Effects of radio transmitters: Review of recent radio-tracking studies

J.D.Godfrey and D.M.Bryant

Institute of Biological Sciences, University of Stirling, Scotland FK9 4LA.

ABSTRACT

An analysis of 1990s radio-tracking literature found that only 10.4% of 836 studies directly addressed the effect of radio tags on their bearers. Studies on mammals were less likely than those on birds or fish to have the impact of tags tested, yet mammals had the highest proportion of tests indicating a significant tag effect. Conservation studies were least likely to assess effects compared with general science or studies based on exploited species. The lower detection of tag effects in conservation studies can be attributed to the use of tags better designed to avoid adverse effects, and to a publication bias whereby deleterious effects are simply not reported.

Keywords: radio-tracking, radio transmitters, tag effects, conservation studies.

Godfrey, J.D.; Bryant, D.M. 2003: Effects of radio transmitters: Review of recent radio-tracking studies. Pp. 83–95 *in*: Williams, M. (Comp.) 2003: Conservation applications of measuring energy expenditure of New Zealand birds: Assessing habitat quality and costs of carrying radio transmitters *Science for Conservation 214*. 95 p.

1. Introduction

During the course of a study on the effects of radio transmitters on energy expenditure of takahe *Porphyrio mantelli* (Godfrey & Bryant 2003), it became apparent that a review of radio-tracking studies was long overdue. We were interested in how (and under what circumstances) the effects of radio-tags on their bearers had been considered. We confine our review to the last decade, partly because of the ease of obtaining the information, and partly because much of pre-1990 radio-telemetry technology is now outdated, and the larger packages then in use would tend to overestimate the current level of tag-bearing effects. We pay particular attention to the use of radio-telemetry in conservation research.

2. Methods

Altogether 836 relevant studies using tracking devices on animals were identified using the BIDS science citation service (http://www.bids.ac.uk). The following search terms were deployed: telemetry; radio; acoustic; transmitters; tag; and track (plus all derivatives and combinations of the above, e.g. transmitter*s*, radio-track*ing,* etc.).

Using only the abstracts supplied by BIDS, studies were classified according to research goal and animal taxa. Research goals were divided into the following categories: Conservation; Economic; and Science. Divisions into these categories were made on the basis of information given in the abstract. Where the goal was not clear from the abstract, work on animals included on IUCN Red List of Threatened Animals (http://www.wcmc.org.uk/species/animals/animal_redlist.html) and/or CITES-listed species (http://www.wcmc.org.uk/CITES/Index.shtml) was classified as 'Conservation', whereas work on non-listed species was regarded as 'Science.' 'Economic' was used to classify research where populations are widely hunted/fished, and not under threat, and also to cover work related to range management, disease transfer between wild and domestic animals, and crop damage. The 'Science' category was used to include work that matched neither 'Conservation' nor 'Economic' criteria.

Within these categories, the bearers of the radio-tags were divided on taxonomic grounds, using the following taxa: Birds; Mammals; Fish; Other vertebrates; and Invertebrates. Within these diverse taxa, further ecological distinctions were drawn, on the basis of locomotion medium (land, air or water), recognising potential differences of effects tags on swimming sea mammals compared to walking/running terrestrial mammals, and between flightless and volant birds. It was not always clear in which category to place amphibious animals (e.g. frog, otter, and semi-aquatic snake), and where there was doubt these were allocated to water rather than land. Since very few radio-tracking studies of invertebrates and amphibians have been published, for the

purpose of statistical analysis 'Other vertebrates' were combined with 'Invertebrates' to form a new category, 'Other taxa.'

Amongst these groups, studies were scored as one of the following: ignoring effects and drawing untagged population-wide conclusions (IGNORE); ignoring effects, but deploying controlled experiments to draw tagged population-limited conclusions (CONTROL); showing differences between tagged and untagged (TvU+); showing no difference between tagged and untagged (TvU-); showing differences between different types of tag (TvT+); and showing no difference between different types of tag (TvT-). In each of the final four classifications 'differences' refer to any considerations of behaviour, survival or breeding success. Many studies made multiple tests of the effects of tags, some of which suggested no effect and others indicating an effect. If any one of these found an effect, then this was scored accordingly, regardless of other tests in that paper detecting no effect. Where both TvU and TvT were assessed in a single paper the TvU rather than the TvT score was assigned, so that each paper contributed just one score. This practice yields a conservative estimate of the effect of tags, since although within a study TvT+ implies that (for at least one type of tag) TvU+ also exists, whereas TvT- does not imply that TvU- is also true. Since the number of TvU and TvT observations was small, for statistical analysis, TvU+ and TvT+ were combined to give an overall effect category, and TvU- was combined with TvT- to give an overall no-effect category. Where studies could not be confidently categorised on the above basis from the abstract alone, they were discarded, unless the abstract suggested the study might fall into TvU or TvT classifications (directly addressing the costs of tags). In this case the full text was used to determine classification. Methodological studies not seeking to draw specific conclusions were excluded.

We would like to make clear that categorising a work as IGNORE does not imply the study is flawed (although some certainly are). In many cases there were compelling reasons for assuming that the tags in question had no relevant effect on the conclusions drawn about the wider population. Furthermore there will undoubtedly be some instances where the potential effects of tags were explored, but not mentioned in the paper's abstract. Finally there are cases where important questions could only be addressed by radio-tracking, and where no feasible methods for testing the effects of tags were available.

A full list of these studies, together with the categories to which each was assigned is available from one of us (D.M.B.). The statistical power of tests purporting to show TvU- or TvT- was calculated from the original data where authors had provided sufficient information (mean, n, s.d., and statistical test) using nQuery Advisor 2.0 software. Where variances differed between means, the unequal-variance t-test power calculator available at http://www.stat.ucla.edu/calculators was used. In every possible case, the power of the test to assess a statistical difference in the given data, and the power to detect a 10% difference in means (given the observed variance) was calculated. In cases where daily survival rates were given for the nestling period (e.g. Hill et al. 1999), a 10% difference in means was taken as a 10% difference in survival over the whole nestling period, not a 10% difference in daily survival. Tests were regarded as 1-tailed for survival and reproduction data, but 2-tailed for behaviour and energy expenditure data.

Binary logistic regression (Minitab release 12) with Logit function was used to explore the data. This procedure uses an iterative-reweighted least squares algorithm to obtain maximum likelihood estimates of the parameters (McCullagh & Nelder 1992).

3. Results

The majority of papers in the review (83.3%) were classified as IGNORE, effectively making a tacit assumption that radio-tags had no significant impact on their bearers. A further 6.3% were CONTROLS, taking no account of an effect of tags, but testing hypotheses relating to tagged individuals only, or seeking information on another matter using two or more groups of tagged individuals in a controlled experiment. The proportion of studies directly addressing the effect of radio-tags on their bearers was just 10.4%, and differed by research goal and by taxa (Table 1).

Between taxonomic groups, mammals were less likely than birds or fish to have the impact of their tags tested (only 4.5% of 399 studies on mammals, 12.6% of 103 fish studies, 19.0% of 269 bird studies, and 7.7% of 65 studies of other taxa: $\chi^2_3 = 37.032$, $P < 0.001$). Despite being the least likely to be tested, mammals had the highest proportion of tests that indicated a significant impact of tags (Table 2).

Conservation workers appear least likely to assess effects (significantly smaller proportion of tests made $\chi^2_2 = 9.167$, $P < 0.010$: 6.6% of 319 tested in conservation, 11.6% of 302 in science, and 14.4% of 215 in economic work). Furthermore, of studies which did investigate tag effects, conservation subjects were the least likely to indicate a significant effect of tag-bearing (33.3 % of 21 conservation, 62.9% of 35 science and 67.8% of 31 economic: $\chi^2_2 = 6.759$, $P = 0.034$).

Conservation subjects might be less likely to show deleterious effects of tags for three reasons: (1) conservation scientists deploy radio-tags better designed to avoid adverse effects on study subjects; (2) the inherent rarity of the subjects of conservation work imposes small sample sizes on researches, and hence lower power to detect tag-bearing effects; and (3) a publication bias exists.

Hypothesis (1) appears to have some credence, because an ANOVA of the percentage body mass that tags represent (arcsine-square root transformed) by research goal indicates a significant effect ($F_{2,60} = 3.15$, $P < 0.05$); conservation work has the lowest mean percentage body mass (conservation, 2.47%; science, 4.72%; economic, 3.38%). However, when controlling for log-transformed body mass (excluding one outlying study, of sperm whales, to achieve normality) this effect disappears: $F_{2,58} = 1.69$, $P < 0.19$.

Hypothesis (2) is not supported by the available data. Taking the log-transformed highest level of calculated statistical power (since there are often several tests per study) achieved to detect a >10% difference in means, ANOVA

TABLE 1. COUNTS FOR A CLASSIFICATION OF 1990S PAPERS CONSIDERING RADIO-TRACKING.

IGNORE: those not considering the effect of tag-bearing on study subjects. CONTROL: those comparing two or more tagged groups with respect to some other variable. TAG v. NOTAG: studies testing for an effect of tag-bearing. TAG v. TAG: studies testing for differences in tag-bearing effects between two or more types of tags. Counts are divided by research goal: science, economic, and conservation.

	NO. OF STUDIES	IGNORE	CONTROL	TAG V. NOTAG		TAG V. TAG	
				EFFECT	NO EFFECT	EFFECT	NO EFFECT
Science							
Volant bird	82	64	5	5	8*	0	0
Aquatic bird	8	3	0	5	0	0	0
Flightless bird	1	1	0	0	0	0	0
Terrestrial mammal	139	112	17	7	2	1	0
Aquatic mammal	15	14	0	1	0	0	0
Volant mammal	8	8	0	0	0	0	0
Fish	20	16	1	1	2	0	0
Other terrestrial vertebrates	13	12	1	0	0	0	0
Other aquatic vertebrates	14	10	1	2	1	0	0
Invertebrates	2	2	0	0	0	0	0
TOTAL	302	242	25	20	13	2	0
Economic							
Volant bird †	80	51	8	8	5	7	1
Aquatic bird	0	0	0	0	0	0	0
Flightless bird	0	0	0	0	0	0	0
Terrestrial mammal	60	57	3	0	0	0	0
Aquatic mammal	0	0	0	0	0	0	0
Volant mammal	0	0	0	0	0	0	0
Fish	73	57	6	6	4	0	0
Other terrestrial vertebrates	0	0	0	0	0	0	0
Other aquatic vertebrates	0	0	0	0	0	0	0
Invertebrates	1	1	0	0	0	0	0
TOTAL	203	160	16	14	5	7	1
Conservation							
Volant bird	87	74	1	2	7	2	1
Aquatic bird	2	2	0	0	0	0	0
Flightless bird	9	8	1	0	0	0	0
Terrestrial mammal	130	118	9	2	0	0	1
Aquatic mammal	27	24	0	0	1	0	0
Volant mammal	19	18	0	0	1	0	0
Fish	10	10	0	0	0	0	0
Other terrestrial vertebrates	14	13	0	0	1	0	0
Other aquatic vertebrates	20	19	0	1‡	0	0	0
Invertebrates	1	1	0	0	0	0	0
TOTAL	319	287	11	5	12	2	2

* One study (Sedinger et al. 1990) measured DEE in confined (incapable of locomotion) geese.

† One study (Olsen et al. 1992) presents no comparative data to support (in abstract) claim of 'TAG v. NOTAG no effect,' therefore classified as 'IGNORE.'

‡ One study involving a model of an animal only (Watson & Granger 1998).

TABLE 2. CLASSIFICATION OF 1990S PAPERS USING RADIO-TRACKING, CONSIDERING DIFFERENT TAXONOMIC DIVISIONS AND LOCOMOTION MEDIA.

IGNORE indicates studies which did not consider the possible impacts of tags on study animals, CONTROL studies compared between tow or more groups of tagged individuals with respect to another variable, the % tested column refers to those studies which directly investigated the effect of tag-bearing. The final column shows the percentage of these tests of tag-bearing effects that detected a significant impact of tags. All figures are percentages.

	NO. OF STUDIES	IGNORE	CONTROL	TESTED	EFFECT
Taxon					
Volant bird	249	75.90	5.62	18.47	52.2
Aquatic bird	10	50	0	50	100
Flightless bird	10	90	10	0	-
Terrestrial mammal	329	87.23	8.81	3.95	76.9
Aquatic mammal	43	90.70	0	9.30	25.0
Volant mammal	27	96.30	0	3.70	0
Fish	103	80.58	6.80	12.62	53.9
Other terrestrial vertebrate	27	92.59	3.70	3.70	0
Other aquatic vertebrate	34	85.29	2.94	11.76	75.0
Invertebrate	4	100	0	0	-
Summary groups					
Birds	289	75.46	5.58	18.96	56.9
Mammals	399	88.22	7.27	4.51	61.1
Fish	103	80.58	6.80	12.62	53.9
Other taxa	65	89.23	3.08	7.69	60.0
Locomotion media					
Air	276	77.90	8.33	17.03	51.1
Land	367	87.74	1.09	3.81	71.43
Water	193	82.38	5.18	13.47	61.54
TOTAL	836	83.25	6.34	10.41	57.47

reveals no influence of research goal ($F_{2,29} = 0.27$, $P < 0.762$, Fig. 1). It should be noted that the geometric mean power ($1 - \beta$) of tests that failed to detect an effect of tags across all research goals was only 17.3%: less than a quarter of the conventional power. It should further be noted that the power of this test of power is, by coincidence, also 17%.

Hypothesis (3) receives some support, because as a proportion of all studies, tests indicating deleterious effects of tags were scarcest amongst conservation work (7/319 in conservation, 22/302 in science, and 21/215 in economic: $\chi^2_2 = 14.529$, $P < 0.001$), whereas studies finding no effect of tags formed a similar proportion of the work amongst all three goals (14/319 in conservation, 13/302 in science, and 10/215 in economic: $\chi^2_2 = 0.037$, $P < 0.982$). Together these facts imply that a publication bias exists.

Binary logistic regression models were used to explore the reviewed data. In Model 1, the likelihood of a test of tag-effects being made was considered. The response variable was how the paper was scored ('test' v. 'no test'), with three factorial predictors: research goal (conservation, science, economic) taxon (bird, mammal, fish, other taxa), and locomotion medium (air, water, land).

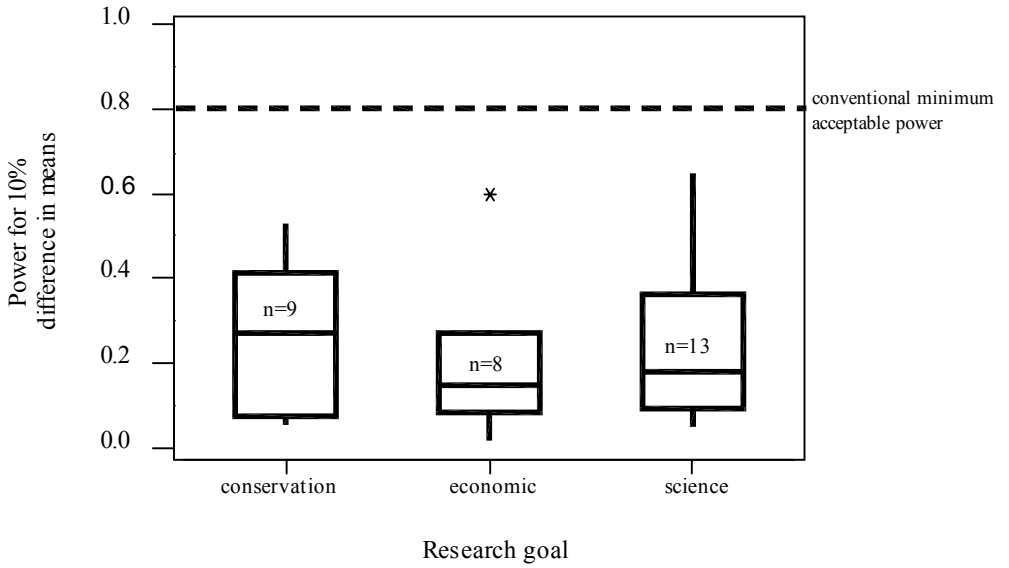

Figure 1. Calculated statistical power to detect a > 10% difference in means (tagged v. untagged) for papers which reported no effect of tag-bearing on study animals (n = 3 7). Where several tests were made within a study, the one with the highest power was used. Power could not be calculated for seven papers, one of which (Kenward et al. 1999: conservation) appeared to meet the conventional power level. The box extends from the first to the third quartile, and the central line represents the median. Vertical lines extend to the full range of the data except where these fall outwith 1.5 × the inter-quartile range, in which case an asterisk marks the point. There was no evidence of an effect of research goal on power achieved (ANOVA $F_{2,27} = 0.00$, $P < 0.998$).

A second logistic regression model, Model 2, was used to explore the impact of research goal, taxon, locomotion and body size on the likelihood of those tests which were made finding an effect or finding no effect of tag-bearing. Initially taxon comprised 'bird', 'mammal', 'fish' and 'other taxa', but Goodness-of-fit tests of the model's reliability indicated an unacceptable uncertainty (Pearson and Hosmer-Lemeshow tests both $P < 0.05$). Accordingly taxon was modified to assess the role of homeothermy. The response variable was the paper's score ('effect' v. 'no effect'), with research goal (conservation, science, economic), taxon (homeotherm, poikilotherm) and locomotion medium (air, land, water) entered as factors, and body mass of subject ($\log_{10}$ transformed) as a covariate. One study was excluded from this analysis, involving sperm whales, whose outlying body mass gave it undue influence in the model.

For both models, overall confidence that all slopes were not equal to zero was high: Model 1, log-likelihood = -253.332, $G_7 = 51.665$, $P < 0.001$; Model 2, log-likelihood = -52.361, $G_5 = 12.211$, $P < 0.032$. Nor was there evidence that the model was an insufficient fit (Goodness of fit tests: (Model 1, Pearson $P < 0.335$, Hosmer-Lemeshow $P < 0.116$; Model 2, Pearson $P < 0.306$, Hosmer-Lemeshow $P < 0.830$).

Model 1 (Table 3) reveals that, when controlling for taxon and locomotion, tests were just under twice as likely (log odds = 1.89, $P < 0.33$) to be made in science research, and just over twice as likely (log odds = 2.06, $P < 0.028$) in economic research, than they were in the field of conservation. Further, the model shows that birds were approximately five times more likely, when controlling for research goal and locomotion, to have a test of tag-bearing effects performed on them than on mammals ($P < 0.007$), fish ($P < 0.007$) and other taxa combined ($P < 0.028$). When controlling for taxon and research goal, animals with water

as their chief locomotion medium were almost three times more likely to have a test of tag effects performed on them than were flying animals.

Model 2 (Table 4) showed that, where tests of tag-bearing effects were carried out, economic work was over seven times more likely to detect an effect of tags than was conservation work ($P < 0.004$), and science nearly four times more likely, although this did not quite meet the $\alpha = 0.05$ criterion ($P < 0.063$). (If both economic and science categories were combined into a single 'non-conservation category, the combined group was over five times more likely to detect significant effects than was conservation work (log odds = 5.33, $P < 0.009$)). Body mass did not emerge as significant predictor in the model, but ground-based animals were probably ($P < 0.054$) just over four times more likely to show a significant impact of tag-bearing than volant animals, when controlling for other factors. If thermoregulatory costs are a significant component of the cost of tag-bearing we predict that homeotherms would be more likely to be affected by tags than are poikilotherms. Although the trend was in the expected direction (homeotherms were 2.71 times more likely to show an effect of tag-bearing) the confidence in the significance of the slope was low ($P < 0.357$), and the 95% confidence intervals ranged from three times *less* likely to over 22 times *more* likely).

TABLE 3. BINARY LOGISTIC REGRESSION MODEL EXPLORING THE PROBABILITY OF A TEST OF TAG-BEARING EFFECTS ON STUDY ANIMALS BEING MADE.

Logit link function of TEST ($n = 87$) (the event) v. NO TEST ($n = 749$) was used. Research goal (reference = conservation), taxon (reference = bird), and locomotion medium (reference = air) entered as factors.

PREDICTOR	ODDS RATIO	95% CONFIDENCE INTERVAL LOWER	UPPER	*Z*	*P*
Constant				–7.50	0.000
Goal					
Conservation	1				
Science	1.89	1.05	3.38	2.13	0.033
Economic	2.06	1.08	3.93	2.19	0.028
Taxon					
Bird	1				
Mammal	0.23	0.08	0.68	–2.67	0.007
Fish	0.18	0.05	0.63	–2.71	0.007
Other taxa	0.22	0.06	0.85	–2.20	0.028
Locomotion					
Air	1				
Land	0.74	0.24	2.30	–0.52	0.601
Water	2.91	1.01	8.37	1.98	0.048

MODEL: Log-likelihood = –253.3. Test that all slopes equal to zero: $G_7 = 51.7$, $P < 0.001$. Goodness of fit tests: Pearson $\chi^2_{15} = 16.736$, $P > 0.33$; Hosmer-Lemeshow $\chi^2_5 = 8.832$, $P > 0.11$. Relationship between response variable and predicted probabilities: 67.0% of observed pairs were concordant with the model, 23.6% discordant, and 9.5% tied (Somers' $D = 0.43$; Goodman-Kruskal $\gamma = 0.48$).

TABLE 4. BINARY LOGISTIC REGRESSION MODEL EXPLORING THE PROBABILITY OF AN EFFECT OF TAG-BEARING EFFECTS ON STUDY ANIMALS BEING DETECTED WHEN SOUGHT.

Logit link function of EFFECT (n = 36) (the event) v. NO EFFECT (n = 50) was used. Research goal (reference = conservation), taxon (reference = poikilotherm), and locomotion medium (reference = air) entered as factors, and body mass of study animal ($\log_{10}$ transformed) entered as a covariate.

PREDICTOR	ODDS RATIO	95% CONFIDENCE INTERVAL LOWER	UPPER	Z	P
Constant				–1.61	0.107
Goal					
Conservation	1				
Science	3.71	0.93	14.80	1.86	0.063
Economic	7.53	1.87	30.28	2.84	0.004
Taxon					
Poikilotherm	1				
Homeotherm	2.71	0.33	22.45	0.92	0.357
Locomotion					
Air	1				
Land	4.28	0.98	18.72	1.93	0.054
Water	2.50	0.32	19.46	0.87	0.382
Log_{10} Body mass	1.52	0.79	2.93	1.26	0.206

MODEL: Log-likelihood= –51.569. Test that all slopes equal to zero: $G_6 = 13.794$, $P < 0.032$. Goodness of fit tests: Pearson $\chi^2_{72} = 77.575$, P > 0.306; Hosmer-Lemeshow $\chi^2_8 = 4.287$, $P > 0.830$. Relationship between response variable and predicted probabilities: 72.2% of observed pairs were concordant with the model, 26.8% discordant, and 1.1% tied (Somers' $D = 0.45$; Goodman-Kruskal $\gamma = 0.46$).

4. Discussion

4.1 REPORTING BIAS

This review has revealed that tests of the effects of tags on their bearers are relatively scarce (less than 11% of studies using radio-tags), and particularly so in conservation-related work (less than 7%). Furthermore, it is evident that the likelihood of a test finding a significant impact of tags varies with research goal, with effects being least likely to be reported in the conservation field. Nor is any detected effect obviously attributable either to tag design (since, when body size (log mass) is controlled for tag size (log tag mass), it does not vary with research goal), or to statistical power (since, where calculable, the power to detect a > 10% difference in means of published tests NOT finding an effect of tags did not vary with research goal). On the other hand, the scarcity of published tests in the field of conservation finding deleterious effects compared to those not doing so (relative to other research fields) suggests a publication bias. The cause of the reduced number of tests of tag-bearing effects in

conservation work might be an unwillingness to publish deleterious results (or a tendency to abandon work where severe deleterious effects are immediately apparent). Additionally there may be some unwillingness to carry out tests where adverse effects are anticipated.

Birds are by far the best-researched taxon with regards to tag-effects (although no examination of effects of tags in flightless birds (excluding penguins) in wildlife biology has been published in the last decade). This may result from an intuitive sense that tag-bearing must be most costly amongst flying animals. However, this review does not support such intuition, because there was almost ($P < 0.054$) evidence to suggest that land-based animals were more likely to show a deleterious impact of tags (Table 4). This may partly reflect the different severity of stress-related effects in birds and mammals, rather than locomotion medium *per se*.

4.2 PHYLOGENETIC BIAS

An underlying weakness of the review is that it could not correct for phylogenetic biases in the sample. For example, all the water-based birds were penguins, and might, conceivably, have responded in the way they did to radio-tags (e.g. Gales et al. 1990, Bannasch et al. 1994) not because they were birds with water-based locomotion, but because of some feature specific to penguins. This problem compromises the predictive power of the findings of this review. Nevertheless, there are often compelling reasons for assuming that many effects of radio-tags are general rather than particular. To return to the penguins, the increased drag that external transmitters impose would be expected to affect all similar-sized swimming animals to a similar degree. Costs of supporting extra loads on land are also likely to be general. Other features may be specific to particular tag-designs: tail-based tags on birds may avoid thermoregulatory costs, but generate extra flight costs because of effects on the centre of gravity (Orbrecht 1988); implanted tags may be associated with post-surgery traumas, but not with drag-costs. Similarly there are features of particular ecologies that might render some groups more susceptible to tag-effects than others: a reduction of take-off angle consequent on increased mass might make small passerines more vulnerable to predation (Metcalfe & Ure 1995), but have little impact on mortality in large carrion-feeders.

4.3 POWER OF TESTS

The statistical power of a test is particularly relevant when the null hypothesis cannot be rejected, but where an acceptance of the null hypothesis is published as a result supporting a particular course of action (for example, the continued use of radio-tags on a population). The conventional level of power ($1 - \beta$) is $P = 0.8$ (Cohen 1988), which means that false null hypotheses are accepted 20% of the time (a Type II Error, β). This implies that making a Type II error is four times less important than making a Type I Error (α), since the conventional level of probability at which biologists accept making the latter is 0.05. We

query the equity of this convention, or rather, suggest that authors accepting a null hypothesis should consider explicitly the applicability of the convention to their data. Nothing should be more clear than that the statistical acceptance of a null hypothesis does not equate to a demonstration that the converse hypothesis is false. Thus, for example, finding no evidence for an effect of tags on survival does not prove that tags do not affect survival. Under the system of logical positivism (Popper 1963), a failure to falsify (potential Type II error) is a rather trivial matter, whereas an incorrect falsification of the null hypothesis (Type I error) is a serious mistake. In the biological sciences, however, most hypotheses tested are *statistical* rather than *scientific* (Quinn & Dunham 1983), dealing with the properties of populations and particular systems. Under such circumstances, we can see no reason why a conclusion that 'there is no effect' should routinely require a different probability of being true than a conclusion that 'there is an effect.' Indeed Toft & Shea (1983) have argued that in ecological science, the seriousness of making a Type II error is often greater than that of making a Type I error.

Imagine a test of the effect of tag-bearing on survival of a threatened species. Imagine that the data collected allows us to conclude 'there was no effect of tag-bearing on survival, $P = 0.20$', but imagine also that the sample size and variation gives a statistical power of $P = 0.80$ to detect, say, a 10% effect of tags on survival. It is at least equally problematic to argue for continued use of tags to aid in the discovery of how mortality occurs, as it would be to recommend that further use of tags be discontinued (on the basis that there was a probability of 0.2 of the null hypothesis being incorrect). Ideally we require α and β to be equal, and at a low level (perhaps $P < 0.05$). Once in the literature, a study entitled 'No effect of radio-transmitters on …' is available for citation, and will be used in a discussion, such as this one, to support the notion that tags are generally acceptable. Yet this review has shown that power is generally very much lower than the conventional 0.8, and that many studies purporting to demonstrate no effect of tags had power of < 0.1 to detect a 10% difference (Fig. 1). Assuming that about half of null hypotheses (< 10% difference) are in fact false, then researchers using such tests would be five times less likely to identify the correct answer than if they were to toss a coin. In fact, only a single study (Kenward et al. 1999) which appears to have achieved the conventionally accepted level of power to detect a 10% difference in survival, energy expenditure, or reproductive success between experimental and control individuals, has failed to find one. This amounts to a large body of evidence demonstrating that tags can have a deleterious effect on their bearers, and just one study to support the notion that any such effect is either absent or trivial.

We argue that making the assumption that there is no effect of tag-bearing because a study has failed to demonstrate one is the worst possible conclusion. Instead we favour equalising α and β, since we believe that making a Type I or a Type II error is equally serious in this instance. We suggest stepping away from the convention of $P = 0.05$, which appears to give scientific weight to what is essentially obfuscation, and which provides no guidance for managers. Instead we suggest that admitting that a particular piece of research in fact offers no clear answer is both more honest and more useful. To say for example that 'no effect of tags was detectable at the $P = 0.05$ level' gives much less information than, say, to observe that 'an effect size of [say] 12% was detected, with α and β

at [say] 0.19'. In the second instance, accepting the reality of the effect size comes with just under a 1 in 5 risk of being wrong. In the case of many of the studies reviewed here, a more likely scenario would be a 1 in 2 chance of mistakenly rejecting the observed effect size. So instead of one form of words implying confidence that there is no significant impact of tags, we suggest an alternative which admits that the data truly yield almost no improvement over coin-tossing about a particular effect size. This may not be convenient, but it might be the best advice on the basis of current information.

Another flawed but widespread practice is making the assumption that the absence of a tag-bearing cost in one species implies its absence in another. This might be justifiable within close phylogenetic groups and where a powerful test has shown that only a trivial effect of tags cannot be ruled out in the first species. Yet examination of the literature shows that studies with a less than 20% chance of finding a substantial significant difference between tag-bearers and controls in one species are routinely cited as partial justification for the use of transmitters on an unrelated species.

Johnson (1999) argues that statistical hypothesis-testing is generally more subjective than is often realised, and puts forward a strong case for placing more emphasis, not on the probability that two populations differ in some measure, but on the degree by which they differ—the so-called 'effect size'. Johnson (1999) suggests that 'questions about the likely size of true effects can be better addressed with confidence intervals than with retrospective power analysis', and quotes Shaver (1993), who described power analysis as 'a vacuous intellectual game'. Whilst we do not disagree with the former point, but merely observe that confidence intervals are rarely published and so do not lend themselves to meta-analysis, we believe that the latter point is mistaken at least in the ecological sciences. Used properly, power analysis can focus attention on both the risk of a Type II error, and on the effect size that researchers are interested in. Strong evidence for a lack of even a modest effect of tag-bearing on some measure of behaviour, survival or reproduction is rare amongst those animals tested. By contrast, a number of studies have indicated quite severe effects. It is therefore important that some attempt to investigate the costs of tag-bearing is made with each new animal (or each new tag-type) being used. This is the more the case where scarcity makes the fate of individual animals important to species survival. The costs of tag-bearing should be given full consideration by researchers, even where it is not possible to measure them, and extrapolation from a tagged-population to the general population should be done only with considerable circumspection.

5. References

Bannasch, R.; Wilson, R.P.; Culik, B. 1994: Hydrodynamic aspects of design and attachment of a back-mounted device in penguins. *Journal of Experimental Biology 194*: 83–96.

Cohen, J. 1988: Statistical Power Analysis for the Behavioural Sciences. (2nd edn) Lawrence Erlbaum Associates, New Jersey.

Gales, R.; Williams, C.; Ritz, D. 1990: Foraging behaviour of the little penguin, *Eudyptula minor*: initial results and assessment of instrument effect. *Journal of Zoology 220*: 61–85.

Hill, I. F.; Cresswell, B.H.; Kenward, R.E. 1999: Field-testing the suitability of a new back-pack harness for radio-tagging passerines. *Journal of Avian Biology 30*: 135–142.

Johnson, D.H. 1999: The insignificance of statistical significance testing. *Journal of Wildlife Management 63*: 763–772.

McCullagh, P.; Nelder, J.A. 1992: Generalized Linear Models. Chapman & Hall, London.

Metcalfe, N.B.; Ure, S.E. 1995: Diurnal variation in flight performance and hence potential predation risk in small birds. *Proceedings of the Royal Society of London, Series B 261*: 395–400.

Orbrecht, H.H.; Pennycuick, C.J.; Fuller, M.R. 1988: Wind tunnel experiments to assess the affect of back-mounted radio-transmitters on bird body drag. *Journal of Experimental Biology 135*: 265–273.

Popper, K. 1963: The Logic of Scientific Discovery. Harper & Row, New York.

Quinn, J.F.; Dunham, A.E. 1983: On hypothesis testing in ecology and evolution. *American Naturalist* 122: 602–617.

Shaver, J.P. 1993: What statistical significance testing is, and what it is not. *Journal of Experimental Education 61*: 293–316.

Toft, C.A.; Shea, P.J. 1983: Detecting community-wide patterns: estimating power strengthens inference. *American Naturalist 122*: 618–625.