Invited Paper:

THE IMPORTANCE OF REPLICATION IN WILDLIFE RESEARCH

DOUGLAS H. JOHNSON,[1] U.S. Geological Survey, Northern Prairie Wildlife Research Center, Jamestown, ND 58401, USA

Abstract: Wildlife ecology and management studies have been widely criticized for deficiencies in design or analysis. Manipulative experiments—with controls, randomization, and replication in space and time—provide powerful ways of learning about natural systems and establishing causal relationships, but such studies are rare in our field. Observational studies and sample surveys are more common; they also require appropriate design and analysis. More important than the design and analysis of individual studies is metareplication: replication of entire studies. Similar conclusions obtained from studies of the same phenomenon conducted under widely differing conditions will give us greater confidence in the generality of those findings than would any single study, however well designed and executed.

JOURNAL OF WILDLIFE MANAGEMENT 66(4):919–932

Key words: control, experiment, metareplication, observational study, pseudoreplication, randomization, replication, sample survey, science.

Wildlife researchers seem to be doing everything wrong. Few of our studies employ the hypothetico–deductive approach (Romesburg 1981) or gain the benefits from strong inference (Platt 1964). We continually conduct descriptive studies, rather than the more effective manipulative studies. We rarely select study areas at random, and even less often do the animals we study constitute a random sample. We continue to commit pseudoreplication errors (Hurlbert 1984, Heffner et al. 1996). We confuse correlation with causation (Eberhardt 1970). Frequently we measure the wrong variables such as indices to things we really care about (Anderson 2001). And we may measure them in the wrong places (convenience sampling; Anderson 2001). We often apply meaningless multivariate methods to the results of our studies (Rexstad et al. 1988). We test null hypotheses that not only are silly but are known to be false (Cherry 1998, Johnson 1999, Anderson et al. 2000). We rely on nonparametric methods that are neither necessary nor appropriate (Johnson 1995).

Such problems permeate our field. In my early years as a hypercritical statistician, I read many articles in *The Journal of Wildlife Management* and related journals. In virtually every article, I found problems—often serious ones—in the methods used to analyze data. That experience was repeated later in a class in evolutionary ecology. During that class, we critically reviewed many key papers in evolutionary ecology. Some students were assigned to attack, others to defend those articles. We identified substantial problems in the design, analysis, or interpretation in nearly all of those influential and highly regarded studies.

Despite all our transgressions, we must be doing something right. We have brought some species back from the brink of extinction. The bald eagle (*Haliaeetus leucocephalus*), whooping crane (*Grus americana*), Aleutian Canada goose (*Branta canadensis leucopareia*), and gray wolf (*Canis lupus*) were extremely rare over much or all of their range only a few years ago; now they are much more common. Many of us had given up on the black-footed ferret (*Mustela nigripes*) and California condor (*Gymnogyps californianus*), species that, while still at risk, appear to be recovering. And we can manage for abundance if we want to, such as we have done for white-tailed deer (*Odocoileus virginianus*) and mallards (*Anas platyrhynchos*). Recently, Jack Ward Thomas spoke of the "tremendous record of success" in our field (Thomas 2000:1).

Why this apparent inconsistency between our error-prone methods and the successes of our profession? I hope to address that question here by discussing what truly is important in scientific research. I first discuss causation, then manipulative experimentation as a powerful way of learning about causal mechanisms. The 3 cornerstones of experimentation are control, randomization, and replication. These features also are integral to observational studies and sample surveys, which are more common in our field. For those types of studies especially, I argue that the most

[1] E-mail: Douglas_H_Johnson@usgs.gov

important feature is replication. Further, I expand this concept to the level of metareplication—replication of entire studies—and suggest that this is the most reliable method of learning about the world. It is a natural way of human thinking and is consistent with a Bayesian approach to statistical inference. Metareplication allows us to exploit the values of small studies, each of which individually may be unable to reach definitive conclusions. Metareplication provides us greater confidence that certain relationships are general and not specific to the circumstances that prevailed during a single study.

CAUTION ABOUT CAUSATION

The "management" in "wildlife management" implies causality. We believe we can perform some management action that will produce a predictable response by wildlife. Even if the causes cannot be manipulated, it is useful to know the mechanisms that determine certain outcomes, such as that spring migration of birds is a response to increasing day length, or that drought reduces the number of wetland basins that contain water.

The concept of causation is most readily adopted in the physical sciences, where models of the behavior of atoms, planets, and other inanimate objects are applicable over a wide range of conditions (Barnard 1982) and the controlling factors are few (e.g., pressure and temperature are sufficient to determine the volume of a gas). In the physical sciences, causality implies lawlike necessity. In many fields, however, notions of causality reduce to those of probability, which suggests exceptions and lack of regularity. Here, causation means that an action "tends to make the consequence more likely, not absolutely certain" (Pearl 2000:1). This is so in wildlife ecology because of the multitude of factors that influence a system. For example, liberalizing hunting regulations for a species tends to increase harvest by hunters. In any specific instance, liberalization may not result in an increased harvest because of other influences such as population size of the species, weather conditions during the hunting season, and the cost of gasoline as it affects hunter activity.

Suppose you want to determine the effect on squirrel abundance of some treatment (= putative cause), for example, selective logging in a woodlot by removing all trees greater than 45 cm diameter at breast height (dbh). The treatment effect on some woodlot can be defined as

$$T = Y_t(u) - Y_c(u), \quad (1)$$

where $Y_t(u)$ is the number of squirrels in woodlot u after the treatment, and $Y_c(u)$ is the number of squirrels in that woodlot if the treatment had not been applied (I follow Rubin [1974] and Holland [1986] here). If the woodlot is logged, then you can observe $Y_t(u)$ but not $Y_c(u)$. If the treatment is not applied, then you can observe $Y_c(u)$ but not $Y_t(u)$. Thus arises the fundamental problem of causal inference: you cannot observe the values of $Y_t(u)$ and $Y_c(u)$ on the same unit. That is, any particular woodlot is either logged or not.

Holland (1986) described 2 solutions to this problem. With the first, one has 2 units (u_1 and u_2, here woodlots) and assumes they are identical. Then the treatment effect T is estimated to be

$$T = Y_t(u_1) - Y_c(u_2), \quad (2)$$

where u_1 is treated and u_2 is not. This approach is based on the very strong assumption that the 2 woodlots, if not logged, would have the same number of squirrels, that is, $Y_c(u_2) = Y_c(u_1)$. That assumption is not testable, of course, because 1 woodlot had been logged. It can be made more plausible by matching the 2 units as closely as possible or by believing that the units are identical. That latter belief comes more easily to physicists thinking about molecules than to ecologists thinking about woodlots, however.

Holland (1986) termed the other solution statistical. One gets an expected, or average, causal effect T over the units in some population:

$$T = \mathrm{E}(Y_t - Y_c), \quad (3)$$

where, unlike with the other solution, different units can be observed. The statistical solution replaces the causal effect of the treatment on a specific unit, which is impossible to observe, by the *average* causal effect in the population, which is possible to estimate.

This discussion reflects the need for a control, something to compare with the treated unit, which is required for either approach. To follow the statistical approach, we often invoke randomization. If, for example, we are to compare squirrel numbers on a treated woodlot and an untreated one, we might get led astray if the woodlots were of very different size, or if one contained more mast trees, or if one was rife with predators of squirrels and the other was not. One way—but not the only way—to protect against this possibly misleading outcome is to determine at random which woodlot receives the treatment

and which does not. This can be done if the researcher has tight control over the experiment; it is impossible in many "natural experiments" and observational studies.

But even if you select at random a woodlot for treatment and another as a control, you still may end up by chance comparing a large woodlot that has numerous mast trees and few predators with a woodlot with opposite characteristics. This leads to the third important criterion for determining causation: replication. Repeating the randomization process and treatments on several woodlots reduces the chance that woodlots in any group consistently are more favorable to squirrels. In summary, then, assessing the effect of some treatment with a manipulative experiment requires a control, randomization, and replication (Fisher 1926).

One might attempt to determine the effect of selective logging on squirrels by comparing woodlots that have trees greater than 45 cm dbh with woodlots that lack such large trees. But such a comparison is not as definitive as a manipulative experiment. The 2 types of woodlots might differ in numerous ways, other than the presence or absence of large trees, that influence squirrel abundance. If variables that are known or suspected to be influential are measured, careful statistical analysis may account for their effects, but large samples may be necessary, and it is possible that an important variable went unmeasured.

An ideal design might involve a number of woodlots on which squirrel density is measured both before and after the treatment is applied. Then, instead of comparing the density of squirrels on treated versus untreated woodlots, one could compare the change in density (before and after treatment) between the 2 groups. Crossover designs also provide a powerful way to reduce the influence of inherent differences among experimental units. Under a crossover design, for a certain time period, some units receive treatments and other units serve as controls. Then the roles of the units switch: control units receive treatments and formerly treated units are left alone to serve as controls. An obvious concern with crossover designs is that treatment effects may persist. One remedy is to have a time period between the 2 phases of the study sufficient to allow treatment effects to dissipate. A crossover design would not be appropriate for the squirrel–woodlot example because the effect of logging would persist for decades, if not longer. Crossover designs were used by Balser et al. (1969) and Tapper et al. (1996) to estimate the effects of predator reduction on prey species. In these studies, predators were removed from 1 study area for 3 years, while another area served as a control; after 3 years, the treatments were switched.

Correlation versus Causation: the Importance of Mechanisms

It is always useful to have an understanding of the mechanisms that influence phenomena of interest and to distinguish causation from correlation. We might be able to relate mallard production to precipitation (Boyd 1981), but more useful is the understanding that precipitation affects the condition of wetlands where mallards breed, which in turn influences breeding propensity, clutch size, and survival of young (Johnson et al. 1992). We can have greater confidence in our findings if they are consistent with mechanisms that are both reasonable and supported by other evidence. The presence of such mechanisms gives credibility that the correlational smoke may in fact represent causational fire (Holland 1986). Romesburg (1981) argued that causation may be invoked if the correlational evidence is accompanied by, for example, the elimination of other possible causes, demonstration that the correlation occurs under a wide variety of circumstances, and the existence of a plausible dependence between the putative cause and the outcome. In a similar vein, mechanistic models are more useful than descriptive models for understanding systems (Johnson 2001*a*, Nichols 2001).

MANIPULATE, IF YOU CAN

Manipulative experimentation is a very effective way to determine causal relationships. One poses questions to nature via experiments such as selective logging. By manipulating the system yourself, you reduce the chance that something other than your treatment causes the results that are observed. Further, as emphasized by Macnab (1983), little can be learned about the dynamics of systems at equilibrium. Manipulation is helpful to understand how the systems respond to changes. Experimentation also forms the basis of what has been termed strong inference (Platt 1964), in which alternative hypotheses are devised and crucial experiments are performed to exclude 1 or more of the hypotheses.

Wildlife ecologists sometimes face severe difficulties meeting the needs of control, randomization, and replication in manipulative experiments. Many systems are too large and complex for ecol-

ogists to manipulate (Macnab 1983). Often "treatments"—such as oil spills—are applied by others, and wildlife ecologists are called in to evaluate their effects. In such situations, randomization is impossible and replication undesirable. Methods for conducting environmental studies, other than experiments with replications, are available (Smith and Sugden 1988, Eberhardt and Thomas 1991); among these are experiments without replications, observational studies, and sample surveys.

Replication is particularly difficult with experiments at the ecosystem level, which are more complex but also more meaningful than experiments at microcosm or mesocosm levels, where replication is more feasible (Carpenter 1990, 1996; Schindler 1998). Experiments lacking replications can be, and indeed often have been, analyzed by taking multiple measurements of the system and treating them as independent replicates. This practice was criticized by Eberhardt (1976) and Hurlbert (1984), the latter naming it pseudoreplication. I address this topic more fully below.

Observational studies lack the critical element of control by the investigator, although they can be analyzed similarly to an experimental study (Cochran 1983). One is less certain that the presumed treatment actually caused the observed response, however. In lieu of controlled experimentation, one can (1) reduce the influence of extraneous effects by restricting the scope of inference to situations similar to the one under observation; (2) employ matching, by which treated units are compared with units that were not treated but in other regards are as similar as possible to the treated units; or (3) adjust for the effects of other variables during analysis, with methods such as analysis of covariance (Eberhardt and Thomas 1991).

Longitudinal observational studies, with measurements taken before and after some treatment, generally are more informative than cross-sectional observational studies, in which treated and untreated units are studied only after the treatment (Cox and Wermuth 1996). (Of course, measurements on experimental and control units before and after treatments are highly desirable in experimental studies, as well as observational studies.) Intervention analysis is a method used to assess the effect of some distinct treatment (intervention) that has been applied to a system. The intervention was not assigned by the investigator and cannot reasonably be replicated. One approach is to model the system as a time series and look for changes subsequent to the intervention. That approach was taken with air-quality data by Box and Tiao (1975), who sought to determine how ozone levels might have responded to events such as a change in the formulation of gasoline.

Sometimes it is known that a major treatment will be applied at some particular site such as a dam to be constructed on a river. It may be feasible to study that river before as well as after the dam is constructed. That simple before-and-after comparison suffers from the weakness that any change that occurred coincidental with dam construction, such as a decrease in precipitation, would be confounded with changes resulting from the dam, unless the changes were specifically included in the model. To account for the effects of other variables, one can study similar rivers during the same before-and-after period. Ideally, these rivers would be similar to and close enough to the treated river so to be equally influenced by other variables but not influenced by the treatment itself. This design has been called the BACI (before–after, control–impact) design (Stewart-Oaten et al. 1986, Stewart-Oaten and Bence 2001, Smith 2002) and is used for assessing the effects of impacts.

It is difficult for investigators to manipulate large and complex systems such as ecosystems. But wildlife managers, as well as those who manage ecosystems for other objectives such as timber production, do so frequently. This disparity between investigators and managers led Macnab (1983) to recommend that management activities be viewed as experiments that offer opportunities to learn about large systems. Actions taken for management benefits generally lack controls, randomization, and replication; such shortcomings can be remedied by incorporating these features into the experiment. Key assumptions should be identified and stated as hypotheses, rather than treated as facts. The results of management actions, even if they show no effect, should be measured and reported.

The adaptive resource management approach blends the idea of learning about a system with the management of the system (Walters 1986, Williams et al. 2002). The key notion, which moves the concept beyond a "try something and if it doesn't work try something else" attitude, is that knowledge about the system becomes one of the products of the system that is to be optimized.

Sample surveys differ from experiments in that one endeavors either to estimate some characteristic over some domain—such as the number of mallards in the major breeding range in North

America—or to compare variables among groups—such as the median age of hunters compared with nonhunters.

CONTROLS

The term control, confusingly, has at least 3 different meanings in experimental design. The first meaning, which is more general and not specifically addressed here, involves the investigator's role. In a controlled study, the treatment (cause) is assigned by the investigator; the study is an experiment. In an uncontrolled study, the treatment is determined to some extent by factors beyond the investigator's control; the study is observational (Holland 1986). The second meaning, design control, implies that, while some experimental units receive a treatment, others (the "controls") do not. The third meaning, statistical control, means that other variables that may influence the response are measured so that we may estimate their effects and attempt to eliminate them statistically.

The major benefit of design control is to provide a basis for comparison between treated and untreated units. It reduces the error; our measured response is likely to reflect only the treatment rather than a variety of other things. Statistical control usually is less effective in reducing error and is applied after treatments are applied.

Sometimes strict design controls are not possible. Intervention analysis and BACI designs can demonstrate that some variables may have changed subsequent to the intervention or impact, but one will be less confident from that analysis that the intervention caused that change. Confidence will increase if potential confounding variables are measured and their effects are accounted for during analysis—that is, through statistical control.

Controls should be distinguished from reference units. The latter are units that represent some ideal that management actions are intended to approach. Reference sites are especially useful in restoration ecology, when evaluating the effectiveness of alternative management activities for restoring degraded areas to conditions embodied in the reference sites (Provencher et al. 2002).

RANDOMIZATION

Randomization can occur at 2 levels. In both experiments and sample surveys, randomization means that the objects to be studied are randomly selected from some population (called a target population) for which inference is desired. Accordingly, each member of that population has some chance of being included in the sample. Chances may be the same for all members, but that is not necessary. At a second level, in a manipulative experiment, randomization means that the treatment each unit receives is randomly determined.

Randomization makes variation among sample units, due to variables that are not accounted for, act randomly, rather than in some consistent and potentially misleading manner. Randomization thereby reduces the chance of confounding with other variables. Instead of controlling for the effects of those unaccounted-for variables, randomization makes them tend to cancel one another out, at least in large samples. In addition, randomization reduces any intentional or unintentional bias of the investigator. It further provides an objective probability distribution for a test of significance (Barnard 1982).

While randomizing the assignment of treatments to units is crucial in experimentation, I suggest that randomization in selecting the units in an experiment or sample survey is less important than control or replication. First, the intended benefits of randomization apply only conceptually. Randomly sampling from a population does not ensure that the resulting sample will represent that population, only that, if many such samples are taken, the average will be representative. But in reality only a single sample is taken, and that single sample may or may not be representative. Randomization does make variation act randomly, rather than systematically. However, this property is only conceptual, applying to the notion that samples were repeatedly taken randomly. The single sample that was taken may or may not have properties that appear systematic.

Randomization ostensibly reduces hidden biases of, or "cheating" by, an investigator. But, if an investigator wishes to cheat, why not do so but say that randomization was employed (Harville 1975)?

What Does a Sample Really Represent?

Any sample, even a nonrandom one, can be considered a representative sample from some population, if not the target population. What is the population for which the sample is representative? Extrapolation beyond the area from which any sample was taken requires justification on nonstatistical bases. For example, studies of animal behavior (or physiology) based on only a few individuals may reasonably be generalized to entire species if the behavior patterns (or physiological processes) are relatively fixed (i.e., the units are homogeneous with respect to that fea-

ture). In contrast, traits that vary more widely, such as habitat use of a species or annual survival rates, cannot be generalized as well from a sample of comparable size. Consistency of a feature among the sampled and unsampled units is more critical than the randomness of a sample. Can one comfortably draw an inference to a population from a sample, even if that sample is nonrandom? In reality, most useful inferences require extrapolation beyond the sampled population. For example, if we want to predict the consequences of some action carried out in the future based on a study conducted in the past, we are extrapolating forward in time.

Is Randomization Always Good?

Suppose you want to assess the characteristics of vegetation in a 10-ha field. You decide to place 8 quadrats in the field and measure vegetation within each of those quadrats. Results from those 8 samples will be projected to the entire field. You can select the 8 points entirely at random. It is possible that all 8 quadrats will be within the same small area of the field, however, and be very different from most of the field. Choosing points at random ensures that, if you repeat the process many times, on average you will have a representative sample. But in actuality you have only 1 of the infinitely many possible samples; randomness tells you nothing about your particular sample. It might be perfectly representative of the entire field, or it might be very deviant. The chance that it is representative increases with sample size, so the risk of a random sample not being representative is especially troublesome in small samples.

There are methods for taking samples to increase the chance that they better represent the entire field. One method is to stratify, if there is prior knowledge of some variable likely to relate to the variable of interest. Another method is to take systematic rather than random samples. Hurlbert (1984) emphasized the importance of interspersion in experimental design, having units well distributed in space; this serves 1 goal of randomization, often more successfully. Such balanced designs diminish the errors of an experiment (Fisher 1971).

What Is Independence, and Is It Necessary?

Randomization provides a basis for probability distributions because the observations in a random sample generally are statistically independent. Independence is a mathematically wondrous property, since it facilitates the definition of distributional properties, such as the variance, test statistics, and *P*-values. But what is independent? Consider, as did Millspaugh et al. (1998), the assessment of habitat preference of animals that occur together. If the animals are inextricably tied together, such as a mother and her dependent offspring, then the locations of each certainly are not independent. If the animals occur together simply because they favor the same habitats, Millspaugh et al. (1998) argued that the individual animals are independently making habitat choices and thus should be treated as independent units. Then there are intermediate situations, such as a mother and her not-quite-dependent offspring. Ascertaining independence is not a simple matter; statistical independence can be evaluated only in reference to a specific data set and a specified model (Hurlbert 1997).

But what is the problem if data are not independent? Suppose you have 100 observations, but only 50 of them are independent, and for each of those there is another observation that is identical to it. So the apparent sample size is 100, but only 50 of those are independent. If you estimate the average of some characteristic of the individuals, the mean in fact will be a good estimator. But the standard error will be biased low. And a test statistic, say, for comparing the mean of that group with another, will be inflated and will tend to reject the null hypothesis too often (e.g., Erickson et al. 2001).

This is a fundamental problem for any test statistic from an individual study. There are ways to correct for the disparity between the number of observations and the number of independent observations. Dependencies among observations sometimes can be modeled explicitly, such as with generalized estimating equations (Liang and Zeger 1986). Dependencies, such as sampling from clusters of units, often result in overdispersion, in which the sample variance exceeds the theoretical value; in such cases certain adjustments to the theoretical variance can be made (McCullagh and Nelder 1989, Burnham and Anderson 2002). A similar issue arises with respect to temporally or spatially correlated observations.

I argue later that problems caused by a lack of independence, while affecting inferences from individual studies, are less consequential than they appear.

Independence and the Scope of Inference

Suppose you are investigating area sensitivity in grassland birds. That is, you wonder whether certain species prefer larger patches of grassland to

smaller patches. Area sensitivity might be manifested by reduced densities (not just total abundance) in smaller habitat patches or by the avoidance of habitat edges (Faaborg et al. 1993, Johnson and Winter 1999, Johnson 2001*b*). Avoidance of edge means that birds are restricted to the interior portions of a patch, which results in reduced densities for the patch as a whole. To determine whether certain species are area-sensitive, you might compare densities of the species in patches of similar habitat but different size. Alternatively, you might examine the locations of birds (let's say nests, but we could consider song perches, etc.) within a habitat patch and determine whether there is evidence that densities of nests are reduced near edges compared to interiors.

For comparing densities, the sample units are patches. Those are the units to which a "treatment" (patch size) pertains; all birds in that patch have the same patch size. For examining edge avoidance, in contrast, the sample units are nests because each has its own, possibly unique, value of the "treatment" (distance to edge). Logically, then, the latter approach would be more powerful because a single patch might produce dozens of sample units (nests), resulting in much larger sample sizes.

Assuming there is no free lunch, what is happening here? The disparity is the scope of inference. If we study densities in patches, the studied patches can be considered a sample from some target population of patches, and inferences should apply to that population. If we study nests within a patch and examine distances to an edge, the inference is only to that single patch. You might conclude that birds avoid locating their nests near a habitat edge, but that conclusion applies only locally.

Replication Is Necessary for Randomization to Be Useful

The properties of randomization in the selection of units to study are largely conceptual; that is, they pertain hypothetically to some long-term average. For example, randomization makes errors act randomly, rather than in a consistent direction. But in any single observation, or any single study, the error may well be consistent. It is only through replication that long-term properties hold.

REPLICATION

Replication requires that a sample consist of more than 1 member of a population, or that treatments be applied independently to more than 1 unit. Replication provides 2 benefits. First, it reduces error because an average of independent errors tends to be smaller than a single error. Replication serves to ensure against making a decision based on a single, possibly unusual, outcome of a treatment or measurement of a unit. Second, because we have several estimates of the same effect, we can estimate the error, as the variation in those estimates reflects error. We then can determine whether the value of the treated units are unusually different from those of the untreated units. The validity of that estimate of error depends on the experimental units having been drawn randomly; thus, the validity is a joint property of randomization and replication.

Is Replication Always Necessary?

Imagine yourself cooking a stew. You want to see if it needs salt. You dip a teaspoon into the kettle and take a taste. If it's not salty enough, you add more salt. Notice that you did not take replicate samples. Only one. (Further, you probably didn't randomly select where in the kettle to sample; you most certainly took it from the surface and most likely near the center of the kettle.) Cooks have been using this sampling approach for probably centuries, without evident problem. Why?

The single, nonrandomly selected sample generally suffices because the stew is fairly homogeneous with respect to salt. A teaspoon from 1 location will be about as salty as a teaspoon from another. This is because the stew has been stirred. Note that the same approach would not work for sampling meat, which is distributed less uniformly throughout a stew. Replication may not be necessary if all the members of the universe are identical, or nearly enough so.

OTHER LEVELS OF REPLICATION

I find it useful to think of replication occurring at 3 different levels (Table 1). The fundamental notion is of ordinary replication in an experiment: treatments are applied independently to several units. In our squirrel–woodlot example, we would want several woodlots to be logged and several to be left as controls. (Comparable considerations apply to observational studies or sample surveys.) As mentioned above, replication serves to ensure against making a decision based on a single, possibly unusual, outcome of the treatment. It also provides an estimate of the variation associated with the treatment. Other levels of replication are pseudoreplication and metareplication.

Table 1. The types of replication differ in what actions are repeated, what scope of inference is valid, and the role of *P*-values.

Term	Repeated action	Scope of inference	*P*-value	Analysis
Pseudoreplication	Measurement	Object measured	Wrong	Pseudo-analysis
Ordinary replication	Treatment	Objects for which samples are representative	"OK"	Analysis
Metareplication	Study	Situations for which studies are representative	Irrelevant	Meta-analysis

Pseudoreplication

At a lower level than ordinary replication is what Hurlbert (1984) called pseudoreplication. Often couched in analysis of variance terms (using the wrong error term in an analysis), typically it arises by repeating measurements on units and treating such measurements as if they represented independent observations. The treatments may have been assigned randomly and independently to the units, but repeated observations on the same unit are not independent. This was what Hurlbert (1984) called simple pseudoreplication and what Eberhardt (1976) had included in pseudodesign. Pseudoreplication was common when Hurlbert (1984) surveyed literature on manipulative ecological experiments, mostly published during 1974–1980, and estimated that about 27% of the experiments involved pseudoreplication. Heffner et al. (1996:2561) found that the frequency of pseudoreplication in more recent literature (1991–1992) had dropped but was still "disturbingly high." Stewart-Oaten (2002) provided some keys for recognizing pseudoreplication, which is not always straightforward.

Metareplication

At a higher level than ordinary replication is what I term metareplication. Metareplication involves the replication of *studies*, preferably in different years, at different sites, with different methodologies, and by different investigators. Conducting studies in different years and at different sites reduces the chance that some artifact associated with a particular time or place caused the observed results; it should be unlikely that an unusual set of circumstances would manifest itself several times or, especially, at several sites. Conducting studies with different methods similarly reassures us that the results were not simply due to the methods or equipment employed to get those results. And having more than 1 investigator perform studies of similar phenomena reduces the opportunity for the results to be due to some hidden bias or characteristic of that researcher. Just as replication within individual studies reduces the influence of errors in observations by averaging the errors, metareplication reduces the influence of errors among studies themselves.

Youden (1972) provided a classic example of the need for metareplication. He described the sequence of 15 studies conducted during 1895–1961 to estimate the average distance between Earth and the sun. Each scientist obtained an estimate, as well as a confidence interval for that estimate. Every estimate obtained was outside the confidence interval for the previous estimate! The confidence each investigator had in his estimate thus was severely overrated. The critical message from this saga is that we should have far less confidence in any individual study than we are led to believe from internal estimates of reliability. This also points out the need to conduct studies of any phenomenon in different circumstances, with different methods, and by different investigators. That is, to do metareplication.

Allied to this reasoning is Levins' notion of truth lying at the "intersection of independent lies" (Levins 1966:423). He considered alternative models, each of which suffered from 1 or more simplifying assumptions (and all models involve some simplification of the system being modeled) that made each model unrealistic in some way or another. He suggested that if the models—despite their differing assumptions—lead to similar results, we have a robust finding that is relatively free of the details of each model. In the context of metareplication, although independent studies of some phenomenon each may suffer from various shortcomings, if they paint substantially similar pictures, we can have confidence in what we see.

The idea of robustness in data analysis is analogous to robustness among studies. Robustness in the analysis of data from a single study means that the conclusions are not strongly dependent on the assumptions involved in the analysis (Mallows 1979). Similar inferences would be obtained from statistical methods that differ in their assumptions. For example, conclusions might not vary even if the data do not follow the assumed distribution, such as the Normal, or if outliers are present in the data. Analogously, robustness in metareplication

means that similar interpretations about phenomena are reached from studies that differ in methods, investigators, locations, times, etc.

The notion that studies should be replicated certainly is not new. Replication, in the form of repetition of key experiments by others, has been conventional practice in science far longer than statistics itself has been (Carpenter 1990). Fisher (1971) observed that conclusions are always provisional, like progress reports, interpreting the evidence so far accrued. Tukey (1960) proposed that conclusions derive from the assessment of a series of individual results, rather than a particular result. Eberhardt and Thomas (1991:57) observed that "truly definitive single experiments are very rare in any field of endeavor, progress is actually made through sequences of investigations." Cox and Wermuth (1996:10) noted that, "Of course, deep understanding is unlikely to be achieved by a single study, no matter how carefully planned." Hurlbert and White (1993:149) suggested that, although serious statistical errors were rampant in at least 1 area of ecology, principal conclusions, "those concerning phenomena that have been studied by several investigators, have been unaffected." And Catchpole (1989:287) stated that, "Most hypotheses are tested, not in the splendid isolation of one finely controlled 'perfect' experiment, but in the wider context of a whole series of experiments and observations. Surely a much more valuable form of validity comes from the independent repetition of experiments by colleagues in different parts of the world." As summarized by Anderson et al. (2001:312), "In the long run, science is safeguarded by repeated studies to ascertain what is real and what is merely a spurious result from a single study."

MORE ON METAREPLICATION

What's *P* Got to Do With It?

P-values resulting from statistical tests of null hypotheses often are used to judge the significance of findings from a study. A small *P*-value suggests either that the null hypothesis is not true or that an unusual result has occurred. *P*-values often are misinterpreted as: (1) the probability that the results were due to chance, (2) an indication of the reliability of the result, or (3) the probability that the null hypothesis is true (Carver 1978, Johnson 1999). Small *P*-values are taken to represent strong evidence that the null hypothesis is false, but in reality the connection between P and $\Pr\{H_0 \text{ is true}|\text{data}\}$ is nebulous (Berger and Sellke 1987).

R. A. Fisher was an early advocate of *P*-values, but he actually recommended that they be used opposite to the way they are mostly used now. Fisher viewed a significant *P*-value as providing reason to continue studying the phenomenon (recalling that either the hypothesis was wrong or something unusual happened). In stark contrast, modern researchers often use nonsignificant *P*-values as reason to continue study; many investigators, when faced with nonsignificant results, argue that, "a larger sample size [i.e., further research] is needed to reach significance."

The Importance of Consistent Methods in Replication

Scientists are encouraged to replicate studies using the same methods as were used in the original studies. This practice eliminates variation due to methodology and, if different results are obtained, suggests that the initial results may have been an accident (Table 2). That is, they did not bear up under metareplication. Obtaining the same results when using the same methods, however, allows for the possibility that the results were specific to the method, rather than a general truth.

Replication with different methods is critical to determine whether results are robust with respect to methodology and not an artifact of the methods employed. When we get consistent results with different methods, we have greater confidence in those results; the results are robust

Table 2. There are both advantages and disadvantages to replicating a study with the same or different methods as the original study.

Methods of original and replicated studies	Results from original and replicated studies	
	Same	Different
Same	Results may have been specific to method, rather than a general truth	Original results may have been accidental, not bearing up under metareplication
Different	Results are robust with respect to method	Results may have been an artifact of the method used

with respect to method. Should we get different results when different methods are used, the original results may have been artifacts of the methods (Table 2).

What to Do With Surprises?

A cogent argument has been made that only well-thought-out hypotheses should be tested in a study. Doing so avoids "fishing expeditions" and the chance of claiming that accidental findings are real (Johnson 1981, Rexstad et al. 1988, Anderson et al. 2001, Burnham and Anderson 2002). I think that surprise findings in fact should be considered, but not as confirmed results from the study so much as prods for further investigations. They generate hypotheses to test. For example, suppose you conduct a regression analysis involving many explanatory variables. If you use a stepwise procedure to select variables, results from that analysis can give very misleading estimates of effect sizes, *P*-values, and the like (Pope and Webster 1972, Hurvich and Tsai 1990). Variables deemed to be important may or may not actually have major influence on the response variable, and conclusions to that effect should not be claimed. It is appropriate, on the other hand, to use the results in a further investigation, focusing on the explanatory variables that the analysis had suggested were influential. It is better to conduct a new study (i.e., to metareplicate), but at a minimum cross-validation will be useful. In that approach, a model is developed with part of the data set and evaluated on the remaining data. This is not to suggest that a priori hypotheses are not important, or that carefully designed studies to evaluate those hypotheses are not a highly appropriate way to conduct science. Only that a balance between exploratory and confirmatory research is needed. Studies should be designed to learn something, not merely to generate questions for further research. Apparent findings need to be rigorously confirmed. If scientists look only at variables known or suspected to be influential, however, how would we get new findings?

Meta-analysis

Meta-analysis essentially is an analysis of analyses (Hedges and Olkin 1985, Osenberg et al. 1999, Gurevitch and Hedges 2001). The units being analyzed are themselves analyses. Meta-analysis dates back to 1904, when Karl Pearson grouped data from various military tests and concluded that vaccination against intestinal fever was ineffective (Mann 1994). Often studies of comparable effects are analyzed by vote counting: of the studies that looked for the effect, this many had statistically significant results and that many did not. One problem with the vote-counting approach is that, if the true effect is not strong and sample sizes are not large, most studies will not detect the effect. So a critical review of the studies would conclude that most studies found no effect, and the effect would be dismissed.

In contrast, meta-analysis examines the full range of estimated effects (not *P*-values), whether or not they were individually statistically significant. From the resulting pattern may emerge evidence of consistent effects, even if they are small. Mann (1994) cited several instances in which meta-analyses led to dramatically different conclusions than did expert reviews of studies that used vote-counting methods. Meta-analysis does have a serious danger, however, in publication bias (Berlin et al. 1989). A study that demonstrates an effect at a statistically significant level is more likely to be written for publication, favorably reviewed by referees and editors, and ultimately published than is a study without such significant effects (Sterling et al. 1995). So the published literature on an effect may give a very biased picture of what the research in toto demonstrated. (The medical community worries that ineffective and even harmful medical practices may be adopted if positive results are more likely to be published than negative results [Hoffert 1998]. Indeed, an on-line journal, the *Journal of Negative Results in Biomedicine*, is being launched to correct distortions caused by a general bias against null results [Anonymous 2002].) Even if results from unpublished studies could be accessed, much care would be needed to evaluate them. Bailar (1995) observed that quality meta-analysis requires expertise in the subject matter reviewed. A question always looms about unpublished studies (Hoffert 1998): Was the study not published because it generated no statistically significant results or because it was flawed in some way? Further, could it be that the study was not published because it was contrary to the prevailing thinking at the time? Yet, despite the concerns with meta-analysis, it does provide a vehicle for thoughtfully conducting a synthesis of the studies relevant to a particular question.

Weak Studies May Be OK, But ...

Statisticians, including myself (Johnson 1974), regularly advise against conducting studies that

lack sufficient power. Observations in those studies are too few to yield a high probability of rejecting some null hypothesis, even if the hypothesis is false. While large samples are certainly preferable to small samples, I no longer believe that it is appropriate to condemn studies with small samples. Indeed, it may be preferable to have the results of numerous small but well-designed studies rather than results from a single "definitive" investigation. This is so because the single study, despite large samples, may have been compromised by an unusual happenstance or by the effect of a "lurking variable" (a third variable that induces a correlation between 2 other variables that are otherwise unrelated). Numerous small studies, due to the benefits of metareplication, are less at jeopardy of yielding misleading conclusions.

One danger of a small study is that the sampled units do not adequately represent the target population. Lack of representation also can plague larger studies, however. I suspect that the greatest danger of a small study is the tendency to accept the null hypothesis as truth, if it is not rejected. Concluding that a hypothesis is true simply because it was not rejected in a statistical test is folly. Nonetheless, it is done frequently; Johnson (1999, 2002) cited numerous instances in which authors of *The Journal of Wildlife Management* articles concluded that null hypotheses were true, even when samples were small and test statistics were nearly significant.

Metareplication protects against situations in which there is an effect, but it is small and therefore not statistically significant in individual studies, and thus is never claimed. Hence, small studies should not be discouraged, as long as the investigators acknowledge that they are not definitive. Studies should be designed to address the topic as effectively and efficiently as possible. If the scope has to be narrow and the scale has to be small, or if logistic constraints preclude large samples, results still may be worthwhile and should be published, with their limitations acknowledged. Without meta-analysis or a similar strategy, any values of small studies will not be realized.

Should Authors Avoid Management Recommendations?

This journal encourages authors to present management implications deriving from the studies they describe. That practice may not always be appropriate. Results from a single study, unless supported by evidence from other studies, may be misleading. The fact that a study is the only one dealing with a certain species in a particular state is no reason to base management recommendations solely on that single study. Recommendations should be based on a larger body of knowledge. Similarly, manuscripts should be considered for publication even if they are not "groundbreaking," but instead provide support for inferences originally obtained from previous studies.

What about "management studies"? These seem to be studies conducted by others than scientists or graduate students. They also are claimed to be in less need of quality (good design, adequate sample size, etc.) than are "research studies." I would argue that the reverse may in fact be true: Management studies should at least equal research investigations in quality. If an erroneous conclusion is reached in a research study, the only negative consequence is the publication of that error in a journal. And, hopefully, further investigation will demonstrate that the published conclusion was unwarranted. In contrast, an erroneous conclusion reached in a management study may well lead to some very inappropriate management action being taken, with negative consequences to wildlife and their habitats.

Metareplication and the Bayesian Approach

The Bayesian philosophy offers a more natural way to think about metareplication than does the traditional (frequentist) approach. In concept, a frequentist considers only the likelihood function, acting as if the only information about a variable under investigation derives from the study at hand. A Bayesian accounts for the context and history more explicitly by considering the likelihood in conjunction with the prior distribution. The prior incorporates what is known or believed about the variable before the study was conducted. I think people naturally tend to be Bayesians. They have what might be termed mental inertia: they tend to continue in their existing beliefs even in the face of evidence against those beliefs. Only with repeated doses of new evidence do they change their opinions. Sterne and Smith (2001) suggested that the public, by being cynical about the results of new medical studies, were exhibiting a subconscious Bayesianism.

CONCLUSIONS

Any imaginative wildlife biologist can easily list a dozen or more variables that could influence a response variable of interest, be it the number of squirrels in a woodlot, the nest success rate of bobolinks (*Dolichonyx oryzivorus*) in a field, or the survival rate of mallards in a particular year. An

investigation of such a response variable will adequately determine the influences of only a few of the multitude of explanatory variables. The remainder will not be under the investigator's control and indeed may not even be known to the investigator, or may be known but not measured.

The extent to which these other variables influence the response variable confounds the observed relationship between the response variable and the explanatory variables under study. In addition, those unknown influences may restrict the scope of inference for the relationships that are discovered.

Consider again our example of estimating the effect on squirrel density of selectively logging woodlots. Suppose that, in general, such logging does reduce squirrel density. In any particular situation, however, that result might not follow because of the effects of other (possibly unmeasured) variables. Predators of squirrels in a logged woodlot might have been reduced, offsetting any population decline associated with logging. Or an outbreak of disease in the squirrels might have reduced their numbers in the unlogged woodlot, erasing any difference between that woodlot and the one that was logged.

Design control (restricting the range in variation of potentially confounding variables) reduces the influence of such variables, but that practice is not always feasible. Randomization tends to make variables that are not studied act, well, randomly, rather than in some consistent direction. With replication, those variables then contribute to variance in the observed relationship, rather than a bias. Nonetheless, in any single study, those unobserved relationships may give us a misleading impression of the true relationship between the response variable and the explanatory variables under study.

Metareplication provides us greater confidence that certain relationships are general. Obtaining consistent inferences from studies conducted under a wide variety of conditions will assure us that the conclusions are not unique to the particular set of circumstances that prevailed during the study. Further, by metareplicating studies, we need not worry about *P*-values, issues of what constitute independent observations, and other concerns involving single studies. We can take a broader look, seeking consistency of effects among studies. Consistent results suggest generality of the relationship. Inconsistency will lead us either not to accept the results as truth or to determine conditions under which the results hold and those under which they do not. That approach will lead to understanding the mechanisms.

If, indeed, most individual wildlife studies are flawed to some degree, why have we any confidence whatsoever in the science? Perhaps the errors are inconsequential. Or, possibly we don't really believe in those single studies anyway, and don't take action until a clear pattern emerges from disparate studies of the phenomenon. Our innate Bayesianism may be weighting results from an individual study with our prior thinking, based on other things we know or believe.

To conclude, we certainly should use the best statistical methods appropriate for a given data set to maximize the value of those data. However, as Hurlbert (1994:495) wisely noted, "lack of understanding of basic principles and simple methods by practising ecologists is a serious problem, while under-use of advanced statistical methods is not." More important than the methods used to analyze data, we should collect the best data we can. We should use the principles of design—controls, randomization, and replication in manipulative experiments; matching and measuring appropriate covariates in observational studies. And, most critically, studies themselves need to be replicated to have confidence in the findings and their generality. Metareplication exploits the value of small studies, obviates concerns about *P*-values and similar issues, protects against claiming spurious effects to be real, and facilitates the detection of small effects that are likely to be missed in individual studies.

ACKNOWLEDGMENTS

I am grateful to the many statisticians and biologists who have taught me so much as teachers, colleagues, or students. B. R. Euliss helped prepare the manuscript. Special thanks for valuable comments on this manuscript to D. R. Anderson, L. A. Brennan, K. P. Burnham, L. L. Eberhardt, S. H. Hurlbert, W. P. Kuvlesky, Jr., J. D. Nichols, M. R. Riggs, G. A. Sargeant, A. Stewart-Oaten, and an anonymous referee, all of whom nonetheless deserve to be held blameless for remaining faults. Thanks also to L. A. Brennan for the invitation to write the paper.

LITERATURE CITED

ANDERSON, D. R. 2001. The need to get the basics right in wildlife field studies. Wildlife Society Bulletin 29:1294–1297.

———, K. P. BURNHAM, W. R. GOULD, AND S. CHERRY. 2001. Concerns about finding effects that are actually spurious. Wildlife Society Bulletin 29:311–316.

———, ———, AND W. L. THOMPSON. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

ANONYMOUS. 2002. Getting null results into print. Science 296:2137.

BAILAR, J. C., III. 1995. The practice of meta-analysis. Journal of Clinical Epidemiology 48:149–157.

BALSER, D. S., H. H. DILL, AND H. K. NELSON. 1968. Effect of predator reduction on waterfowl nesting success. Journal of Wildlife Management 32:669–682.

BARNARD, G. A. 1982. Causation. Pages 387–389 *in* S. Kotz and N. L. Johnson, editors. Encyclopedia of statistical sciences. Volume 1. Wiley, New York, USA.

BERGER, J. O., AND T. SELLKE. 1987. Testing a point null hypothesis: the irreconcilability of *P* values and evidence. Journal of the American Statistical Association 82:112–122.

BERLIN, J. A., C. B. BEGG, AND T. A. LOUIS. 1989. An assessment of publication bias using a sample of published clinical trials. Journal of the American Statistical Association 84:381–392.

BOX, G. E. P., AND G. C. TIAO. 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association 70:70–79.

BOYD, H. 1981. Prairie dabbling ducks, 1941–1990. Canadian Wildlife Service Progress Notes 119, Ottawa, Ontario, Canada.

BURNHAM, K. P., AND D. R. ANDERSON. 2002. Model selection and multi-model inference: a practical information–theoretic approach. Second edition. Springer, New York, USA.

CARPENTER, S. R. 1990. Large-scale perturbations: opportunities for innovation. Ecology 71:2038–2043.

———. 1996. Microcosm experiments have limited relevance for community and ecosystem ecology. Ecology 77:677–680.

CARVER, R. P. 1978. The case against statistical significance testing. Harvard Educational Review 48:378–399.

CATCHPOLE, C. K. 1989. Pseudoreplication and external validity: playback experiments in avian bioacoustics. Trends in Ecology & Evolution 4:286–287.

CHERRY, S. 1998. Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin 26:947–953.

COCHRAN, W. G. 1983. Planning and analysis of observational studies. Wiley, New York, USA.

COX, D. R., AND N. WERMUTH. 1996. Multivariate dependencies—models, analysis and interpretation. Chapman & Hall, London, United Kingdom.

EBERHARDT, L. L. 1970. Correlation, regression, and density dependence. Ecology 51:306–310.

———. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management 4:27–70.

———, AND J. M. THOMAS. 1991. Designing environmental field studies. Ecological Monographs 61:53–73.

ERICKSON, W. P., T. L. MCDONALD, K. G. GEROW, S. HOWLIN, AND J. W. KERN. 2001. Statistical issues in resource selection studies with radio-marked animals. Pages 209–242 *in* J. J. Millspaugh and J. M. Marzluff, editors. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

FAABORG, J., M. BRITTINGHAM, T. DONOVAN, AND J. BLAKE. 1993. Habitat fragmentation in the temperate zone: a perspective for managers. Pages 331–338 *in* D. M. Finch and P. W. Stangel, editors. Status and management of neotropical migratory birds. U.S. Forest Service General Technical Report RM-229.

FISHER, R. A. 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture 33:503–513.

———. 1971. The design of experiments. Hafner, New York, USA.

GUREVITCH, J. A., AND L. V. HEDGES. 2001. Meta-analysis: combining the results of independent experiments. Pages 347–369 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Second edition. Oxford University Press, Oxford, United Kingdom.

HARVILLE, D. A. 1975. Experimental randomization: Who needs it? American Statistician 29:27–31.

HEDGES, L. V., AND I. OLKIN. 1985. Statistical methods for meta-analysis. Academic Press, San Diego, California, USA.

HEFFNER, R. A., M. J. BUTLER, IV, AND C. K. REILLY. 1996. Pseudoreplication revisited. Ecology 77:2558–2562.

HOFFERT, S. P. 1998. Efforts increase to boost validity of meta-analyses. Scientist 12:7–8.

HOLLAND, P. W. 1986. Statistics and causal inference. Journal of the American Statistical Association 81:945–960.

HURLBERT, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

———. 1994. Old shibboleths and new syntheses. Trends in Ecology & Evolution 9:495–496.

———. 1997. Experiments in ecology [book review]. Endeavour 21:172–173.

———, AND M. D. WHITE. 1993. Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. Bulletin of Marine Science 53:128–153.

HURVICH, C. M., AND C.-L. TSAI. 1990. The impact of model selection on inference in linear regression. American Statistician 44:214–217.

JOHNSON, D. H. 1974. Estimating survival rates from banding of adult and juvenile birds. Journal of Wildlife Management 38:290–297.

———. 1981. The use and misuse of statistics in wildlife habitat studies. Pages 11–19 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. U.S. Forest Service General Technical Report RM-87.

———. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

———. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

———. 2001*a*. Validating and evaluating models. Pages 105–119 *in* T. M. Shenk and A. B. Franklin, editors. Modeling in natural resource management: development, interpretation, and application. Island Press, Washington, D.C., USA.

———. 2001*b*. Habitat fragmentation effects on birds in grasslands and wetlands: a critique of our knowledge. Great Plains Research 11:211–231.

———. 2002. The role of hypothesis testing in wildlife science. Journal of Wildlife Management 66:272–276.

———, J. D. NICHOLS, AND M. D. SCHWARTZ. 1992. Population dynamics of breeding waterfowl. Pages 446–485 *in* B. D. J. Batt, A. D. Afton, M. G. Anderson, C. D. Ankney, D. H. Johnson, J. A. Kadlec, and G. L. Krapu,

editors. Ecology and management of breeding waterfowl. University of Minnesota Press, Minneapolis, USA.

———, AND M. WINTER. 1999. Reserve design for grasslands: considerations for bird populations. Proceedings of the George Wright Society Biennial Conference 10:391–396.

LEVINS, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

LIANG, R. I., AND S. L. ZEGER. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

MACNAB, J. 1983. Wildlife management as scientific experimentation. Wildlife Society Bulletin 11:397–401.

MALLOWS, C. L. 1979. Robust methods—some examples of their use. American Statistician 33:179–184.

MANN, C. C. 1994. Can meta-analysis make policy? Science 266:960–962.

MCCULLAGH, P., AND J. A. NELDER. 1989. Generalized linear models. Second edition. Chapman & Hall, London, United Kingdom.

MILLSPAUGH, J. J., J. R. SKALSKI, B. J. KERNOHAN, K. J. RAEDEKE, G. C. BRUNDIGE, AND A. B. COOPER. 1998. Some comments on spatial independence in studies of resource selection. Wildlife Society Bulletin 26:232–236.

NICHOLS, J. D. 2001. Using models in the conduct of science and management of natural resources. Pages 11–34 *in* T. M. Shenk and A. B. Franklin, editors. Modeling in natural resource management: development, interpretation, and application. Island Press, Washington, D.C., USA.

OSENBERG, C. W., O. SARNELLE, AND D. E. GOLDBERG, editors. 1999. Meta-analysis in ecology: concepts, statistics, and applications. Ecology 80:1103–1167.

PEARL, J. 2000. Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, United Kingdom.

PLATT, J. R. 1964. Strong inference. Science 146:347–353.

POPE, P. T., AND J. T. WEBSTER. 1972. The use of an *F*-statistic in stepwise regression procedures. Technometrics 14:327–340.

PROVENCHER, L., N. M. GOBRIS, L. A. BRENNAN, D. R. GORDON, AND J. L. HARDESTY. 2002. Breeding bird response to midstory hardwood reduction in Florida sandhill longleaf pine forests. Journal of Wildlife Management 66:641–661.

REXSTAD, E. A., D. D. MILLER, C. H. FLATHER, E. M. ANDERSON, J. W. HUPP, AND D. R. ANDERSON. 1988. Questionable multivariate statistical inference in wildlife habitat and community studies. Journal of Wildlife Management 52:794–798.

ROMESBURG, H. C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45:293–313.

RUBIN, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66:688–701.

SCHINDLER, D. W. 1998. Replication versus realism: the need for ecosystem-scale experiments. Ecosystems 1:323–334.

SMITH, E. P. 2002. BACI design. Pages 141–148 *in* A. H. El-Shaarawi and W. W. Piegorsch, editors. Encyclopedia of environmetrics. Volume 1. Wiley, Chichester, United Kingdom.

SMITH, T. M. F., AND R. A. SUGDEN. 1988. Sampling and assignment mechanisms in experiments, surveys and observational studies. International Statistical Review 56:165–180.

STERLING, T. D., W. L. ROSENBAUM, AND J. J. WEINKAM. 1995. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. American Statistician 49:108–112.

STERNE, J. A. C., AND G. D. SMITH. 2001. Sifting the evidence—what's wrong with significance tests? British Medical Journal 322:226–231.

STEWART-OATEN, A. 2002. Pseudo-replication. Pages 1642–1646 *in* A. H. El-Shaarawi and W. W. Piegorsch, editors. Encyclopedia of environmetrics. Volume 3. Wiley, Chichester, United Kingdom.

———, AND J. R. BENCE. 2001. Temporal and spatial variation in environmental impact assessment. Ecological Monographs 71:305–339.

———, W. W. MURDOCH, AND K. R. PARKER. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67:929–940.

TAPPER, S. C., G. R. POTTS, AND M. H. BROCKLESS. 1996. The effect of an experimental reduction in predation pressure on the breeding success and population density of grey partridges *Perdix perdix*. Journal of Animal Ecology 33:965–978.

THOMAS, J. W. 2000. From managing a deer herd to moving a mountain—one Pilgrim's progress. Journal of Wildlife Management 64:1–10.

TUKEY, J. W. 1960. Conclusions vs decisions. Technometrics 2:423–433.

WALTERS, C. 1986. Adaptive management of renewable resources. Macmillan, New York, USA.

WILLIAMS, B. K., J. D. NICHOLS, AND M. J. CONROY. 2002. Analysis and management of animal populations: modeling, estimation, and decision making. Academic Press, San Diego, California, USA.

YOUDEN, W. J. 1972. Enduring values. Technometrics 14:1–11.