Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 15, April, 2001 ISSN=1531-7714

Replication: A Design Principle for Field Research

William D. Schafer
University of Maryland

This article suggests the routine use of replications in field studies. Since replications are generally independent, it is usually possible to synthesize them quantitatively using meta-analysis, a technique heretofore associated primarily with amalgamating prior work. It is argued that the use of replication as a feature in data collection and quantitative synthesis for data analysis is especially attractive for those investigators whose research paradigm choices are limited because they are working in field environments. Two examples are described briefly.

Control of extraneous variables is a fundamental condition to causal interpretations of research (Johnson, 2001). Randomization of participants to treatment conditions has long been considered a powerful method of control, so much so that this is the distinguishing characteristic between true experimental and other types of research (Campbell & Stanley, 1963). When a researcher uses randomization, it is clear that the basis upon which participants receive treatment conditions is unrelated except by chance to any variable that can be confounded with the treatments.

A great deal of research is done in field settings in education. State-level or district-based researchers, for example, are often interested in practical interventions that can occur naturally in schools. However, randomization is typically unavailable to those who work in field settings because the investigator is not able to manipulate treatment conditions at the level of the individual participant. This often arises because institutions such as schools are reluctant to move participants (e.g., students) from group to group (e.g., class to class) or otherwise assign them to groups according to researcher needs. Similarly, it may not be possible even to determine randomly which group receives which treatment condition, that being decided through other means, such as teacher choice.

Failing randomization, one approach used in the field is to measure extraneous variables and employ statistical control (e.g., analysis of covariance). Pedhazur (1997) describes three common contexts for statistical control with intact groups: attempting to equate them on the outcome variable(s) using one or more pretest(s), attempting to control for other variable(s) in looking at mean differences, and attempting to control for other variable(s) in looking at differences in regressions. He points out that these are usually invalid uses of analysis of covariance.

Because statistical procedures are generally less effective than experimental control, theoretical inferences about relationships observed in field settings are often subject to multiple reasonable internal validity threats. And in many cases it is not even possible to measure extraneous variables effectively, such as when limited time is available, when the number of participants in the research is limited, or when the measurement is too intrusive. Johnson (2001) has recently concluded that there is little that can be gained from a single, non-experimental research study. A feasible alternative that can enhance the ability of field investigators to draw causal inferences in field settings clearly would be an advantage.

In field contexts, there are typically many opportunities available to investigators that are not open to researchers in more controlled settings. Laboratory researchers commonly have small pools of potential participants to select from and may need to expend nontrivial resources to obtain their cooperation. On the other hand, in applied settings such as classrooms and schools, and especially for employees of the institution, students or other participants are often generously available as long as the intrusion of the research is minimal. Many investigators in the field thus have broad feasible research opportunities that laboratory researchers do not enjoy. It is therefore possible in common applied research settings to be able to repeat, or replicate, a study design more than once.

It is argued here that careful planning of replications can enhance the interpretability of applied research. When results are consistent across several studies, there is a stronger basis for observed relationship(s) than the support that is available within each study by itself, since results that have been replicated are considered more likely to generalize (continue to be observed). It is also possible to compare the studies with each other to identify constructs that interact with, or moderate, relationships. Although these advantages exist whether or not the research includes experimental control, the opportunity to replicate a basic study design in multiple field contexts is more likely to be available to the applied researcher and is a technique that can lead to stronger inferences in any setting. Thus, it is recommended that persons who conduct field research try to include replication as a fundamental feature in their studies.

The analysis of the several studies' results should also be addressed. Meta-analysis is an attractive vehicle for combining, or synthesizing, a series of research replications. Although meta-analysis is generally thought of as a means for studying an existing research literature quantitatively, it also may be used to analyze a series of related studies generated within a single project.

In the remainder of this article, pertinent features of meta-analysis are discussed briefly and then two examples are described in which multiple replications of a basic field design have been analyzed using meta-analysis to strengthen the evidence available. The basic designs differ markedly in the two examples. Finally, some design approaches for applied researchers thinking about using replications are discussed.

Meta-Analysis

Meta-analysis is commonly used to synthesize the findings of multiple, but related, research studies. Those who are unfamiliar with meta-analysis can find a brief overview along with a completely analyzed example in Schafer (1999). More extensive discussions on a broad array of topics pertinent to meta-analysis are widely available in Hedges & Olkin (1985) and Cooper & Hedges (1994).

Fundamental to meta-analysis is an effect-size measure calculated within a study. An effect-size measure may be used to compare two groups or to relate two variables. For example, the difference between two group means divided by the pooled standard deviation of the two groups in a study might be the effect-size measure [when adjusted for bias, this is Hedges & Olkin's (1985) d index]. Another might be the correlation between two variables in a study. In general, to be used in a meta-analysis, an effect-size measure must be capable of transformation to a normally distributed statistic with a known variance. Under reasonable assumptions, both the examples here are appropriate.

Techniques are described in the cited sources that allow a researcher to model the size of the effect (the effect-size index) as a result of study characteristics. That is, equations may be written, as in multiple regression, for relationships between study characteristics as predictors and an effect-size index as the criterion. These study characteristics may be descriptive of the participants, of the settings, of the treatment implementations, or of the outcome variables; in other words, virtually anything that can differentiate studies from each other can be used in the analysis as study characteristics.

In one typical approach to meta-analysis, an effect-size index is calculated for each study. The suitably weighted average of the effect sizes is tested against a null hypothesis of zero. Variation of the studies' effect sizes about the average is tested to determine whether it is at a greater-than-chance level and, if it is, then a study characteristic may be entered into a model (equation), so that effect size is then predicted as the sum of a constant (intercept) and a study characteristic scaled with (multiplied by) a slope estimate. The slope estimate is tested against the null hypothesis of zero. The variance of the effect sizes about the model (the residuals) is compared with the chance level. If homogeneity (chance-level variance) is achieved, modeling ceases; otherwise further study characteristics are added to the model. Of course, variations exist, some as solutions to special problems that may arise; only a very (over)simplified treatment is described here.

Example 1: Descriptive Gains for Schools

A descriptive, or non-experimental, design is one in which there is no manipulation of treatments. The research problem studied in Guthrie, Schafer, Von Secker, & Alban (2000) was the relationships between instructional characteristics of schools and the variation they showed in their degrees of gain or loss in student achievement over a year's time (growth). The effect-size index was the bias-corrected difference between school means at a target grade level between year one and year two on a statewide, standardized test, divided by the pooled standard deviation for the two years. The indexes were scaled so a positive difference showed improvement. The study was replicated in all six tested content areas at both tested grade levels in all 33 schools in three volunteer districts for a total of 396 effect sizes.

The independent variables in the meta-analysis were school means for teacher-reports of emphasis devoted to different approaches in reading instruction. All teachers in each school were surveyed on a questionnaire with six subscales that had been developed through factor analysis using data from a fourth volunteer district in an earlier study.

The meta-analyses were used to evaluate the association of the set of six instructional variables to achievement growth, of each variable individually to growth, and of each variable as a unique predictor of growth in a six-predictor model. The six content areas at each of the grade levels were analyzed separately. The results of the syntheses were interpretable and generally consistent with an extensive literature review for these variables.

Although it is statistically possible to compare the two years of data for any one school, that single finding by itself would not have been remarkable. While the school might have developed instructional hypotheses for the direction and degree of growth observed, there would have been far too many plausible competing explanations for the difference, such as teacher turnover, test form calibrations, and student aptitude, for example. While the replicated study cannot entirely substitute for experimental control through randomization, the plausibility of at least some of the rival explanations is decreased if instructional explanations can be observed across replications, as they were in this example study. Indeed, only by replicating the fundamental growth-study design was it possible to study the instructional characteristics of the schools as variables used to explain differences among gains across schools.

Example 2: Static Group Comparisons

A static group comparison design is one in which intact groups are randomly assigned to treatments (Campbell & Stanley, 1963). Schafer, Swanson, Bené, & Newberry (2001) studied the effects on student achievement of a treatment consisting of a workshop for high school teachers centering on an instructional method (use of rubrics). Districts nominated teacher pairs within content areas, with the classes for the two members of a given pair chosen to consist of students with similar abilities. There were 46 teacher pairs who provided complete data, evenly divided among four instructional content areas (92 teachers and 3,191 students supplied useable data in the study).

The two teachers in each pair were randomly assigned by coin flip to treatment or control conditions. The treatment, attended by one teacher from each pair, was the experimental manipulation. At the end of the study's duration, each student received a test consisting of two parts, a selected-response section and a constructed-response section. The nature of the study suggested that these two parts might yield different results and so effect sizes were calculated separately for each of the item formats. Each effect size was the difference between the means of the two classes divided by the pooled standard deviation and scaled such that a positive effect size favored the treatment.

This study was part of a larger study that required more than one form of the test. Accordingly, there were three forms in each content area. They were distributed randomly in each classroom, yielding six effect sizes (two formats on each of three forms) for each of the 46 teacher pairs, or 276 effect sizes across the four content areas.

Although there were six non-equated test forms in each of four distinct content areas, it was possible to synthesize the results of these disparate conditions in one analysis and to differentiate the findings in a planned way by contents and by forms. An interpretable pattern of outcomes was obtained and related to prior literature.

In general, there are too many competing plausible rival explanations for observed achievement differences between the two intact groups for this study's single-replicate design, in isolation, to be interesting as evidence for a difference between the instructional methods. But by using replications it was possible to synthesize findings from multiple parallel studies and thus to enhance the ability to draw inferences from the overall results.

Discussion

Consistent with Johnson's (2001) suggestions for strengthening interpretations of causality from non-experimental research, this article has recommended planning replications in field settings. The examples illustrate ways in which these replicated field designs can be synthesized to enhance the inferences that can be drawn from them. Further, when planned replications are used, it is possible also to plan for the measurement of variables that should prove useful to model effect sizes in a meta-analysis (e.g., the instructional variables in example 1). Fortunately for the researcher, a meta-analysis based on planned replications is far more straightforward to implement than a traditional synthesis of a disparate literature since fewer challenges, such as design differences, inadequate information, and inconsistent reporting of results across studies, exist.

An investigator planning to use replications in field research must make several decisions. Some of these are discussed below.

The basic design. The stronger the basic design, the stronger the inferences that may be made from any one replicate, and thus from the overall meta-analysis. The strongest feasible design should be chosen. Cook and Campbell (1979) provide an overview of designs that are particularly suitable in applied research contexts and discuss their strengths and weaknesses. It is important to be very clear what variable is independent and what is dependent in the basic design. In the two examples here, the independent variable was time (year 1 vs. year 2) in the first and presence or absence of the instruction workshop in the second. In both, the dependent variable was achievement. While year could not be manipulated in the first (the basic design was non-experimental), it was possible to manipulate the workshop in the second. Random assignment of instructors to workshop conditions strengthened that study [the basic design was pre-experimental (Campbell & Stanley, 1963)].

The effect-size measure. Magnitude of effect should be capable of coding as a standardized measure indicating direction and strength of relationship between the independent and dependent variables. Its quantification should yield an index that is is normally distributed and has a known or estimatable variance. Rosenthal (1994) provides a menu of possibilities. Three common examples that differ depending on the scaling of the two variables are: both continuous (the correlation coefficient, r); both dichotomous (the log-odds ratio, L); or, as in the two examples here, the independent variable a dichotomy but the dependent variable continuous (bias-corrected d, discussed above).

Maintaining effect-size independence. The effect sizes are assumed to be independent in a meta-analysis. That is generally the case across studies, but is not always true within studies. In our two examples, each study produced several dependent effect-size indices. Dependencies created by the measurement of six content areas in each school were ignored in the first study by analyzing each grade level and content area separately; in the second study, the six tests were analyzed together at first and a Bonferroni-like correction was applied throughout the analyses (Gleser & Olkin, 1994). Of course, care should be taken in field studies that the sites at which the replications occur maintain separation; sharing of information by participants across replications can threaten effect-size independence.

The variables to be measured. Besides the independent and dependent variables, it is advantageous to capitalize on the opportunity to measure variables that could be related to effect size (study characteristics). To generate a list of these, the researcher might consider how he or she might explain any observed differences that could appear among effect sizes across replicates. Whether substantive or artifactual, those explanations virtually always will be based on variables that should, if possible, be measured. These could be different contexts and dependent variables as in our second example in which effect sizes yielded by four different content areas and two test formats were combined into one meta-analysis. Or they may be descriptive of persons, such as demographics or aptitudes, or settings such as physical features in schools or classrooms. Coding characteristics of the replications that produced the different effect sizes provides data that are analyzed through relating these characteristics as independent variables to the effect sizes as dependent variables in the meta-analysis. The potential for assessing study differences that may be related to magnitude of effect represents an opportunity for creativity in designing robust multiple-study investigations through replication.

Meta-analysis is a relatively new approach to data analysis and the field is changing rapidly. One recent advance has been development of effective methods to conduct random-effects model analyses. Hedges & Vivea (1998) present a straightforward and relatively simple modification that is consistent with the techniques used in the two examples cited here. They also provide a worked example. An advantage of using a random model is that the results generalize to a population of studies not included in the present analysis, whereas in the two examples described here, the conclusions were restricted to the specific replications themselves. Hedges & Vivea (1998) discuss the conditions under which each type of analysis, fixed or random, is more appropriate.

References

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Cooper, H. & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Sage.

Gleser, L. J. & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Sage.

Guthrie, J. T., Schafer, W. D., Von Secker, C., & Alban, T. (2000). Contributions of instructional practices to reading achievement in a statewide improvement program. *Journal of Educational Research*, *93*, 211-225.

Hedges, L. V. & Olkin, I (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. & Vivea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.

Johnson, B. (2001). Toward a new classification of nonexperimental quantitative research. *Educational Researcher*, *30*(2), 3-13.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd Ed.). Orlando, FL: Harcourt Brace.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Sage.

Schafer, W. D. (1999). An overview of meta-analysis. *Measurement and Evaluation in Counseling and Development*, *32*, 43-61.

Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, *14*, 151-170.

William D. Schafer is an Affiliated Professor with Emeritus status, in the Maryland Assessment Research Center for Education Success, H. R. W. Benjamin Building Room 1230, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD 20742-1115.

Descriptors: Meta-analysis; replication; test scores; field research; research methodology

Citation: Schafer, William D. (2001). Replication: a design principle for field research. *Practical Assessment, Research & Evaluation*, 7(15). Available online: http://PAREonline.net/getvn.asp?v=7&n=15.