Accuracy of Pup Classifications and Its Effect on Population Estimates in the Hooded Seal (*Cystophora cristata*)

G. B. Stenson and R. A. Myers

Department of Fisheries and Oceans, Science Branch, P.O. Box 5667, St. John's, Nfld. A1C 5X1

Stenson, G. B., and R. A. Myers. 1988. Accuracy of pup classifications and its effect on population estimates in the hooded seal (*Cystophora cristata*). Can. J. Fish. Aquat. Sci. 45: 715–719.

Information from the classification of age-specific developmental stages has been used to adjust aerial survey estimates of pup production in a number of species of seals, including the hooded seal (*Cystophora cristata*). We test the assumption that hooded seal pups were accurately and consistently classified according to developmental stage and examine the consequences of misclassifications upon adjusted population estimates. We determined overall misclassification rates, the effect of survey height on classifications, and interobserver variability. At ice level, misclassifications rates were low (<3%). From an altitude of 30 m, newborn pups could not be classified correctly and misclassification rates for the two other stages of attended pups varied between 6.4 and 21.3%. There was no evidence of an overall bias in classifications or differences among observers although there was a significant interaction between day and stage. Individual pups appear to have been misclassified independently by each observer. Under actual survey conditions, observers classified a similar proportion of pups into each recognizable stage. The misclassification rates we observed did not significantly alter the previous population estimate. Methods for improving the current survey design include modifying classification criteria, providing observers with a period of on-ice training, and reducing the width of survey transects.

L'information obtenue de la classification des stades de développement selon l'âge a été utilisée pour l'ajustement des estimations par levé aérien de la production de nouveau-nés chez un certain nombre d'espèces de phoques, y compris le phoque à capuchon (*Cystophora cristata*). Les auteurs vérifient l'hypothèse que les nouveau-nés du phoque à capuchon ont été précisément et uniformément classés selon le stade de développement et examinent les conséquences d'une classification erronée sur les estimations ajustées de la population. Ils ont ainsi déterminé les taux globaux de classification erronée, l'incidence de l'altitude des levés sur les classifications et la variabilité entre observateurs. Au niveau de la glace, les taux de classification erronée étaient faibles (<3 %). D'une altitude de 30 m, les nouveau-nés ne pouvaient être classés correctement; les taux de classification erronée des deux autres stades de nouveau-nés accompagnés de leur mère ont varié de 6,4 à 21,3 %. Il n'y avait aucun signe de biais global des classifications ou aucune différence entre les observateurs, bien qu'il y avait une interaction significative entre le jour et le stade. Chaque nouveau-né semble avoir été indépendamment mal classé par chaque observateur. Selon les conditions réelles de levés, les observateurs ont classé une proportion semblable de nouveau-nés selon chaque stade identifiable. Les taux de classification erronée observés n'ont pas nettement modifié l'estimation précédente de la population. Les méthodes d'amélioration de la conception des levés en cours comprennent la modification des critères de classification, une période de formation des observateurs sur la glace et la réduction de la largeur des transects de levés.

Received March 25, 1987
Accepted January 7, 1988
(J9203)

Reçu le 25 mars 1987
Accepté le 7 janvier 1988

Aerial survey techniques are commonly used to estimate pup production in a number of pinnipeds (e.g. Lavigne 1976; Radford et al. 1978; Hay et al. 1985; Bowen et al. 1987). However, determining total pup production is often difficult because in many species, individual pups may leave the whelping patch before all of the births are completed. Therefore, during any single survey, some pups may not be present because they have yet to be born or they have already left the whelping concentration. One method to correct a single survey estimate of pup production is to use a model which estimates the temporal distribution of births in the population. Such a model uses information on the duration of distinct age-specific developmental stages and the changes which occur in the proportion of pups in each stage over the entire period of births. Such models have been used successfully to correct aerial survey estimates of grey seals (*Halichoerus grypys*) (Radford et al. 1978), harp seals (*Phoca groenlandica*) (Myers and Bowen 1986), and hooded seals (*Cystophora cristata*) (Bowen et al. 1987) and are applicable to any species where only part of the population is present to be surveyed at any one time.

In the hooded seal, it is important to correct any estimate of pup production for the distribution of births. In the northwest Atlantic, births may occur on the pack ice from mid- to late March (Kovacs and Lavigne 1986). Individual pups, however, are present for only a short time and may leave the ice after weaning, as early as 4 d after birth (Bowen et al. 1985). Correcting for the temporal distribution of births, Bowen et al. (1987) increased their initial aerial survey estimate by more than

TABLE 1. Error rates of individual observers classifying hooded seal pups from an altitude of 30 m.

Date	Stage	N	Proportion of pups misclassified		
			Obs. A	Obs. B	Obs. C
March 16	2	12	0.083	0.167	0.083
	3	12	0.25	0.25	0.083
March 17	1	2	1.0	0.50	1.0
	2	40	0.05	0.20	0.05
	3	5	0.20	0.20	0.20
March 18	2	23	0.304	0.261	0.261
	3	14	0.071	0.071	0
Total	1	2	1.0	0.50	1.0
	2	75	0.133	0.213	0.120
	3	31	0.161	0.161	0.064

TABLE 2. Number of stage 2 and stage 3 hooded seal pups correctly classified by each observer from an altitude of 30 m.

Date	Stage	Correctly classified by								Total
		ABC	AB	AC	BC	A	B	C	None	
March 16	2	9	1	1				1		12
	3	8		1	1			1	1	12
March 17	2	27	2	8	3					40
	3	3		1			1			5
March 18	2	9	2	4	4	1	2		1	23
	3	12		1	1					14
Total	2	45	5	13	7	1	2	1	1	75
	3	23		3	2		1	1	1	31

50%. Therefore, considering the importance of this correction factor to the estimate of total pup production, it is important that the assumptions of the model used to estimate the temporal distributions of births be tested. One assumption of the model used by Bowen et al. (1987) is that all pups were correctly classified according to developmental stage. This involved classifying pups from a moving helicopter at a height of 30 m (Bowen et al. 1987). The purpose of this paper is to test this assumption experimentally by determining misclassification rates, the effect of survey altitude, and the interobserver variability in stage classification. We also determine if the current model is robust to realistic levels of misclassification.

Methods

Two experiments were conducted in the Gulf of St. Lawrence in March 1986 to examine errors in classifying hooded seal pups. Pups were classified by three observers: two (B and C) with previous field experience with hooded seals and one (A) without. Each observer examined photographs and slides of pups in the laboratory to become familiar with the developmental stages defined by Bowen et al. (1987):

Stage 1: Newborn — skin in loose folds along flanks, fur saturated to wet, entire pelage yellowish, awkward body movements, mother present.

Stage 2: Thin blueback — ventrum white, neck well defined, trunk conical, mother present.

Stage 3: Fat blueback — ventrum white, neck not clearly distinguishable, trunk fusiform, mother present.

Stage 4: Solitary blueback — as in fat blueback but mother not present.

On March 16, a small concentration of whelping hooded seals was located at 46°40′N, 62°12′W, approximately 40 M (1 M = 1.852 km) south of the Magdalen Islands. Aerial observations were conducted from a Bell 206B helicopter in a manner similar to that used by Bowen et al. (1987). Flying altitude was 30 m and flying speed was approximately 65 km/h. Tape marks were placed on each side window to delimit a 100-m strip at an altitude of 30 m.

Before beginning the experiments, a series of pups was examined and the classification criteria used by Bowen et al. (1987) were found to be inappropriate for use from the air. The criteria used in this study were modified slightly; stages 1 and 4 remained the same, but stages 2 and 3 were differentiated solely by the presence or absence of a visible neck.

Results

Error Rates and the Effect of Survey Height

To estimate the overall error rates and the effect of survey height, individual attended pups (stages 1–3) were classified from an altitude of 30 m and again near ice level. Pups were chosen by the pilot who presented the pup to each observer so that it was visible within the marked strip width. After each observer independently classified the pup, it was then observed from the helicopter while hovering just above ice level (1–3 m) and classified a second time. Only after these two classifications were completed was the developmental stage of the pup discussed between the observers and a consensus ("true stage") reached. To obtain a variety of pups in each developmental stage, this experiment was conducted over a 3-d period from March 16 to 18. Unattended pups were not included in this experiment because the absence of an adult female was an unambiguous criterion for stage 4 pups.

Each observer examined 108 pups. Although the percentage of pups in each of the three stages varied by day, in total, 1.8% were in stage 1, 69.4% in stage 2, and 28.7% in stage 3 (Table 1).

The observers agreed on the classifications of nearly all of the pups which were examined at ice level. The two stage 1 pups encountered were classified correctly and error rates for stage 2 and stage 3 pups were 2.6 and 1.1%, respectively. Most of the stage 2 and stage 3 pups misclassified were considered to be transitional between the two stages.

From an altitude of 30 m, stage 1 pups (newborns) were correctly classified only once (Table 1). Because this stage is difficult to correctly classify and was estimated by Bowen et al. (1987) to last for a very short time (approximately 3 h), we recommend that the newborn stage not be used when classifying pups from the air. It can, however, be used if classifications are made at ice level.

Misclassification rates for stage 2 and stage 3 pups were higher at 30 m than those made at ice level. Overall misclassification rates for stage 2 pups ranged between 12.0 and 21.3%, while misclassification rates for stage 3 pups ranged between 6.4 and 16.1% (Table 1). However, the most striking pattern that emerges from an examination of Table 1 is that the misclassification rates for stage 2 pups appear to be less than the misclassification rates of stage 3 for March 16 and 17, but the reverse occurs for March 18. The numbers of stage 2 and 3 pups misclassified were modeled by a logit transformation of the data (McCullagh and Nelder 1983) and estimating a model with stage, observer, day, and interaction effects using maximum likelihood methods. There was no significant effect of stage ($P = 0.78$), indicating that there is no difference in misclassification rates between the two stages. Similarly, there was no significant effect of day ($P = 0.97$) or observer ($P = 0.10$). However, there was a significant interaction between day and stage ($P = 0.02$). Thus, although there was no bias between stage 2 and 3 misclassifications in the sense that there was no stage effect, there was a significant interaction between day and stage due to the reversal in misclassification rates on March 18.

The reason for the increase in stage 2 misclassifications (and decrease in stage 3 misclassifications) on March 18 is puzzling. The increase in stage 2 misclassifications on March 18 may be due to an increase in the proportion of pups which were transitional between the two stages and thus difficult to correctly classify. However, if transitional pups account for a large proportion of those misclassified, then we would expect that certain individual pups would tend to be misclassified more frequently. Thus, there should be a greater proportion of pups that were misclassified by two or more observers than would be expected if misclassifications were independent among observers. The hypothesis was checked by fitting a Poisson model to the data in Table 2 in which days and the two stages were combined (see Bishop et al. 1980, p. 316). The results are as follows:

	Number of observers misclassified			
	0	1	2	3
Observed frequency of seals	68	30	6	2
Poisson expected frequency	67.2	30.6	7.0	1.2

This results in a χ^2 of 0.62, indicating a good fit of the Poisson model. We cannot conclude from this analysis that the tendency of observers to misclassify the same pup is of any importance. (Note that the expected frequency for the last stage is actually for three or more observers.) The analysis was repeated separating the data by stage and day of observation with similar results. Therefore, there is no evidence that the misclassification rates were caused by individual transitional pups being misclassified.

Although stage 4 pups were considered to be readily classified by the absence of an attending female (Bowen et al. 1987), errors in identifying this stage could occur in areas with high densities of seals if a wandering solitary pup was mistaken for the attending stage 3 pup of a nearby adult. Because twinning has not been reported (Kovacs and Lavigne 1986), two pups associated with a single female would not result in an appreciable error; one pup would be considered solitary and the other classified as the attending pup. However, if the solitary pup was in close proximity to an unattended female (i.e. before birth or after weaning), it could result in an increased estimate of the number of stage 3 pups. A similar error could occur if the solitary pup was associated with a male that was mistaken for being female. Because hooded seals are sexually dimorphic, this potential error can be reduced by conducting the surveys at a speed slow enough to allow for the verification of the sex of the attending adults.

TABLE 3. Interobserver variability: numbers of pups classified and percentage of identifiable pups in each developmental stage during 10 transects.

	No. of pups			Stage (%)		
Observer	Total	Unidentified	Identified	2	3	4
A	116	11	105	82.8	16.2	1.0
B	128	32	96	78.1	19.8	2.1
C	127	24	103	78.6	18.4	2.9

Interobserver Variability

A second experiment was conducted on March 17 to determine the variability in pup classifications among observers under actual survey conditions. Using a Loran C navigational system, 10 east–west transects were flown across the entire whelping patch at an altitude of 30 m. All pups observed within the marked strip (100 m) were classified into one of the four stages. Pups that could not be classified were counted as unknown. Each transect was then flown in the opposite direction so that all observers examined the same study area. Observers A and B examined the transect on the same pass and observer C viewed it on the other.

The total number of pups in each developmental stage observed by each observer on the 10 survey transects was similar (Table 3). The greatest variability was in the number of pups which could not be classified (unknown). The classifications of two experienced observers (B and C) were similar, while the inexperienced observer (A) tended to classify more pups as stage 2 and fewer as unknowns.

Only the proportion of pups in a recognizable stage are considered when corrections are made to survey results (Bowen et al. 1987). If the unknown pups are ignored, the three observers classified a similar proportion of pups into each recognizable stage (Table 3).

The consistency of the classifications may be increased by improving the method used to delimit the viewing area and by narrowing the strip. The use of a single set of tape marks to delimit the viewing area resulted in significant changes in the effective strip width when the observer changed his sitting position slightly. This difference in viewing area may account for some of the difference between observers A and B who were examining the field on the same pass. A narrower strip would eliminate the pups which were far away from the observer and difficult to classify owing to the low angle of view. It would also decrease the maximum distance between the animal and the observer, thereby increasing the sightability (Caughley 1974; Caughley et al. 1976; Estes and Gilbert 1978). Finally, a narrower strip would give more time to view individual pups, which increases the accuracy of aerial surveys (Caughley 1974; Wartzok and Ray 1975). Having more time to classify pups may be particularly important in areas such as northeastern Newfoundland (the "Front"), where pup densities are higher than those in the Gulf.

Although reducing the width of the viewing strip would result in fewer pups being classified per transect, it is unlikely to significantly reduce the accuracy of the estimate of stage composition of the whelping patch. This is because hooded seal pups of similar stage tend to be associated (Bowen et al. 1987), and thus the major source of variability in the estimate is due to differences among geographical regions within the whelping patch. Therefore, the accuracy of the stage composition estimate is only slightly improved if a large number of pups are classified on individual transects, providing the entire patch is properly sampled to determine the extent of geographical differences.

Discussion

Accurate classifications of developmental stages are possible at ice level. By descending upon a random series of pups without actually landing, accurate classifications of a number of pups can be made reasonably quickly. On-ice classifications have been found to be the most effective method of obtaining stage composition information for harp seals (Myers and Bowen 1986). However, classifications from a greater altitude provide some advantages over those at ice level: a larger sample size can be obtained; systematic transects can be flown and data collected for an estimate of population abundance; and the mother–pup pairs are not disturbed. Therefore, aerial transects are preferable if misclassification rates are acceptable.

Under the favorable conditions of experiment 1, stage 2 and 3 pups can be identified from an altitude of 30 m with reasonable error rates. However, the results of these experiments suggest that stage 1 should not be used as a category. Also, it is important that classification criteria be easily recognized. The criteria Bowen et al. (1987) used to distinguish stage 2 and stage 3 pups were reduced to the presence or absence of a visible neck. This criterion was chosen in an attempt to make the classification more consistent among observers and was found to work reasonably well. We suggest that this criterion be used in future surveys.

The initial pup classifications made before the experiments began were not consistent among the observers, and consequently, a period of training was needed to standardize the observers before the experiments began. It was necessary to view at least two dozen pups of various developmental stages from the air and at ice level before reasonably consistent classifications could be made. After training, all three observers had similar error rates (experiment 1), indicating that even inexperienced observers can learn quickly and effectively with the proper period of training. Before the training period, the variability in classifications was similar for observers with or without previous experience with hooded seals; experienced observers without current practice were no more accurate than inexperienced observers. A similar problem was recognized by LeResche and Rausch (1974) who found that experienced observers with current experience identified significantly more moose during an aerial survey than experienced observers without current experience. Therefore, before future surveys are undertaken, both experienced and inexperienced observers should have a training period during which pups are examined under field conditions.

Effect of Misclassifications on Estimates of Population Size

There are two effects of pup misclassification that are of concern: bias and variance. There was little bias in the estimates of proportions of pups in stages 2 and 3; overall, this bias was not statistically significant and varied among the observers. Nevertheless, this bias should be considered in future estimates of pup production. The second effect of misclassification is to increase the variance in the estimates of the proportion of pups in each stage. If errors occur in classifying stage 2 and 3 pups only, an observer who mistakes stage 2's and stage 3's at an equal rate of 20% will estimate that a population consisting only of stage 2 pups will contain 20% of stage 3 pups. Thus, if the two stages are not present in equal proportions, this misclassification will result in a biased estimate of the proportions of pups present. The effects of variance are shown in Table 4. Observer B recorded a significantly larger ($P < 0.001$) number of stage 2 pups than were actually present on March 17. Although the error rates for stage 2 and 3 pups were identical for observer B on March 17 (Table 1), the disproportionately large number of stage 2 pups present resulted in this effect. This effect can be corrected for if the misclassification rates of each observer are known. Thus, it is important to perform experiments similar to this prior to future surveys to determine error rates for each observer. However, if misclassification rates for each observer are determined prior to the beginning of the experiment, they may not be valid for all days, since error rates were observed to vary from day to day (Table 1).

One method to modify a model which corrects pup production based on the temporal distribution of births to account for errors in classification is to incorporate errors in stage determination in the manner that Fournier and Archibald (1982) suggested for age misclassification. However, this method assumes that misclassifications occur independently of how long a pup has been in a stage (i.e. are independent of the age of the pup). If pups that are near a stage transition are classified incorrectly more often than other pups, this technique will not be applicable.

A more useful method is to modify estimates of the probabilities of pup stage transition as a function of age (these transition probabilities allow the information on stage composition to be used to estimate the distribution of births over time). However, to do this, we would need observations from the air of known-age pups.

We tested the sensitivity of the conclusions in Bowen et al. (1987) by modifying the relative durations of stages 2 and 3. In the model described in Bowen et al. (1987) the stages were of stochastic duration where the sojourn time in each stage was described by a gamma distribution. For the sensitivity analysis, we reduced the mean duration of stage 2 by 10% and increased the mean duration of stage 3 by 10%. We held the shape param-

TABLE 4. Numbers of stage 2 and stage 3 pups classified by each observer from an altitude of 30 m. *Significantly different from true ($P < 0.05$); ***significantly different from true ($P < 0.001$).

	March 16		March 17		March 18		Total	
Observer	Stage 2	Stage 3	Stage 2	Stage 3	Stage 2	Stage 3	Stage 2	Stage 3
A	14	10	39	6	17	20	70	36
B	13	11	33	12***	18	19	64	42*
C	12	12	39	6	17	20	68	38
True	12	12	40	5	23	14	75	31

eter of the gamma distributions constant. The result was only a 1% increase in the estimate of the total number of pups born off southeastern Labrador in 1984. Similar results were obtained if the mean duration of stage 2 was increased and the mean duration of stage 3 decreased. We also tested the sensitivity of the results to misclassifying stage 1 pups. The data described in Bowen et al. (1987) were reanalyzed using a model that eliminated the newborn stage (stage 1); the observed stage 1 pups were reclassified as stage 2. There was less than a 1% change in the estimates of pup production in this case because of the small numbers of stage 1 pups observed in 1984.

If the error rates we identified are similar to those which occurred during the survey of Bowen et al. (1987), these changes have only a small effect on their estimates of hooded seal pup production. This is because the time pups spend in each stage was assumed to be stochastic. Furthermore, the stochastic duration pups spend in each stage was estimated from data in which the pup age was estimated with some error. This effect will mitigate the greater rate of misclassifying pups from helicopter surveys.

Acknowledgments

We thank D. Wakeham and W. Hoek for their assistance in determining the classifications and their help in conducting the experiments. We also thank our pilot Kim Miller for his skill and cooperation. D. Bowen, M. Kingsley, P. Pepin, and an anonymous reviewer provided valuable suggestions to improve this paper.

References

BISHOP, Y. M. M., S. E. FIENBERG, AND P. W. HOLLAND. 1980. Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, MA. 557 p.

BOWEN, W. D., R. A. MYERS, AND K. HAY. 1987. Abundance estimation of a dispersed, dynamic population: hooded seals (*Cystophora cristata*) in the Northwest Atlantic. Can. J. Fish. Aquat. Sci. 44: 282–295.

BOWEN, W. D., O. T. OFTEDAL, AND D. J. BONESS. 1985. Birth to weaning in four days: remarkable growth in the hooded seal, *Cystophora cristata*). Can. J. Zool. 63: 2841–2846.

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

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

ESTES, J. A., AND J. R. GILBERT. 1978. Evaluation of an aerial survey of Pacific walruses (*Odobenus rosmarus divergens*). J. Fish. Res. Board Can. 35: 1130–1140.

FOURNIER, D., AND C. ARCHIBALD. 1982. A general theory of analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39: 1195–1207.

HAY, K., G. B. STENSON, D. WAKEHAM, AND R. A. MYERS. 1985. Estimation of pup production of hooded seals (*Cystophora cristata*) at Newfoundland during March 1985. Can. Atl. Fish. Sci. Adv. Comm. Res. Doc. 85/96: 10 p.

KOVACS, K. M., AND D. M. LAVIGNE. 1986. *Cystophora cristata*. Mammalian Species No. 258. The American Society of Mammalogists, Pittsburgh, PA. 9 p.

LAVIGNE, D. M. 1976. Counting harp seals with ultra-violet photography. Polar Rec. 18: 269–277.

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

MCCULLAGH, P., AND J. A. NELDER. 1983. Generalized linear models. Chapman and Hall, London. 26 p.

MYERS, R. A., AND W. D. BOWEN. 1986. Analyses of the 1983 aerial survey of harp seals at the Front: corrections for the birthing curve and counting error. Can. Atl. Fish. Sci. Adv. Comm. Res. Doc. 86/7: 25 p.

RADFORD, P. J., C. F. SUMMERS, AND K. M. YOUNG. 1978. A statistical procedure for estimating grey seal pup production from a single census. Mammal Rev. 8: 35–42.

WARTZOK, D., AND G. C. RAY. 1975. A comparison of the accuracy and precision of visual and photographic aerial censuses of ice-inhabiting marine mammals: results of a simulated flight. Proceedings of the Workshop on Remote Sensing of Wildlife, Quebec, Canada, 17–20 November, 1975. 20 p.