Studies in Avian Biology No. 6:353–359, 1981.

THE ROLE OF OBSERVER BIAS IN THE NORTH AMERICAN BREEDING BIRD SURVEY

CRAIG A. FAANES[1] AND DANNY BYSTRAK[2]

ABSTRACT.—Ornithologists sampling breeding bird populations are subject to a number of biases in bird recognition and identification. Using Breeding Bird Survey data, these biases are examined qualitatively and quantitatively, and their effects on counts are evaluated. Differences in hearing ability and degree of expertise are the major observer biases considered. Other, more subtle influences are also discussed, including unfamiliar species, resolution, imagination, similar songs and attitude and condition of observers. In most cases, well-trained observers are comparable in ability and their differences contribute little beyond sampling error. However, just as hearing loss can affect results, so can an unprepared observer. These biases are important because they can reduce the credibility of any bird population sampling effort. Care is advised in choosing observers and in interpreting and using results when observers of variable competence are involved.

The ability of observers to discriminate among individuals of several breeding bird species aurally and visually is of paramount importance to the outcome of most bird population sampling efforts. A factor limiting the accuracy of most such efforts is observer ability to correctly identify songs of each species. Visual identification biases are certainly encountered, but do not appear to be as prevalent nor as important as aural biases. Because identification by song is predominant in most bird surveys, hearing acuity and training would be expected to affect results greatly. This factor should always be considered when selecting observers for bird population sampling.

The North American Breeding Bird Survey (BBS) was established in 1966 to provide an annual index of bird abundance and establish trends in continental and regional populations of most North American bird species (Robbins and Erskine 1975). The techniques of this survey have been discussed by Robbins and Van Velzen (1967). Normally only observers with a keen sense of hearing and a thorough knowledge of bird song are selected for BBS routes. In some instances, however, observers who are unfamiliar with many bird songs, or have a hearing loss are selected. Although data from these observers do not contribute to the intended purpose of the BBS, their results provide useful measures of observer bias.

METHODS

Data from 65 selected BBS routes conducted in the central and eastern U.S. and Canada were subjected to two tests for similarity using the Bray-Curtis (hereafter BC) similarity index (Clifford and Stephenson 1975:57). Many other such indices are available (Huhta 1979) but this one was used because it considers actual numbers of birds as well as presence and absence. One test (26 routes) compares qualified to unqualified observers and the other (39 routes) compares pairs of two comparably qualified observers. For both tests, four consecutive years of data were used from each route, with a change of observer occurring after the second, so that within-observer (internal), as well as between-observer similarities could be calculated. The quartets of years were chosen from the entire 14 years of the BBS, so any yearly biases are minimized.

In addition, results of eight BBS routes conducted in the central and eastern United States were used to examine some typical situations in more detail. Three routes were used to analyze biases resulting from hearing loss, and two each for observer training and song confusion in otherwise well-trained observers. The eighth route was run simultaneously by the authors to provide data on results from observers of equal ability.

RESULTS

UNEQUAL OBSERVERS

In editing BBS routes, it has become obvious that known "underqualified" (either from hearing loss or lack of training) observers record consistently lower species totals on the same routes than do qualified observers. However, qualified observers of equal expertise produce consistently similar species totals. Considering these facts, species total alone was used as the criterion for choosing unequal observers.

To evaluate the influence of underqualified observers on the BBS, we used results of 26 routes on which the change of observer was from qualified to underqualified or vice-versa. Table 1 shows the mean internal similarity, with standard deviations, of qualified and underqualified observers as well as the mean and standard deviation of the between-observer similarity.

In this test, the mean internal similarity of the qualified observers is only 0.80. This is also true in the equal observer test (Table 2). Considering that each index was calculated from data gathered on different runs conducted in adjacent years, the two major factors explaining the re-

[1] U.S. Fish and Wildlife Service, Northern Prairie Wildlife Research Center, Jamestown, North Dakota 58401.

[2] U.S. Fish and Wildlife Service, Migratory Bird and Habitat Research Laboratory, Laurel, Maryland 20811.

TABLE 1
MEAN YEAR-TO-YEAR SIMILARITY OF RESULTS OF 26 BBS ROUTES, EACH CONDUCTED FOR FOUR CONSECUTIVE YEARS, TWO YEARS EACH BY A QUALIFIED AND AN UNDERQUALIFIED OBSERVER

	Underqualified		Qualified
	Internal[a] similarity	Between[b] observer similarity	Internal[a] similarity
$\bar{x}$	.7415	.6360	.7958
σ	.0574	.0866	.0458

[a] Bray-Curtis Similarity Index of results of two runs of the same BBS route in adjacent years by the same observer.

[b] Bray-Curtis Similarity Index of results of two runs of the same BBS route conducted in adjacent years by different observers.

maining 20% are probably sampling error and annual bird-population change. To test for sampling error alone, data from consecutive runs of the same route by the same observer are needed. A Maryland route was run by Bystrak on 3 consecutive days in 1969 and the mean of the three consequent BC indices was 0.84. This is a small sample, but it suggests that as little as 4% is the effect of annual change. Because the internal similarity indices of some of the observers in our sample were as high as 0.90, we feel that annual change contributing less than 4% is no doubt an insignificant consideration in this test. This is supported by the fact that few species ever show a significant population change from one year to the next (Bystrak 1981). Thus, sampling error appears to predominantly explain the 20%. Sampling error is influenced by several factors, such as weather, bird activity, noise, observer alertness and others. Further testing will be helpful in separating these factors and their magnitudes.

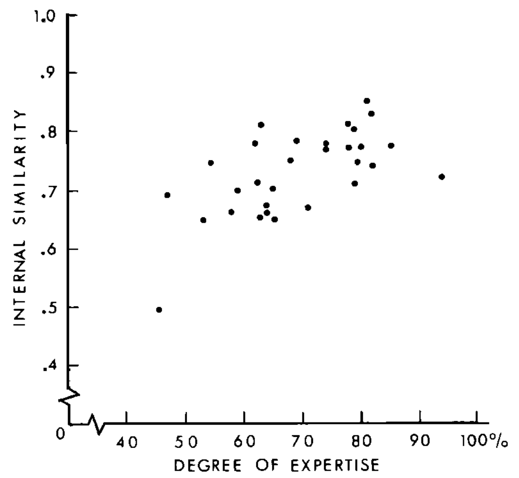

FIGURE 1. Relationship of degree of expertise to observer reliability. Degree of expertise is expressed as a ratio of species totals of underqualified to qualified observers on the same route. Reliability represents internal similarity of results of two separate runs of the same route by the underqualified observers.

The mean internal similarity index of the qualified observers in this test (0.7958) was compared to the mean index of the qualified to underqualified observers (0.6360) and found to differ significantly ($P < 0.01$). This suggests that, in calculating population trends, it would be dangerous to compare an underqualified observer's results with those of a qualified observer. It is often suggested, however, that underqualified observers' results are reliable enough to be used in analyzing annual trends, presuming that they are at least producing a reliable index of those species they are recording. However, the mean internal similarity (cf. reliability) of the 26 underqualified observers (0.7415) was significantly different ($P < 0.01$) from that of the qualified observers (0.7958). The underqualified observers' own results from one year to the next are not even as comparable as those of two different qualified observers in adjacent years (0.7415—Table 1 vs. 0.7537—Table 2). We decided to investigate this further by testing the possibility that reliability increases as a function of expertise. We compared the internal similarity (BC index) of 30 underqualified observers with their apparent degree of expertise (Fig. 1). The correlation is significant ($\tau = .310$, $P = 0.016$) when tested with Kendall's coefficient of rank correlation. This strengthens our belief that it is safer to not include the results of underqualified observers whenever they can be identified as such.

TABLE 2
MEAN YEAR-TO-YEAR SIMILARITY OF RESULTS OF 39 BBS ROUTES, EACH CONDUCTED FOR FOUR CONSECUTIVE YEARS, THE FIRST TWO BY ONE QUALIFIED OBSERVER AND THE NEXT TWO BY A COMPARABLE OBSERVER

	Observer 1		Observer 2
	Internal[a] similarity	Between[b] observer similarity	Internal[a] similarity
$\bar{x}$	.7991	.7537	.7865
σ	.0490	.0707	.0624

[a] Bray-Curtis Similarity Index of results of two runs of the same BBS route in adjacent years by the same observer.

[b] Bray-Curtis Similarity Index of results of two runs of the same BBS route conducted in adjacent years by different observers.

Hearing loss

To explore the effects of hearing loss in more detail, we used results of three BBS routes, each

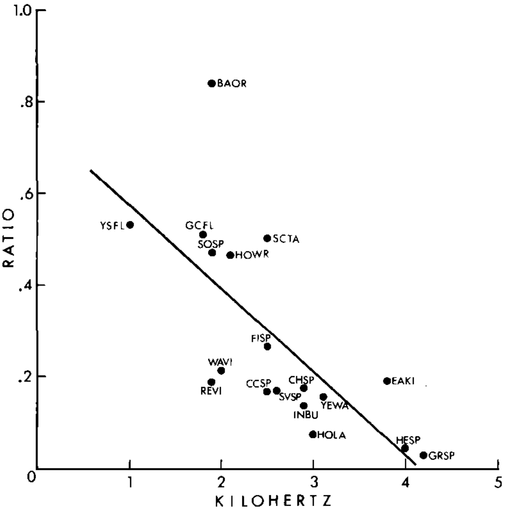

FIGURE 2. Relationship of bird-song frequency to hearing loss on two Wisconsin BBS routes. Ratio = *i*/*u*, where *i* = 3-year mean recorded by impaired observer and *u* = 3-year mean recorded by unimpaired observer ($y = 0.7575 - 0.1826$, $r^2 = 0.4931$). The 18 species are *Y*ellow-*s*hafted *Fl*icker (*Colaptes auratus*), *Ea*stern *Ki*ngbird (*Tyrannus tyrannus*), *G*reat *C*rested *Fl*ycatcher (*Myiarchus crinitus*), *Ho*rned *La*rk (*Eremophila alpestris*), *Ho*use *Wr*en (*Troglodytes aedon*), *R*ed-*e*yed *Vi*reo (*Vireo olivaceus*), *Wa*rbling *Vi*reo (*Vireo gilvus*), *Ye*llow *Wa*rbler (*Dendroica petechia*), *Ba*ltimore *Or*iole (*Icterus galbula*), *Sc*arlet *Ta*nager (*Piranga olivacea*), *In*digo *Bu*nting (*Passerina cyanea*), *S*a*v*annah *Sp*arrow (*Passerculus sandwichensis*), *Gr*asshopper *Sp*arrow (*Ammodramus savannarum*), *He*nslow's *Sp*arrow (*Passerherbulus henslowii*), *Ch*ipping *Sp*arrow (*Spizella passerina*), *C*lay-*c*olored *Sp*arrow (*Spizella pallida*), *Fi*eld *Sp*arrow (*Spizella pusilla*) and *So*ng *Sp*arrow (*Melospiza melodia*). Italicized letters in common names indicate four-letter species codes.

conducted in adjacent series of years by an observer with acute hearing and one without. We examined differences attributable to hearing loss and the progression of the loss as it relates to different song frequencies. The first two routes were run by an unimpaired observer for 3 years, followed by 12 years by an observer with an admitted progressive hearing loss. From these, we compared (Fig. 2) the interobserver ratio of 3-year mean counts of 18 species to the respective lowest frequencies of their typical songs (Robbins et al. 1966). The two observers' song perceptions were most similar at low frequencies, but the impaired observer had great difficulty in perceiving higher frequencies.

On these routes, the impaired observer's results were initially very similar to the unimpaired observer's, and thus offer a prime opportunity to study the effects of the progression of hearing loss during a 12-year period. The 18 species in Figure 2 were equally broken into three frequency categories, and the total annual counts for each plotted against time (Fig. 3). The three graphs imply that high-pitched songs were lost rapidly at first to a point where the birds were being recorded only visually or at close range, mid-frequency songs were lost at a steady rate, and low-pitched songs were stable longer, but also lost rapidly after time.

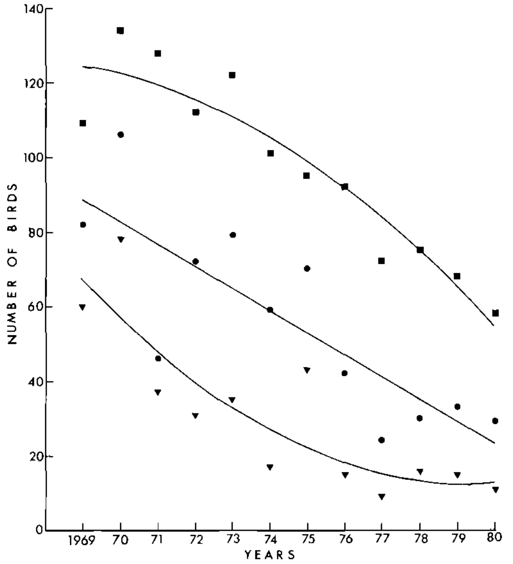

FIGURE 3. Twelve-year progression of hearing loss in a BBS observer, relative to frequency of bird song. Numbers of birds are annual totals in three frequency categories. Squares represent six low-pitched (1–2 kHz) species ($y = -0.44x^2 - 0.67x + 125.27$, $r^2 = 0.8861$). Circles represent six middle-pitched (2.1–2.9 kHz) species ($y = -5.92x + 94.45$, $r^2 = 0.6745$). Triangles represents six high-pitched (3.0–4.2 kHz) species ($y = 0.52x^2 - 11.66x + 78.36$, $r^2 = 0.7399$).

In the second example, we compared differences in mean species totals in several avian families on a BBS route covered by another pair of observers with differing hearing abilities (Table 3). We chose the five families portrayed because they represent a wide range of song frequencies and also vary considerably in conspicuousness. Totals were most similar for the Mimidae and Turdidae whose songs are generally repetitious, multi-syllabic, and of long duration and moderate frequency (Borror 1964), and least similar among the Tyrannidae and Parulidae whose songs are generally brief, simple and of higher frequency. As might be expected, the smallest differences among Tyrannids and

TABLE 3
TOTAL INDIVIDUALS OF FIVE BIRD FAMILIES REPORTED ON A WISCONSIN BBS ROUTE RUN IN SEVEN CONSECUTIVE YEARS BY TWO OBSERVERS WITH DIFFERENT HEARING ACUITY

Family	Good hearing (4-year means)	Poor hearing (3-year means)	% difference
Tyrannidae	55.75	21.33	−61.74
Mimidae	26.25	17.00	−35.24
Turdidae	54.25	55.67	+2.55
Vireonidae	43.00	19.67	−54.26
Parulidae	127.25	49.33	−61.23

Parulids were also in those species with the loudest, most repetitive, longest or lowest-frequency songs. For example, among Tyrannids the smallest difference was in Great Crested Flycatcher (*Myiarchus crinitus*), and among Parulids, Common Yellowthroat and Ovenbird (Table 4) were the least different.

Expertise

Degree of expertise is the second major contributor to observer inequality, hence the observer's ability to correctly identify and discriminate the species involved should always be considered in breeding bird population sampling. In this section, we offer a comparison of results obtained by observers of different levels of expertise. Also, some examples of confusion of similar songs by otherwise well-trained observers are given.

In the first example, a Maryland BBS route was covered for 4 years by qualified and comparable observers. The next 4 years, the route was covered by an observer who compared in age and hearing ability, but not in expertise. Four year mean counts of each species are used in this comparison (Table 5). In this example, three categories of species recognition are recognized: "sight," "song" and "both." The "song" species are further divided by song type: "easily learned," "intermediate" and "difficult to learn." The untrained observer recorded fewer birds overall but was consistent among "sight" species, "both" species and even "easily learned song" species. The "intermediate song" species showed a lower percentage and the "difficult song" species an even lower percentage. In examining BBS data, this low percentage for "difficult song" species appears to be typical for poorly trained observers. There is even some indication that, in many cases, poorly trained observers record the species they are familiar with in higher numbers than do well trained observers on the same route.

In another example, two observers of different levels of expertise conducted the same route in Wyoming three days apart in 1980. The results demonstrate 2 instances of misidentification of two pairs of birds with similar songs, Western and Willow Flycatchers (*Empidonax difficilis* and *E. traillii*) and Warbling Vireo (*Vireo gilvus*) and Cassin's Finch (*Carpodacus cassinii*). The experienced observer recorded 10 Western Flycatchers and 57 Warbling Vireos, both of which are common breeders in the area (Pettingill and Whitney 1965). The poorly trained observer recorded, instead, 12 Willow Flycatchers and 15 Cassin's Finches, neither of which is known to breed there (Pettingill and Whitney 1965). Because these songs are so similar, especially to an untrained observer, we have no doubt that the poorly trained observer misidentified them. The species totals differed on these two runs (52 vs. 34), in keeping with our observation that species total is the best single measure of degree of expertise.

Bird species with similar songs can present an

TABLE 4
TOTAL INDIVIDUAL WOOD WARBLERS REPORTED ON A WISCONSIN BBS ROUTE RUN IN SEVEN CONSECUTIVE YEARS BY TWO OBSERVERS WITH DIFFERENT HEARING ACUITY

Species	Good hearing (4-year means)	Poor hearing (3-year means)	% difference
Black-and-white Warbler (*Mniotilta varia*)	6.0	0.3	94.4
Golden-winged Warbler (*Vermivora chrysoptera*)	12.5	1.0	92.0
Nashville Warbler (*Vermivora ruficapilla*)	4.5	0.7	85.2
Yellow Warbler (*Dendroica petechia*)	20.2	6.3	68.6
Chestnut-sided Warbler (*Dendroica pensylvanica*)	15.7	3.7	76.6
Ovenbird (*Seiurus aurocapillus*)	13.5	10.0	25.9
Mourning Warbler (*Oporornis philadelphia*)	3.7	0.3	91.0
Common Yellowthroat (*Geothlypis trichas*)	43.5	26.3	39.5
Canada Warbler (*Wilsonia canadensis*)	1.2	0	100
American Redstart (*Setophaga ruticilla*)	6.2	0.7	89.2

TABLE 5
COMPARISON OF RESULTS OF QUALIFIED AND UNDERQUALIFIED OBSERVERS

	Species recognition method				
	Song				
	Difficult to learn	Intermediate	Easy to learn	Slight	Both
Qualified[a]	42.25	116.5	189.75	284.0	286.25
Underqualified[a]	6.75	63.25	156.25	210.0	243.0
Ratio	.16	.54	.82	.74	.95

[a] Numbers are 4-year means recorded on the same BBS route.

identification problem to many well-trained observers also, but usually on a small scale. In Maryland, the Ovenbird and the Kentucky Warbler (*Oporornis formosus*) are common nesting species with similar songs. Along a Maryland BBS route, a well trained observer familiar with both songs recorded means of 2.6 Kentucky Warblers and 1.4 Ovenbirds over five years. Another known well-trained observer covered this route for the following four years and recorded no Kentucky Warblers, but did record a mean of 3.5 Ovenbirds. The original observer resumed coverage and recorded 4-year means of 2.5 and 1.5 respectively. It thus appears that the second observer was combining the two. Counting birds by call note can produce similar differences between experienced observers because call notes of many species are confusing. For example, two comparable observers conducted another Maryland BBS route simultaneously, and obtained similar results for most species. However, one observer recorded 4 Common Flickers (*Colaptes auratus*) and 10 Red-bellied Woodpeckers (*Melanerpes carolinus*), while the other recorded 11 Common Flickers and 4 Red-bellied Woodpeckers. The call notes of these species are similar and distant birds can easily be confused.

EQUAL OBSERVERS

Comparable, qualified observers should, logically, produce similar results on an effort such as the BBS. The analysis programs used with the BBS have always used data from all qualified observers, regardless of changes of personnel on specific routes. To examine the results of comparable observers, we selected on the basis of similar species total, 39 routes on which changes of observers had occurred (Table 2). As with unequal observers (Table 1), 4-year periods were chosen, with the change occurring after the second year. Both sets of qualified observers produced mean BC indices of approximately 0.80. Unlike the unequal observers test, there is no significant difference between the internal similarities of the 2 groups (0.7991 vs. 0.7865). Surprisingly, however, the between-observer similarity (0.7537) was significantly different from both of the internal similarities ($P < 0.01$, $P < 0.05$ respectively). Although the differences are significant, they are slight, and because personnel changes are scattered throughout the 14 years of data, their effect is minimized. Robbins and Van Velzen (1969) analyzed BBS results for annual change using results of all qualified observers versus using only results of routes covered by the same observers and concluded that the increased sample size more than compensates for the small additional variability.

In late May 1980, we had the opportunity to conduct simultaneously a BBS route in the Turtle Mountains of North Dakota. We used this run to examine, in detail, the results of two equal observers when all other variables are the same. There were no significant differences between species totals (79 vs. 77, with 75 in common) or total individuals (1061 vs. 1141). Individual totals among species were quite similar, with no apparent difference in perception of most species. These similarities were readily apparent for both common and conspicuous species (e.g., Red-winged Blackbird, *Agelaius phoeniceus,* 97 vs. 96) and for uncommon species (e.g., Common Snipe, *Capella gallinago,* 4 vs. 3). The only major discrepancy was in the counts of Yellow Warbler (*Dendroica petechia*): 85 individuals at 41 stops, and 105 individuals at 48 stops.

Because observer bias did not appear to be a major factor affecting the outcome of this route, we investigated variability between ourselves by examining the raw data for Franklin's Gull (*Larus pipixcan*). There was no significant difference in our individual totals. Bystrak recorded 57 gulls at 16 stops; Faanes 55 gulls at 15 stops. Next, we examined the stops at which Franklin's Gulls were recorded, and found that we were not recording the same individuals at the same stops. The average error rate for individual totals was 3.7%, which suggests that our population figures were the same. The average error rate for each of the 50 stops was 18.9%, and at each stop where Franklin's Gulls were recorded, we usually obtained the same figure, or came within one bird of each other. The greatest amount of variability came from stops where eight or more individuals were observed. At several of these, our totals varied by nearly five birds, and there were several stops where one observer recorded the gulls and the other did not.

DISCUSSION

Hearing

As we and others have demonstrated, hearing is an important bias affecting counts of singing birds. We have found that, depending upon habitat, as high as 95% of the individual birds recorded on BBS routes are detected by hearing. Mayfield (1966) discussed hearing loss as it affects ornithologists, and showed that human males begin to lose perception of higher frequencies at age 32, and females at 37. The first frequencies to be affected are usually those above 4 kHz, the range of most bird songs. It is clear from our results that most bird population sampling is dependent on keen hearing ability for accurate results. An aspect of hearing ability not often considered is that of exceptionally good hearing. A few BBS recruits consistently report species and individual counts higher than those of well-trained observers with no known hearing impairments. In some instances, this may represent the influence of imagination, but in legitimate instances, such observers can innocently produce incomparable data.

Expertise and Training

Poorly trained observers present a problem that is potentially more difficult to deal with than hearing loss. Such observers are usually unreliable because they are inconsistent and given to incorrect identifications in addition to missing an unpredictable array of birds. The ability to learn bird songs is unfortunately highly variable and usually an individual process, yet few skilled observers did not benefit from earlier interaction with others. All observers respond to training and experience differently. Some are well trained in recognizing call notes as well as songs. In efforts such as the BBS, where more than singing males are sampled, this additional knowledge can bias results. This is also true for unusual song variations and song-trading, both of which are common, with examples too numerous to mention here. Even highly skilled observers find that there is much additional information to be learned about unusual bird vocalizations.

Subtle Biases

Although it is relatively easy to recognize observers with similar hearing or expertise, other subtle factors must be considered when striving for uniformity in sampling. Resolution, the ability to distinguish individual birds from a large number, such as in a dawn chorus, can produce higher counts, as can an active imagination that creates a second bird when one turns its head or sings from a different perch. Separating these two influences can often be difficult. Attitudes of observers can contribute to observer bias. Some are unaware of or unwilling to admit their shortcomings, and consider themselves well-trained. They are usually surprised to discover that they are unfamiliar with several species on a study plot or along a survey route. In these situations, data can be gathered incorrectly for years. A similar situation can occur when an observer moves to a new area and is unfamiliar with or unaware of new species or dialects. The condition of the observer can also be important. Most bird population sampling is conducted during early hours, so results are often a function of amount of sleep. It is likewise important to keep the effort reasonable because of individual differences in tolerance levels. A common complaint from BBS participants is that 50 stops is excessive. Proper rest notwithstanding, 4 hours appears too long for many to maintain the necessary high level of alertness.

SOLUTIONS

In order to work with the seemingly hopeless array of observer bias problems, the most important first step in any bird population sampling effort is to consider these problems and to what extent they will affect results. Next, an effort must be made to identify and overcome as many of these problems as possible.

Training is the logical solution in most cases, and should not only include a basic knowledge of identification cues but also as much familiarity with local variations and dialects as possible. If several observers are being used, it is fairly easy to equalize their abilities with concurrent field testing. Kepler and Scott (1981) describe an attempt to offset some observer bias in bird censusing with training sessions prior to initiation of fieldwork, and conclude that training is beneficial in arriving at more precise estimates. Such sessions help to acquaint observers with unknown, confusing, and similar songs, as well as to identify areas of weakness. Because of the large scale nature of the BBS, intensive training of observers would be excessively expensive and met with varying degrees of success. People differ widely in their abilities and speed in learning bird songs. Motivation plays an important role, and unmotivated individuals will probably never become particularly proficient. The best that can be hoped for on the BBS is quality control in choosing observers and in carefully screening results after they are submitted.

To overcome the problem of dawn chorus overloading or to increase the accuracy of most kinds of counts, a division of responsibility is useful. Scott and Ramsey (1981) found that by using two observers together and reducing the responsibility assigned to each, the accuracy of each observer was increased. This approach is, of course, not possible on the BBS because uniformity of coverage is crucial. It would be impossible, with the number of qualified observers available, to ensure two or more observers on every route.

Underqualified observers should not be a problem in small studies where control is easy or training is possible, but these observers can be numerous in large-scale projects such as the BBS. Enemar et al. (1978) postulated that observer variability in large scale census work involving many observers tends to produce an insignificant bias. In the BBS, where predominantly competent, comparable observers are involved, this appears to be true, especially when reports from the most obviously underqualified observers are eliminated. It is hoped that our examples, however, demonstrate the necessity for strict controls in small-scale studies.

ACKNOWLEDGMENTS

This paper benefitted from the constructive criticisms provided by D. H. Johnson, C. S. Robbins, S. D. Robbins and A. B. Sargeant.