Journal of Comparative Psychology
2009, Vol. 123, No. 4, 447–449

0735-7036/09/$12.00 DOI: 10.1037/a0016094

Trade-Offs in the Design of Experiments

R. Haven Wiley
University of North Carolina, Chapel Hill

This comment supplements and clarifies issues raised by J. C. Shank and T. C. Koehnle (2009) in their critique of experimental design. First, the pervasiveness of trade-offs in the design of experiments is emphasized (Wiley, 2003). Particularly germane to Shank and Koehnle's discussion are the inevitable trade-offs in any decisions to include blocking or to standardize conditions in experiments. Second, the interpretation of multiple tests of a hypothesis is clarified. Only when interest focuses on *any*, rather than *each*, of *N* possible responses is it appropriate to adjust criteria for statistical significance of the results. Finally, a misunderstanding is corrected about a disadvantage of large experiments (Wiley, 2003). Experiments with large samples raise the possibility of small, but statistically significant, biases even after randomization of treatments. Because these small biases are difficult for experimenters and readers to notice, large experiments demonstrating small effects require special scrutiny. Such experiments are justified only when they involve minimal human intervention and maximal standardization. Justifications for the inevitable trade-offs in experimental design require careful attention when reporting any experiment.

Keywords: experimental design, blocking, multiple tests, analysis of variance, sample size

Shank and Koehnle (2009) review many issues in experimental design in their discussion of pseudoreplication as characterized by Hurlbert (1984). Whether their specific criticisms of Hurlbert are correct, I plan to leave to others to determine. I presume that few would disagree with their overall conclusions that experiments should always (1) include rigorous attempts to identify and reduce unintended influences of one subject (or other experimental unit) on another and (2) make appropriate use of any multilevel (nested) statistical designs. Furthermore, there can be no argument that (3) treatments must be assigned randomly to experimental units.

Instead, this note focuses on three issues that seem inadequately emphasized by Shank and Koehnle (2009): (1) the pervasiveness of trade-offs in the design of experiments, (2) the inappropriate and appropriate uses of multiple tests of a hypothesis, and (3) the dangers that counteract the benefits of large samples.

Trade-Offs Are Pervasive in Experiments

My previous note on behavioral experiments (Wiley, 2003) emphasized that any decision about experimental design (other than correct application of statistics and randomization) involves trade-offs. To avoid repeating that discussion, here I mention two examples germane to Shank and Koehnle's (2009) discussions.

First, the decision to include blocking in an experimental design requires repetition of a treatment on any one experimental unit. It might involve measuring the response of each individual more than once or obtaining measurements in each location or in each block of time more than once. The advantage of this blocking is the information it yields about variation among blocks (individuals, localities, times). This information can be used to test what might be called the secondary hypotheses of the experiment, the sources of variation that affect the responses to the treatment. The disadvantage is the lower sample size available for testing the primary hypothesis, the overall effect of the treatment on the response. Whenever there are limitations on the overall number of individuals, localities, or other units to be studied (the overall sample size), the decision to include blocking in an experiment can lower the chance of reaching a decision about the primary hypothesis.

There is another advantage of blocking by obtaining repeated measures of each experimental unit or subject. Even if the units were not expected to differ intrinsically in response to treatments, there remains the possibility of errors of measurement. If the accuracy or precision of measurements were low, it could happen that variation in repeated measurements of any one unit were greater than the variation in means between units under any one treatment or even between treatments. In this case, repeated measurements of each experimental unit, treated as a block, would produce more reliable estimates of the means within each block and thus improve the chances of detecting differences between treatments. Here, too, blocking requires a trade-off whenever there are limitations on the number or cost of measurements. The advantage of more accurate and precise characterization of each experimental unit is then offset by the disadvantage of fewer degrees of freedom for testing the primary hypothesis.

Trade-offs are also prominent in any attempt to standardize the features of experimental units to reduce the influence of confounding variables. Shank and Koehnle recognize that standardizing conditions is necessary in most experiments to reduce the effects of irrelevant variables on responses. In practice, all experiments must include some standardization of experimental units, simply as a result of identifying the pool of available units or subjects and the locations and times of the experiment. Shank and Koehnle, how-

R. Haven Wiley, Department of Biology, University of North Carolina, Chapel Hill.

Correspondence concerning this article should be addressed to R. Haven Wiley, Department of Biology, CB-3280, University of North Carolina, Chapel Hill, Chapel Hill, NC 27599-3280. E-mail: rhwiley@email.unc.edu

ever, do not emphasize enough that standardization of conditions requires imperfect human intervention.

Furthermore, standardization often involves a trade-off with the generality of the results. To the extent that conditions and subjects are restricted to particular criteria, the results become constrained by those conditions. The advantage of reducing variation in the responses of subjects is thus offset by a disadvantage in the generality of the results. The issue is thus not so much whether or not to standardize conditions or subjects but to define explicitly the criteria for standards and to justify the trade-off between precision and generality.

Adjustments for Multiple Measures of Response

This issue in experimental design has attracted repeated attention. When studies include multiple tests of the same hypothesis, it seems plausible to adjust the criteria for a decision to accept the hypothesis. After all, if we adopt a criterion of 5% for Type I errors, we expect that 1 in every 20 tests will erroneously fail to reject the null hypothesis; so 20 tests of the same alternative hypothesis have a high chance that one will confirm it.

On the other hand, why should 19 negative results invalidate a single positive one? The history of science must include many instances when a single positive test, after many failures, finally measured the right response and showed the way forward. As a result, some have advocated abandoning any corrections to the criteria for decisions to accept hypotheses. Each test should instead be evaluated on its own merits.

The resolution of these conflicting views comes from attention to the rationale for the multiple measurements. The question is: Is the experimenter interested in *each of N* separate possible responses to the treatment? Or is the experimenter interested in *any of N* possible responses to the treatment?

In the first case, following *N* different experiments, each of which measured a different response to the same treatment, it would make no sense to adopt more stringent criteria than usual for accepting each of the *N* hypotheses. It makes no more sense to do so after measuring the same response to *N* different treatments or after measuring *N* responses to *N* treatments.

Contrast these experiments to one that measured *N* responses to the same treatment but then accepted *any one* positive response as evidence that the treatment had *any* effect. For instance, an epidemiological study of the risks of exposure to some agent might compare two groups of subjects, one exposed to the agent and the other not, and measure 20 aspects of each subject's health (e.g., the occurrence of a number of carcinomas, a number of birth defects in offspring, and a number of indices of physical condition). If the study concluded that the agent had an effect on certain features of health but not on others, a single positive result would suggest a subsequent study to address that particular effect of the agent. If, in contrast, the study concluded that the agent influenced health nonspecifically, provided that any one of the 20 measures met the criterion, then there would be reason to think that a weaker criterion had in fact been adopted for this conclusion. Whether a conclusion is based on *any of N* results or on *each of N* results makes a difference in how we design the next experiment. For the two kinds of conclusions to have the same footing, different criteria must be adopted.

Large Samples Risk Small Biases

Shank and Koehnle cite the discussion in my earlier article about one consequence of sample size:

> It is possible to deliberately avoid detection of small block effects, for instance by using small sample sizes (Wiley, 2003), but researchers should aim for the detection of all biases imposed by an experimental design, small or large, to facilitate progress in the testing of theories. (Shank & Koehnle, 2009, p. 430)

This summary misses the point. Of course, investigators should aim to detect all possible biases, and of course they should aim to detect the smallest effects of the treatment possible. There is a clear trade-off when selecting a sample size between the advantage of detecting a smaller effect and the disadvantage of the time and expense required. Shank and Koehnle, however, missed the point in my discussion that there is also another disadvantage of large samples: they risk small biases, ones difficult to detect.

Large samples allow detection of small effects because of the influence of sample size on the estimation of sample means. For samples drawn from the same population, characterized by some standard deviation, the standard error of the mean of a sample varies approximately inversely with the sample size. Small samples have more variation among samples, and this variation is a source of bias. Statistical tests take account of the sample size for just this reason.

Bias is some initially unsuspected systematic difference between the experimental units (subjects) for different treatments. Randomization of treatments among subjects reduces bias. Nevertheless, in the real world it does not usually remove all systematic differences between treatments, because randomization does not assure that all features of experimental units are equally distributed among different treatment groups. In any real experiment, with a finite sample size, randomization cannot assure that all possible confounding variables are equalized in different treatment groups.

If the sample size is small, an experiment can only detect a large effect of the treatment. Likewise, only a large difference in a confounding variable can produce an apparent effect of the treatment. A large difference in a confounding variable (a large bias) is likely to be noticed by the investigator or by reviewers.

A large experiment is also subject to bias remaining after randomization of a finite sample. Of course, large random samples are less likely to have large differences. Small differences remain likely, however. Small differences from bias in large samples can reach a criterion for statistical significance, just as small differences from treatments can. In other words, large samples are both more likely to show small but statistically significant effects of a treatment and more likely to have small but statistically significant biases. The trouble arises because small unintentional biases are less likely to attract the attention either of the investigator or of reviewers. Large experiments are thus at greater risk from unanticipated unintentional bias than are small experiments. Large experiments leading to conclusions about small effects need extra vigilance.

The choice of sample size thus involves more than one trade-off. There is the trade-off between the possibility of discovering small effects and the necessary time and expense. There is also a trade-off between the size of an experiment and the risk of small biases. Experiments that minimize the risks of unintentional biases, those with minimal human intervention (mechanical or double-blinded) and with maximal standardization of subjects and conditions,

might justify large samples. Experiments under conditions that require more human intervention and face intrinsic difficulties in standardization, such as experiments in the field, cannot so easily justify searches for small effects in large samples.

Conclusion

Shank and Koehnle (2009) make reasonable cases for multilevel (nested) experimental designs and standardization of conditions in experimental designs. However, they fail to emphasize the trade-offs required for any experiment. In particular, there is a trade-off between blocking or nesting experimental units and degrees of freedom for testing the primary hypothesis. There is also a trade-off between standardizing conditions and generalizing conclusions. Because Shank and Koehnle fail to emphasize the importance of trade-offs in experimental design, they fail to emphasize the importance of justifying these trade-offs in any report of an experiment.

This note has also attempted to clarify the use and misuse of multiple tests of hypotheses. The important distinction is between conclusions based on *each of N* tests and those based on *any of N* tests. Finally, this note has reemphasized an overlooked disadvantage of large experiments. Large samples risk detecting small biases, ones inherently difficult for investigators and readers to spot. Only experiments with minimal human intervention and maximal standardization can justify large samples to seek small effects. The advantage of large samples for detecting small effects is balanced by the disadvantage of discovering small biases.

All experiments involve multiple trade-offs. All require clear justifications for these trade-offs.

References

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs, 51,* 187–211.

Shanke, J. C., & Koehnle, T. C. (2009). Pseudoreplication is a pseudoproblem. *Journal of Comparative Psychology, 123,* 421–433.

Wiley, R. H. (2003). Is there an ideal behavioural experiment? *Animal Behaviour, 66,* 585–588.

Received October 31, 2008
Revision received March 9, 2009
Accepted March 9, 2009 ■