THE CONDOR
Ornithological Applications
AmericanOrnithology.org

Volume 121, 2019, pp. 1–13
DOI: 10.1093/condor/duy017

RESEARCH ARTICLE

Effects of individual misidentification on estimates of survival in long-term mark–resight studies

Anna M. Tucker,[1*] Conor P. McGowan,[2] Robert A. Robinson,[3] Jacquie A. Clark,[3] James E. Lyons,[4] Audrey DeRose-Wilson,[5] Richard du Feu,[6] Graham E. Austin,[3] Philip W. Atkinson,[3] and Nigel A. Clark[3]

[1] School of Forestry and Wildlife Sciences, Auburn University, Alabama, USA
[2] U.S. Geological Survey, Alabama Cooperative Fish and Wildlife Research Unit, Auburn University, Alabama, USA
[3] British Trust for Ornithology, The Nunnery, Thetford, Norfolk, United Kingdom
[4] U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, Maryland, USA
[5] Delaware Division of Fish and Wildlife, Smyrna, Delaware, USA
[6] Information Systems Services, Lancaster University, Lancaster, United Kingdom
* Corresponding author: annamtucker@gmail.com

Submitted April 28, 2018; Accepted November 28, 2018; Published February 19, 2019

ABSTRACT

All ecological measurements are subject to error; the effects of missed detection (false negatives) are well known, but the effects of mistaken detection (false positives) are less understood. Long-term capture–recapture datasets provide valuable ecological insights and baselines for conservation and management, but where such studies rely on noninvasive re-encounters, such as field-readable color bands, there is the potential to accumulate detection errors as the length of the study and number of tags deployed increases. We investigated the prevalence and effects of misreads in a 10-yr dataset of Red Knots (*Calidris canutus rufa*) marked with field-readable leg flags in Delaware, USA. We quantified the effects of misreads on survival estimation via a simulation study and evaluated whether removal of individuals only reported once in a year (potential misreads) influenced survival estimation from both simulated datasets and our case study data. We found overall apparent error rates of 0.31% (minimum) to 6.6% (maximum). Observer-specific error rates and the variation among observers both decreased with the number of flags an observer recorded. Our simulation study showed that misreads lead to spurious negative trends in survival over time, particularly for long-term studies. Removing all records in which a flag was only recorded once in a sampling occasion reduced bias and eliminated spurious negative trends in survival but also reduced precision in survival estimates. Without data filtering, we found a slight decrease in Red Knot annual survival probability from 2008 to 2018 ($\beta = -0.043 \pm 0.03$), but removing all single-observation records resulted in no apparent trend ($\beta = -0.0074 \pm 0.02$). Spurious trends in demographic rates could influence inference about population trajectories and resultant conservation decision-making. Data filtering could eliminate errors, but researchers should carefully consider the tradeoff between precision obtained by larger sample sizes and potential bias due to misreads in their data.

Keywords: capture–recapture, citizen science, error rate, false positives, individual misidentification, mark–resight

Efectos de la identificación errónea de individuos sobre las estimaciones de supervivencia en estudios de largo plazo de marca y re-avistamiento

RESUMEN

Todas las mediciones ecológicas están sujetas a error; los efectos de la falta de detección (falso negativo) son bien conocidos, pero los efectos de la detección incorrecta (falso positivo) son menos entendidos. Las bases de datos de largo plazo de captura-recaptura brindan información ecológica valiosa y líneas de base para conservación y manejo, pero en los casos en que estos estudios se sustentan en reencuentros no invasivos, como anillos de color legibles a campo, existe el potencial de acumular errores de detección a medida que aumenta la duración del estudio y el número de marcas colocadas. Investigamos la frecuencia y los efectos de las malinterpretaciones en una base de datos de 10 años de *Calidris canutus rufa* marcados en las patas con anillos legibles a campo en Delaware, USA. Cuantificamos los efectos de las malinterpretaciones en las estimaciones de supervivencia a través de un estudio de simulación y evaluamos si la remoción de individuos solo reportados una vez en el año (malinterpretaciones potenciales) influenciaron la estimación de supervivencia considerando tanto las bases de datos simuladas como los datos de nuestro caso de estudio. Encontramos tasas globales de error aparente de 0.31% (mínimo) hasta 6.6% (máximo). Las tasas de error específico observadas y la variación entre observadores ambas disminuyeron con el número de anillos registrados por un mismo observador. Nuestro estudio de simulación mostró que las malinterpretaciones llevaron a tendencias negativas espurias en la supervivencia a lo largo del tiempo, particularmente para estudios de largo plazo. La remoción de todos los registros en los cuales un anillo fue registrado solo una vez en una ocasión de muestreo redujo el sesgo y eliminó las tendencias negativas espurias en la supervivencia, pero también redujo la precisión en las estimaciones de supervivencia. Sin el filtrado de los datos, encontramos una leve disminución en la probabilidad de supervivencia anual de *C. c. rufa* de 2008 a 2018 (β = –0.043 ± 0.03), pero la remoción de todos los registros de observaciones únicas aparentemente no generó ninguna tendencia (β = –0.0074 ± 0.02). Las tendencias espurias en las tasas demográficas podrían influenciar las inferencias sobre las trayectorias poblacionales y las decisiones de conservación resultantes. El filtrado de los datos podría eliminar los errores, pero los investigadores deben considerar con cuidado el balance entre la precisión obtenida con tamaños de muestreo más grandes y el sesgo potencial debido a las malinterpretaciones en sus datos.

Palabras clave: captura–recaptura, ciencia ciudadana, falso positivos, malinterpretación individual, marca y re-avistamiento, tasa de error

INTRODUCTION

Observation error is a well-studied problem in ecology. While the effect of missed detections (false negatives) is largely understood, the effects of incorrect detections (false positives) are more complex. False positives can occur in different ways depending on the data type and sampling methods, for example via species misidentification in occupancy surveys (Miller et al. 2011, Yu et al. 2014) or via genotyping error in noninvasive genetic sampling (Wright et al. 2009, Yoshizaki et al. 2009). The resulting effects on estimation vary among different types of errors. This issue is central to capture–recapture analyses, which rely on a core assumption that individuals are correctly identified when encountered and that there are no false positive detections (Lebreton et al. 1992, Williams et al. 2002). Many types of individual marks can be "encountered" without physical recapture, facilitating data collection on large spatial and temporal scales, including both artificial marks (e.g., dye marks, color bands, leg flags, patagial or ear tags; Silvy et al. 2012) and natural marks (e.g., skin or pelage patterns, permanent scars [Beck et al. 2004], or genetic markers [Yoshizaki 2011]). However, with noninvasive encounters, also termed mark–resight studies, the risk of misidentifying individuals is greater than with physical recaptures because there is limited opportunity to confirm individual ID upon encounter.

Long-term datasets are valuable to ecology and conservation biology, and often rely on a large number of observers with varying experience and who spend varying lengths of time with the project (Bildstein 1998, Newman et al. 2003, Cohn 2008, Magurran et al. 2010, Conrad and Hilchey 2011, Tulloch et al. 2013). Both observer training and data quality control protocols are necessary to ensure accurate data collection. In some cases it may be relatively straightforward to remove impossible observations from the dataset before analysis; however, as both the study scale (spatial or temporal) and the number of deployed tags available for observation increases, initial filtering of the data becomes more difficult. Additionally, even if error rate is low and constant over time, the absolute number of errors may accumulate as the length of the study increases. These concerns apply particularly to data collected via citizen science, volunteer-based surveys, or by seasonal interns and field technicians, but even skilled professionals can make errors. Regardless of data collection protocol, as the number of observers and length of study increases, the potential for errors to appear in the dataset also increases. Determining the observer-specific attributes that are associated with misreads can inform training protocols and aid in vetting data collected by observers with varying experience or training.

The potential for misidentifications to occur in mark–resight studies is understood by field researchers, and many marking protocols are designed to reduce the probability of misidentifications, such as excluding easily confused letters from alphanumeric codes (Clark et al. 2005) or avoiding deployment of similar codes on individuals with nearby territories or home ranges. Several researchers have also estimated error rates for reading color band combinations and alphanumeric codes using both experimental methods and double-marking studies (Weiss et al. 1991, Burton 2000, Milligan et al. 2003, Lavers and Jones 2008, Mitchell and Trinder 2008, Roche et al. 2014), resulting in estimated error rates from 1% to 16%. These errors could bias demographic estimates if not accounted for during either data processing or analysis (Schwarz and Stobo 1999, Bearhop et al. 2003, Morrison et al. 2011).

Some modeling approaches have been developed to explicitly account for individual misidentification during estimation, but most of these models were developed specifically for data collected via noninvasive genetic sampling or photographs (Link et al. 2010, Morrison et al. 2011, Yoshizaki et al. 2011). In this context, misidentification results in the first detection of a new "ghost" individual that can only appear once in the data because it is not known to the researchers whether individuals are correctly identified upon first encounter (for example, the first photograph of a new individual or first sample of a unique genotype). However, with mark–resight studies it is almost certain that marks are perfectly identified upon first capture but that subsequent resightings may occur with error. Additionally, reports of nonexistent marks are often easily identified and removed from the database, but false detection of existing marks may occur multiple times after the initial capture. We note, however, that in migratory stopover systems individuals marked elsewhere in the flyway are often encountered and it may not be known to all researchers which marks have and have not been deployed. When misidentification upon first encounter leads to ghosts, resulting survival estimates are negatively biased by as much as 25% (Morrison et al. 2011). If misidentification errors result in the false detection of a real individual that is no longer alive, however, annual survival estimates from earlier in the time series could be inflated, leading to an apparent negative trend in survival over time. The effects of these errors—incorrect observations of marks that are valid but may not be truly present because the individual has since died or permanently emigrated—has not been as thoroughly explored in the existing literature.

To address this type of false positive error, we evaluated the effect of misreads on analysis of long-term mark–resight datasets using a simulation study. Many monitoring programs that use field-readable marks involve systematic searching for marked individuals throughout a defined sampling period, which often results in multiple observations of the same individuals within a sampling occasion. Multiple observations allow for more opportunities for misreads to occur and for the ability to confirm presence of individuals reported more than once. We propose a simple data filtering protocol that removes individuals recorded only once in a sampling period, and we evaluate the accuracy and precision of models that estimate survival using a both the raw and filtered dataset under varying levels of misread error.

We evaluated misread errors in the context of a long-term monitoring program for Red Knot (*Calidris canutus rufa*) during migratory stopover. The objectives of this study were first to estimate flag reading error rate in our dataset and determine whether observer experience was associated with misread errors. Second, we evaluated the effect of those misreads on estimation of apparent annual survival probability using a simulation study. Last, we evaluated the effect of our data filtering protocol, which removes potential misreads, on survival estimates by comparing estimates from both simulated data and Red Knot mark–resight data with and without data filtering.

METHODS

Flag Deployment and Resighting

We estimated error rate in mark–resight observations of Red Knot in Delaware, USA, from 2008 to 2016. The Delaware Shorebird Project is a long-term volunteer-based research program designed to monitor the population status of migratory shorebirds that use Delaware Bay during spring stopover. Red Knot are long-distance migratory shorebirds that stop in Delaware Bay en route to their Arctic breeding grounds each year (Baker et al. 2001). The mark–recapture monitoring program in Delaware Bay has been an important component of numerous research and management studies since the 1990s (e.g., Atkinson et al. 2007, Gillings et al. 2009, McGowan et al. 2015).

Throughout the study, shorebirds were captured in mixed-species foraging flocks using cannon nets. At the time of capture, we collected biometric data for each individual and applied a USGS band along with an individually identifiable field-readable flag (Figure 1). Inscribed plastic leg flags were first deployed in Delaware Bay in 2004 and have been used thereafter. Flags used in our study were made of lime green Darvic PVC and were laser engraved with a unique 3-character alphanumeric code, filled with black acrylic paint (RAL 9005 Avkote KS Satin [AVCO; www.avko.co.uk]) and varnished. Leg flag design and manufacturing are described in detail by Clark et al. (2005). With the intent of estimating future rates of misidentification in this system, before the 2008 field season 20% (280 individual flags) of the flags manufactured for that year were haphazardly selected and withheld from circulation.

During a 3-week monitoring season in May each year (typically May 10 to June 1), trained observers visit beaches in Delaware to count the numbers of shorebirds and scan flocks with spotting scopes to record observations of Red Knot with leg flags. Most observers are volunteers who have widely varying backgrounds and experience with biological fieldwork in general and flag resighting specifically. Flag resighting occurs throughout the day in 30-min time increments that are considered independent observation occasions, allowing for repeated detections of individuals throughout the day. After returning from the field, observers transcribe their resightings to data sheets, which are then entered into a Microsoft Access database. After entry, resight records are printed from the database and compared to the transcribed data sheets to ensure accurate data entry. These protocols minimize transcription errors from field data to digital data entry, but do not protect from

FIGURE 1. Picture of a 3-character lime green leg flag on a Red Knot in Delaware, USA. This flag was deployed in 2014 and the photograph was taken in 2017. Photograph by Jean Hall.

transcription errors from field notebooks to data sheets, and do not eliminate observation errors that occur in the field.

Quantifying Error Rate

Because a subset of alphanumeric codes were randomly withdrawn from use in 2008, for all analyses described below, we only used resighting data from 2009 to 2018 to estimate misread rates. We defined misidentification rate in 2 ways to establish a possible range of error rates. First, we determined the proportion of flag resightings from 2009 to 2018 that were either one of the withheld combinations or had not yet been deployed (i.e. known false detections). For this species and many other long-distance migrants, marks are deployed by researchers throughout the range and therefore the deployment date was not known for all flags resighted. Of the 8,135 unique flags resighted by observers on our project, 4,558 were deployed by our project and therefore had a known deployment date. The proportion of known false detections in the dataset served as our minimum error rate since we knew that reports of these flag codes were errors. To establish a maximum probable error rate we calculated the proportion of flags recorded by Delaware Shorebird Project volunteers that were only recorded once in a season (i.e. single-observation events). Because flags are frequently observed more than once in a season, but it is unlikely that the same incorrect flag is recorded more than once, these single-observation records were considered as possible misreads. This maximum probable error rate will likely be an overestimate since some individuals may have been transient during migration and only available to observe once in that year. However our focus in this study was annual survival probability and not within-year stopover dynamics. We calculated confirmed (impossible resightings) and possible (single-observation) error rates at the population level as the proportion of all resightings in our resighting database and at the observer level as the proportion of each observer's total resightings.

Observer-specific Misread Rates

Identifying flag- and observer-specific attributes associated with flag misreads can inform data filtering and observer training to minimize errors; therefore, we were interested in observer-specific factors that might be associated with a greater proportion of potential misreads. We modeled observer-specific misread rates as an additive effect of (1) number of years they had spent on the project and (2) the total number of resightings they have contributed to the database across the whole study. Although those metrics are correlated within observers (ρ = 0.63), there is considerable variation because some volunteers participate for many years, but only for 1 or 2 days (many years, few resightings), while others might be new but join the project for the entire season (few years, many resightings). We used a beta-binomial generalized linear model to analyze apparent misread rates while accounting for overdispersion and unequal variance in our data. Overdispersion and unequal variance in the data are likely caused by observers coming to the project with widely varying previous background and training. Among-observer variance in misread rate decreased with observer experience on the project (Figure 2), so we modeled the variance in the probability of misreads (represented by the overdispersion parameter θ) as a function of the log-transformed total number of resightings. Beta-binomial models were specified in R and fit via maximum-likelihood estimation using the bbmle package (R Core Team 2016, Bolker and R Development Core Team 2017).

We used the beta-binomial model to estimate the logit-linear relationship between observer-specific misread rate (m) and the logged total number of resightings each observer contributed to the project:

$$\text{logit}\,(m) = \text{intercept} + \beta * \log\,(\text{total resightings})\,.$$

Based on this relationship, we were interested in determining the number of resightings after which we would expect an observer's misread rate to be equal to the

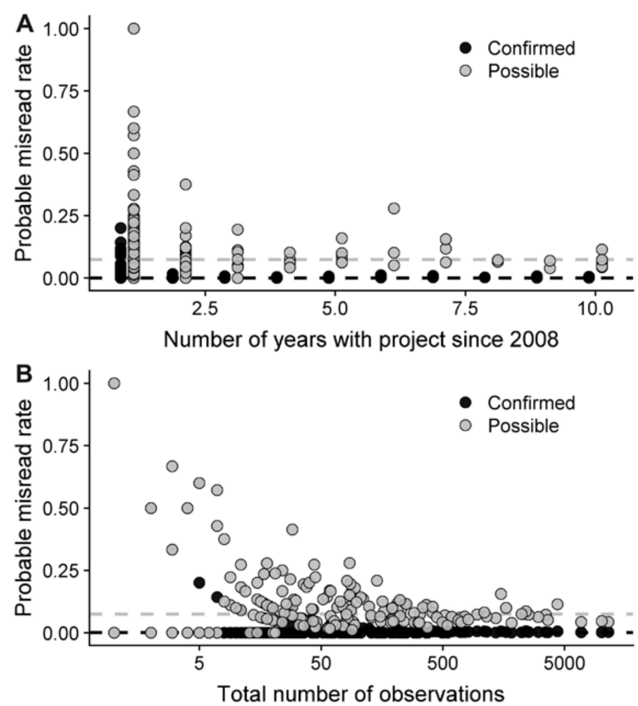

FIGURE 2. For each observer, the proportion of their total resightings that are confirmed or possible misreads as a function of the number of years they have spent on the project (**A**) and the proportion of their total resightings that are single observer-day as a function of their total number of resightings (**B**). Total number of observations are plotted on the log scale. Horizontal dashed lines indicate among-observer median.

median rate among all observers, expecting observers with fewer observations than this threshold to have an above-average misread rate. We used median error rate instead of mean because distributions of observer-specific error rates were right-skewed, and therefore the median is a better metric of central tendency. We derived this metric as a data screening tool to define observer experience with the project. We calculated this threshold by setting m equal to the median error rate and rearranging the above equation to solve for the total number of resightings.

$$\text{threshold resightings} = e^{\left(\frac{logit(m)\ -\ intercept}{\beta}\right)}$$

Observers below this threshold were considered "inexperienced" and those above this threshold were considered "experienced" for the purposes of data filtering described below.

Effect of Flag Misreads on Survival Estimation

We used a simulation study to assess the effect of false detections arising from flag misreads on estimates of annual apparent survival probability. Capture histories were randomly generated to represent encounters of individual birds over a 5-yr, 10-yr, or 20-yr study. We assumed all detections after first capture were field resightings and not physical captures, and that 250 individuals were newly marked each year. We generated capture histories with annual survival probability of $\phi = 0.8$ and detection probability of $p = 0.5$, which are broadly representative of our study system (McGowan et al. 2011, Méndez et al. 2018), by simulating a series of Bernoulli trials for survival and detection for each individual in each year following Kéry and Schaub (2012). To represent the Red Knot flag observation process in Delaware Bay, we allowed each individual to be detected multiple times in a given year. The number of times an individual was detected in a given year was randomly drawn from a Poisson distribution with a sampling intensity of $\lambda = 0.7$, which corresponds to a probability of being detected at least once in that year of 0.5 ($p = 1 - e^{-\lambda}$). Misidentification was a Binomial process with the number of trials equal to the number of detections of each individual in each year with probability of misidentification of 0, 0.005, 0.01, 0.05, or 0.1. Allowing for multiple detections in a year meant that misidentifications rarely resulted in non-detections of the real individual, which typically only occurred if an individual was only seen once and misidentified on that occasion.

There are 2 types of false positive errors that could result from flag misreads. The first type occurs when one flag is incorrectly recorded as another flag that exists in the database and the second occurs when a flag is incorrectly recorded as a flag that does not exist. For our simulations, we assumed that the second type of error would be scrubbed from the database before analysis and only considered the first type of error. We introduced errors in our simulated capture histories by randomly changing a subset of detections to non-detections and reassigning them to individuals that had been marked prior to the year of reassignment but were not otherwise detected in that year. Therefore only real individuals had the opportunity to be reported as a result of a misread and additional fake individuals (i.e. marks that have not been deployed) cannot appear in the data. This approach results in false positive detections of both individuals that were alive but not detected and those that had died or emigrated from the study area. Errors were introduced at 1 of 5 error rates: 0, 0.005, 0.01, 0.05, or 0.1. These rates were chosen to represent a range from low error rate (0.5%) to very high error rate (10%), and correspond with observed minimum and maximum apparent misread rates in our dataset. We simulated 1,000 data sets under each of 15 scenarios (3 possible study lengths, 5 possible error rates). We estimated apparent annual survival for each simulated data set using a Cormack-Jolly-Seber (CJS) model implemented in Program MARK using the RMark package for R (Laake et al. 2013, R Core Team 2016). For each dataset, we fit a model that estimated a linear trend in survival probability (ϕ) over time and a time-constant detection probability (p). We quantified the effects of misreads on resulting estimates by calculating the root mean squared error (RMSE) and relative bias (%) between model estimates and true data-generating values of ϕ and p, and by comparing estimated slope parameters of the trend in survival probability over time. RMSE was estimated as:

$$RMSE(\widehat{\phi}) = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\phi}_i - \phi)^2}{n-1}}$$

Where $\widehat{\phi}_i$ is the survival estimate from a single replicate, ϕ is the true survival probability, and n is the number of replicates. Relative bias was estimated as:

$$Bias\left(\widehat{\phi}\right) = \frac{\sum_{i=1}^{n} {(\widehat{\phi}_i - \phi)}/{\phi}}{n}$$

Data Filtering to Minimize Errors

We evaluated the effect of 4 proposed data filtering methods on estimates of Red Knot apparent annual survival and estimates from simulated datasets. We estimated apparent annual survival for Red Knot first captured from 2008 to 2017 and resighted from 2009 to 2018 in Delaware, USA. All first encounters were physical captures during which the flag was deployed, and all subsequent encounters were field resightings. We first filtered the data to remove all resightings of flags not deployed by researchers in Delaware, which removed all impossible records of withheld flags and those that were observed before their deployment date. Second, we removed both impossible flags and all records from inexperienced observers. We determined observers to be "inexperienced" based on the threshold method described above. Third, we removed both impossible flags and all observations of flags that were only observed once in a given year. Lastly we removed all 3 types of known or potential errors.

We evaluated the fit of the Red Knot data to the fully time-dependent CJS model using the R2ucare package for R (Gimenez et al. 2018), which indicated potential transience and trap response. To account for these, we included 2 time-varying individual covariates. The first was a dummy variable that indicated whether each individual was detected in the previous year (0 = not seen, 1 = seen). This was included as a covariate on detection probability to account for a type of trap-response likely resulting from nonrandom temporary emigration (individuals skipping stopover in Delaware in some years). We accounted for transience by assigning each individual to an age class (first capture = 1, all subsequent captures = 2) and estimating the effect of this dummy age class on survival probability. This method of estimating survival after first capture separately from survival after subsequent encounters has been used to account for the presence of transients (individuals only present for one year) in capture–recapture models (Pradel et al. 1997).

We used RMark to fit a CJS model that estimated a linear trend and the effect of the dummy age class on annual apparent survival probability. Detection probability was modeled as a fixed effect of year and whether the individual was seen in the previous year. We compared the estimated slope of survival over time among the 4 data filtering methods to determine whether they produced different ecological inferences. To provide some context for interpreting these results and more objectively evaluate the effects of removing observations from the data, we also fit the model to simulated datasets with and without data filtering. For each simulated dataset, we fit the CJS model both without data filtering and after removing all instances in which a flag was only detected once in a given year. We compared estimated trends in survival over time among scenarios and calculated RMSE and relative bias of estimates with and without removing single-observations.

RESULTS

Quantifying Error Rate

Our flag resighting dataset contained 80,880 total recorded observations by 201 observers of 8,135 individual Red Knot from 2009 to 2018 (Table 1). The number of observers in a given year ranged from 36 to 53, with an average of 42 observers per year. The intensive survey effort results in many flags being observed more than once each year, with the average flag seen 5 times in total by 3 different observers in a year and at maximum seen 58 times by 23 different observers. There were 136 reports of withheld flags (one of the 280 flags removed from circulation in 2008) by 46 different observers and 116 reports of flags in a year before they were deployed by 37 different observers, giving a total of 252 impossible observations (0.31%). There were 5,374 observations that occurred only once in a given year in the database (6.6%). Therefore the range of potential misread errors in our resighting data was 0.31% at minimum and 6.6% at maximum.

Observer-specific Misread Rates

Distributions of observer-specific misread rates were all right-skewed, so we report among-observer medians and interquartile range (IQR). Withheld and not-yet-deployed flags were observed rarely. The median observer-specific rate of recording impossible flags was 0% (IQR: 0%, 0.24%), but the observer mean was 0.62%. The median observer-specific rate of single-observations was 7.4% of all observations (IQR: 4.2%, 12%).

TABLE 1. Summary of resightings of 3-character lime Red Knot flags each year from 2009-2018, including the number of observers, number of unique flags seen, and total resightings recorded. Confirmed misreads are records of withheld or not-yet-deployed flags. Possible misreads are flags that were only recorded a single time in that year. Percentage of total resightings is given in parentheses..

Year	Number of observers	Individual flags	Total resightings	Average resightings per flag	Confirmed misreads	Possible misreads
2009	42	1,917	10,888	5.7	27 (0.25%)	581 (5.3%)
2010	41	1,217	4,144	3.4	13 (0.31%)	445 (10.7%)
2011	41	1,897	8,261	4.4	38 (0.46%)	634 (7.7%)
2012	39	1,409	6,781	4.8	26 (0.38%)	517 (7.6%)
2013	36	939	3,879	4.1	19 (0.49%)	368 (9.5%)
2014	39	1,336	6,540	4.9	30 (0.46%)	422 (6.5%)
2015	42	2,265	13,755	6.1	49 (0.36%)	738 (5.4%)
2016	43	726	2,346	3.2	6 (0.26%)	320 (13.6%)
2017	53	1,831	11,066	6.0	18 (0.16%)	622 (5.6%)
2018	46	2,098	13,220	6.3	26 (0.19%)	727 (5.5%)

The total number of resightings made by an observer was a significant predictor of both minimum and maximum possible misread rate (Figure 2). Both the probability of misreads and the among-observer variation in misread probability decreased as the number of total resightings logged increased ($\beta^{total}_{minimum}$ = −0.21 ± 0.08, $\beta^{total}_{maximum}$ = −0.19 ± 0.04), but the number of years an observer worked on the project was not a significant predictor of misread rates ($\beta^{years}_{minimum}$ = 0.039 ± 0.04, $\beta^{years}_{maximum}$ = 0.036 ± 0.02). Among-observer variation in rates of single-observation resightings decreased as the total number of resightings increased, as indicated by a positive relationship between the overdispersion parameter θ and the total number of resightings logged (α = 1.9 ± 0.53).

Using the predicted relationship between number of observations and apparent misread rate, we calculated the number of observations after which observers converged on the median error rate. Predicted maximum error rates for observers in our study converged on the median after 307 resightings. The median maximum error rate calculated only from experienced observers above this threshold was 6.4% (IQR: 4.3%, 7.6%) possible misreads per observation, while inexperienced observers below this threshold had a median single-observation rate of 8.8% (IQR: 3.1%, 15.3%).

Effect of Flag Misreads on Survival Estimation

Introducing random flag misreads into our simulated capture histories resulted in apparent negative trends in survival probabilities over time, particularly when the error rate was high (≥0.05; Figure 3). The effects of misreads were most pronounced in longer time series, with the RMSE of survival probability estimates for a 20-yr study ranging from 0.007 to 0.066 and relative bias from 0.5% to 7.6% (RMSE and relative bias for all scenarios are listed in Appendix Table 2). With low misread rates (0, 0.005, 0.01), precision of model estimates increased with study length; however for the higher error rates (0.05, 0.1), longer study lengths resulted in decreasing accuracy and precision of model estimates as errors had more chance to accumulate (Figure 4).

Data Filtering to Minimize Errors

We analyzed capture histories of 2,594 individual Red Knot from 2008 to 2018. Of the 25,226 total recorded resightings of these individuals, 2,850 were from inexperienced observers (defined for our study as observers with fewer than 300 total resightings) and 941 were single-observations. When only impossible resightings were removed, we found evidence of a negative trend in apparent annual survival probability over time (β = −0.043, CI: −0.094 to 0.0081; Figure 5). Removing records from inexperienced observers had little effect on estimated trend (β = −0.034, CI: −0.084 to 0.015), but removing single-observations resulted in no evidence of a trend in annual survival probability over the past 10 yr (β = −0.0074, CI: −0.047 to 0.032). Removing both inexperienced observers and single-observations had a similar effect as removing single-observations alone (β = −0.0059, CI: −0.046 to 0.034). The differences in estimated survival probability with and without removal of single-observation records is most pronounced when comparing estimates for the first and last years included in this analysis. Without data filtering, estimated survival probabilities apparently declined from ϕ = 0.87 (95% CI: 0.84, 0.89) in 2008 to ϕ = 0.82 (95% CI: 0.77, 0.86) in 2017. After removing single-observation records, estimated survival probabilities were lower overall, with ϕ = 0.81 (95% CI: 0.78, 0.84) in 2008 and ϕ = 0.80 (95% CI: 0.76, 0.84) in 2017.

Removing single-observation records from the simulated datasets eliminated bias caused by misreads (Figure 4), but also decreased precision of estimates. When data filtering was applied to simulated data, no annual trend was detected in survival probabilities (Figure 3). For the 20-yr study length, RMSE was reduced to 0.01 and 0.009 for error rates of 0.005 and 0.1, respectively, and relative bias ranged from −0.02% to 0.05%. However, for the 5-yr study length, removing single-observations decreased the precision of survival estimates, with RMSE of ~0.09 for all error rate scenarios (Figure 4).

DISCUSSION

We estimated the rate of individual misidentification in the 10-yr mark–resight dataset and found a minimum error rate of 0.31% and maximum of 6.6%. Our simulation study showed that introducing misreads into long-term mark–resight data results in spurious negative trends in annual survival probability when in reality survival is constant over time, but that those effects can be mitigated by removing all single-observation records from the data. The bias in survival estimates caused by misread errors increased with both the simulated error rate and the length of the study. Longer studies allow for more opportunities for misreads to occur, and individuals marked early in the study have a greater propensity to be falsely detected after they have died, inflating estimates of survival from previous years and leading to apparent negative trends.

Without data filtering, analysis of the Red Knot data indicated a slight decline in apparent annual survival from 2008 to 2018, but that trend was no longer detected when all single-observation records were removed. Our post-filtering estimates of Red Knot annual survival probability agree with a recent meta-analysis of global shorebird survival rates that synthesized the existing literature and reported an average Red Knot annual survival probability of

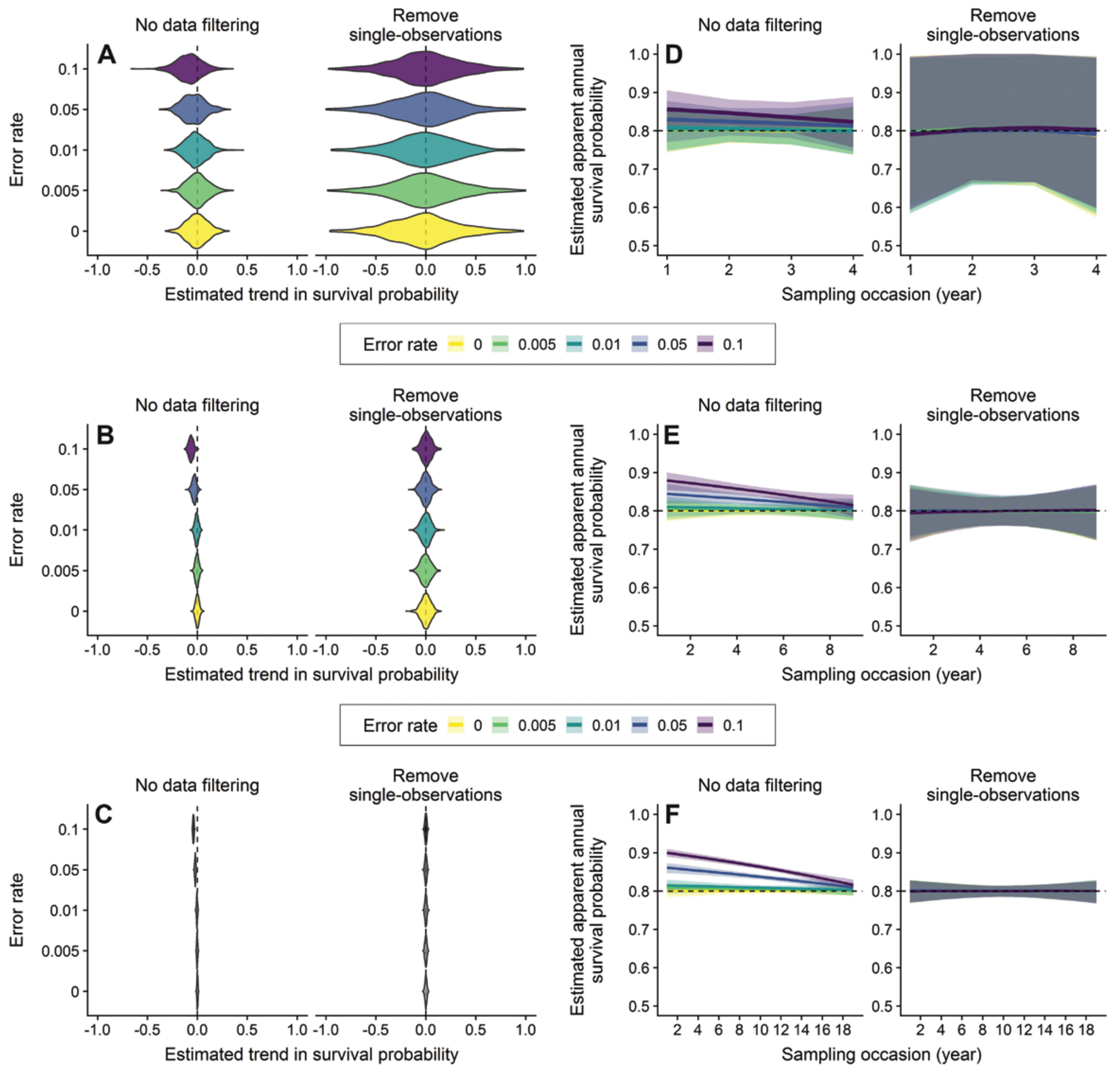

FIGURE 3. Results of simulation study evaluating the effect of individual misidentification on estimates of annual survival probability for a 5-yr study (**A**, **D**), 10-yr study (**B**, **E**), and 20-yr study (**C**, **F**). Misreads led to apparent negative trends in survival probability over time (**A**–**C**), particularly for long-term studies. Removing all single-observation records eliminated bias caused by misreads but also decreased precision of estimates, especially for short time series (**A**, **D**).

0.801 ± 0.011 (Méndez et al. 2018). Along with many Arctic-breeding shorebirds, the *rufa* Red Knot is a species of conservation concern; accurate estimates of demographic rates are important both for understanding the ecological drivers of declines and for implementing effective conservation actions. Most mortality for long-distance migrants likely occurs during the migratory or nonbreeding period (Newton 2006). Survival estimation is a core component of full annual cycle population modeling (e.g., Flockhart et al. 2015, Rushing et al. 2017); positively biased survival estimates and spurious trends in those estimates could obscure our understanding of relative importance of each phase of the annual cycle for population stability.

Simulation Study Extensions

Our simulation study indicated that for misread error rates ≥0.05, survival estimates were positively biased for earlier years in the time series, resulting in apparent negative trends in survival probability over time. Other studies have also demonstrated the potential for errors in

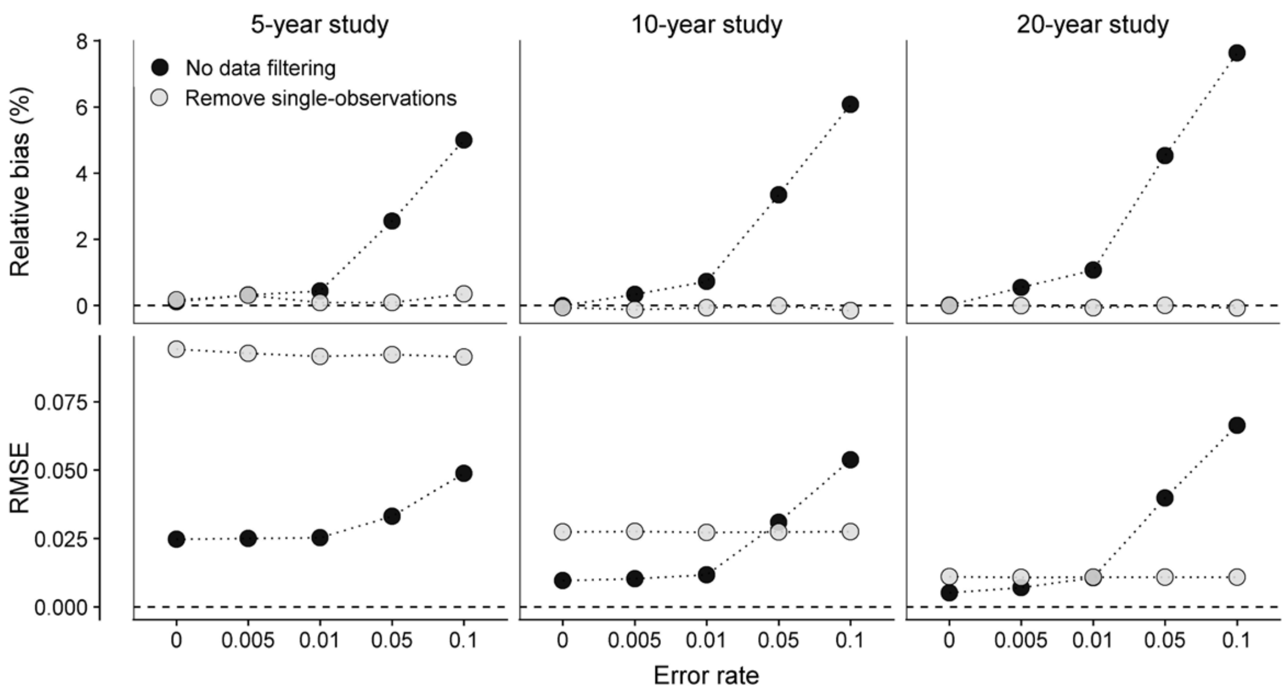

FIGURE 4. Effect of misreads on the precision (root mean squared error, RMSE) and bias (%) of annual survival estimates for studies of varying lengths. Without data filtering (black), bias induced by misread errors increased with study length. Filtering the data by removing all single-observations (gray) effectively reduced bias but resulted in less precision of estimates, particularly for shorter time series.

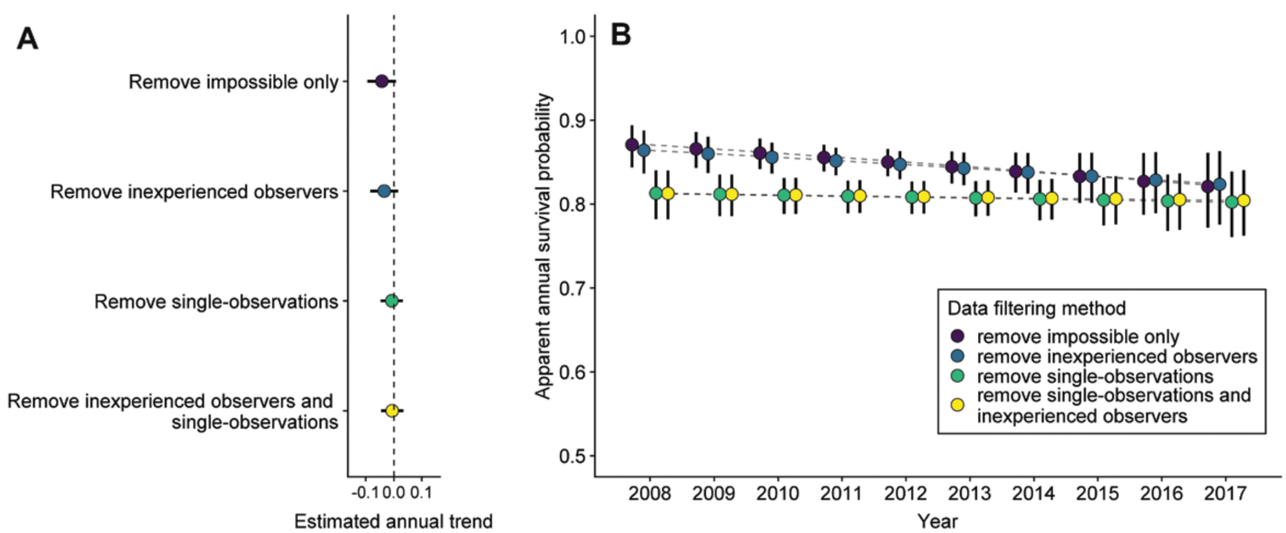

FIGURE 5. Effects of data filtering methods to reduce the effect of potential flag misread error on estimated trends (**A**) and annual apparent survival probability (**B**) of Red Knot marked and resighted in Delaware during spring stopover. Removing only impossible observations or those from inexperienced observers led to an apparent negative trend in survival probability over the past 10 yr, while removing single-observation records resulted in no trend in survival probability.

mark–recapture data to bias estimates of survival (Schwarz and Stobo 1999, Morrison et al. 2011), but here we also show that errors could result in spurious ecological inference about trends or, potentially, drivers of survival probability over time.

The inputs of our simulation could be tailored to the specifics of any study to empirically evaluate the effect of various levels of misidentification errors on inference about apparent annual survival. We encourage others to use a simulation approach to compare observed trends in survival estimates with those that would be expected given high error rates alone. Our simulation code is freely available online for this purpose. This exploratory analysis could assuage concerns about suspected low misread rates, or alert researchers to potential bias in their estimates if error rates are presumed to be higher.

The way we simulated misreads was simplistic for the purposes of assessing general bias. In reality misreads are likely nonrandom and occur in several ways that can be difficult to explicitly model. The simplest type of misread occurs when a single character is misread and replaced by another, incorrect, character (e.g., AT2 recorded as 4T2, or vice versa). This type of misread is unlikely to occur randomly, as certain pairs or groups of characters are more likely to be mistaken for each other (e.g., A and 4, K and X). Another type of misread could occur where all individual characters are recorded correctly, but transcribed in the wrong order (e.g., AT2 recorded as A2T). This incorrect transcription could affect encounter histories in a more random way. The goal of our simulation work was not to capture and explore all the specific misread cases that apply to our system, but rather to generally explore the potential consequences of misread errors on estimation and ecological inference.

Minimizing Errors during Data Collection and Processing

The first, and most obvious, approach to reduce the effect of misreads is to take measures to reduce the probability of misreads occurring during data collection (Bearhop et al. 2003). Observer training and emphasis on data quality over quantity is key to ensuring robust mark–recapture data. For inscribed flags or bands, several papers have detailed the combinations of letters and numbers that are more likely to be misread, as well as the field conditions most likely to lead to errors (Burton 2000, Bearhop et al. 2003, Milligan et al. 2003, Clark et al. 2005, Mitchell and Trinder 2008). Here we also stress the importance of long-term participants instead of rapid turnover of new observers (whether they be volunteers or paid interns). In our study, the number of resightings recorded in the field, but not necessarily the number of years with the project, was a good predictor of misread rates. Regardless of background or previous experience, all observers in our study converged on the median apparent misread rate after ~300 total observations. However, although estimated error rates from inexperienced observers were higher (8.8% compared to 6.4% from experienced observers), removing records from inexperienced observers had little to no effect on estimates of Red Knot survival. This may be a function of the larger variation in error rates among "inexperienced" observers, who have widely varying personal backgrounds and experience.

A second way to reduce the effect of misreads is through more extensive data filtering, as demonstrated in this paper. Analyzing observer-specific misread rate as a function of some metric for experience for any study (number of years, number of observations), as we did here, could help determine the point at which observers tend to the median, and therefore the appropriate study-specific threshold of excluding observations. We found that removing records from "inexperienced" observers had little effect on survival estimates, but estimating survival with and without those records could be a useful exercise in diagnosing whether observer experience is an important factor to consider.

Due to intensive resighting effort that occurs in Delaware during spring migration, we were able to identify instances when a flag was recorded only once in a year as possible misreads. Because the probability of the same false positive detection occurring more than once in the same year is low, most false detections will also be single observations. Removing those from the data will inevitably remove detections of real individuals, but will also remove all or nearly all false positives. Similar to the ad hoc method of dealing with transients first proposed by Pradel et al. (1997), this will reduce precision but ensure unbiased estimates of survival. We do not propose removing all single-observation records as a rule of thumb, but as a potential data filtering option that could be considered if the occurrence of misreads is a concern and the primary goal of the analysis is to estimate annual survival probability. Researchers should carefully consider the tradeoff between precision obtained by larger sample sizes and the number of potential errors due to misreads in their data.

Modeling Misidentification Errors

The Cormack-Jolly-Seber (CJS) model for estimating annual survival is the basis for a large suite of extensions to deal with more complex questions and data structures (see Williams et al. 2002). We investigated the effect of misreads on a simple application of the CJS model, and expect that biases identified here would carry through to more complex extensions. We only investigated the effect of misreads on estimation of apparent annual survival, but capture–recapture data are used to estimate a variety of other demographic parameters. Additionally, we simulated only one type of misread, where real marks are mistakenly observed, as opposed to the type of misread that results in the observation of a new, nonexistent mark. In most mark–resight studies, the latter type is likely to be filtered from the data before analysis, and therefore is of less concern.

The type of misidentification that results in new, nonexistent individuals is the type frequently encountered with noninvasive genetic sampling and photographic capture–recapture, and several model-based methods have been proposed to deal with these false positives in genetic and photographic mark–recapture. The $M_{t,\alpha}$ model, introduced by Lukacs and Burnham (2005) and extended by Link et al. (2010) and Yoshizaki et al. (2011), uses a latent multinomial to determine the probability that a given capture history is real. A key assumption of the $M_{t,\alpha}$ models is that false identifications always create new nonexistent individuals and that these ghost individuals can only be observed once. To deal with this problem of ghost capture histories, Morrison et al. (2011) described a method of "conditioning" natural mark capture–recapture data by removing all first detections of new individuals, which is conceptually similar to the ad hoc method to remove transients proposed by Pradel et al. (1997) and to our process of removing all single-observation events. Even though some real observations will be discarded in the process, it is likely that doing so will also eliminate all false detections, therefore reducing bias. For long-term mark–resight datasets, it is likely that both possible and impossible marks could be recorded incorrectly, and may be recorded more than once in the dataset. If filtering impossible marks from the data is not feasible, for example in a migratory flyway with multiple projects marking individuals, modeling the misread probability for these types of data becomes incredibly complex.

By modifying the likelihood of the existing CJS family of models, it could be possible to design an analysis framework that could directly estimate misread probability. Multi-event models that directly account for uncertainty in state assignment (Pradel 2005, Choquet et al. 2009) could provide a basis for explicitly modeling for the probability of false positive detections. These complex models add additional parameters and associated variances, however, which could obfuscate ecological research and management support. Therefore addressing the problem first through data collection and screening protocols is preferable, and modeling false positives directly may only be necessary for very long studies with suspected high rates of false positive detections.

All human data recording is prone to error, and errors due to individual misidentifications can accumulate in mark–resight studies as the study length and number of deployed marks increases. Long-term mark–resight studies may be especially vulnerable to biased estimates of apparent survival as a result of these errors if misread rates are high. Observer training and retention protocols may help reduce error during data collection, but ecologists working with historical mark–resight datasets should also consider potential effects of misread errors when designing analyses. Exploratory data analyses and simulation studies such as those presented here can aid in estimating probable error rates in a dataset and evaluating potential bias with the planned modeling approach. If the potential for misreads is not considered in analysis of long-term datasets collected by many observers, spurious trends caused by errors could influence ecological inference about population demographics and resulting conservation decision-making.

ACKNOWLEDGMENTS

We thank the many individuals who support the Delaware Shorebird Project, including J. Woods, G. Breese, and the large numbers of volunteers who carried out the Red Knot capture and resighting activities on which this paper is based. We thank C. Lepczyk, R. Lanctot, and 2 anonymous reviewers for comments that improved the quality of this manuscript. We also thank J. Hall for the use of her photograph. The findings and conclusions in this article do not represent the views of the U.S. Fish and Wildlife Service or U.S. Geological Survey. Use of product or firm names does not imply endorsement by the U.S. Government. We also acknowledge that data collection for this study was conducted on Lenape and Nanticoke land.

Funding statement: Support for the monitoring project was provided by U.S. Fish and Wildlife Service, Delaware Division of Fish and Wildlife, and the British Trust for Ornithology. AMT received support from Auburn University. No funders provided input regarding the content of the manuscript or required approval before submission.

Ethics statement: All capture and marking of birds was conducted by Delaware Division of Fish and Wildlife following their animal care protocols under permit #23332 from the Bird Banding Lab.

Author contributions: NAC and JAC conceived the study idea and withheld flag codes from circulation. AMT, RAR, CPM, and PWA conceived and designed the simulation study. AMT performed the analyses. ADW, RdF, and GEA oversaw data collection. AMT wrote the initial draft with significant input from JEL, CPM, RAR, and JAC. All authors provided comments and feedback on subsequent drafts.

Code availability: All R code used in this study is available at https://github.com/annamtucker/misreads

LITERATURE CITED

Atkinson, P. W., A. J. Baker, K. A. Bennett, N. A. Clark, J. A. Clark, K. B. Cole, A. Dekinga, A. Dey, S. Gillings, P. M. González, et al. (2007). Rates of mass gain and energy deposition in Red Knot on their final spring staging site is both time- and condition-dependent. Journal of Applied Ecology 44:885–895.

Baker, A. J., P. M. Gonzalez, R. I. G. Morrison, and B. A. Harrington (2001). Red Knot (*Calidris canutus*). In The Birds of North America (A. F. Poole, Editor). Cornell Lab of Ornithology, Ithaca, NY, USA. doi:10.2173/bna.563

Bearhop, S., R. M. Ward, and P. R. Evans (2003). Long-term survival rates in colour-ringed shorebirds – Practical considerations in the application of mark–recapture models. Bird Study 50:271–279.

Beck, A., B. Ackerman, L. Barton, and C. Hartley (2004). Survival estimates for Florida manatees from the photo-identification of individuals. Marine Mammal Science 20:438–463.
Bildstein, K. L. (1998). Long-term counts of migrating raptors: A role for volunteers in wildlife research. Journal of Wildlife Management 62:435–445.
Bolker, B., and R Development Core Team (2017). bbmle: Tools for general maximum likelihood estimation. R package version 1.0.20.
Burton, N. H. K. (2000). Variation in sighting frequencies of colour-ringed Redshanks *Tringa totanus* according to ringing-scheme and ring colour. Wader Study Group Bulletin 91:50–53.
Choquet, R., L. Rouan, and R. Pradel (2009). Program E-Surge: A software application for fitting multievent models. In Modeling Demographic Processes in Marked Populations (D. L. Thompson, E. G. Cooch, and M. J. Conroy, Editors). Springer, New York, NY, USA. pp. 845–865.
Clark, N., S. Gillings, A. J. Baker, P. M. Gonzalez, and R. Porter (2005). The production and use of permanently inscribed leg flags for waders. Wader Study Group Bulletin 108:38–41.
Cohn, J. P. (2008). Citizen science: Can volunteers do real research? BioScience 58:192–197.
Conrad, C. C., and K. G. Hilchey (2011). A review of citizen science and community-based environmental monitoring: Issues and opportunities. Environmental Monitoring and Assessment 176:273–291.
Flockhart, D. T. T., J. B. Pichancourt, D. R. Norris, and T. G. Martin (2015). Unravelling the annual cycle in a migratory animal: Breeding-season habitat loss drives population declines of monarch butterflies. Journal of Animal Ecology 84:155–165.
Gillings, S., P. W. Atkinson, A. J. Baker, K. A. Bennett, N. A. Clark, K. B. Cole, P. M. González, K. S. Kalasz, C. D. T. Minton, L. J. Niles, et al. (2009). Staging behavior in Red Knot (*Calidris canutus*) in Delaware Bay: Implications for monitoring mass and population size. The Auk 126:54–63.
Gimenez, O., J. D. Lebreton, R. Choquet, and R. Pradel (2018). R2ucare: An R package to perform goodness-of-fit tests for capture–recapture models. Methods in Ecology and Evolution 9:1749–1754.
Kéry, M., and M. Schaub (2012). Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective. Academic Press, Waltham, Massachusetts, USA.
Laake, J. L., D. S. Johnson, and P. B. Conn (2013). marked: An R package for maximum-likelihood and MCMC analysis of capture–recapture data. Methods in Ecology and Evolution 4:885–890.
Lavers, J. L., and I. L. Jones (2008). Assessing the type and frequency of band resighting errors for Razorbill *Alca torda* with implications for other wildlife studies. Marine Ornithology 36:19–23.
Lebreton, J. D., K. P. Burnham, J. Clobert, and D. R. Anderson (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. Ecological Monographs 62:67–118.
Link, W. A., J. Yoshizaki, L. L. Bailey, and K. H. Pollock (2010). Uncovering a latent multinomial: Analysis of mark–recapture data with misidentification. Biometrics 66:178–185.
Lukacs, P., and K. Burnham (2005). Estimating population size from DNA-based closed capture–recapture data incorporating genotyping error. Journal of Wildlife Management 69:396–403.
Magurran, A. E., S. R. Baillie, S. T. Buckland, J. M. Dick, D. A. Elston, E. M. Scott, R. I. Smith, P. J. Somerfield, and A. D. Watt (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. Trends in Ecology & Evolution 25:574–582.
McGowan, C. P., J. E. Hines, J. D. Nichols, J. E. Lyons, D. R. Smith, K. S. Kalasz, L. J. Niles, A. D. Dey, N. A. Clark, P. W. Atkinson, et al. (2011). Demographic consequences of migratory stopover: Linking Red Knot survival to horseshoe crab spawning abundance. Ecosphere 2. doi:10.1890/ES11-00106.1
McGowan, C. P., D. R. Smith, J. D. Nichols, J. E. Lyons, J. Sweka, K. Kalasz, L. J. Niles, R. Wong, J. Brust, M. Davis, and B. Spear (2015). Implementation of a framework for multi-species, multi-objective adaptive management in Delaware Bay. Biological Conservation 191:759–769.
Méndez, V., J. A. Alves, J. A. Gill, and T. G. Gunnarsson (2018). Patterns and processes in shorebird survival rates: A global review. Ibis 160:723–741.
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.
Milligan, J. L., A. K. Davis, and S. M. Altizer (2003). Errors associated with using colored leg bands to identify wild birds. Journal of Field Ornithology 74:111–118.
Mitchell, C., and M. Trinder (2008). On reading colour rings. Ringing and Migration 24:11–14.
Morrison, T. A., J. Yoshizaki, J. D. Nichols, and D. T. Bolger (2011). Estimating survival in photographic capture–recapture studies: Overcoming misidentification error. Methods in Ecology and Evolution 2:454–463.
Newman, C., C. D. Buesching, and D. W. Macdonald (2003). Validating mammal monitoring methods and assessing the performance of volunteers in wildlife conservation—"*Sed quis custodiet ipsos custodies?*" Biological Conservation 113:189–197.
Newton, I. (2006). Can conditions experienced during migration limit the population levels of birds? Journal of Ornithology 147:146–166.
Pradel, R. (2005). Multievent: An extension of multistate capture–recapture models to uncertain states. Biometrics 61:442–447.
Pradel, R., J. E. Hines, J. D. Lebreton, and J. D. Nichols (1997). Capture–recapture survival models taking account of transients. Biometrics 53:60–72.
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Roche, E. A., C. M. Dovichin, and T. W. Arnold (2014). Field-readable alphanumeric flags are valuable markers for shorebirds: Use of double-marking to identify cases of misidentification. Journal of Field Ornithology 85:329–338.
Rushing, C. S., J. A. Hostetler, T. S. Sillett, P. P. Marra, J. A. Rotenberg, and T. B. Ryder (2017). Spatial and temporal drivers of avian population dynamics across the annual cycle. Ecology 98:2837–2850.
Schwarz, C. J., and W. T. Stobo (1999). Estimation and effects of tag-misread rates in capture–recapture studies. Canadian Journal of Fisheries and Aquatic Sciences 56:551–559.
Silvy, N. J., R. R. Lopez, and M. J. Peterson (2012). Techniques for marking wildlife. The Wildlife Techniques Manual Volume 1: Research (N. J. Silvy, Editor). Johns Hopkins University Press, Baltimore, MD, USA. pp. 230–257.

Tulloch, A. I. T., H. P. Possingham, L. N. Joseph, J. Szabo, and T. G. Martin (2013). Realising the full potential of citizen science monitoring programs. Biological Conservation 165:128–138.

Weiss, N. T., M. D. Samuel, D. H. Rusch, and F. D. Caswell (1991). Effects of resighting errors on capture–resight estimates for neck-banded Canada Geese. Journal of Field Ornithology 62:464–473.

Williams, B., J. Nichols, and M. Conroy (2002). Analysis and management of wildlife populations. 1st edition. Academic Press. San Diego, CA, 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.

Yoshizaki, J., C. Brownie, K. H. Pollock, and W. A. Link (2011). Modeling misidentification errors that result from use of genetic tags in capture–recapture studies. Environmental and Ecological Statistics 18:27–55.

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.

Yu, J., R. A. Hutchinson, and W. K. Wong (2014). A latent variable model for discovering bird species commonly misidentified by citizen scientists. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, Palo Alto, CA, USA. pp. 500–507.

APPENDIX TABLE 2. RMSE and bias of CJS model estimates under varying error rates and study lengths.

		Survival probability		Detection probability	
Study length (years)	Error rate	RMSE	Relative bias	RMSE	Relative bias
No data filtering					
5	0	0.025	0.113	0.018	0.231
5	0.005	0.025	0.320	0.019	−0.245
5	0.01	0.025	0.433	0.019	0.216
5	0.05	0.033	2.55	0.021	0.411
5	0.1	0.049	5.00	0.020	0.564
10	0	0.010	−0.006	0.008	0.004
10	0.005	0.010	0.334	0.009	−0.222
10	0.01	0.012	0.723	0.011	−0.950
10	0.05	0.031	3.344	0.022	−4.04
10	0.1	0.054	6.073	0.034	−6.56
20	0	0.005	0.008	0.005	−0.037
20	0.005	0.007	0.547	0.008	−1.40
20	0.01	0.011	1.070	0.016	−3.00
20	0.05	0.040	4.53	0.058	−11.5
20	0.1	0.066	7.63	0.092	−18.3
Remove single-observations					
5	0	0.094	0.164	0.425	−84.4
5	0.005	0.093	0.306	0.423	−84.1
5	0.01	0.092	0.086	0.424	−84.3
5	0.05	0.092	0.091	0.425	−84.5
5	0.1	0.091	0.352	0.424	−84.4
10	0	0.027	−0.062	0.425	−84.6
10	0.005	0.028	−0.126	0.425	−84.6
10	0.01	0.027	−0.067	0.426	−84.7
10	0.05	0.027	−0.004	0.426	−84.9
10	0.1	0.027	−0.157	0.425	−84.5
20	0	0.011	−0.006	0.425	−84.6
20	0.005	0.011	−0.006	0.425	−84.6
20	0.01	0.011	−0.070	0.425	−84.6
20	0.05	0.011	0.001	0.426	−84.7
20	0.1	0.011	−0.076	0.425	−84.6