J. Field Ornithol., 53(2):159–167

OBSERVER VARIABILITY IN ESTIMATING NUMBERS: AN EXPERIMENT

By R. Michael Erwin

Census estimates of bird populations provide an essential framework for a host of research and management questions. However, with some exceptions (Matthews 1960, LeResche and Rausch 1974, Caughley 1974, Caughley et al. 1976), the reliability of numerical estimates and the factors influencing them have received insufficient attention. Independent of the problems associated with habitat type, weather conditions (Stott and Olson 1972), cryptic coloration, etc., estimates may vary widely due only to intrinsic differences in observers' abilities to estimate numbers (LeResche and Rausch 1974, Prater 1979, Erwin 1979, 1980).

Lessons learned in the field of perceptual psychology may be usefully applied to "real world" problems in field ornithology. Based largely on dot discrimination tests in the laboratory, it was found that numerical abundance (Jerons 1871, Kaufman et al. 1949, Krueger 1972, Indow and Ida 1977), density of objects (Horne and Allee 1971, Class 1972), spatial configuration (Lechelt and Tanne 1976), color, background, and other variables influence individual accuracy in estimating numbers.

The primary purpose of the present experiment was to assess the effects of observer, prior experience, and numerical range on accuracy in estimating numbers of waterfowl from black-and-white photographs. By using photographs of animals rather than black dots, I felt the results could be applied more meaningfully to field situations. Further, reinforcement was provided throughout some experiments to examine the influence of training on accuracy.

METHODS

Fifty 23 cm^2 black-and-white vertical aerial photographs of rafting Canvasbacks (*Aythya valisineria*) were selected with total counts ranging from 40 to 3100 birds. Only those photographs were used which had high clarity and uniform (water) background. Most of the photographs had been slightly overexposed, reducing the contrast differences of male and female ducks in mixed rafts. The photographs were divided into 5 groups of 10 so that each "observer" could be tested over 5 consecutive days. Each group of 10 photographs contained a roughly even distribution of numbers in each of the following size ranges: <100, 100–250, 251–500, 501–750, 1000–2000, >2000. Because of the limited number of photographs, only one photograph per day was used in the 501–750 size range. All other size ranges usually had 2 photographs per day. Presentation time allowed for each photograph varied from 15 s (for <250 birds) to one min (>1000 birds).

Three observers were chosen in each of three experience categories: inexperienced (no previous experience in estimating bird numbers),

past experience (3 seasons of experience in counting or estimating waterfowl from aircraft and/or aerial photographs, but no practice in the past 3 years), and recent experience (either current aerial estimation work or within the past year). All observers were male professional wildlife biologists or managers. Each observer was tested alone during 20-min morning sessions conducted over 5 consecutive days in the same location.

As one type of training, reinforcement was given in which the observer would first estimate the number of birds on the photograph. The experimenter then would reveal the correct number. Estimates were compared with actual counts for each photograph, day (totals from 10 photographs), and the overall experiment (totals from 50 photographs) for each observer.

As a separate experiment, 20 photographs (days 4 and 5) were used during one session of non-reinforcement in which correct counts were not revealed. The 3 observers with recent experience were tested one week prior to the reinforcement tests. Three additional inexperienced observers were used in the non-reinforcement session.

Data were analyzed by parametric and non-parametric tests (Siegel 1956). The Statistical Analysis System (SAS) (Barr et al. 1979) was used for ANOVA procedures.

RESULTS

Reinforcement tests.—The results of the experiments involving reinforcement are summarized and divided into those based on mean accuracy for individual photographs (Table 1) and those based on daily totals of 10 photographs (Table 2). The daily means from Table 1 were analyzed by analysis of variance (SAS - General Linear Model (GLM), Barr et al. 1979), using a split plot design. The effects of experience level, training (5 daily trials), and observers (nested within experience level) were examined. None of the factors singly (or interactions) accounted for a significant fraction of the variance (for experience, $F = 0.12$, df = 2, 6, $P = .89$; daily training, $F = 3.66$, df = 1, 33, $P = .10$).

Reinforcement as a training method apparently achieved little improvement in accuracy during the 5-day period (Table 1). In only one of 9 trials was the fifth day the "best" (=most accurate). Conversely, the first day was the worst for only 3 of the 9.

The above results reveal that experience level has little bearing on accuracy in estimating numbers. Because of small sample size and variability, a power analysis (Steel and Torrie 1960, Cohen 1977) showed that differences between 2 specified treatments of 77% or greater could be detected with 95% probability (power) when testing at the 95% confidence level (Steel and Torrie 1960:154–156).

Despite the high error rates found (Tables 1 and 2) by all observers, the overall deviations (50 photographs) were surprisingly low for most observers (Table 2). With one exception, all observers were within 10% of the total figure.

TABLE 1. Summary of results showing daily mean % error[1] of numerical estimates of waterfowl on photographs (with reinforcement).

Experience level	Observer	Day					Overall		
		1	2	3	4	5	Daily $\bar{x}$	S^2	CV^2
Inexperienced	1	34.8	18.7	22.2	11.5	15.7	20.6	78.7	43
	2	20.4	12.1	12.6	20.6	19.7	17.1	18.8	25
	3	38.4	18.7	19.5	18.1	25.9	24.1	73.5	36
Past experience	1	20.6	11.9	13.3	13.7	13.9	14.7	11.6	23
	2	12.0	19.3	19.3	14.9	13.2	15.7	11.6	22
	3	32.2	35.3	16.3	21.6	23.9	25.9	60.7	30
Recent experience	1	23.5	20.6	27.4	27.1	13.4	21.9	27.8	24
	2	15.8	20.8	21.7	21.6	20.1	20.0	5.9	12
	3	28.9	25.5	55.2	28.5	29.6	33.5	149.1	36

[1] Each daily trial consisted of 10 photographs.
[2] Coefficient of variability, $S/\bar{x} \times 100$.

The effect of numerical magnitude on estimation accuracy and direction of deviation can best be examined by partitioning the numerical range into 3 broad classes (Table 3). Each photograph was placed in the appropriate size category and scored; + indicating overestimation, − indicating underestimation. Two null hypotheses were tested. The first is that equal proportions of over- (+) and underestimates (−) are expected across the 3 size ranges. Combining responses of the 3 observers in each experience level, results indicated that only those observers with recent experience were significantly affected by numerical range in their tendency to over- or underestimate. A second hypothesis was simply that, within each size category, equal proportions of overestimates and underestimates are expected. Chi-square tests conducted on the combined responses for each experience level revealed that both inexperienced observers ($\chi^2 = 11.26$, df = 3, $P < .05$) and those with recent experience ($\chi^2 = 8.76$, df = 3, $P < .05$) were significantly skewed while observers with past experience showed no departure from a 50:50 probability ($\chi^2 = 2.98$, df = 3, $P > .90$).

To briefly summarize these two tests, two points are clear: (1) inexperienced observers underestimate across all numerical categories, (2) observers with recent experience consistently underestimated only in the smaller size category. It is noteworthy that of all 6 experienced observers, only one consistently underestimated across all size ranges. Least squares regression lines are compared for all 9 observers in Fig. 1.

Reinforcement after each photograph in each daily trial allowed me to examine the pattern of overestimation and underestimation by the observers using a runs test (Siegel 1956). The sequence of + (over) and − (under) signs was analyzed daily for each observer. For no observers did the pattern deviate significantly from random.

TABLE 2. Summary of results showing overall % error[1] of numerical estimates of waterfowl.

Experience level	Observer	Day 1	2	3	4	5	Overall deviation (%)[2]	Daily $\bar{x}$	S^2	CV
Inexperienced	1	28+	27−	11−	5−	5+	0	15.2	132.3	76
	2	9−	10−	3+	19−	9+	6−	10.0	32.5	57
	3	6+	9−	5+	2−	26−	6−	9.6	90.3	99
Past experience	1	10−	4+	3−	13+	7+	2+	7.4	17.6	57
	2	11+	18+	15+	1+	6−	2+	10.2	46.2	67
	3	22+	4+	3+	14+	5−	9+	9.6	67.2	85
Recent experience	1	3+	17−	14−	14+	1−	2−	9.8	53.3	74
	2	5−	11−	3+	7−	17−	7−	8.6	30.3	64
	3	11+	12−	31+	37+	26−	20+	23.4	132.3	49

[1] Based on summed counts from 10 photographs each day; + = overestimate, − = underestimate.
[2] Based on summing counts from all 50 photographs.

TABLE 3. Effect of numerical magnitude on tendency to over- (+) and underestimate (−) numbers[1] in reinforcement tests.

		Numerical range category					
		40–300		301–750		1000–3100	
Experience level	Observer	+	−	+	−	+	−
Inexperienced	1	11	9	5	8	5	12
	2	9	9	3	9	6	11
	3	6	13	4	9	6	11
Totals		26	31	12	26	17	34
					$\chi^2 = 2.50, P > .05$		
Past experience	1	9	11	8	5	8	9
	2	9	11	7	6	8	7
	3	6	14	6	7	10	6
Totals		24	36	21	18	26	22
					$\chi^2 = 2.83, P > .05$		
Recent experience	1	7	10	8	5	7	9
	2	5	14	5	8	4	11
	3	5	15	7	5	13	3
Totals		17	39	20	18	24	23
					$\chi^2 = 6.34, P < .05$		

[1] Where accuracy was within 1%, the photograph was not used in the table.

Non-reinforcement test.—To compare the accuracy of reinforced vs. non-reinforced estimates, I compared results from the same set of 20 photographs (non-reinforced the first week, reinforced the second) for the 3 experienced (recent) observers (Table 4). There was no recollection by any observers that the photographs had been seen the previous week; therefore, the trials were assumed to be independent. For only one (#1) of the three observers was there a significant difference in accuracy because of reinforcement (Table 4, 3rd line for each observer). One observer (#1) was more accurate due to reinforcement, one showed no significant difference (#3), and one (#2) was actually less accurate. A fixed-effects ANOVA test using observers as one effect verified a significant difference among observers ($F = 4.06$, $df = 2, 56$, $P = .02$).

In non-reinforced tests, as with reinforcement (above), the 3 additional inexperienced observers showed a much stronger tendency to underestimate (59 of 60 photographs) than did the observers with recent experience (39 of 60).

Unlike the above tests with reinforcement, there was a detectable difference between the mean accuracy of the inexperienced and recently experienced observers. Inexperienced observers had mean % errors of 35, 45, 53, whereas experienced observers had 24, 29, and 34% errors (Mann-Whitney $U = 0$, $P = .05$).

Density effects.—The density of birds on the photographs varied sufficiently (6–35 birds/cm^2) to compare density effect on accuracy and

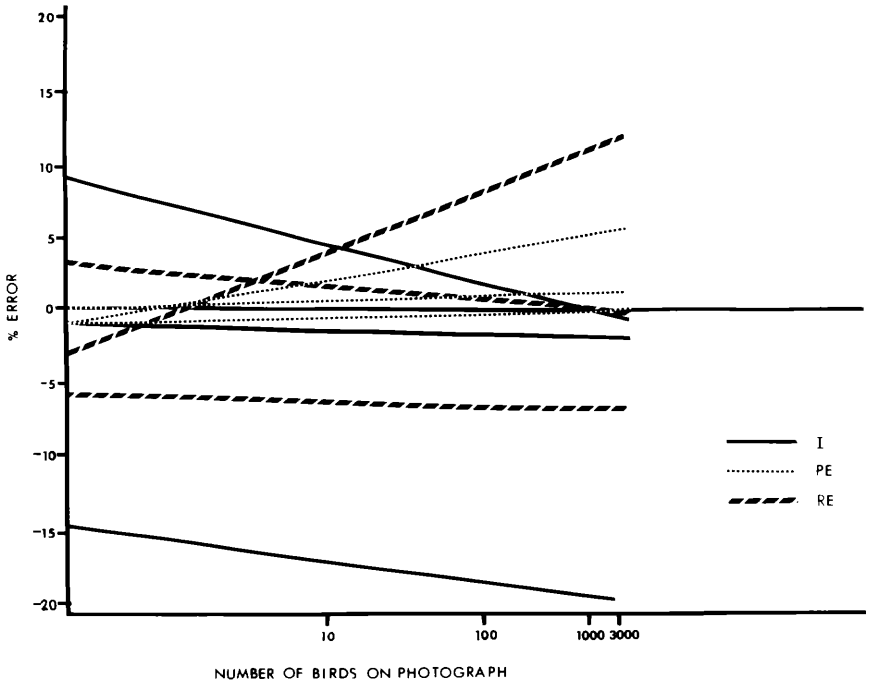

FIGURE 1. The effect of numerical range on estimation accuracy of 3 groups of observers, inexperienced (I), those with past experience (PE), and those with recent experience (RE), based on 50 photographs. Lines are derived from least squares regression for each individual. Numerical range is shown on a log scale.

tendencies to over- or underestimate. Five photographs with sparse density ($\leq$18 birds/cm^2, range 6.7–18.0) and 5 "dense" photos (>23 birds/cm^2, range 23.3–32.7) were used in the size range 400–700. I attempted to minimize other differences between the groups such as background contrast, spatial configuration (circular vs. linear array), etc.

Using the mean % error of all observers for each photograph, no difference was found in accuracy between sparse and dense groups (Mann-Whitney $U = 7$, $P = .155$). There was a tendency, however, for the dense photographs to be underestimated by a greater amount than the sparse photographs. Pairing mean % deviations of the dense vs. sparse photos for each of the 9 observers showed that, in 7 of 9 cases, estimates were lower for dense photos than for sparse (Sign test, $P = .09$), an important (although not statistically significant) difference.

DISCUSSION

As with earlier studies, the results revealed that variation among observers was substantial even within the same experience level (LeResche and Rausch 1974, Caughley et al. 1976, Prater 1979). Although there

TABLE 4. Paired-*t* comparison of reinforced (RE) vs. non-reinforced[1] (NR) estimates of photographic counts (CT) of waterfowl by 3 experienced observers using 20 photographs.

Observer	Variable	CV	T	*P*
1	NR – CT[2]	653	−0.69	0.50
	RE – CT[3]	229	1.95	0.07
	RE – NR[4]	211	−2.12	0.05
2	NR – CT	268	1.67	0.11
	RE – CT	150	2.99	0.01
	RE – NR	1714	0.26	0.80
3	NR – CT	164	−2.72	0.01
	RE – CT	290	−1.54	0.14
	RE – NR	339	−1.32	0.20

[1] Non-reinforced estimates made one week prior to the experiment using reinforcement.

[2] Comparison of non-reinforced estimate with actual count.

[3] Comparison of reinforced estimate with actual count.

[4] Comparison of reinforced estimate with non-reinforced estimate.

was little relationship between experience and accuracy (except when not reinforced), the underestimation tendency was consistent for 6 inexperienced observers. This tendency is believed to be nearly universal once the numerical range exceeds 10 (Kaufman et al. 1949). The relation between actual and estimated numbers can be expressed as: $R = kS^a$, where R = estimated number, S = actual number, k = constant, and *a* may vary between .85 (Krueger 1972) and 1.34 (Stevens 1957) with other studies confirming the .85 range (Indow and Ida 1977). The form of this equation indicates that, beyond the 100–200 range, the slope changes only slightly. However, the psychological experiments from which these relationships were derived were mostly confined to the numerical range 0 to 300.

Results from this experiment run counter to those reported above. Although 5 of the "reinforced" observers showed an overall tendency to underestimate, only 2 did so consistently (as psychological tests would predict) across all numerical ranges (Tables 2 and 3). Two experienced observers strongly over-estimated large numbers but underestimated small ones. McLandress (1979) found that he overestimated the numbers of Ross' Geese (*Anser rossii*) when flock size exceeded 1500 birds. Second, in the range 300 to 750, the response behavior of 5 of the 9 observers appeared to shift (Table 3) indicating that perhaps a threshold of some type was exceeded. This finding is consistent with some empirical data on Snow Geese (*Anser caerulescens*) (H. Lumsden, unpubl. data cited in Ferguson and Kuck 1979) in which flock sizes exceeding 500 seem to introduce significant bias in aerial surveys. Additional data are required in more finely-divided size range categories before any threshold size can be accurately defined. At the lower end of the scale, a perceptual switch has been shown to occur in the n = 6 range (Miller

and Baker 1968). It may be that there are a number of thresholds in the range 500 to 25,000, ranges that biologists may often encounter when conducting waterfowl surveys or seabird inventories.

The relative accuracy in the overall (50 photos) estimate for 8 of the 9 observers (Table 2) should encourage field ornithologists who conduct regional censuses. In many cases, the accuracy in estimating a single aggregation or flock is much less essential than is determining a reliable cumulative estimate.

In addition to numerical range, density appeared to have some effect on the tendency to underestimate. These results are consistent with those reported by psychologists (Horne and Allee 1971, Class 1972).

As an epilogue, I would comment that this experiment, while an improvement in realism over dot tests, still does not simulate the field situation. Nonetheless, the single problem of numerical estimation is far greater for many migratory birds (e.g., open-water wintering waterfowl, roosting blackbirds, shorebirds, and colonially nesting seabirds) than are problems associated with dense habitats and/or small, cryptic animals (Isakov 1963).

SUMMARY

The effects of observer differences, prior experience, training, and numerical magnitude on accuracy in estimating numbers of birds from photographs were examined. Groups of 10 vertical photographs of waterfowl were shown on 5 consecutive days to 3 observers in each of 3 experience groups: inexperienced, those with past experience, and those with recent experience. Results from reinforcement tests showed that, because of marked individual differences, the effects of experience level and training on estimation accuracy were not statistically significant. Without reinforcement, however, experienced observers were more accurate than inexperienced observers. The most apparent pattern was for inexperienced observers ($n = 6$) to underestimate across all numerical ranges, but most strongly when $N > 1000$. Observers with recent experience ($n = 3$) only underestimated when numbers were small (<300). Despite large errors made on individual photographs by all observers, the overall deviations (summed over 50 photos) were very low. Eight of the 9 observers' estimates were within 10% of the total count when reinforcement was given. Density of the birds on the photographs appeared to have a very limited effect on both accuracy and tendency to underestimate. The results are discussed in relation to findings by perceptual psychologists and to applications for bird censusing.

ACKNOWLEDGMENTS

I thank all anonymous observers for their participation and their willingness to humble themselves. Dr. N. Anderson, Psychology Department, University of Maryland, helped me in designing the experiments. P. Geissler, L. Moyer, and K. Williams provided statistical assistance. M. Haramis and J. Goldsberry kindly provided waterfowl photographs. D.

Brown, B. Dowell, S. Gniadek, and K. Hall provided technical support. W. Blandin, P. Geissler, and J. Nichols commented on the manuscript.

LITERATURE CITED

BARR, A., J. GOODNIGHT, J. SALL, W. BLAIR, AND D. CHILKO. 1979. SAS user's guide. Raleigh, North Carolina, SAS Institute, Inc.

CAUGHLEY, G. 1974. Bias in aerial survey. J. Wildl. Manage. 38:921–933.

———, R. SINCLAIR, AND D. SCOTT-KEMMIS. 1976. Experiments in aerial survey. J. Wildl. Manage. 40:290–300.

CLASS, P. 1972. Display density and judgments of number. Perception and Motor Skills 34:531–534.

COHEN, J. 1977. Statistical power analysis for the behavioral science. New York, Academic Press.

ERWIN, R. M. 1979. Coastal waterbird colonies: Cape Elizabeth, Maine to Virginia. Office of Biological Services, U.S. Fish & Wildlife Service FWS/OBS-79/10.

———. 1980. Censusing waterbird colonies: some sampling experiments. Trans. Linn. Soc. New York 9:77–86.

FERGUSON, E., AND T. KUCK. 1979. A photographic survey of Canada Geese (*Branta canadensis*) compared with ocular counts on a portion of the Missouri River, South Dakota. Report of Technical Committee of the Central Flyway Waterfowl Council, U.S. Fish & Wildlife Service.

HORNE, E., AND M. ALLEE. 1971. Estimation as a function of density and contrast. J. Psychol. 78:87–94.

INDOW, T., AND M. IDA. 1977. Scaling of dot numerosity. Perception and Psychophysics 22:265–276.

ISAKOV, V. 1963. Organization and methods of censusing terrestrial vertebrate faunal resources. Israel Program for Scientific Translations, Jerusalem.

JERONS, W. 1871. The power of numerical discrimination. Nature 3:281–282.

KAUFMAN, E., M. LORD, T. REESE, AND J. VOLKMANN. 1949. The discrimination of visual number. Am. J. Psychol. 62:498–525.

KRUEGER, L. 1972. Perceived numerosity. Perception and Psychophysics 11:5–9.

LECHELT, E., AND G. TANNE. 1976. Visual orientational anisotropy and stimulus surround effects in discrimination of spatial numerosity. Perception and Motor Skills 43:431–438.

LERESCHE, R., AND R. RAUSCH. 1974. Accuracy and precision of aerial moose censusing. J. Wildl. Manage. 38:175–182.

MATTHEWS, G. V. T. 1960. An examination of basic data from wildfowl counts. Proc. Int. Ornithol. Congr. 12:483–491.

MCLANDRESS, M. R. 1979. Status of Ross' Geese in California. Pp. 255–265, *in* Management and biology of Pacific flyway geese. R. Jarvis and J. C. Bartonek (eds.). OSU Book Stores, Inc., Corvallis, Oregon.

MILLER, A., AND R. BAKER. 1968. The effects of shape, size, heterogeneity and instructional set on the judgment of visual number. Am. J. Psychol. 81:83–91.

PRATER, A. 1979. Trends in accuracy of counting birds. Bird Study 26:198–200.

SIEGEL, S. 1956. Non-parametric statistics. McGraw-Hill, New York.

STEEL, R., AND J. TORRIE. 1960. Principles and procedures of statistics. McGraw-Hill, New York.

STEVENS, S. 1957. On the psychophysical law. Psychol. Rev. 64:153–181.

STOTT, R., AND D. OLSON. 1972. An evaluation of waterfowl surveys on the New Hampshire coastline. J. Wildl. Manage. 36:468–477.

Patuxent Wildlife Research Center, Migratory Nongame Bird Section, U.S. Fish and Wildlife Service, Laurel, Maryland 20708. Received 27 May 81; accepted 4 Jan. 1982.