EVALUATION STUDIES REVIEW ANNUAL
Volume 8

Evaluation Studies

EDITORIAL ADVISORY BOARD

Mark A. Abramson, *Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services*
Richard A. Berk, *Social Process Research Institute, University of California Santa Barbara*
Robert P. Boruch, *Department of Psychology, Northwestern University*
Seymour Brandwein, *Director, Office of Program Evaluation, Manpower Administration, U.S. Department of Labor*
Donald T. Campbell, *Maxwell School of Citizenship & Public Affairs, Syracuse University*
Francis G. Caro, *Institute for Social Welfare Research, Community Service Society, New York*
Thomas D. Cook, *Department of Psychology, Northwestern University*
Thomas J. Cook, *Research Triangle Institute, North Carolina*
Joseph dela Puente, *National Center for Health Service Research Hyattsville, Maryland*
Howard E. Freeman, *Institute for Social Science Research, University of California, Los Angeles*
Irwin Garfinkel, *Institute for Social Research on Poverty, University of Wisconsin, Madison*
Gene V. Glass, *Laboratory of Educational Research, University of Colorado*
Ernest R. House, *CIRCE, University of Illinois, Urnana*
Michael W. Kirst, *School of Education, Stanford University*
Henry M. Levin, *School of Education, Stanford University*
Robert A. Levine, *System Development Corporation, Santa Monica, California*

Review Annual

Richard J. Light, *Kennedy School of Government, Harvard University*
Katherine Lyall, *Director, Public Policy Program, Johns Hopkins University*
Laurence E. Lynn, Jr., *Kennedy School of Government, Harvard University*
Trudi C. Miller, *Applied Research on Public Management and Service Delivery, National Science Foundation*
David Mundell, *Education and Manpower Planning, Congressional Budget Office, Washington, D.C.*
Henry W. Riecken, *School of Medicine, University of Pennsylvania*
Peter H. Rossi, *Department of Sociology, University of Massachusetts, Amherst*
Susan E. Salasin, *National Institute of Mental Health, Rockville, Maryland*
Frank P. Sciolo, Jr., *Division of Advanced Production Research, National Science Foundation*
Lee Sechrest, *Director, Center for Research on Utilization of Scientific Knowledge, Institute for Social Research, University of Michigan*
Sylvia Sherwood, *Social Gerontological Research, Hebrew Rehabilitation Center for the Aged. Boston, Massachusetts*
Stephen M. Shortell, *Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle*
Ernst W. Stromsdorfer, *School of Public Health, Columbia University*
Michael Timpane, *Teacher's College, Columbia University*
Carol H. Weiss, *Graduate School of Education, Harvard University*

Evaluation Studies Review Annual

Volume 8 1983

Edited by

Richard J. Light

SAGE PUBLICATIONS
Beverly Hills / London / New Delhi

此为试读,需要完整PDF请访问: www.ertongbook.com

Copyright © 1983 by Sage Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photo-copying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

For information address:

SAGE Publications, Inc.
275 South Beverly Drive
Beverly Hills, California 90212

SAGE Publications India Pvt. Ltd.
C-236 Defence Colony
New Delhi 110 024, India

SAGE Publications Ltd
28 Banner Street
London EC1Y 8QE, England

Printed in the United States of America

International Standard Book Number-8039-1987-5

International Standard Series Number 0364-7390

Library of Congress Catalog Card Number 76-15865

FIRST PRINTING

CONTENTS

ABOUT THE EDITOR

RICHARD J. LIGHT is Professor at the Graduate School of Education and the Kennedy School of Government at Harvard University. His Ph.D. in statistics was taken at Harvard, and his work in program evaluation emphasizes methodological developments. In addition to his teaching and research, Professor Light has been director of faculty studies at the Institute of Politics, a consultant to the President's Commission on Federal Statistics, a consultant to the President's Commission on Federal Statistics, and a member of the Board of Directors of the Council on Applied Research and the Evaluation Research Society. He has recently served on National Advisory Panels for evaluation projects at the World Bank, General Accounting Office, and National Academy of Sciences. Professor Light has recently coauthored *Data for Decisions*, with D. Hoaglin, B. McPeek, F. Mosteller, and M. Stoto, a book describing how different forms of statistical evidence can be used to inform policy decisions. In his spare time, he particularly enjoys taking walks on windswept beaches with the three women in his life.

此为试读,需要完整PDF请访问: www.ertongbook.com

Introduction

Richard J. Light

This volume of the *Evaluation Studies Review Annual* looks at an interesting growth area in the behavioral sciences. For years, doctoral theses, journal articles, and research reports have had a standard format. First, state the problem. Second, do a literature review. Third, explain what new work is being done, and finally report the results.

It is the second step, the review, that for years got short shrift. Why? Probably for several reasons. Some scientists feel it is more exciting to develop new findings than to reexamine the old ones. Others feel that professional rewards rarely come to the fellow who "simply" pulls together what other people have done.

But I would put my chips on a different reason. I think most of us don't emphasize reviews because we simply don't know how to do them very well. Basic principles for designing a good single study are constantly being developed, debated, and refined. But what are principles for designing a good review?

Paul Smith and I took a crack at this question (Light and Smith, 1971), laying out solutions to some dilemmas that face any scientist preparing a review. This includes dealing with studies that measure outcomes differently, studies with dramatically different sample sizes, and findings that seem to conflict. In recent years, many others have put forward good ideas, and a body of techniques is beginning to crystallize. A big step forward was the work of Gene Glass and colleagues in the late 1970s (Glass 1976, 1977; Smith and Glass 1980; Glass, McGaw, and Smith 1981). Glass developed the notion of quantitative "meta-analysis" in a robust way. Also, his work was designed to help policymakers reach practical conclusions from large masses of studies. He especially popularized an outcome measure called "effect size" that enables a scientist to compare findings across many studies. The growth in methodological sophistication has stimulated a large number of reviews, especially in the last three years, using systematic methods. Indeed, two books have appeared (Glass et al., 1981; Hunter, Schmidt, and Jackson, 1982) to facilitate these efforts.

It seem constructive to stand back, then, and see what has been learned. That is the goal of this *Annual.* The material in the *Annual* is divided into two parts. The first set of essays emphasize methodology, and help us to think about how specific

quantitative methods for reviewing expand or limit inferential possibilities. They offer *concrete suggestions* for carrying out reviews, and they suggest assumptions and caveats that a reviewer sometimes forgets in the excitement of discovering treatment effects or a significant relationship between variables.

The second set of papers are exemplary reviews. Some are previously unpublished. I chose them because they are convincing. Each review either puts forth an interesting finding, sheds new light on a controversy, or demonstrates a nice application of an analytic technique.

I believe that, taken together, the two parts of this collection offer some clear messages. These messages are useful to an evaluator organizing a synthesis, and also to a research manager asking whether it is worth commissioning a new study. I have pulled out six such messages.

WHAT THE COLLECTION TELLS US

(1) Most Evaluations Find Small Effects

This is not an extraordinary finding. Earlier work (Gilbert, Light, and Mosteller, 1975) reports a similar result. Its importance comes, I think, from having managers of programs understand that they shouldn't expect large, positive findings to emerge routinely from new programs. Indeed, *any* positive findings are good news. I say this not in a political sense, but rather in a statistical one. There are several reasons why, even when an innovation works, an evaluation might not notice it.

One possible explanation is low power. A sample size may not be big enough to detect a positive program effect even if it really exists. A second is errors in variables. If both a program's features and its outcomes are measured with error, then the chance of detecting small effects can drop dramatically. A third explanation for missing positive effects is that with multisite innovations, only some sites or places *really will have* the positive effects. With new programs in particular, it would be surprising if *all* sites, and all program variants, work well. Indeed, it would be extraordinary. If it is so easy to mount new programs to solve social, educational, and health problems that have persisted for so many years, we would live in an engineer's ideal world. The more likely reality is that new programs (whether CETA, Head Start, or a new hospital emergency room procedure), differ from place to place in their early format, work well in a few places, and aren't much value is others. Evaluating outcomes at a few sites may lead to just one or two showing positive findings, while other sites show nothing special.

What does "small effects" mean? It depends upon the outcome measure's form. Many of the reviews use Glass's "average effect size" as the key summary measure. For example, Devine and Cook find an average effect size of .48 in their analysis of how interventions can reduce the length of hospital stay. Wortman and Yeaton find an average effect size of approximately 10 percent in their studies of coronary artery bypass graft surgery. Williams et al. find an average effect size of -.05 for studies relating television watching to school performance.

Some years ago, Jacob Cohen suggested rough guidelines for interpreting such effect sizes. His rules of thumb were that a .2 effect size was small, a .5 was moderate, and a .8 was large. I see no reason to modify these, except to remind us all that programs having an average effect of .8 are very rare.

A different outcome measure is the simple Pearson correlation coefficient. For example, White's paper finds an average correlation of .32 between socioeconomic status and academic achievement. Stock et al. find an average correlation of .03 between age and sense of well-being. It is reasonable to wonder, what is a "large" value for a correlation coefficient?

Rosenthal and Rubin's work makes a real contribution here. They reformulate the standard correlation coefficient into a comparison between two proportions. This is easily displayed in a simple two by two contingency table. The display gives nonstatisticians (and maybe some statisticians) a much better feel for the practical meaning of an average correlation. For example, using Rosenthal and Rubin's suggestion we find that an average R^2 of .04, rarely large enough to create tremendous excitement in evaluation circles, is equivalent to a new treatment's cutting a failure rate, or death rate, or dropout rate, by one-third (say from 60 percent to 40 percent). I consider such an accomplishment worth noticing! Another example: We may ask how large an R^2 is necessary to describe a 50 percent reduction in, say, death rates. The answer is an R^2 of only .10. Such analogies are not intuitive. I know that in future research reviews, small average correlations will command new respect, at least from me.

(2) Research Design Matters

Several of the syntheses drive home a point that has been speculated about in many essays: Research design matters, and sometimes matters a lot. One example is the review by DerSimonian and Laird. They find that coaching for SAT exams can help a lot, a modest amount, or hardly at all. It depends primarily on the research design of the evaluation. Observational studies generally find that coaching helps a lot; matched designs turn up a smaller positive value; randomized designs find coaching is hardly effective. A second example is Wortman and Yeaton's review of heart bypass surgery. It appears *far* more effective than drug treatment when examined with observational designs. The difference between treatments shrinks noticeably when the comparison uses randomly assigned groups.

Should we conclude from these two examples that randomized trials always lead to less positive findings about innovations? Indeed, this idea is broadly consistent with the discussion in Hoaglin and associates (1982). They cite a research review by Chalmers (1982) of portacaval shunt surgery. It found a strong negative relationship between how well controlled evaluations were, and how well the surgery fared. Controls were adequate in seven evaluations: *None* of the seven report big enthusiasm for the surgery. For the 67 where controls were absent, 50 led to big enthusiasm. This is consistent with Hugo Muench's rule, "Results can always be improved by omitting controls" (from Bearman et al., 1974).

Yet some reviews suggest this rule is not universal. For example, in this collection, Stock and associates find no relationship between research design and outcomes in studies of age and mental well-being. Similarly, Straw finds no relationship in his review of effects of deinstitutionalization in mental health. Finally, in a thorough review done some years ago, Yin and Yates (1974) find the opposite relationship. They report that for innovations in urban government, the better controlled research designs tended to find innovations *more* effective. They suggest as an explanation that innovators who are well trained and competent enough to evaluate their effort with a strong research design are more likely than average to have also developed a thoughtful innovation, which in turn is reasonably likely to be successful.

The point here is not that any rule applies in a predictable direction. From a modest size collection of reviews, it is difficult to create a general rule about the relationship between specific research designs and positive or negative outcomes in evaluations. The point, rather, is that for many reviews, a clear relationship exists between research design and probability of a positive finding. Overall, then, this collection strengthens the hypothesis that design matters. This is a valuable principle for evaluators to remember when designing new studies. Whatever the field, some effort should be made to search for a relation between design and outcome. Finding such a relation should enrich readers' interpretations of results from any one particular study.

(3) Good Syntheses Examine Treatment Implementation and Control Group Comparability

A big contribution reviews can make is to suggest, based on an aggregation of findings, what specific features of a broad program are especially likely (or unlikely) to work. An illustration comes from Giaconia and Hedges's synthesis of open education programs. In the 1960s and early 1970s, a major movement developed to reduce rigidities in precollege education. Innovations such as multi-age grouping, open architecture, and team teaching were introduced. Any review of evaluations of such programs reporting an "on the average" finding about open education has little value to policymakers. Open education involves so many components that it is unlikely *all* of them have positive, or negative, effects. It is more likely that just a few components will matter. Identifying those few is a key contribution of a review. Giaconia and Hedges, for example, identify three features of open education that lead to clearly positive outcomes: diagnostic evaluation of children, availability of manipulative materials, and individualized instruction. Aspects of open education that generate the most publicity, such as mixed age grouping and open spaces, do not distinguish more from less effective programs.

Similar distinctions should be made among *who* is being investigated. For example, the Kulik and Kulik article reviews studies of ability grouping for high school students. Some of the studies they include examine ability grouping of

particularly able students. Others focus on low achievers. Others include a broad range of achievement. Any broad conclusion about ability grouping should be tempered with careful statements about *subgroups.* Kulik and Kulik do an especially fine job dividing their analyses by type of ability grouping.

Just as the detailed nature of a *treatment* must be clear, *control groups* in comparative studies must also be carefully investigated. If a treatment's effectiveness is generally estimated by comparing it with a control group, then a reviewer must see whether control groups are comparable across studies. If not, conflicting outcomes can easily arise because controls differ.

Devine and Cook's review illustrates how to do it well. In their review of 34 interventions designed to reduce length of hospital stay, they found at least three different kinds of control groups. Some studies compared an intervention with a "usual care" group. This eliminates the possibility of a Hawthorne effect. Others used "placebo controls." Devine and Cook found that such studies reported patients in the control groups received as much attention from researchers as patients in the treatment group. Indeed, in one case, they received even more. A third group of studies included *both* of these control group types. By separating studies that use different kinds of control groups, Devine and Cook found that type of control group matters when one wants to assess the intervention. Studies using usual care for controls found a noticeably bigger treatment effect than similar efforts with placebo controls. This is strong evidence that evaluations with placebo controls underestimate treatment impact because of a Hawthorne effect.

The general point is the important one. Just because each in a group of evaluations reports is examining a certain treatment, reviewers should not casually assume that either the treatments or the controls are in fact identical across studies. Careful research reviews should specifically analyze the comparability of treatments and of controls.

(4) Publication Bias Seems to Exist

The collection of reviews demonstrates convincingly that evaluations in refereed journals report, on the average, more significant findings about a program or treatment's effectiveness than similar unpublished work. This is not an extraordinary finding. It has been speculated about for some time (Greenwald, 1975; Rosenthal, 1979). Yet the consistency with which journal reports show stronger program effects than other evaluation sources suggests that any review should involve a serious effort to track down findings from sources other than journal articles.

The approximately two dozen reviews in this collection give a clear indication of where these other, nonjournal evaluations come from. The three main sources are

(1) chapters in books often invited by an editor,
(2) research reports, produced by contract research organizations such as Abt Associates or Rand or SRI or Mathematica, or produced by government agencies such as NIE, NIH, or NICHD,

此为试读,需要完整PDF请访问: www.ertongbook.com