REPLICATION RESEARCH IN THE SOCIAL SCIENCES

Edited by James W. Neuliep

REPLICATION RESEARCH IN THE SOCIAL SCIENCES

Edited by James W. Neuliep

Originally published as a special issue of the
Journal of Social Behavior and Personality

SAGE PUBLICATIONS
The International Professional Publishers
Newbury Park London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Newbury Park, California 91320

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Main entry under title:

Replication research in the social sciences / edited by James W. Neuliep.
p. cm.
"Original published as a special issue of the Journal of social behavior and personality."
Includes bibliographical references and index.
ISBN 0-8039-4091-2. — ISBN 0-8039-4092-0 (pbk.)
1. Social sciences—Research. 2. Social sciences—Methodology.
I. Neuliep, James William, 1957-
H62.R443 1991
300'.72—dc20 90-24581
CIP

FIRST SAGE PRINTING, 1991

Sage Production Editor: Astrid Virding

REPLICATION RESEARCH IN THE SOCIAL SCIENCES

Contents

REPLICATIONS IN OTHER DISCIPLINES

Replication in Behavioral Research

Robert Rosenthal

Department of Psychology, Harvard University
William James Hall, 33 Kirkland Street, Cambridge, MA 02138

The two major questions addressed here are (1) how to evaluate the importance of one or more replications and (2) how to define the success of one or more replications. Among the factors affecting the importance of a replication are when, how, and by whom the replication was conducted. Procedures for weighting the results of specific replications are described. It is suggested that the older view of replication success, defined by dichotomous significance testing decisions, be replaced by the newer view of replication success defined by degree of agreement of effect sizes obtained in the original study and its replication. Some metrics of the success of replication are described and suggestions are offered as to what should be reported in a replication study.

On April 13, 1989, the daily newspapers and the television network news programs were filled with the theme of replication in science. The day before there had been an extraordinary symposium with an audience of 7,000 attendees of the Dallas meeting of the American Chemical Society, at which the replicability of "cold fusion" was a major topic. Professors B. Stanley Pons of the University of Utah and Martin Fleischmann of the University of Southampton reported a large excess of energy produced at room temperature from a simple electrolytic cell. There had

Author's Notes: I want to thank James W. Neuliep, Special Editor of this issue of the *Journal of Social Behavior and Personality*, and Rick Crandall, General Editor of the *Journal*, for having given me this opportunity to organize some ideas about replication on which I have been working for nearly a quarter of a century. Some of the ideas presented here were first presented in the following references: Rosenthal, 1966; 1979b; 1984; 1986a; 1989. Part of this paper was presented as a portion of an EPA Distinguished Lecture at the meeting of the Eastern Psychological Association, Boston, April 2, 1989. Preparation of this paper was supported in part by the National Science Foundation while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. I am grateful for financial support provided by the John D. and Catherine T. MacArthur Foundation, and for improvements suggested by Lynn Gale, Deanna Knickerbocker, Harold Luft, and Lincoln Moses.

been a number of failures to replicate this "cold fusion" effect but, by the time of this symposium, successful replications had apparently been conducted at Texas A&M and at Moscow University.

THE IMPORTANCE OF REPLICATION

Scientists of all disciplines have long been aware of the importance of replication to their enterprise (e.g., Campbell & Jackson, 1979). Now, thanks to the efforts of science writers for daily newspapers and reporters for network news, even the general public has become aware of the importance of replication.

The undetected equipment failure, the rare and possibly random human errors of procedure, observation, recording, computation, or report are known well enough to make scientists wary of the unreplicated experiment. When we add the possibility of the random "fluke" common to all sciences, the fact of individual organismic differences and the possibility of systematic experimenter effects, the importance of replication looms larger still to the behavioral scientist (Rosenthal, 1976).

What shall we mean by "replication"? Clearly the *same* experiment can never be repeated by a different worker. Indeed, the *same* experiment can never be repeated by even the same experimenter (Brogden, 1951). At the very least, the subjects and the experimenters themselves are different over series of replications. The subjects are usually different individuals and the experimenter changes over time, if not necessarily dramatically. But to avoid the not very helpful conclusion that there can be no replication in the behavioral sciences, we can speak of relative replications. We can rank order experiments on how close they are to each other in terms of subjects, experimenters, tasks, and situations. We can usually agree that *this* experiment, more than *that* experiment, is like a given paradigm experiment.

Replications may be crucial but some replications are more crucial than others. Three of the variables affecting the value, or utility, of any particular replication are:

(a) *when* the replication is conducted.
(b) *how* the replication is conducted.
(c) *by whom* the replication is conducted.

When the Replication is Conducted

Replications conducted early in the history of a particular research question are usually more useful than replications conducted later in the history of a particular research question. Weighting all replications equally, the first replication doubles our information about the research issue, the fifth replication adds 20% to our information level, and the fiftieth replication adds only 2% to our information level.

Once the number of replications grows to be substantial, subsequent investigators' felt need for further replication is likely to be due not to a real need for further replication but for a real need for the more adequate evaluation and summary of the replications already available. That was the situation for the research area of psychotherapy. Despite the availability of scores of studies on the effects of psychotherapy, investigators continued to cite the well-known conclusion drawn by Eysenck (1952, 1960) that psychotherapy was ineffective. It was not until Glass (1976) and Smith and Glass (1977) conducted their superb analyses of several hundred replications of studies of psychotherapy outcome, that we were able to reap the benefit of having so many replications available. It was this effort by Glass and his colleagues to deal quantitatively with a large number of replications that gave rise to the widespread and growing use of quantitative methods referred to collectively as meta-analytic procedures.

The File Drawer Problem

Once the number of replications grows to be substantial we find ourselves in the fortunate position of being able to assess the seriousness of the "file drawer problem."

Both behavioral researchers and statisticians have long suspected that the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out (Bakan, 1967; McNemar, 1960; Sterling, 1959). The extreme view of this problem, the "file drawer problem," is that the journals are filled with the 5% of the studies showing Type I errors while the file drawers back at the lab are filled with the 95% of the studies showing nonsignificant (e.g., $p > .05$) results.

In the past there has been very little we could do to assess the net effect of studies tucked away in file drawers that did not make the magic .05 level (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963; 1964). Now, however, we can establish reasonable boundaries on the problem and estimate the degree of damage to any research conclusion that could be done by the file drawer problem (Rosenthal, 1979; 1984; Rosenthal & Rubin, 1988).

The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error can be just brought to any desired level of significance, say .05. This number of filed studies, or the tolerance for future null results, is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the reviewer. If the overall level of significance of

the research review will be brought down to the level of "just significant" by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

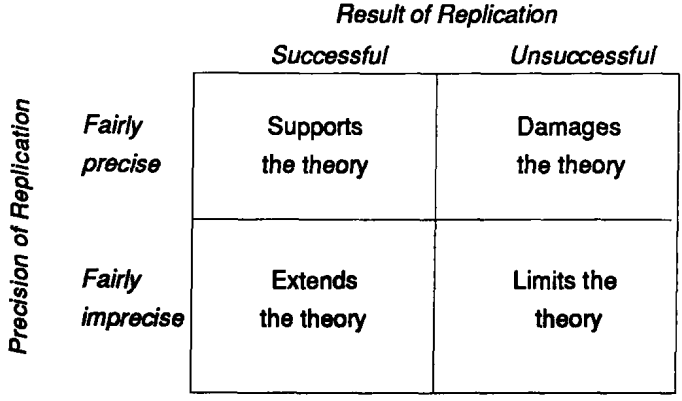
The computational procedures for addressing the "file drawer problem" are presented elsewhere (Rosenthal, 1979; 1984). Happily, the computations require little time and little effort. There is both a sobering and a cheering lesson to be learned from careful study of the "file drawer problem." The sobering lesson is that small numbers of studies, even when their combined p is significant, if they are not very significant, may well be misleading in that only a few studies filed away could change the combined significant result to a nonsignificant one. Thus, 15 studies averaging a Z of +0.50, p = .31) have a combined p of .026; but if there were only 6 studies tucked away showing a mean Z of 0.00 p = .50), the tolerance level for null results would be exceeded, and the significant result would become nonsignificant (i.e., p > .05). Or, if there were two studies averaging a Z of +2.00 (p = .023), the combined p would be about .002; but uncovering four new studies averaging a z of 0.00 would bring p into the "not significant" region.

The cheering lesson is that when the number of studies available grows large and/or the mean directional Z grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out. If 345 studies are found averaging a Z of +1.22 (p = .111), it would take 65,122 studies to bring the new combined p to a nonsignificant level; that many file drawers full are simply too improbable. These were the results obtained in our review of 345 studies of interpersonal expectancy effects (Rosenthal & Rubin, 1978).

At present no firm guidelines can he given as to what constitutes an unlikely number of unretrieved and/or unpublished studies. For some areas of research 100 or even 500 unpublished and unretrieved studies may be a plausible state of affairs, while for others even 10 or 20 seems unlikely. Probably any rough and ready guide should be based partly on k, the number of replications retrieved, so that as more studies are known it becomes more plausible that other studies in that area may be in those file drawers. Perhaps we could regard as resistant to the file drawer problem any combined results for which the tolerance level (X) reaches 5k + 10. That seems a conservative but reasonable tolerance level; the 5k portion suggests that it is unlikely that the file drawers have more than five times as many studies as the reviewer, and the +10 sets the minimum number of studies that could be filed away at 15 (when k=1).

It appears that more and more reviewers of research literatures will be estimating average effect sizes and combined p's of the studies they summarize. It would be very helpful to readers if for each combined p

FIGURE 1 Effects on Theory of the Success of Replication and the Precision of Replication



they presented, reviewers also gave the tolerance for future null results associated with their overall significance level.

How the Replication is Conducted

It has already been noted that replications are possible only in a relative sense. Still, there is a distribution of possible replications in which the variance is generated by the degree of similarity to the original study that characterizes each possible replication. If we choose our replications to be as similar as possible to the study being replicated, we may be more true to the original idea of replication but we also pay a price; that price is external validity.

If we conduct a series of replications as exactly like the original as we can, and if their results are consistent with the results of the original study, we have succeeded in "replicating" but not in extending the generality of the underlying relationship investigated in the original study. The more imprecise the replications, the greater the benefit to the external validity of the tested relationship if the results support the relationship. If the results do not support the original finding, however, we cannot tell whether that lack of support stems from the instability of the original result or from the imprecision of the replications. Figure 1 summarizes the consequences for the theory that is derived from the initial study of (a) successful versus unsuccessful replications and (b) the precise versus the imprecise nature of the replications.

FIGURE 2 Effects on Original Investigator of the Success of Replication and the Precision of Replication

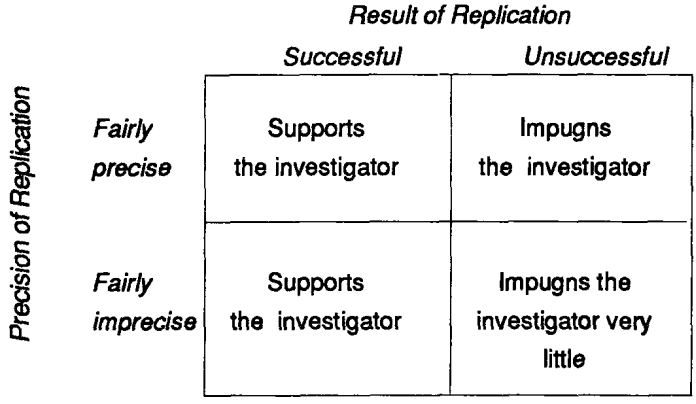


Figure 2 summarizes the consequences for the original investigator of (a) successful versus unsuccessful replications and (b) the precise versus imprecise nature of the replications. Compared to the theory tested by the replication the investigator has much to lose if a fairly precise replication is unsuccessful, since such failure is often associated with ascriptions to the original investigator of having been careless, incompetent, and in some cases, even dishonest.

The Replication Battery

Whenever we conduct a single replication (and that *is* how we conduct most replications) that fails to support the results of the original study, we are in a very serious dilemma. We can never be sure that the "failure to replicate," i.e., obtain consistent results, is due to the "non-replicability" of the original result or to the necessary inexactness of the replication procedure. It is this dilemma that leads me to suggest the employment of a *replication battery*.

The simplest form of replication battery requires two replications of the original study. One of these replications is as similar as we can make it to the original study, the other is at least moderately dissimilar to the original study. Suppose the size of the effect investigated in the original study was .80 standard deviation units (d). If our replication battery showed that both replications obtained consistent results (say around .60), we would be inclined to believe a bit more both in the reliability of the basic result and in its robustness in the face of moderate procedural variation.

If our replication battery showed that neither of the replications obtained consistent results (say, .10 and -.10), we would be inclined to believe a bit less in the reliability of the basic result with or without procedural variation.

If our replication battery showed that only the more exact replication obtained consistent results (say, .60) while the more dissimilar replication obtained noticeably less consistent results (say, .20), we would be inclined to believe a bit more that the basic result is reliable but that its reliability depends substantially on procedural consistency. Such a result suggests less external validity for the relationship investigated than would a result in which both replications yielded results consistent with the original results.

Less simple forms of the replication battery require more than two replications of the original study. For example, several replications can be ordered on the degree of similarity of procedure to the original study. Outcomes showing relatively homogeneous effect sizes for all studies would inform us as to the robustness or external validity of our results while marked diminution of effect sizes as procedures become more dissimilar would indicate systematic sensitivity to procedural variations.

More complex forms of the replication battery would also be useful and instructive. For example, one can envisage a three-dimensional design in which a battery of replications varying in type of task, type of subject, and type of instructions could be employed to address more precise questions of the factors serving to increase or decrease the magnitude of the effect being replicated.

Quality of Procedures

So far we have discussed the issue of the homogeneity of procedures employed in our replications. In this section we examine briefly the issue of the quality of those procedures. Implicit in all our discussion of replications is the idea that the original study is worth replicating. We can all conjure up studies in principle and, sad to say, studies in fact, that are so poorly designed and/or so poorly executed that they are not worth replicating because they were not worth doing in the first place. Assuming a constant level of scientific importance, those studies that were better done merit replication more than do those that were more poorly done. Studies that were very poorly done in the first place are not so much "done again" as they are "done right" when replicated.

Thus there is no virtue to replicating as precisely as possible a study that was so invalid internally as to render all inference attempts useless. The replicator's role in these cases is to do the study right while acknowledging that the idea of examining the particular relationship came from the original study.

Suppose we have collected a set of replications that we believe to vary substantially in their degree of internal validity or inferential quality. Shall we weight the better-done studies more heavily? Glass (1978) has pointed out that in real life it sometimes happens that the better studies and the worse studies all yield about the same results. In that happy situation we have no need to try to weight the better studies more heavily.

But suppose we find that the better, more carefully designed and conducted studies tend to show effects that are larger or smaller than the effects obtained in the less-well-done studies? The purist response might be to discard the poorer studies; but that seems an uneconomical solution. Studies cost in effort, time, and money—even poor studies. We can probably do better to assess the state of a research finding by weighting studies in proportion to the excellence of their design and procedures.

The seminar of methodologists who would assign the ratings of excellence, say on a scale of .00 to 1.00, should, of course, be blind to the outcome of the studies whose quality they are assessing. Such weighting might result in ten poorly done studies (e.g., mean quality rating of .10) counting for no more in our overall assessment than a single very well done study (e.g., quality rating of 1.00).

By Whom the Replication is Conducted

So far in our discussion of replications we have assumed that the replications are independent of one another. But what does independence mean? The usual minimum requirement for independence is that the subjects of the replications be different persons. But what about the independence of the replicators? Are ten replications conducted by a single investigator as independent of one another as ten replications each of which is conducted by a different investigator? This issue of potentially correlated replicators bears additional comment.

The Problem of Correlated Replicators

To begin with, an investigator who has devoted her life to the study of vision, or of psychological factors in somatic disorders, is less likely to carry out a study of verbal conditioning than is the investigator whose interests have always been in the area of verbal learning or interpersonal influence processes. To the extent that (a) experimenters with different research interests are different kinds of people, and to the extent that (b) it has been shown that different kinds of people, experimenters, are likely to obtain different data from their subjects, we are forced to the conclusion that within any area of behavioral research the experimenters come precorrelated by virtue of their common interests and any associated characteristics. Immediately, then, there is a limit placed on the degree of

independence we may expect from workers or replications in a common vineyard. But for different areas of research interest the degree of correlation or of similarity among its workers may be quite different. Certainly we all know of workers in a common area who obtain data quite opposite from that obtained by colleagues. The actual degree of correlation, then, may not be very high. It may, in fact, even be negative, as with investigators holding an area of interest in common but holding opposite expectancies about the results of any given experiment (Rosenthal, 1966).

A common situation in which research is conducted nowadays is within the context of a team of researchers. Sometimes these teams consist entirely of colleagues; often they are composed of one or more faculty members and postdoctoral students, and one or more predoctoral students at various stages of progress toward the Ph.D. Experimenters within a single research group may reasonably be assumed to be even more highly intercorrelated than any group of workers in the same area of interest who are not within the same research group. And perhaps students in a research group are more likely than a faculty member in the research group to be more correlated with their major professor. There are two reasons for this likelihood. The first is a selection factor. Students may elect to work in a given area with a given investigator because of their perceived and/or actual similarity of interest and associated characteristics. Colleagues are less likely to select a university, area of interest, and specific project because of a faculty member at that university. The second reason why students may be more correlated with their professor than another professor might be is a training factor. Students may have had a large proportion of their research experience under the direction of a single professor. Another professor, though collaborating with colleagues, has most often been trained in research elsewhere by another person. Although there may be exceptions, even frequent ones, it seems reasonable, on the whole, to assume that student researchers are more correlated with their adviser than another adviser might be.

The correlation of replicators that we have been discussing refers directly to a correlation of *attributes* and indirectly to a correlation of *data* these investigators will obtain from their subjects. The issue of correlated experimenters or observers is by no means a new one. Nearly 90 years ago Karl Pearson spoke of "the high correlation of judgments... [suggesting] an influence of the immediate atmosphere, which may work upon two observers for a time in the same manner" (1902, p. 261). Pearson believed the problem of correlated observers to be as critical for the physical sciences as for the behavioral sciences, as did Collins (1985) and Nye (1986) more recently.