneity that induces bias in standard occupancy estimators (Royle 2006). Little empirical evidence is available about effects of distance on detection of calling anurans, but Simons et al. (2007) demonstrated that songbird observation error increases with distance and ambient noise. Simons et al. (2009) also found substantial measurement error when observers attempted to estimate the distance to a sound source without visual cues. These findings suggest that potential effects of distance on observation error may be difficult to incorporate properly into occupancy estimators.

The false-positive detection process, including potential impacts of false-positive errors on inferences about occupancy, colonization, and extinction, is poorly understood. Based on a volunteer mail-in questionnaire accompanied by audio recordings of calling anurans, Genet and Sargent (2003) reported an overall 3.9% false-positive error rate (as a proportion of all detections) among respondents for 13 species commonly occurring in Michigan, USA. Rates for individual species ranged from 0.8% for the bullfrog (*Rana catesbeiana*) to 6.0% for the wood frog (*R. sylvatica*). Simultaneously using digital sound recorders and observers during field surveys, Lotz and Allen (2007) assumed recorded calls represented truth and concluded that 18.3% and 10.7% of all samples included false positives on 2 study areas in Nebraska, USA, with >50% of these being for the northern cricket frog (*Acris crepitans*). Even with these levels of false-positive errors, neither Genet and Sargent (2003) nor Lotz and Allen (2007) found significant evidence of observer experience impacting false-positive error rates. Simons et al. (2007) reported an increase in false-positive errors with increasing levels of ambient noise for avian point count surveys, but no such information is available for anurans.

Despite this emerging evidence to the contrary, researchers often assume that no false-positive errors exist in data sets used to model occupancy (MacKenzie et al. 2006). However, Royle and Link (2006) demonstrated via simulation that even low levels of these errors can result in substantial underestimation of detection probability and overestimation of occupancy. Those authors found false-positive detection probabilities of 5% and 10% yielded average percent relative biases in occupancy of 17% and 32%, respectively. Although largely overlooked up to this

point, there is clearly a potential for false-positive detections to induce severe biases in standard occupancy estimators.

To better understand auditory observation error in anuran call surveys, we conducted a series of replicated field experiments with the following objectives: 1) to identify potential factors that may influence the detection process in auditory surveys, and 2) to quantify effects of these factors on false-negative and false-positive detection probabilities. Using a simulated calling anuran system, we evaluated effects of distance, competing species, ambient noise, and observer ability on detection probabilities for 7 anuran species common to the Piedmont region of the eastern United States. We hypothesized that distance, competing species, and ambient noise would negatively impact detection probabilities and increase the probability of false-positive errors. We also predicted that observer ability index scores, such as those utilized by NAAMP to evaluate volunteers, would be consistent with observer performance in the field.

STUDY AREA

We conducted our experiments in an open agricultural field owned by North Carolina State University, Raleigh, North Carolina, USA. The 600-m² site consisted of mown grass bordered by wooded lots. We conducted experiments 24–25 October 2008 during morning and afternoon. Although wetlands were located <1 km from the site, local anurans were not calling during our experiments and are not known to call at this time of year in similar temperatures.

METHODS

Call Survey Field Simulation

We modified a birdsong simulation system (Simons et al. 2007) to simulate simple anuran call surveys in the Piedmont Region of North Carolina. Briefly, the system uses a laptop computer and playlist software to control up to 64 remote players capable of broadcasting a user-specified series of sound clips (technical details are in Simons et al. 2007). The playlist specifies the species, timing, location, and direction of calls played during a simulated survey. We limited simulated surveys to a simple design to focus on a reasonable number of hypotheses using this system.

We first identified a few anuran species, native to the North Carolina Piedmont, that we expected to have different detection probabilities based on their respective call durations, call volumes, and tendencies to aggregate. We selected the southern leopard frog (*Rana sphenoccephala*) and pickerel frog (*R. palustris*) as low-detection species, the wood frog (*R. sylvatica*) and upland chorus frog (*Pseudacris feriarum*) as medium-detection species, and the spring peeper (*P. crucifer*) as a high-detection species (Crouch and Paton 2002, Genet and Sargent 2003, de Solla et al. 2005). We selected *R. sphenoccephala* in part because its call can be confused with those of both *R. sylvatica* and *R. palustris*, and we selected *P. crucifer* in part because its aggression call can be confused with calls of chorus frogs (Genet and Sargent 2003, de Solla et al. 2005). We searched the NAAMP Frog Quiz database (Weir 2009) and

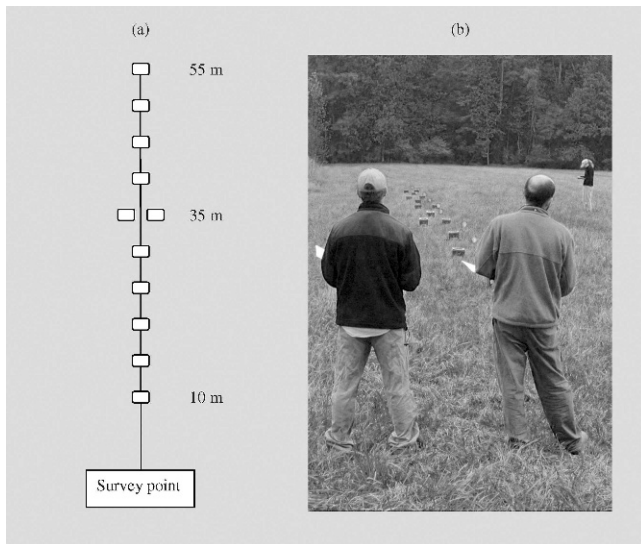


Figure 1. (a) Schematic diagram and (b) photograph of the simulated calling anuran system we used in experiments conducted 24–25 October 2008 in Raleigh, North Carolina, USA. We placed speakers along a straight line at 5-m intervals between 10 m and 55 m from observers stationed at a survey point. We placed a second speaker broadcasting competing species treatments at 35 m. A technician positioned midway along the survey line confirmed the system functioned properly for each simulated call.

commercially available compilations (e.g., Dorcas et al. 2007) for recordings from which to produce representative 20-second clips of one calling individual for each species. To simulate a period with no species calling, we also produced a 20-second clip of silence (NONE).

To test the hypothesis that observation error increases with distance, we simulated species calling from different distances under manipulated noise treatments. Volunteer observers were positioned together at a survey point on one end of a straight line, and 10 speakers facing the observers were positioned on the ground beginning at 10 m and continuing at 5-m increments up to 55 m from observers (Fig. 1). We selected 3 noise treatments in the form of competing species calls. The first was a control treatment consisting of no competing species. The second treatment was a continuous spring peeper chorus, consisting of both advertisement and aggression calls. The third treatment consisted of one southern leopard frog calling continuously. We played these treatments from a speaker positioned on the survey line 35 m from observers.

We composed a playlist of randomly selected call clips played from 1 of the 10 distances along the survey line. For control treatment playlists, we played each of the 6 species call clips (*P. crucifer*, *P. feriarum*, *R. palustris*, *R. sphenoccephala*, *R. sylvatica*, NONE) twice at each distance. For the spring peeper chorus and southern leopard frog competing species treatments, we played each of 5 species call clips (*P. feriarum*, *R. palustris*, *R. sphenoccephala*, *R. sylvatica*, NONE) twice at each distance. Playlists for the control treatment therefore consisted of 120 call clip trials in total, and playlists for the other 2 treatments consisted of 100 trials.

After we compiled and tested the playlists, we assembled our simulated call survey system in the North Carolina State

University agricultural fields. We conducted experiments during morning and afternoon when ambient noise was low (<50 dB). Audible motor traffic occurred within 2 km of the site, but this was not considered unusual because anuran surveys are commonly performed along roadside routes (Weir and Mossman 2005). We broadcast calls at 1-m sound pressure levels of 84–85 dB for *P. crucifer*, 76–77 dB for *P. feriarum*, 79–80 dB for *R. sphenoccephala*, 81–82 dB for *R. sylvatica*, and 80–81 dB for *R. palustris*, which were slightly below published accounts found in Gerhardt (1975) and Pough et al. (1992). Noting potential effects of uncontrolled fluctuations in ambient noise, we recorded a measurement of ambient noise (dB) from the survey point, near observers, immediately prior to each 20-second trial.

We recruited 5 expert observers, 3 of whom previously collected data for NAAMP, to participate in the experiment. Four of these observers completed 5 randomly ordered playlists (2 control, 2 spring peeper, and 1 southern leopard treatment) on the first day, and 2 of these observers completed 4 randomly ordered playlists (1 control, 1 spring peeper, and 2 southern leopard treatments) on the second day. Each of the 9 playlists required approximately 1 hour to complete. Before the experiment, we instructed observers to record any species identified during each 20-second trial, but we did not inform observers about the number of species or that some trials were silent. We provided volunteers a list of potential Piedmont species that included 11 species, but we played only the 5 species mentioned above. The additional phantom species included the American toad (*Bufo americanus*), Fowler's toad (*B. fowleri*), northern cricket frog (*Acris crepitans*), gray treefrog (*Hyla chrysocelis* or *H. versicolor*), bullfrog (*R. catesbeiana*), and green frog (*R. clamitans*). We gave observers an automated auditory cue to signify the transition between each successive 20-second trial. We recorded $N = 3,000$ responses to 960 trials from 4 observers on the first day and 2 observers on the second day of experiments. On the first day, 4 instances occurred when the intended species clip failed to play (1 *P. crucifer*, 1 *P. feriarum*, 2 *R. sylvatica*), resulting in 4 additional NONE trials where we played no species.

Each observer also completed an online exam modeled after the NAAMP Frog Quiz (Weir 2009) as an evaluation of the observer's ability to detect and correctly identify the 11 potential Piedmont species. Each exam consisted of 15 randomly selected sound files containing calls of 1–5 species, with the total number of calls in exams ranging from 37 to 46. Each of the 11 species appeared ≥ 1 in each exam. Quiz scores account for both false-negative and false-positive detections, where $\text{Score} = 100 \times (\text{no. correct} - \text{no. false positives}) / (\text{no. correct} + \text{no. false negatives})$. Under the standard NAAMP protocol, a score of ≥ 65 is satisfactory for volunteers.

Empirical Data Analysis

We used multinomial logistic regression models (e.g., Agresti 2002) to predict a polytomous dependent variable using both categorical and continuous independent variables. We categorized the dependent variable 3 ways

(correct, false positive, or false negative) according to an observer's response for each call clip. We considered a response correct if a species clip was played and detected (true-positive detection) or if no species was played and none was detected (true-negative detection). We considered a response a false positive if a species was incorrectly detected, regardless of whether any species was actually played. We considered a response a false negative if a species clip was played and no species was detected. We did not use detections of *P. crucifer* during the spring peeper chorus treatment to categorize the response. We did not use detections of *R. sphenoccephala* during the southern leopard frog treatment unless the random call played was also that of *R. sphenoccephala*. We note that a false negative was impossible when no species was played or when the *R. sphenoccephala* call was played during the southern leopard frog treatment.

We considered several categorical variables for predicting probabilities of a correct, false-positive, and false-negative response (p_c , p_{fp} , and p_{fn} , respectively), which included up to 6 levels for species (*P. crucifer*, *P. feriarum*, *R. palustris*, *R. sphenoccephala*, *R. sylvatica*, NONE), 3 levels for treatment (control, spring peeper chorus, southern leopard frog), and 5 levels for observer (observers 1–5). We also investigated 3 additional categories for species based on call rate, call volume, and call duration (Corn et al. 2000, Crouch and Paton 2002, de Solla et al. 2005). For call rate, levels were low (*R. sphenoccephala*, *R. palustris*), medium (*R. sylvatica*, *P. feriarum*), and high (*P. crucifer*). For call volume, levels were low (*R. palustris*), medium (*R. sphenoccephala*, *R. sylvatica*), and high (*P. feriarum*, *P. crucifer*). For call duration, levels were low (*R. sylvatica*, *P. crucifer*) and medium (*P. feriarum*, *R. sphenoccephala*, *R. palustris*). We examined effects of time in 2 ways, one with 2 levels (day 1, day 2) and another with 3 levels (day 1 morning [AM], day 1 afternoon [PM], day 2). In the latter case, we did not divide the second day because inclement weather in the morning and late afternoon restricted these surveys to midday. Continuous independent variables included linear, quadratic, and cubic terms for distance and ambient noise (dB). We included 2-way interactions for species category (by treatment, distance, or ambient noise), time (by distance or ambient noise), observer (by distance or ambient noise), and distance (by treatment or ambient noise). As a potentially more parsimonious alternative to the categorical observer variable, we also employed the observer Frog Quiz scores and individual components of the scores for false-negative or false-positive errors as continuous covariates.

We conducted our analysis using a multiple-state occupancy model (Nichols et al. 2007) in Program MARK (White and Burnham 1999). This required a slight reparameterization of the model, with parameters $p = p_c / (1 - p_{fp})$ and $\delta = 1 - p_{fn}$, where δ is the conditional probability of a true-positive detection (given a call clip was played and detected) or probability of a true-negative detection (given no call clip was played). We then derived the original parameters of interest as $p_c = p\delta$, $p_{fp} = 1 - \delta$, and $p_{fn} = \delta(1 - p)$ with variances approximated via the delta

method (Casella and Berger 2002), noting that $p_c = \delta$ when $p_{fn} = 0$. An additional advantage of this approach was that it allowed covariates describing detection in general (i.e., any positive detection, p) to be investigated separately from covariates relating to false-positive detections.

With so many covariates and potential models to examine, we simplified model fitting and selection to a 2-step process using Akaike's Information Criterion adjusted for small sample sizes (AIC_c; Burnham and Anderson 2002). In step 1, we examined logit-linear combinations of explanatory variables for p while maintaining the most general structure for δ . Once we identified the best-supported AIC_c structure for p , step 2 proceeded by maintaining this structure and investigating logit-linear combinations of variables for δ . We also examined 6 additional models combining the 3 AIC_c best-supported structures for p and δ from both steps. This process limited our investigation to 126 models.

RESULTS

Frog Quiz scores for our 5 observers were 100, 95, 100, 98, and 80. As a percentage of the total number of calls played in each quiz, observers correctly detected the species for 100%, 100%, 100%, 98%, and 85% of calls, respectively. As a percentage of all positive detections, observers falsely identified 0%, 5%, 0%, 0%, and 5% of calls in each quiz.

With 61% of all model weight, the 65-parameter minimum AIC_c model included effects on detection probability for species call volume, treatment, call volume by treatment, observer, day, distance, distance by observer, distance by day, distance by call volume, ambient noise, ambient noise squared (dB²), and ambient noise by call volume. For false-positive errors, the model included effects for species, treatment, species by treatment, observer, observer by treatment, distance, distance by species, distance by treatment, ambient noise, distance by ambient noise, and ambient noise by species (Appendix).

Ignoring all other variables, the severity of observation error was different among species by treatment (Table 1). False-negative detection probabilities were low for *P. crucifer* and *P. feriarum* but were considerably greater for *R. palustris*, *R. sylvatica*, and *R. sphenoccephala*. Across species, false-negative errors were less likely, but not significantly so, under spring peeper and southern leopard frog treatments (Table 2). False-positive errors made up 5% of all positive detections, with individual observers exhibiting false-positive percentages of 14, 2, 1, 0, and 11. Most of these errors occurred during *R. sylvatica* calls, but many also occurred when no species or the *R. sphenoccephala* call clip was played. Five of the 6 phantom species were detected at least once, with *A. crepitans* and *B. americanus* falsely identified most frequently (Fig. 2). False-positive errors were lower for the southern leopard frog treatment (Table 2), but these effects varied by species. For example, the *R. palustris* call clip had a 4% chance of being misidentified under the control treatment but was never misidentified under the spring peeper chorus or southern leopard frog treatments. Under the control or spring peeper treatments, *R. sylvatica* exhibited the highest false-positive

Table 1. Mean probabilities of correct detection (p_c), false-negative detection (p_{fn}), and false-positive detection (p_{fp}) from field experiments simulating calling anuran species conducted 24–25 October 2008 in Raleigh, North Carolina, USA. Species calls included no species (NONE), upland chorus frog (*Pseudacris feriarum*), pickerel frog (*Rana palustris*), southern leopard frog (*R. sphenoccephala*), wood frog (*R. sylvatica*), and spring peeper (*P. crucifer*). Competing species treatments included no competing species (NO), spring peeper chorus (SP), and a repeated southern leopard frog call (SL). We only examined *P. crucifer* under the control treatment.

| Species | Treatment | p_c | SE | p_{fn} | SE | p_{fp} | SE |
|--------------------------|----------------------|-------|------|----------|------|----------|------|
| NONE | NO | 0.97 | 0.01 | | | 0.03 | 0.01 |
| | SP | 0.96 | 0.01 | | | 0.04 | 0.01 |
| | SL | 0.96 | 0.02 | | | 0.04 | 0.02 |
| | Overall | 0.96 | 0.01 | | | 0.04 | 0.01 |
| <i>P. feriarum</i> | NO | 0.99 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| | SP | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | SL | 0.98 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | Overall | 0.99 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| <i>R. palustris</i> | NO | 0.75 | 0.03 | 0.21 | 0.03 | 0.04 | 0.02 |
| | SP | 0.83 | 0.03 | 0.17 | 0.03 | 0.00 | 0.00 |
| | SL | 0.86 | 0.03 | 0.14 | 0.03 | 0.00 | 0.00 |
| | Overall | 0.81 | 0.02 | 0.18 | 0.02 | 0.01 | 0.01 |
| <i>R. sphenoccephala</i> | NO | 0.62 | 0.03 | 0.33 | 0.03 | 0.05 | 0.02 |
| | SP | 0.61 | 0.03 | 0.33 | 0.03 | 0.05 | 0.02 |
| | SL | 0.96 | 0.02 | | | 0.04 | 0.02 |
| | Overall ^a | 0.62 | 0.02 | 0.33 | 0.02 | 0.05 | 0.01 |
| <i>R. sylvatica</i> | NO | 0.58 | 0.03 | 0.29 | 0.04 | 0.13 | 0.03 |
| | SP | 0.55 | 0.03 | 0.27 | 0.04 | 0.18 | 0.03 |
| | SL | 0.63 | 0.04 | 0.37 | 0.04 | 0.01 | 0.01 |
| | Overall | 0.58 | 0.02 | 0.30 | 0.02 | 0.11 | 0.02 |
| <i>P. crucifer</i> | NO | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

^a Does not include SL treatment.

probabilities of any species (0.13 and 0.18, respectively), but these fell to <0.01 under the southern leopard frog treatment. False-positive probabilities for *P. feriarum*, *R. sphenoccephala*, or when no species was played were affected little by treatment (Table 1).

Distance and ambient noise both proved to be important predictors, with increases in either variable associated with decreasing detection probabilities. Ignoring all other factors, average effects of distance and ambient noise on detection were best explained by a cubic effect for distance accompanied by an interaction of quadratic terms for distance and ambient noise (Fig. 3). Negative effects of distances >35 m on true-positive detection became much more pronounced as ambient noise increased. We also found a cubic distance effect for overall false-positive detection

Table 2. Mean probabilities of correct detection (p_c), false-negative detection (p_{fn}), and false-positive detection (p_{fp}) from field experiments simulating calling anurans under 3 competing species treatments. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA. Treatments included no competing species (NO), spring peeper chorus (SP), and a repeated southern leopard frog call (SL). For comparative purposes, means are over the 3 species (*Pseudacris feriarum*, *Rana palustris*, and *R. sylvatica*) common to each treatment. We omitted *R. sphenoccephala* because false-negative detections are not possible for this species under the SL treatment.

| Treatment | p_c | SE | p_{fn} | SE | p_{fp} | SE |
|-----------|-------|------|----------|------|----------|------|
| NO | 0.78 | 0.02 | 0.18 | 0.02 | 0.05 | 0.01 |
| SP | 0.80 | 0.02 | 0.16 | 0.01 | 0.05 | 0.01 |
| SL | 0.82 | 0.02 | 0.17 | 0.02 | 0.01 | 0.00 |

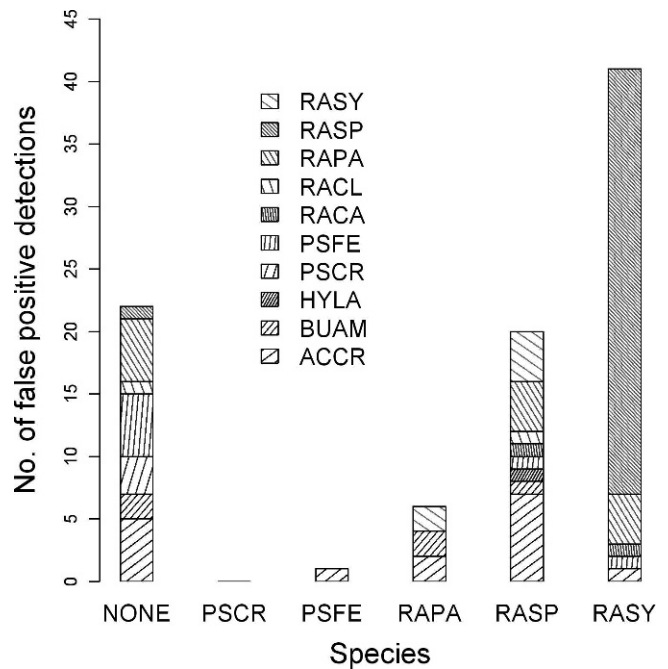


Figure 2. Number of false-positive detections (out of y positive detections from n observer trials) for simulated calls of no species (NONE, $y = 22$, $n = 576$), *Pseudacris crepitans* (PSCR, $y = 196$, $n = 196$), *P. feriarum* (PSFE, $y = 553$, $n = 556$), *Rana palustris* (RAPA, $y = 460$, $n = 560$), *R. sphenoccephala* (RASP, $y = 421$, $n = 560$), and *R. sylvatica* (RASY, $y = 363$, $n = 552$) during experiments we conducted 24–25 October 2008 in Raleigh, North Carolina, USA. The species falsely identified included 5 phantom species for which no calls were simulated: *Acris crepitans* (ACCR), *Bufo americanus* (BUAM), *Hyla* spp. (HYLA), *R. catesbeiana* (RACA), and *R. clamitans* (RACL).

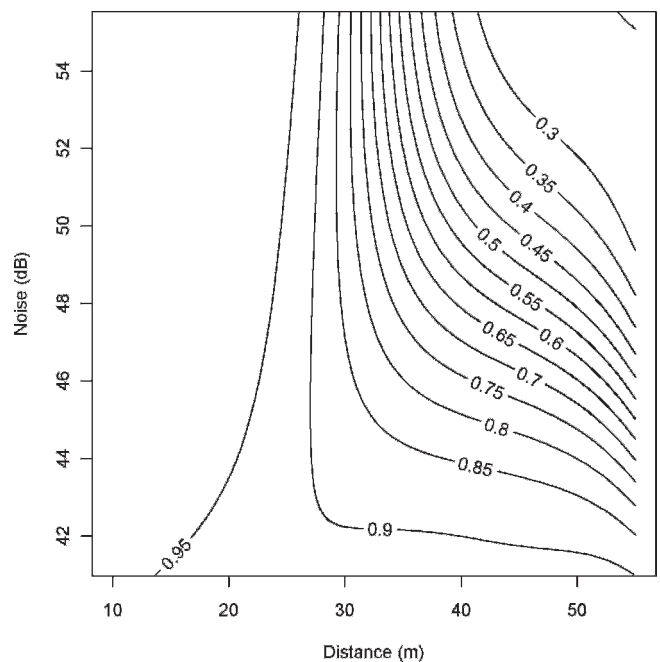


Figure 3. Contour for overall true-positive detection probability, p_c , as a function of ambient noise and distance between calling anurans and observers. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA.

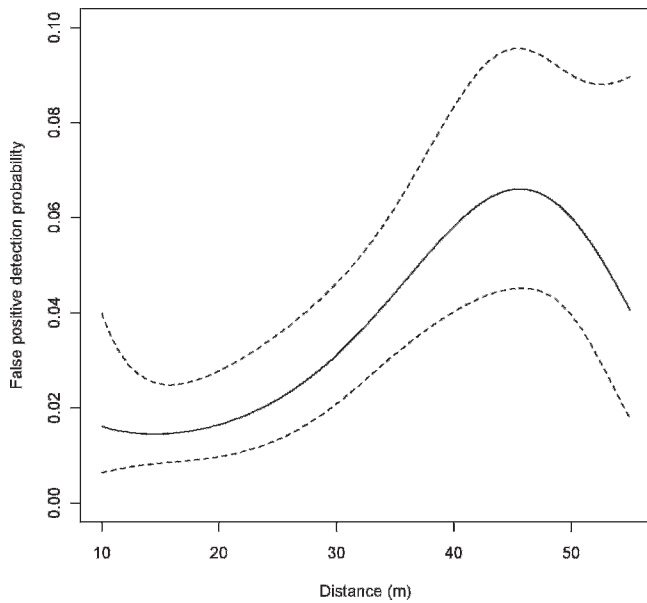


Figure 4. Overall false-positive detection probability, p_{fp} , as a function of distance between calling anurans and observers. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA. Dashed lines represent 95% confidence bands.

probability, but without inclusion of species and treatment interactions, we found no general effect for ambient noise (Fig. 4). This cubic effect suggests a rapid increase in false-positive errors from 15 m to 45 m, followed by a decrease up to 55 m.

Averaging minimum AIC_c model parameters over observers and days revealed diverse effects of distance and ambient noise on true-positive (Fig. 5) and false-positive (Fig. 6) detections under various species and treatment combinations. *Rana sylvatica*, *R. sphenoccephala*, and *R. palustris* were generally the most negatively affected by distance and ambient noise, but these effects varied by treatment. *Rana sylvatica* detections exhibited a complicated trend, with true-positive detections tending lower (and false-positive detections tending higher) at the lower and upper extremes for both distance and ambient noise. We found simpler trends of decreasing true-positive detection and increasing false-positive detection as both distance and ambient noise increased for *R. sphenoccephala* and *R. palustris*. We found some differences in effects of distance and ambient noise between the control and spring peeper chorus treatments, but false-positive errors were less affected for *R. sylvatica* and *R. sphenoccephala* during the southern leopard frog treatment. We detected no effects of distance or ambient noise for *P. crucifer* and *P. feriarum*, presumably because these species had such high true-positive detection probabilities (Table 1). We found little evidence (averaged among observers and days) of ambient noise or treatment affecting overall false-positive probabilities when no species call clips were played (Fig. 7). The 65-parameter minimum AIC_c model had 2.6 times more support, based on the AIC_c weight evidence ratio of the 65-parameter model to the 66-parameter model including ambient noise effects when no species call clips were played (Appendix).

After accounting for species, treatment, distance, and ambient noise, remaining variation in detection probability was best explained by day and observer effects. Conditions were quieter and less variable on the second day, with daily ambient noise measurements averaging 48.5 dB (SD = 3.7) and 44.6 dB (SD = 1.6), respectively. When averaged by day, overall false-negative probabilities on the first day were 0.20 (SE = 0.02) for the control, 0.18 (SE = 0.02) for the spring peeper chorus, and 0.25 (SE = 0.02) for the southern leopard frog treatments. For the second day, these were 0.08 (SE = 0.01), 0.07 (SE = 0.01), and 0.10 (SE = 0.01), respectively. Overall true-positive detection was lower on the first day, with average $p_c = 0.75$ (SE = 0.01) on day 1 and $p_c = 0.86$ (SE = 0.01) on day 2. Negative effects of distance on detection were also more pronounced on the first day, but we found no evidence of an increased effect of ambient noise. We found no significant differences in false-positive error rates between days. Although false-negative and false-positive detection probabilities varied among observers, Frog Quiz scores were a poor predictor of observer performance (Table 3). When we used any of the quiz score covariates in lieu of the categorical observer variable, weight of the minimum AIC_c model fell to 0%.

DISCUSSION

Our experiments suggest that the detection process in anuran call surveys varies as a result of numerous factors. As expected, distance, day, ambient noise, and observer abilities were the most important factors explaining false-negative detections. Distance and observer ability were the best overall predictors of false-positive errors, but ambient noise and competing species also affected error rates for some species. Even under the deliberately simple conditions simulated in our experiment, we found many factors identified a priori affected observation error, but often in a manner contrary to our hypotheses.

We found large differences among species, but the least detectable species were *R. sphenoccephala* and *R. sylvatica* (not *R. palustris*, as hypothesized). This is consistent with results of Crouch and Paton (2000) and de Solla et al. (2005). As expected, detectability was highest for the *P. crucifer*, but we found *P. feriarum* to be almost as detectable out to our maximum distance of 55 m. Under similar environmental conditions, experiments encompassing distances >55 m would be required to detect a difference in observation error between these 2 species. The hypothesized species groupings by call volume best fit the detection data, but estimated detection probabilities for the low- and medium-volume groupings were reversed in magnitude.

None of the species call groupings (rate, volume, or duration) adequately explained false-positive detection, but these error levels did vary by species. Consistent with Genet and Sargent (2003), the largest proportion of false-positive errors occurred during *R. sylvatica* calls, and the species most often misidentified were *R. sphenoccephala*, *A. crepitans*, *R. palustris*, and *P. feriarum*. Alarming, the number of false-positive errors that occurred when no species were played was higher than expected, and every phantom species except

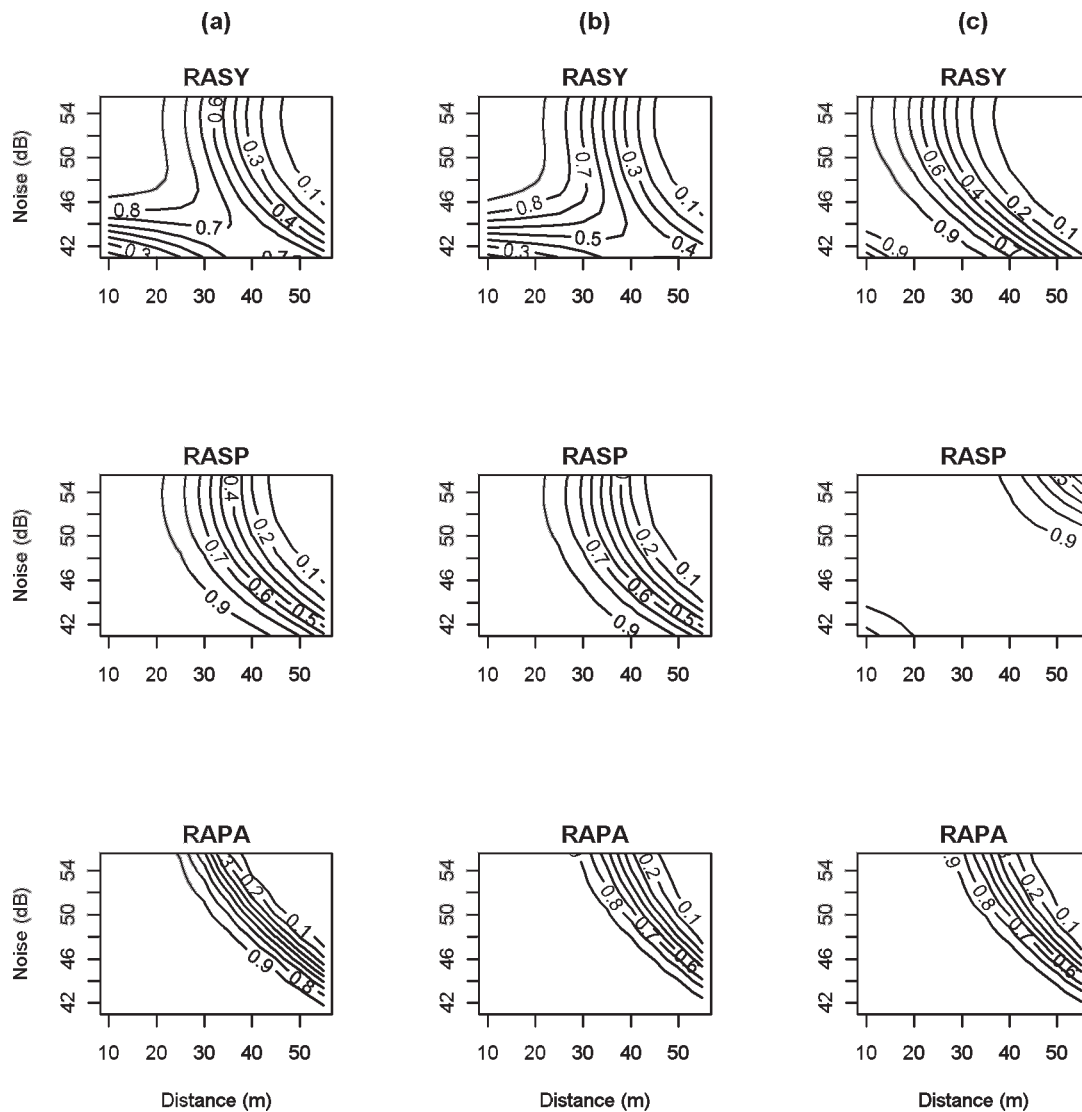


Figure 5. Contours for probability of correct detection, p_c , as a function of distance (m) and ambient noise (dB) for simulated calls of *Rana sylvatica* (RASY), *R. sphenoccephala* (RASP), and *R. palustris* (RAPA) under 3 competing species treatments: (a) no competing species, (b) spring peeper chorus, and (c) repeated southern leopard frog call. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA.

B. fowleri was detected, even by experienced observers. Consistent with Lotz and Allen (2007), *A. crepitans* was the most commonly misidentified phantom species, further suggesting this species may be particularly prone to misclassification in anuran call surveys. True crickets (Gryllidae) were heard during our surveys, but most *A. crepitans* misidentifications occurred when a species call was played. We are therefore unable to determine whether crickets or anurans were the potential source of these misclassifications.

Effects of distance, ambient noise, and treatment were largely species dependent. Our hypotheses about observation error increasing with distance and ambient noise were supported for the less detectable species (*R. palustris*, *R. sylvatica*, and *R. sphenoccephala*), although we were surprised that ambient noise appeared to have little effect on false-positive detections when no species were played. We suspect the decline in overall false-positive detection probabilities for distances >45 m (Fig. 4) was attributable to the lower

overall detection probabilities for those species that were most likely to be misidentified (e.g., *R. sylvatica*). Interestingly, Alldredge et al. (2007b) found a similar error pattern when observers were asked to estimate distances to detected songbirds.

We found little evidence supporting our hypothesis that a competing spring peeper chorus would increase observation error, but the chorus appears to have increased the number of false-positive errors for *R. sylvatica*. Contrary to results of de Solla et al. (2005), we observed no instances of the *P. crucifer* aggression call being misidentified as *P. feriarum* during the spring peeper chorus treatment. We also found little evidence supporting the hypothesis that a competing southern leopard frog would increase observation error. In fact, a competing southern leopard frog had either no effect or even reduced observation error. This result was most notable for *R. sylvatica* and *R. palustris*, where false-positive errors were nearly eliminated with a competing southern leopard frog. Clearly, observers were able to distinguish

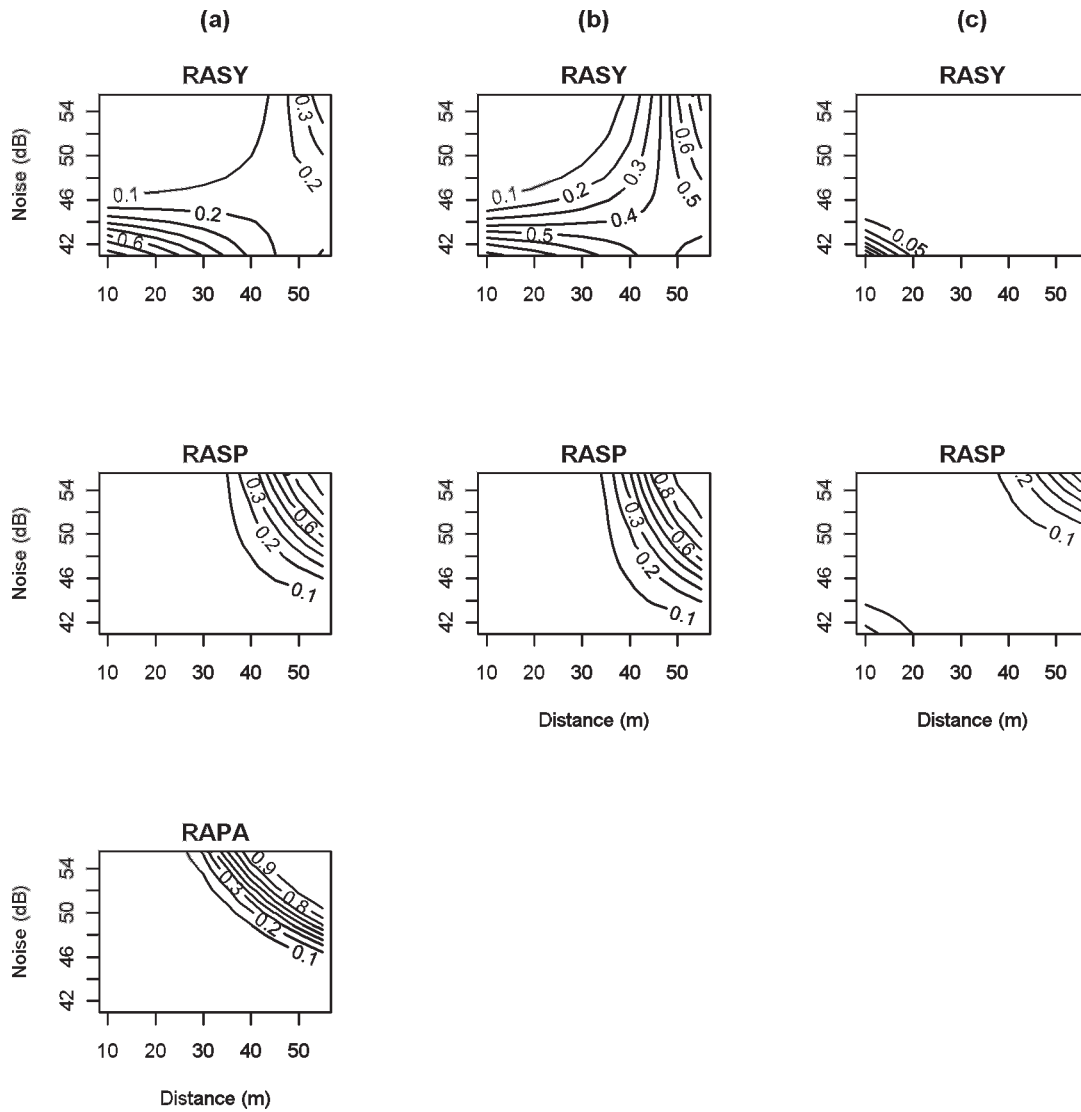


Figure 6. Contours for probability of false-positive detection, p_{fp} , as a function of distance (m) and ambient noise (dB) for simulated calls of *Rana sylvatica* (RASY), *R. sphenoccephala* (RASP), and *R. palustris* (RAPA) under 3 competing anuran species treatments: (a) no competing species, (b) spring peeper chorus, and (c) repeated southern leopard frog call. False-positive errors only occurred for *R. palustris* under the no-competing-species treatment. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA.

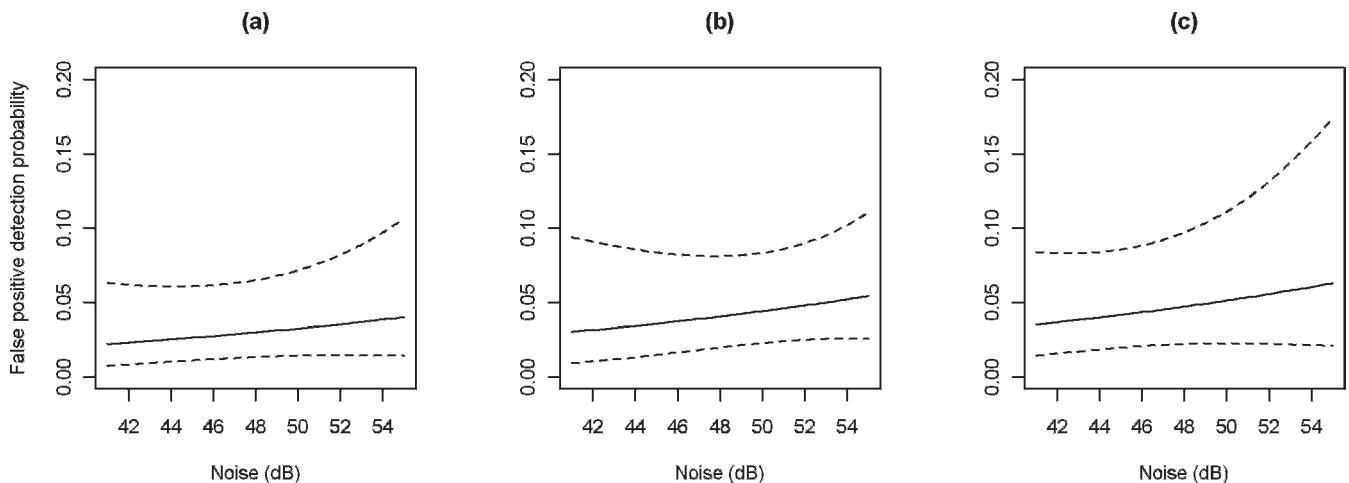


Figure 7. Probability of false-positive detection, p_{fp} , as it related to ambient noise when no additional anuran species was calling (NONE) under 3 competing species treatments: (a) no competing species, (b) spring peeper chorus, and (c) repeated southern leopard frog call. We conducted experiments 24–25 October 2008 in Raleigh, North Carolina, USA. Dashed lines represent 95% confidence bands.

Table 3. Comparison of estimated daily probabilities of correct detection (p_c), false-negative detection (p_{fn}), and false-positive detection (p_{fp}) for 5 observers (Obs.) in simulated anuran call survey experiments conducted 24–25 October 2008 in Raleigh, North Carolina, USA. Means are for simulated calls of no species (NONE) or over all species^a combined (ALL) across 3 competing species treatments. Neither the overall Frog Quiz score nor its individual component scores for false-negative (FN) or false-positive (FP) detections were useful predictors of the survey data.

| Species | Day | Obs. | p_c | SE | p_{fn} | SE | p_{fp} | SE | Observer Frog Quiz score | | |
|---------|-----|------|-------|------|----------|------|----------|------|--------------------------|----|----|
| | | | | | | | | | Overall | FN | FP |
| ALL | 1 | 1 | 0.78 | 0.02 | 0.15 | 0.02 | 0.07 | 0.01 | 100 | 0 | 0 |
| | | 2 | 0.73 | 0.02 | 0.25 | 0.02 | 0.02 | 0.01 | 95 | 0 | 5 |
| | | 3 | 0.77 | 0.02 | 0.22 | 0.02 | 0.01 | 0.01 | 100 | 0 | 0 |
| | | 4 | 0.74 | 0.02 | 0.26 | 0.02 | 0.00 | 0.00 | 98 | 2 | 0 |
| | 2 | 4 | 0.95 | 0.01 | 0.04 | 0.01 | 0.01 | 0.00 | 98 | 2 | 0 |
| NONE | 1 | 5 | 0.77 | 0.02 | 0.13 | 0.02 | 0.10 | 0.02 | 80 | 15 | 5 |
| | | 1 | 0.79 | 0.04 | | | 0.21 | 0.04 | 100 | 0 | 0 |
| | | 2 | 1.00 | 0.00 | | | 0.00 | 0.00 | 95 | 0 | 5 |
| | | 3 | 1.00 | 0.00 | | | 0.00 | 0.00 | 100 | 0 | 0 |
| | 2 | 4 | 1.00 | 0.00 | | | 0.00 | 0.00 | 98 | 2 | 0 |
| | 2 | 4 | 1.00 | 0.00 | | | 0.00 | 0.00 | 98 | 2 | 0 |
| | | 5 | 0.99 | 0.01 | | | 0.01 | 0.01 | 80 | 15 | 5 |

^a We omitted *Rana sphenoccephala* calls under the southern leopard frog treatment.

between species more readily when >1 species was heard calling. We were surprised that both competing species treatments appeared to reduce false-negative and false-positive errors for *R. palustris*, but it is possible that both competing species provided a contrast that made detection easier. To the credit of the observers, competing species treatments appeared to have little effect on false-positive errors when no species clips were played. However, we observed an overall 4% false-positive error rate when no species were played and when the *R. sphenoccephala* call clip was played over the southern leopard frog treatment, which may be an indication of an observer tendency (or desire) to detect species, even when unavailable for detection.

As expected, the degree of observation error varied among observers, and this variation was not particularly well explained by the predictors identified a priori. Our hypothesis that Frog Quiz score would serve as an index for observer ability was unequivocally rejected, suggesting such scores may not be reliable measures of observer ability in the field. Genet and Sargent (2003) and Lotz and Allen (2007) also failed to find measures of observer experience to be reliable predictors of actual performance (but see Weir et al. 2005). However, our observers were all experts, and different results may be found if quizzes are administered to observers encompassing a broader range of abilities prior to the onset of field work.

Distance affected observers differently, with detection probabilities of one observer less impacted, whereas those of another observer were more impacted. We note that these same 2 observers had the highest overall false-positive error rates, with the former responsible for most of the false positives when no species call was played. The other observer regularly misidentified *R. sylvatica* as *R. sphenoccephala* except during the southern leopard frog treatment. This observer detected (and misclassified) *R. sylvatica* from shorter distances at lower levels of ambient noise, resulting in the counterintuitive contours of observation error for this species (Figs. 4, 5). This counterintuitive result may be an artifact of the few participants in our experiments. However,

because all of our observers were considered experts in the field, we do not attribute these differences in performance to experience or knowledge. More likely are differences in hearing ability and level of conservativeness when positively identifying a faint or partially detected call. We recommend that observer ability indices include components intended to measure these qualities as well.

Lower detectability on the first day of experiments was not adequately explained by ambient noise or observer effects. Wind speeds were variable, and we suspect wind speed and direction played a role in improved detectability on the second day. However, we did not identify wind as a factor a priori and failed to record adequate wind data for inclusion as a covariate in our predictive model. Ambient noise is positively correlated with wind speed, but wind direction may be the more important factor because headwinds (or tailwinds) will increase (or decrease) detectability, which emphasizes the importance of standardizing wind conditions acceptable for anuran call surveys (Weir and Mossman 2005). Because the second day of experiments only included 2 observers (one of whom participated in both days), this unexplained variation could also be attributable to the confounded effects of observer learning or different observers being used between days. Although overall false-negative detection probabilities were lower on the second day, overall false-positive error rates were similar between days, which suggests false-positive errors may be less susceptible to temporal variation than are false-negative errors.

We limited our experiments to a simple design to investigate a manageable number of hypotheses about observation error in anuran call surveys. This deliberate simplification came at the cost of some biological realism, and we acknowledge that this simulation does not represent the full reality of an actual survey. Our experiments focused on a few North Carolina Piedmont species and took place along a grass strip in an agricultural field. As calls propagate, attenuation and detectability will vary widely by species and habitat characteristics. For example, we would expect call attenuation to be reduced if traveling along the surface of a

pond. Had our experiments been conducted under these conditions, adjustments would surely be needed to detect similar patterns of the effects of distance, but we would not expect our relative findings to change.

We placed our speakers on flat ground along a straight line, but we do not expect calling anurans to assemble in a similar manner. Unlike the typical nighttime survey, our simulations took place between morning and afternoon when different non-anuran competing species calls could be heard, such as Canada geese (*Branta canadensis*), American crows (*Corvus brachyrhynchos*), and eastern towhees (*Pipilo erythrophthalmus*). Our observers were unable to use additional information, such as geographic location (e.g., eastern or western Piedmont), Julian date, or specific habitat characteristics (e.g., grassland, woodland, wetland) commonly used to anticipate which species are more likely to be present and thus calling. Our observers were also required to identify a species based on calls lasting 20 seconds. In general, our study simulated a shorter calling period and a smaller population size than observers would typically encounter in a field survey.

Despite these departures from reality, we contend that these experiments at their essence capture many fundamental properties of the observation process in anuran call surveys. Further, we suggest that these experiments may represent a best-case scenario because we controlled many factors and only expert volunteers participated. As the typical survey will often have less experienced observers coupled with the added complexity of diverse habitats and competing species, it is reasonable to suspect that observation error may be more severe in reality.

Similar to Allredge et al. (2007a), we chose to ignore the hierarchical repeated-measures nature of the experimental design in our analysis. We did so for simplicity because specification of an error structure incorporating correlations within species, treatments, observers, distances, and days would be complicated and impractical for our purposes of identifying the main factors affecting the auditory detection process. The consequence of this omission could be a slight understatement of uncertainty about parameter estimates, but we do not believe this to have had any significant impact on our general results and conclusions. We note that standard occupancy models (e.g., MacKenzie et al. 2002, 2003) explicitly assume binomially distributed errors with no additional covariance structure within nested factors.

With changes in anuran abundance difficult to assess based on calling surveys, occupancy methods are quickly becoming the preferred alternative for monitoring changes in population status (Hall and Langtimm 2001, Weir et al. 2005). As more resources are continually invested into monitoring programs that capitalize on the relative ease of these surveys (e.g., NAAMP), it becomes increasingly important that we understand the process by which these data are obtained. Given the complexity of the detection process and the mounting evidence of high levels of observation error, it is fortunate that standard occupancy models (MacKenzie et al. 2006) readily handle false-negative errors. However, this aspect of occupancy models

is conditional on sufficient identification and incorporation of factors explaining the false-negative detection process (including detection probability heterogeneity among sites) at the design and analysis phase. There are certainly many more factors driving species availability and calling behavior that we did not investigate here.

Despite the obvious importance of distance on false-negative errors, distance is subject to observer measurement error without visual cues (Simons et al. 2009), thereby making reliable incorporation of this information difficult. However, species calling from different distances to observers between sites could potentially induce site heterogeneity and bias assessments of species occurrence. Some aspects of study design, such as delineating sites of similar size and careful positioning of survey points, may help reduce this form of site heterogeneity.

With respect to false-positive detections, 2 issues of primary concern are 1) mounting evidence that false-positive errors may be more problematic than previously thought, and 2) the current lack of methods addressing this form of observation error. Because call surveys rely on auditory cues, they appear to be more susceptible to false-positive errors than methods relying on captures or sightings of individuals. Not only can auditory cues be more difficult to reliably identify, but they also present additional difficulties because auditory cues can combine to produce new cues, or one cue can mask all or part of another cue, in ways that visual cues cannot. We found false-positive detections to be most attributable to differences in species, observers, and distance, but as with false-negative detections, there are many other potential factors that remain to be investigated. Overall false-positive error rates we observed were greater than those found by Genet and Sargent (2003), but their audio surveys were not conducted in field conditions. Lotz and Allen (2007) found even higher false-positive error rates in their field study, but these were likely inflated because species availabilities and identities were not known, and those authors instead relied on digital recorders and expert opinion for validation. Our simulated field experiment is unique in that the true identification of calling species was known and controlled.

Just as false-negative errors have received considerable recent attention, further research into effects of false-positive errors on inferences about occupancy seems warranted. Despite not knowing true occupancy states of their sites, Lotz and Allen (2007) identified some potential dangers of using standard occupancy models when false-positive errors occur, but no clear or predictable trends were apparent using their methodology. In their limited simulation study, Royle and Link (2006) demonstrated that false-positive error probabilities of 5% and 10% resulted in average percent relative biases in occupancy of 17% and 32%, respectively, which suggests that the level of observation error found in our simulated call surveys would induce substantial positive biases in standard occupancy models. Similar to false-negative errors, as more is learned about the potentially deleterious consequences of false-positive errors, we anticipate the eventual acknowledgment

that these errors also need to be incorporated into estimation to make reliable inferences about occupancy.

MANAGEMENT IMPLICATIONS

When species detections are based on auditory cues alone, managers should consider designing occupancy monitoring protocols that minimize the potential for false-negative and false-positive detections. Both types of error may likely be reduced by establishing protocols that discount detections beyond some distance threshold, which will vary by species, habitat, weather conditions, and ambient noise. Similar to NAAMP, managers should also consider adopting protocols that discount surveys conducted when ambient noise is excessive. The use of handheld decibel meters will help establish when ambient noise thresholds are exceeded. We suggest such thresholds may be 45–55 dB, but these will vary with local conditions. Ambient noise data are also likely to be useful covariates for estimation of detection probability. Focusing surveys and observer training on a small number of target species may help reduce the potential for false-positive errors. Observer training may help reduce species misidentification, but we found that even the most experienced observers committed false-positive errors. Perhaps equally important is training observers of all ability levels to be conservative when positively identifying a faint or partially detected call. Standard occupancy models readily handle false-negative detections, and a non-detection is preferable to a potential false-positive detection when an observer is unsure about species identity. We do not believe this point is stressed enough under current survey protocols and recommend that this be emphasized in any songbird or anuran occupancy monitoring program relying on auditory cues for detections.

ACKNOWLEDGMENTS

This project was funded by the United States Geological Survey Amphibian Research and Monitoring Initiative. We thank our volunteer observers, A. Braswell, E. Corey, J. Hall, J. Humphries, and L. Weir, for their time and support. We thank J. Nichols and L. Weir for helpful discussions.

LITERATURE CITED

- Agresti, A. 2002. *Categorical data analysis*. Second edition. Wiley-Interscience, Hoboken, New Jersey, USA.
- Allredge, M. W., T. R. Simons, and K. H. Pollock. 2007a. Factors affecting aural detections of songbirds. *Ecological Applications* 17:948–955.
- Allredge, M. W., T. R. Simons, and K. H. Pollock. 2007b. A field evaluation of distance measurement error in auditory avian point count surveys. *Journal of Wildlife Management* 71:2759–2766.
- Bailey, L. L., W. L. Kendall, D. R. Church, and H. M. Wilbur. 2004. Estimating survival and breeding probability for pond-breeding amphibians: a modified robust design. *Ecology* 85:2456–2466.
- Blankenhorn, H.-J. 1972. Meteorological variables affecting onset and duration of calling in *Hyla arborea* L. and *Bufo calamita calamita* Laur. *Oecologia* 9:223–234.
- Brander, S. M., J. A. Royle, and M. Eames. 2007. Evaluation of the status of anurans on a refuge in suburban Maryland. *Journal of Herpetology* 41:52–60.
- Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: an information-theoretic approach*. Second edition. Springer-Verlag, New York, New York, USA.
- Casella, G., and R. L. Berger. 2002. *Statistical inference*. Second edition. Duxbury Press, Pacific Grove, California, USA.
- Corn, P. S., E. Muths, and W. M. Iko. 2000. A comparison in Colorado of three methods to monitor breeding amphibians. *Northwestern Naturalist* 81:22–30.
- Crouch, W. B., III, and P. W. C. Paton. 2000. Using egg-mass counts to monitor wood frog populations. *Wildlife Society Bulletin* 28:895–901.
- Crouch, W. B., III, and P. W. C. Paton. 2002. Assessing the use of call surveys to monitor breeding anurans in Rhode Island. *Journal of Herpetology* 36:185–192.
- 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.
- Dorcas, M. E., S. J. Price, J. C. Beane, and S. S. Cross. 2007. *The frogs and toads of North Carolina*. North Carolina Wildlife Resources Commission, Raleigh, USA.
- 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.
- Hall, R. J., and C. A. Langtimm. 2001. The U.S. National Amphibian Research and Monitoring Initiative and the role of protected areas. *George Wright Forum* 18:14–25.
- 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, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* 84:2200–2207.
- MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83:2248–2255.
- MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Academic Press, New York, New York, USA.
- Mazerolle, M. J., L. L. Bailey, W. L. Kendall, J. A. Royle, S. J. Converse, and J. D. Nichols. 2007. Making great leaps forward: accounting for detectability in herpetological field studies. *Journal of Herpetology* 41:672–689.
- Nichols, J. D., J. E. Hines, D. I. MacKenzie, M. E. Seamans, and R. J. Gutiérrez. 2007. Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology* 88:1395–1400.
- 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.
- Pierce, B. A., and K. J. Gutzwiller. 2004. Auditory sampling of frogs: detection efficiency in relation to survey duration. *Journal of Herpetology* 38:495–500.
- 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.
- Royle, J. A. 2006. Site occupancy models with heterogeneous detection probabilities. *Biometrics* 62:97–102.
- Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* 87:835–841.
- Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. *Ecology* 84:777–790.
- Scherer, R. D., E. Muths, B. R. Noon, and P. S. Corn. 2005. An evaluation of weather and disease as causes of decline in two populations of boreal toads. *Ecological Applications* 15:2150–2160.
- Simons, T. R., M. W. Allredge, K. H. Pollock, and J. M. Wettröth. 2007. Experimental analysis of the auditory detection process on avian point counts. *Auk* 124:986–999.
- Simons, T. R., K. H. Pollock, J. M. Wettröth, M. W. Allredge, 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*. Springer, New York, New York, USA.

Weir, L. A. 2009. USGS frog quizzes. <<http://www.pwrc.usgs.gov/frogquiz/>>. Accessed 2 Jul 2009.

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

Weir, L. A., J. A. Royle, P. Nanjappa, and R. E. Jung. 2005. Modeling anuran detection and site occupancy on North American Amphibian

Monitoring Program (NAAMP) routes in Maryland. *Journal of Herpetology* 39:627–639.

White, G. C., and K. P. Burnham. 1999. Program MARK: survival estimation from populations of marked individuals. *Bird Study* 46:120–139.

Associate Editor: Kroll.

Appendix. Akaike's Information Criterion (AIC_c) weights (w_i), ΔAIC_c , and number of parameters (K) for selected models from which we derived the multinomial logistic regression parameters. We examined species (S), distance (D), competing species treatment (T), observer (O), ambient noise (A), ambient noise squared (A2), day of experiment (E), and 2-way interactions of these factors (×). We included model-specific covariates for p and δ (•), p only, or δ only. We grouped the best-supported species effects (S) by species call volume on p but included separate effects for all species on δ . We conducted simulated calling anuran experiments 24–25 October 2008 in Raleigh, North Carolina, USA.

| Model | | | | | | | | | | | | | | | | w_i | ΔAIC_c | K |
|-------|---|---|---|----------------|-----|-----|-------|-------|----------------|----------|----------|------------|-------|----------|-------|-------|----------------|-----|
| S | D | T | O | A | A2 | E | S × T | S × D | S × A | D × T | D × O | D × A | D × E | T × O | | | | |
| • | • | • | • | • ^a | p | p | • | • | • ^a | δ | p | δ^a | p | δ | 0.611 | 0.00 | 65 | |
| • | • | • | • | • | p | p | • | • | • | δ | p | δ | p | δ | 0.231 | 1.95 | 66 | |
| • | • | • | • | • | p | p | • | • | • | δ | | δ | p | δ | 0.111 | 3.41 | 62 | |
| • | • | • | • | • | p | p | • | • | • | δ | | δ | p | δ | 0.036 | 5.66 | 63 | |
| • | • | • | • | • | p | p | • | • | • | δ | p | δ | p | | 0.006 | 9.38 | 58 | |
| • | • | • | • | • | p | p | • | • | • | δ | | δ | p | | 0.002 | 11.04 | 54 | |
| • | • | • | • | • | p | p | • | p | • | δ | p | δ | p | | 0.001 | 12.74 | 55 | |
| • | • | • | • | • | p | p | • | • | • | δ | | • | p | | 0.001 | 13.24 | 55 | |
| • | • | • | • | • | p | p | • | p | • | δ | | δ | p | | 0.000 | 14.47 | 51 | |
| • | • | • | • | • | p | p | • | p | • | | | δ | p | | 0.000 | 14.78 | 49 | |
| • | • | • | • | • | p | p | • | p | • | δ | | • | p | | 0.000 | 16.66 | 52 | |
| • | • | • | • | • | p | p | • | p | • | δ | δ | δ | p | | 0.000 | 18.10 | 55 | |
| • | • | • | • | • | p | p | • | • | • | δ | p | | p | δ | 0.000 | 18.69 | 65 | |
| • | • | • | • | • | p | p | • | • | • | δ | | δ | p | | 0.000 | 30.05 | 53 | |
| • | • | • | • | • | p | p | p | • | • | δ | | δ | p | | 0.000 | 43.26 | 46 | |
| • | • | • | • | • | p | p | • | • | p | δ | | δ | p | | 0.000 | 66.03 | 50 | |

^a Does not include an effect of ambient noise on δ when we played the NONE species (no species calling) call clip.