The Many Faces of Replication

Douglas H. Johnson*

ABSTRACT

Replication is one of the three cornerstones of inference from experimental studies, the other two being control and randomization. In fact, replication is essential for the benefits of randomization to apply. In addition to ordinary replication, the repetition of treatments within a study, two other levels of replication have been identified. Pseudoreplication, a termed coined by Stuart Hurlbert, generally involves making multiple measurements on experiment units (which is commendable) and treating them as if they reflected independent responses to treatment (which is erroneous). Metareplication is a higher level of replication in which entire studies are repeated. Scientists are too much concerned about analysis of data within studies and too little concerned about the repeatability of findings from studies conducted under a variety of conditions. Findings that are consistent among studies performed at different locations at different times with different investigators using different methods are likely to be robust and reliable.

Fundamental to science is the notion of causation, which is less obvious than it might appear. In the physical sciences, causation is a straightforward concept that implies a law-like consistency between antecedents and consequences. Models of the behavior of atoms, planets, and other inanimate objects are applicable over a wide range of conditions (Barnard, 1982), and there are few factors that control the system (e.g., pressure and temperature are sufficient to determine the volume of a gas). In many nonphysical sciences, however, notions of causality reduce to those of probability, which allows exceptions and lack of regularity. Here, causation means that an action "tends to make the consequence more likely, not absolutely certain" (Pearl, 2000). In wildlife management, a discipline familiar to me, many factors influence a system. For example, liberalizing hunting regulations for a species tends to increase harvest by hunters. In any specific instance, however, liberalization may not cause an increased harvest because of other influences such as the number of animals in the population, weather conditions during the hunting season, and the cost of gasoline as it affects hunter activity. In an agronomic setting, adding P fertilizer may generally be expected to increase the yield of a crop, but it may not happen in a specific instance because another nutrient is limiting, P is already in adequate supply, or moisture is insufficient for the plants to utilize the added P.

U.S. Geological Survey, Northern Prairie Wildlife Research Center, Dep. of Fisheries, Wildlife, and Conservation Biology, 204 Hodson Hall, 1980 Folwell Ave., Univ. of Minnesota, St. Paul, MN 55108. Received 30 Apr. 2006. *Corresponding author (douglas_h_johnson@usgs.gov).

Published in Crop Sci. 46:2486–2491 (2006).
Analysis of Unreplicated Experiments (Symposium)
doi:10.2135/cropsci2006.04.0277

677 S. Segoe Rd., Madison, WI 53711 USA

The objective of this paper is to discuss the role of various kinds of replication in scientific studies. The material is not new; the key points were outlined by R.A. Fisher early in the previous century. This paper also draws freely from Hurlbert (1984), Johnson (2002), Shaffer and Johnson (2007, unpublished data), and other cited sources.

Consider an example (Shaffer and Johnson, 2007, unpublished data). Suppose you want to determine how the yield of a crop is affected by a treatment such as the addition of a certain fertilizer. The treatment effect (T) on a particular plot (u) can be defined as

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

where $Y_t(u)$ is the yield in plot u after the treatment, and $Y_c(u)$ is the yield in that plot if the treatment had not been applied. If the plot is fertilized, then you can observe $Y_t(u)$ but not $Y_c(u)$. If the fertilizer is not applied, then you can observe $Y_c(u)$ but not $Y_t(u)$. This leads to what has been termed the fundamental problem of causal inference: one cannot observe the values of $Y_t(u)$ and $Y_c(u)$ on the same unit (Rubin, 1974; Holland, 1986). That is, any particular plot is either fertilized or not.

Two solutions to this problem have been identified (Holland, 1986). The first requires two units (u_1 and u_2; here, plots) and the assumption that 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 two plots, if not fertilized, would have identical yield, that is, $Y_c(u_2) = Y_c(u_1)$, and, if fertilized, then $Y_t(u_2) = Y_t(u_1)$. We cannot test these assumptions, because one plot was fertilized and the other was not. The assumption can be made more plausible by matching the two units as closely as possible or with evidence that the units are identical. Physicists are more likely to believe that two molecules are identical than agronomists are to consider two plots the same, however.

The second solution has been termed statistical (Holland, 1986). We can consider an expected, or average, causal effect T over all units in some population:

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

where, unlike with the first 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 of units, which is possible to estimate.

It is clear that a control, something to compare with the treated unit, is needed for either approach. In the statistical approach, randomization is often invoked. If, for example, we are to compare yield on a treated plot and an untreated one, we could reach an erroneous

conclusion if the plots had different soils, or had grown different crops the previous year. One way to protect against such possibly misleading outcomes is to decide at random which plot is to be treated and which is not. Random assignment can be done in a controlled experiment but not in most observational studies.

Suppose in our example that there are four plots in our area of interest. And suppose, following the first solution, that they are identical: each would yield 300 kg if it were fertilized and 100 kg if not (Table 1). Then, no matter which plot was selected for treatment and which was the comparison, we would estimate the treatment effect to be 200 kg, which is just right. But suppose that the plots themselves varied; in this example we will maintain the unrealistic but convenient simplifying assumption that the treatment effect would be 200 kg no matter which plot was treated (Table 2). Then, if we treat one plot and observe another as a control, there are 12 possible combinations that could constitute our sample (Table 3). Our estimate of the treatment effect would vary, depending on which plots were selected. For example, if Plot 1 was fertilized and Plot 3 served as a control, we would estimate our treatment effect as $300 - 0 = 300$ kg. The 12 possible estimates of treatment effect range from -200 to 600 kg. The average is 200 kg, the correct value, but no possible sample would yield exactly that value.

So, even if treatments are assigned at random, it may just happen that one plot has better soils (possibly Plot 4 in our example), and the other does not (Plot 3). And, such a sample would generate an estimated treatment effect (600 kg) very different from the correct value (200 kg). This consideration leads to the third important criterion for determining causation: replication. Repeating the randomization process and treatment application on several plots makes it unlikely that plots in either group are consistently more favorable. If we take a sample of Size 2 for both treatment and control groups, there are six possible samples, with estimated treatment effects ranging from -100 to 500 kg (Table 4). Note that samples with one plot each in the treatment and in the control group yield estimates that vary around the true value, but are very spread out (Fig. 1, top), whereas samples of Size 2 in each group gives estimates that cluster somewhat more closely around the true value (Fig. 1, bottom). The smaller the sample, the more likely is a wildly misleading result.

These then form the cornerstones for assessing the effect of some treatment with a manipulative experiment: a control, randomization, and replication (Fisher, 1926). The need for a control is obvious and will not be discussed further here. We will explore the functions of randomization and replication more fully.

Table 1. Example of four field plots, the value (yield in kilograms) each would have if it were treated (fertilizer applied), and the value it would have if it were not treated. Note that in this example all plots have identical values under each scenario, and the effect of the treatment is 200 for all plots.

Plot	Value if treated	Value if not treated
1	300	100
2	300	100
3	300	100
4	300	100

Table 2. Example of four field plots, the value (yield in kilograms) each would have if it were treated, and the value it would have if it were not treated. Note that plots vary in values irrespective of treatment, but the effect of the treatment is 200 for all plots.

Plot	Value if treated	Value if not treated
1	300	100
2	500	300
3	200	0
4	600	400

Randomization serves three roles. One is to make 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. Second, randomization reduces any intentional or unintentional bias of the investigator. And third, because all outcomes are equally likely, randomization provides an objective probability distribution that can provide the basis for a test of significance (Barnard, 1982).

But, randomization by itself is not enough; 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. Randomization, for example, makes errors act randomly, rather than in some consistent fashion. However, 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 provides two important 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 if 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.

Table 3. All possible samples of Size 1 each, treated and untreated, from the population of field plots described in Table 2.

Treated plot	Untreated plot	Difference
1	2	0
1	3	300
1	4	−100
2	1	400
2	3	500
2	4	100
3	1	100
3	2	−100
3	4	−200
4	1	500
4	2	300
4	3	600

Table 4. All possible samples of Size 2 each, treated and untreated, from the population of field plots described in Table 2.

Treated plots	Untreated plots	Difference
1, 2	3, 4	200
1, 3	2, 4	−100
1, 4	2, 3	300
2, 3	1, 4	100
2, 4	1, 3	500
3, 4	1, 2	200

Manipulative experimentation is a very effective way to determine causal relationships. One poses questions to nature via experiments, such as fertilization of plots. By manipulating the system, an investigator reduces the chance that something other than the treatment caused the results that were observed. Scientists in many disciplines, however, face severe difficulties meeting the requirements of control, randomization, and replication associated with manipulative experiments. Many systems are too large and complex to be manipulated (Macnab, 1983). Often, *treatments* such as natural or human-caused disasters are applied by others, and scientists attempt to evaluate their effects. In such situations, randomization is impossible and replication undesirable. Some 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. Although observational studies lack the critical element of control by the investigator, 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.

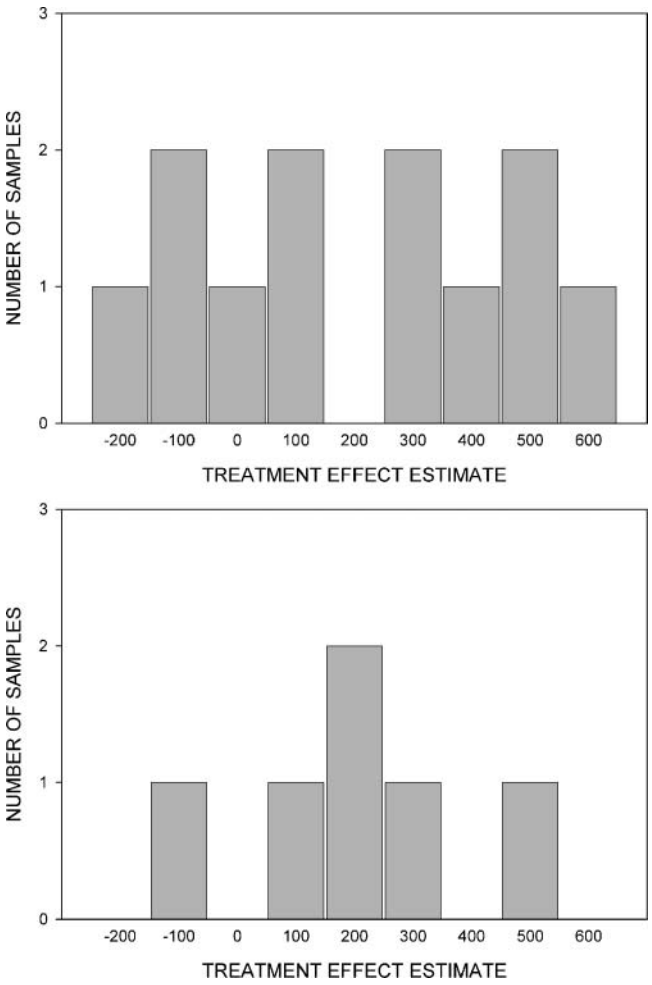

Fig. 1. Numbers of possible samples of Size 1 (top) and 2 (bottom) from plots described in Table 2 and their estimated treatment effects.

Longitudinal observational studies, with measurements taken before and after some treatment, are generally more informative than cross-sectional observational studies, in which treated and untreated units are studied only after the treatment (Cox and Wermuth, 1996). Intervention analysis is one method applicable for assessing the effect of a distinct treatment (intervention) that has been applied to a system. The intervention is 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 following 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 time 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.

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 plant physiology involving 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 feature). In contrast, features 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 important than the randomness of a sample. Can one comfortably draw an inference to a population from a sample, even if that sample is nonrandom? Most useful inferences involve extrapolation beyond the sampled population. Often we want to predict the consequences of some action that will be performed in the future, based on a study conducted in the past; in such a situation, we are extrapolating forward in time.

OTHER LEVELS OF REPLICATION

Three levels of replication have been identified (Johnson, 2002) (Table 5). The basic notion is of ordinary replication in an experiment: treatments are applied independently to several units. In our fertilization example, we would want several plots to be fertilized and several to be left as controls. (Comparable considerations apply to observational studies or sample surveys.) Such replication helps to ensure against making a decision based on a single outcome of the treatment, which may have been atypical. It also provides an estimate of the variation associated with the treatment. Other levels of replication are pseudoreplication and metareplication.

Pseudoreplication

At a lower level than ordinary replication is what Hurlbert (1984) called pseudoreplication. Pseudoreplication often is discussed in terms associated with the ANOVA (using the wrong error term in an analysis), but it usually 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) included in what he termed pseudodesign. Pseudoreplication was common when Hurlbert (1984) surveyed the literature on manipulative ecological experiments, mostly published during 1974 to 1980, and estimated that about 27% of the experiments involved pseudoreplication. It continues to be a problem (Heffner et al., 1996). Stewart-Oaten (2002) offered some keys for recognizing pseudoreplication, which is not always obvious.

Metareplication

At a higher level than ordinary replication is what I have termed metareplication (Johnson, 2002). Metareplication involves the replication of entire studies, preferably in different years, at different sites, with different methodologies, or by different investigators. Conducting studies under a variety of circumstances reduces the chance that some artifact associated with a particular situation caused the observed results; it is unlikely that an unusual set of circumstances would manifest itself repeatedly in time or, especially, at multiple sites. Conducting studies with different methods similarly reassures us that the results were not simply due to the methods or equipment used to produce those results. And having more than one investigator perform studies of similar phenomena reduces the chance that some hidden bias or characteristic of a researcher influenced the results. 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.

A classic example of the need for metareplication was provided by Youden (1972), who described the sequence of 15 studies conducted during 1895–1961 to estimate the average distance between Earth and the sun. Each study yielded an estimate of that distance and a confidence interval for the estimate. As it turned out, rather surprisingly, each estimate was outside the confidence interval for the previous estimate. Clearly each investigator had more confidence in his estimate than was warranted. The key point is that we should have less confidence in any individual study than internal estimates of reliability would lead us to believe. The example also emphasizes the need to conduct studies of any phenomenon under different circumstances, with different methods, and by different investigators. That is what I called metareplication. Independent studies of some phenomenon each may suffer from their own shortcomings, but if they point to substantially similar conclusions, we can have confidence in them.

Clearly, the idea to replicate studies is not new. Repetition of key experiments by others, in fact, has been standard practice in science far longer than statistics itself (Carpenter, 1990). Tukey (1960) argued that conclusions develop from considering a series of individual results, rather than a particular result. Eberhardt and Thomas (1991) wrote: "truly definitive single experiments are very rare in any field of endeavor, progress is actually made through sequences of investigations." Hurlbert and White (1993) suggested that, although serious statistical errors were rampant in at least one discipline, the principal conclusions, "those concerning phenomena that have been studied by several investigators, have been unaffected." And Catchpole (1989) 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." And, "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" (Anderson et al., 2001).

Table 5. The types of replication differ in what actions are repeated, what scope of inference is valid, and the role of *P* values (Johnson, 2002).

Term	Repeated action	Scope of inference	*P* value	Analysis
Pseudoreplication	measurement	object measured	wrong	pseudoanalysis
Ordinary replication	treatment	objects for which samples are representative	ok	analysis
Metareplication	study	situations for which studies are representative	irrelevant	meta-analysis

MORE ON METAREPLICATION

Meta-Analysis

Meta-analysis fundamentally is an analysis of analyses, in which the units being analyzed are themselves analyses (Hedges and Olkin, 1985; Osenberg et al., 1999; Gurevitch and Hedges, 2001). Meta-analysis dates back at least to 1904, when Karl Pearson grouped data from several military tests to conclude that vaccination against intestinal fever was ineffective (Mann, 1994). A proper meta-analysis considers the full range of estimated effects, regardless of their individual statistical significance. The resulting pattern may show evidence of consistent effects, even if the effects are small. Mann (1994) cited several instances in which meta-analyses led to markedly different conclusions than did reviews of studies based on significance levels. A serious danger with meta-analysis, however, is publication bias (Berlin et al., 1989): a study that demonstrates a statistically significant effect is more likely to be submitted for publication, positively viewed by referees and editors, and ultimately published than is a study without such significant effects (Sterling et al., 1995). The published literature on an effect may not offer an unbiased view of what the collective body of research on the effect actually demonstrated.

Metareplication and the Bayesian Approach

The Bayesian philosophy offers a more natural way to think about metareplication than does the frequentist approach traditionally adopted. 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 analysis accounts for the context and history more explicitly by considering the likelihood in conjunction with the prior distribution. The prior incorporates what was known or believed about the variable before the study was conducted.

Replication and the Scope of Inference

The level of replication, as described here, is closely associated with the scope of inference of a study (Table 5). If measurements are repeated within a unit (pseudoreplication), inferences are appropriately drawn only to that unit. If treatments within a study are repeated (ordinarily replication), the scope of inference can validly be considered the population for which the units are representative. If entire studies are repeated (metareplication), then the appropriate scope of inference consists of all situations for which those studies are representative. And, the broader the range of situations, the broader the scope of inference.

CONCLUSIONS

Metareplication provides us greater confidence that identified 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 (Johnson, 2002). We can take a broader look, seeking consistency of effects among studies. Consistent results suggest generality of the relationship. Inconsistency will lead us either to not 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 a better understanding of the mechanisms. Metareplication exploits the value of small studies, protects against claiming spurious effects to be real, and facilitates the detection of small effects that are likely to be missed in individual studies (Johnson, 2002).

ACKNOWLEDGMENTS

I am grateful to R.R. Cox, Jr., D.L. Larson, and M.M. Rowland for comments on an earlier version of this paper, and to S. Machado for organizing the symposium that prompted it.

REFERENCES

Anderson, D.R., K.P. Burnham, W.R. Gould, and S. Cherry. 2001. Concerns about finding effects that are actually spurious. Wildl. Soc. Bull. 29:311–316.

Barnard, G.A. 1982. Causation. p. 387–389. *In* S. Kotz and N.L. Johnson (ed.) Encyclopedia of statistical sciences. Vol. 1. John Wiley & Sons, New York.

Berlin, J.A., C.B. Begg, and T.A. Louis. 1989. An assessment of publication bias using a sample of published clinical trials. J. Am. Stat. Assoc. 84:381–392.

Box, G.E.P., and G.C. Tiao. 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc. 70:70–79.

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

Catchpole, C.K. 1989. Pseudoreplication and external validity: Playback experiments in avian bioacoustics. Trends Ecol. Evol. 4: 286–287.

Cochran, W.G. 1983. Planning and analysis of observational studies. John Wiley & Sons, New York.

Cox, D.R., and N. Wermuth. 1996. Multivariate dependencies—models, analysis and interpretation. Chapman & Hall, London.

Eberhardt, L.L. 1976. Quantitative ecology and impact assessment. J. Environ. Manage. 4:27–70.

Eberhardt, L.L., and J.M. Thomas. 1991. Designing environmental field studies. Ecol. Monogr. 61:53–73.

Fisher, R.A. 1926. The arrangement of field experiments. J. Ministry Agric. Great Britain 33:503–513.

Gurevitch, J.A., and L.V. Hedges. 2001. Meta-analysis: Combining the results of independent experiments. p. 347–369. *In* S.M. Scheiner and J. Gurevitch (ed.) Design and analysis of ecological experiments. 2nd ed. Oxford Univ. Press, Oxford, U.K.

Hedges, L.V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, San Diego, CA.

Heffner, R.A., M.J. Butler, IV, and C.K. Reilly. 1996. Pseudoreplication revisited. Ecology 77:2558–2562.

Holland, P.W. 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81:945–960.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54:187–211.

Hurlbert, S.H., and M.D. White. 1993. Experiments with freshwater invertebrate zooplanktivores: Quality of statistical analyses. Bull. Mar. Sci. 53:128–153.

Johnson, D.H. 2002. The importance of replication in wildlife research. J. Wildl. Manage. 66:919–932.

Macnab, J. 1983. Wildlife management as scientific experimentation. Wildl. Soc. Bull. 11:397–401.

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

Osenberg, C.W., O. Sarnelle, and D.E. Goldberg (ed.). 1999. Meta-analysis in ecology: Concepts, statistics, and applications. Ecology 80:1103–1167.
Pearl, J. 2000. Causality: Models, reasoning, and inference. Cambridge Univ. Press, Cambridge, U.K.
Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66:688–701.
Smith, E.P. 2002. BACI design. p. 141–148. *In* A.H. El-Shaarawi and W.W. Piegorsch (ed.) Encyclopedia of environmetrics. Vol. 1. Wiley, Chichester, U.K.
Smith, T.M.F., and R.A. Sugden. 1988. Sampling and assignment mechanisms in experiments, surveys and observational studies. Int. Stat. Rev. 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. Am. Stat. 49:108–112.
Stewart-Oaten, A. 2002. Pseudo-replication. p. 1642–1646. *In* A.H. El-Shaarawi and W.W. Piegorsch (ed.) Encyclopedia of environmetrics. Vol. 3. Wiley, Chichester, U.K.
Stewart-Oaten, A., and J.R. Bence. 2001. Temporal and spatial variation in environmental impact assessment. Ecol. Monogr. 71:305–339.
Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67:929–940.
Tukey, J.W. 1960. Conclusions vs. decisions. Technometrics 2:423–433.
Youden, W.J. 1972. Enduring values. Technometrics 14:1–11.