

A re-assessment of data on waterfowl behaviour in response to jet overflights at  
Snegamook Lake, Labrador

John W. Chardine  
Research Scientist, Marine Ecosystems  
Canadian Wildlife Service  
Environment Canada, Atlantic Region  
P.O. Box 6227  
Sackville, NB E4L 1G6

Revised January 2002

## Introduction

Bateman et al. (1999) conducted a study of the effects of jet overflights on the behaviour of Black Ducks (*Anas rubripes*) and Canada Geese (*Branta canadensis*) at Snegamook Lake, Labrador, in the summers of 1995-1997. They instantaneously recorded the behaviour of groups of birds within view of two blinds once every fifteen minutes during observation periods, and reported results as the percent of birds observed in each of five activities: in-flight, swimming, feeding, loafing, preening. Observation periods were categorized as "control" or "experimental" based on jet overflight activity during the 15 minute period. "Control" observations occurred before an overflight and "experimental" observations occurred up to one hour after. ANOVAs within year, period of year and time of day revealed no statistically significant differences in the percent of birds in each activity, before compared to up to one hour after overflights.

In 2000, a review of the above report was commissioned by the (Labrador) Institute for Environmental Monitoring and Research (contract no. 2423) and conducted by Société Duvetnor Ltée, Rivière du Loup, Québec. The review made several suggestions for improvement of the data analysis, the main one being a call for a comparison of paired observation periods before and after each overflight, which would function to eliminate behavioural variation in the data due to time of day, time of year, and year, and focus on the effects of overflights themselves. This report outlines the results of a paired analysis I conducted on the same data set reported in Bateman et al. (1999).

## Methods

In all I identified only 13 occasions for Black Ducks and 16 occasions for Canada Geese in the Bateman et al dataset, when behaviour was recorded both immediately before and after overflights at Snegamook Lake in 1995-97 (only observations of groups of 10 birds or more at a time were used). I assumed that the timing of overflights at

Snegamook Lake was random in relation to the timing of behavioural observations, and that observations were completed soon after each 15 minute period started (confirmed by M. Bateman and A. Hicks). On average, overflights would have occurred halfway through a particular 15 minute period between observations. Therefore, the observation immediately prior to the overflight would have occurred on average 7.5 minutes before the overflight with a range of 0-15 minutes. The observation previous to this would have occurred on average  $15+7.5 = 22.5$  minutes before the overflight (range 15-30 minutes before). These observations are referred to as "Before-1" (22.5 minutes before) and "Before-2" (7.5 minutes before) and both should be of behaviour unaffected by overflight. The observation immediately after the overflight would have occurred an average of 7.5 minutes after (range 0-15 minutes), and is referred to as the "After" observation.

For each species and activity I made Before1-After and Before2-After paired comparisons by subtracting the percent of birds engaged in the activity after the overflight from before, and then taking the mean difference (referred to as empirical mean difference). Under the null hypothesis of no effect of overflights, the mean difference for each activity should not differ significantly from zero. The standard appropriate test here is a paired t-test, however, in order to avoid problems associated with the small sample sizes in this study, and uncertain shape of the error distribution, I decided to test the null hypothesis using a resampling or Monte Carlo simulation (Simon 1997). Here, I took the set of paired percents in each of the two before-after comparisons, randomly re-assorted the data within each pair, and calculated a new mean difference. This was repeated in 10,000 simulations and resulted in a set of 10,000 means generated under the null hypothesis of no difference. The probability of obtaining the empirical mean difference under the null hypothesis was then easy to calculate as the proportion of random mean differences that were greater than the absolute value of the empirical mean or less than the negative absolute value (2-tailed test). The method is illustrated in Appendix 1.

I conducted a retrospective power analysis on the Before2-After data presented here using the power calculator available at:

<http://www.health.ucalgary.ca/~rollin/stats/ssize/n1.html>.

The power of Before1-After comparisons was not checked because sample sizes were a little higher for Black Ducks and so analyses would have been less conservative. The web-based calculator computes statistical power given sample size or vice versa, for the test of a mean compared to an hypothesised value. This is analogous to a paired t-test where the mean difference between paired values is compared to the value zero, under the null hypothesis. For this power analysis to be relevant here, the p values obtained using the randomisation technique described above would have to agree with those found using a traditional paired t-test. This was checked (Appendix 2) and close agreement was found between the p-values obtained using the two approaches.

Retrospective power analysis of the observed effect sizes (in this case the mean differences in the percent of birds observed in each behaviour, before and after the overflight) is not recommended because it does not calculate power correctly, and no new information is provided beyond the original analysis and p values (Thomas 1997, Hayes and Steidl 1997). Hayes and Steidl (1997) point out that estimates of retrospective power are meaningful when conducted on specific alternate hypotheses. That is to say, power can be retrospectively calculated on hypothesised effect sizes (but not on the empirical effect size). I used the above web-based calculator to determine power for various effect sizes from 1 to 30%, given an alpha level of 0.05, and the empirical standard deviation and sample size for each Before2-After comparison. I also calculated the sample size required to provide an acceptable power of 80%, given the empirical standard deviation and alpha level of 0.05.

## Results

### 1. Effect of overflight on waterfowl behaviour

The results of the analysis are presented in Table 1. The table shows the mean and standard deviation of the difference in the percent of birds engaged in each activity for the two comparisons: Before1-After and Before2-After. P-values for each mean are also shown as well as the 95% confidence intervals for the mean. All the mean differences were small. Of the 20 mean differences calculated (2 species x 5 activities x 2 comparisons), 45% were within 1% of zero, and all but one (95%) were within 5% of zero. On average the percent of birds engaged in each of the activities changed little before and after overflight. The probabilities for obtaining the observed mean differences under the null hypothesis were all relatively large and none was statistically significant (i.e.,  $< 0.05$ ).

Although all mean differences were relatively small, the 95% confidence intervals for the majority of means were large (Table 1). All confidence intervals encompassed zero but the intervals sometimes ranged as high as over 20% above and below the mean.

### 2. Power analysis

Figure 1 shows the results of the power analysis. The power for in-flight activity in Canada Geese was not calculated due to zero variation in the mean difference. For feeding, loafing and swimming activities, the power to detect a 10% mean difference ranged from about 40 to 90% for Black Ducks and about 29-40% for Canada Geese. Adequate power is considered to be 80% or greater and this was only achieved for swimming in Black Ducks. For in-flight and preening activities, the power to detect a difference remained above 80% for all but the smallest effect sizes.

Figure 2 shows the results of a "reverse" power analysis in that the sample size required to detect a given effect size with a power level of 80% was calculated. The

Table 1. Mean difference in the percent of birds engaged in each activity. Comparisons are shown between observations made before and after a jet overflight.

Activity	Species <sup>2</sup>	Mean difference in percent (SD) <sup>1</sup> 95% CI of mean (lower, upper)	
		Before1 <sup>3</sup> -After	Before2-After
In flight	ABDU	-1.5 (5.3) -4.7, 1.7 p = 0.50 <sup>4</sup>	1.4 (4.1) -1.5, 4.3 p = 0.25
	CAGO	0.0 (0) 0-0 p = 1	0.0 (0) 0-0 p = 1
Swimming	ABDU	3.1 (33.6) -17.2, 23.4 p = 0.77	0.4 (10.6) -7.2, 8.0 p = 0.94
	CAGO	3.4 (38.2) -17.8, 24.6 p = 0.73	0.3 (36.4) -20.8, 21.4 p = 0.94
Feeding	ABDU	-5.2 (36.7) -27.4, 17.0 p = 0.63	-0.3 (18.1) -13.2, 12.6 p = 0.96
	CAGO	1.5 (40.7) -21.0, 24.0 p = 0.89	4.5 (22.8) -8.7, 17.7 p = 0.60
Loafing	ABDU	3.6 (22.1) -9.8, 17.0 p = 0.54	-2.5 (12.8) -11.7, 6.7 p = 0.55
	CAGO	-5.7 (43.1) -29.6, 18.2 p = 0.64	-4.9 (29.2) -21.8, 12.0 p = 0.55
Preening	ABDU	0 (1.2) -0.7, 0.7 p = 1	0.9 (2.3) -0.7, 2.5 p = 0.31
	CAGO	0.8 (3.4) -1.1, 2.7 p = 0.63	0.1 (2.4) -1.3, 1.5 p = 0.87

1. Sample sizes: ABDU Before1-After: n = 13, Before2-After: n = 10; CAGO Before1-After: n = 15, Before2-After: n = 15.
2. ABDU = Black Duck, CAGO = Canada Goose.
3. Before1 observation occurred on average 22.5 minutes before overflight; Before2 observation occurred on average 7.5 minutes before overflight.
4. Two-tailed p-values represent probability of obtaining the observed mean difference under the null hypothesis (no difference Before1 vs. After or Before2 vs. After). P's calculated using resampling techniques (see methods and Appendix 1).

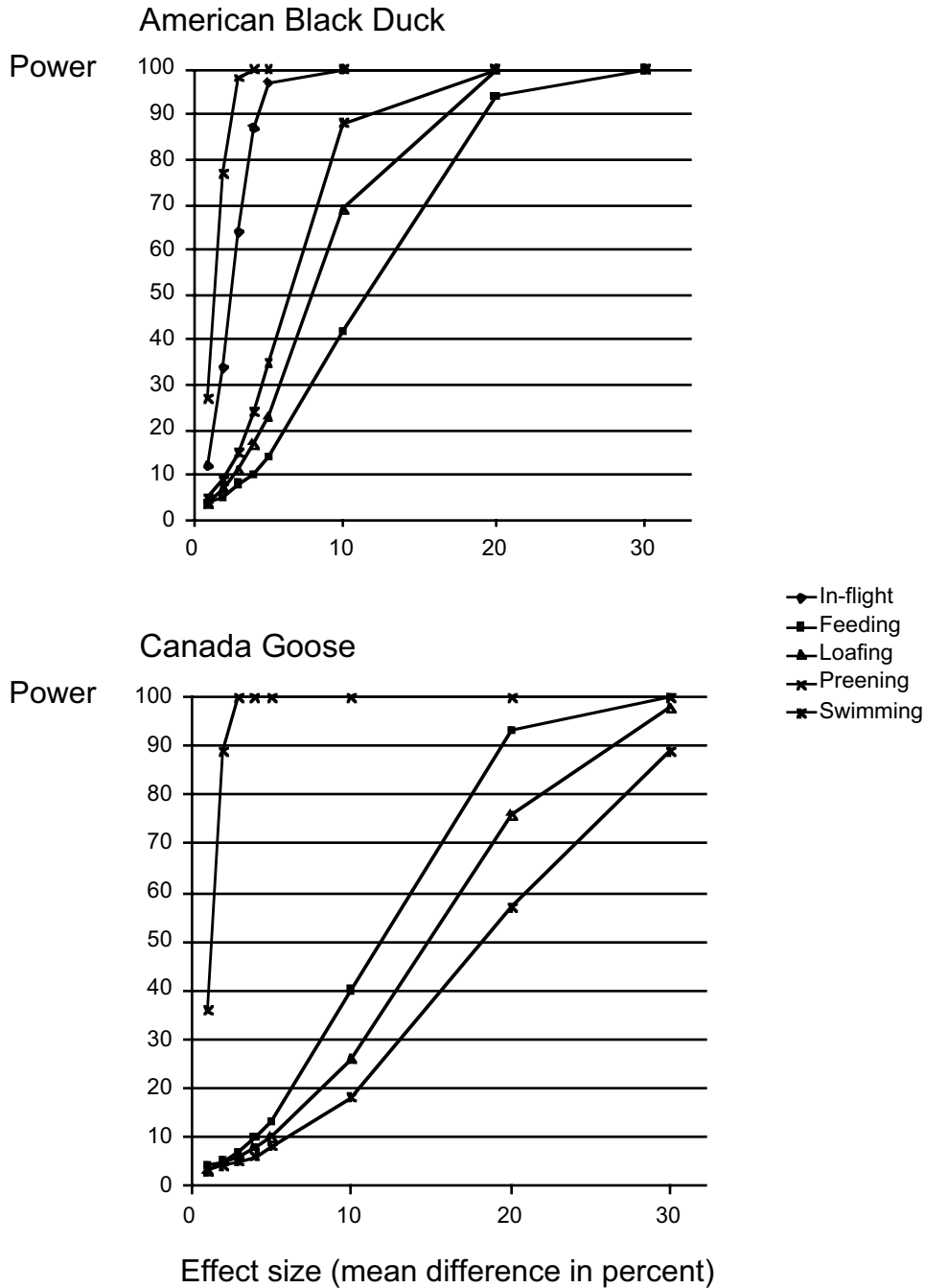


Figure 1. The statistical power obtained from a given effect size, using the empirical standard deviation of the mean difference and sample sizes in the study. Effect size is defined as the mean absolute difference in the percent of birds engaged in each activity, before compared to after the overflight. In-flight activity of Canada Geese was not considered because the variance around the mean difference was zero. Only Before2-After comparisons were considered.

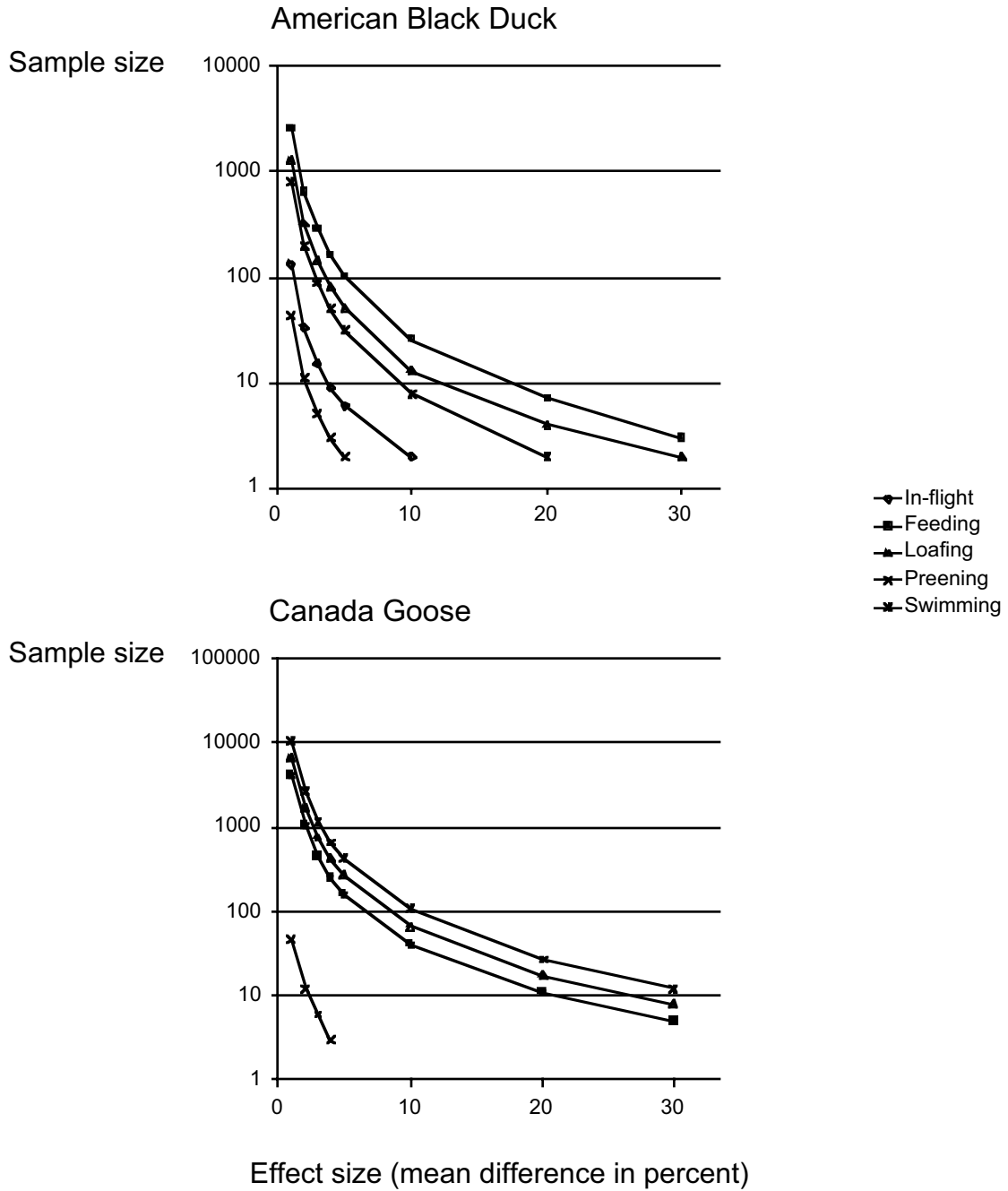


Figure 2. The sample size required to detect a given effect size with statistical power of 80%, using the empirical standard deviation and alpha = 0.05. Effect size is defined as the mean absolute difference in the percent of birds engaged in each activity, before compared to after the overflight. In-flight activity of Canada Geese was not considered because the variance around the mean difference was zero. Only Before2-After comparisons were considered.

results indicate that to detect a 10% effect size, samples of <10 to about 40 data points (before-after pairs) were needed for Black Ducks, and about 50 to >100 for Canada Geese. For small effect sizes of a few percent, large samples of 100s or even 1000s of data points would be required to detect the difference.

## Discussion

The results of this re-assessment of data presented in Bateman et al. (1999) support the conclusions of the report: overflights did not appear to affect the proportion of birds engaged in the behaviours as scored in the study (in flight, swimming, feeding, loafing and preening). All mean differences in behaviour found between paired observations taken before and after overflight were close to zero, which would be expected under the null hypothesis of no effect. The paired analysis of the data presented here eliminated from the comparisons the wide variability in waterfowl behaviour that was found between years, within a season, or within a day (Bateman et al. 1999), and therefore contributed some extra power to the analysis. However, this was offset by the relatively few paired observations that could be identified in the Bateman et al. datasets.

A power analysis of the results of this study indicated that given the variability about the mean differences, and the sample sizes used, only relatively large effect sizes could be detected with an adequate level of power (80%). Assuming that the observed variability was inherent in waterfowl behaviour (albeit lower due to the paired design), the feature of the study design that could have been changed was sample size. The power analysis suggested that for most behaviours a sample size of about 100 would be sufficient to detect differences of about 10% in the mean proportion of birds performing each behaviour, with adequate power. Smaller effect sizes would require larger samples; and these become impossibly large for very small effect sizes.

The methods used to assess waterfowl behaviour in the Bateman et al. study were designed to detect longer-term behavioural effects of jet overflight, and not those that

lasted seconds or a few minutes. The protocol of making an observation every 15 minutes meant that the "After" observations referred to in this study occurred anywhere from 0-15 minutes after overflight. Thus, there was less than a 1/15 (7%) chance of detecting a reaction to overflight that lasted less than 1 minute, 2/15 (13%) chance of detecting a reaction lasting less than 2 minutes *et cetera*. Justification for this approach was not made explicit in the Bateman et al. study but presumably was related to the (reasonable) assumption that shorter-term reactions (e.g., a flight response that may last a few seconds) may have much less significant impacts on the overall well-being of the birds, and thus be of less interest from the standpoint of ecological impacts of jet overflights.

The results of the Bateman et al. study and those reported here do not provide evidence that jet overflight affects the behaviour of Black Ducks or Canada Geese. However, some caution should be used in interpreting the results presented here because the power to detect mean differences of less than 10% was relatively low for most activities measured. This is also evidenced by the wide 95% confidence intervals around the mean differences in the percent of birds engaged in each activity before vs. after overflight. Future studies should ensure that sample sizes are in the order of 100 before-after observations to obtain adequate power to detect effects on waterfowl behaviour.

#### Literature cited

- Bateman, M.C., A.H. Hicks and S.M. Bowes. 1999. Waterfowl behaviour in response to jet overflights at Snegamook Lake, Labrador. Report to Goose Bay Office, National Defence Headquarters, Ottawa, Canada. 139 pp.
- Hayes, J.P. and R.J. Steidl. 1997. Statistical power analysis and amphibian population trends. *Conserv. Biol.* 11:273-275.

Simon, J.L. 1997. Resampling: the new statistics. Resampling Stats Inc., Arlington, VA.

436 pp.

Thomas, L. 1997. Retrospective power analysis. *Conserv. Biol.* 11:276-280.

Appendix 1. Illustration of the resampling technique used in this study

Step 1: Calculate mean difference from raw data:

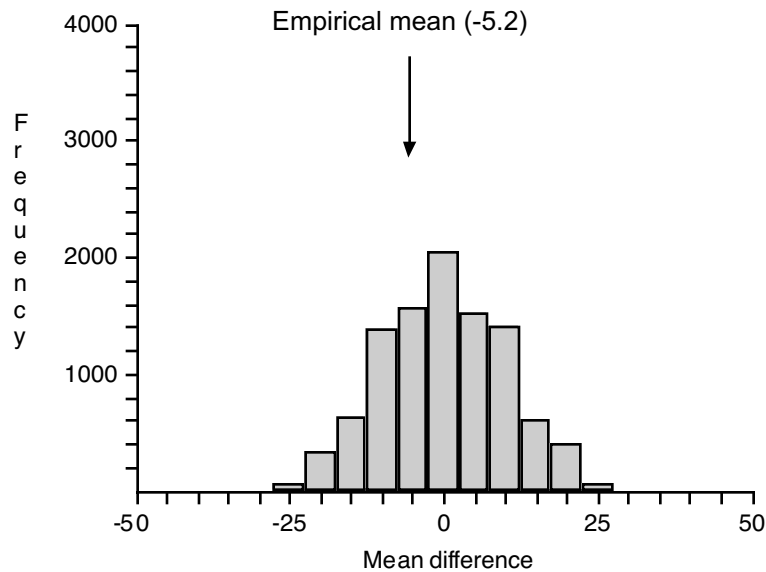
Occasion	Percent of birds feeding		Difference
	Before1	After	
1	60	0	60
2	0	0	0
3	35	56	-21
4	81	83	-2
5	100	53	47
6	76	92	-16
7	6	14	-8
8	30	86	-56
9	85	94	-9
10	80	77	3
11	100	92	8
12	0	81	-81
13	8	0	8
Mean difference			-5.2

Step 2. Randomly re-assort data within pairs to simulate data under the null hypothesis of no difference between Before and After:

Occasion*	Percent of birds feeding		Difference
	Datum 1	Datum 2	
1	60	0	60
2	0	0	0
3*	56	35	21
4	81	83	-2
5*	53	100	-47
6	76	92	-16
7	6	14	-8
8	30	86	-56
9*	94	85	9
10	80	77	3
11	100	92	8
12*	81	0	81
13*	0	8	-8
Mean difference			3.5

\* denotes that data pair was randomly switched

Step 3. Repeat step 2, 10,000 times and generate a distribution of mean differences under the null hypothesis:

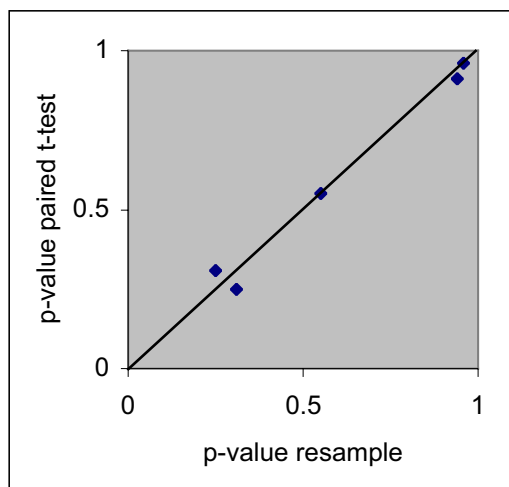


Step 4. Note that the empirical mean difference of -5.2 was not a rare event in relation to the 10,000 random mean differences generated under the null hypothesis. The proportion of random mean differences that were less than the empirical mean or greater than its absolute value (2-tailed test) was 0.63. That is to say, 63 % of the random mean differences were more extreme than the empirical difference. Thus the probability of obtaining the empirical mean difference under the null hypothesis of no difference Before-After was 0.63. The results of a paired t-test of the same data yielded  $p = 0.62$ , so in this case there was close agreement between the two methods.

Appendix 2. Comparison of p-values obtained using the randomisation technique and using a traditional paired t-test. The comparison was made only on Before2-After analyses because power analysis was restricted to these analyses only.

	p value	
	Resampling	paired t-test
<b>Black Duck</b>		
In flight	0.25	0.31
Swimming	0.94	0.91
Feeding	0.96	0.96
Loafing	0.55	0.55
Preening	0.31	0.25
<b>Canada Goose</b>		
In flight	1	not avail.
Swimming	0.94	0.98
Feeding	0.6	0.46
Loafing	0.55	0.53
Preening	0.87	0.83

Black Duck



Canada Goose

