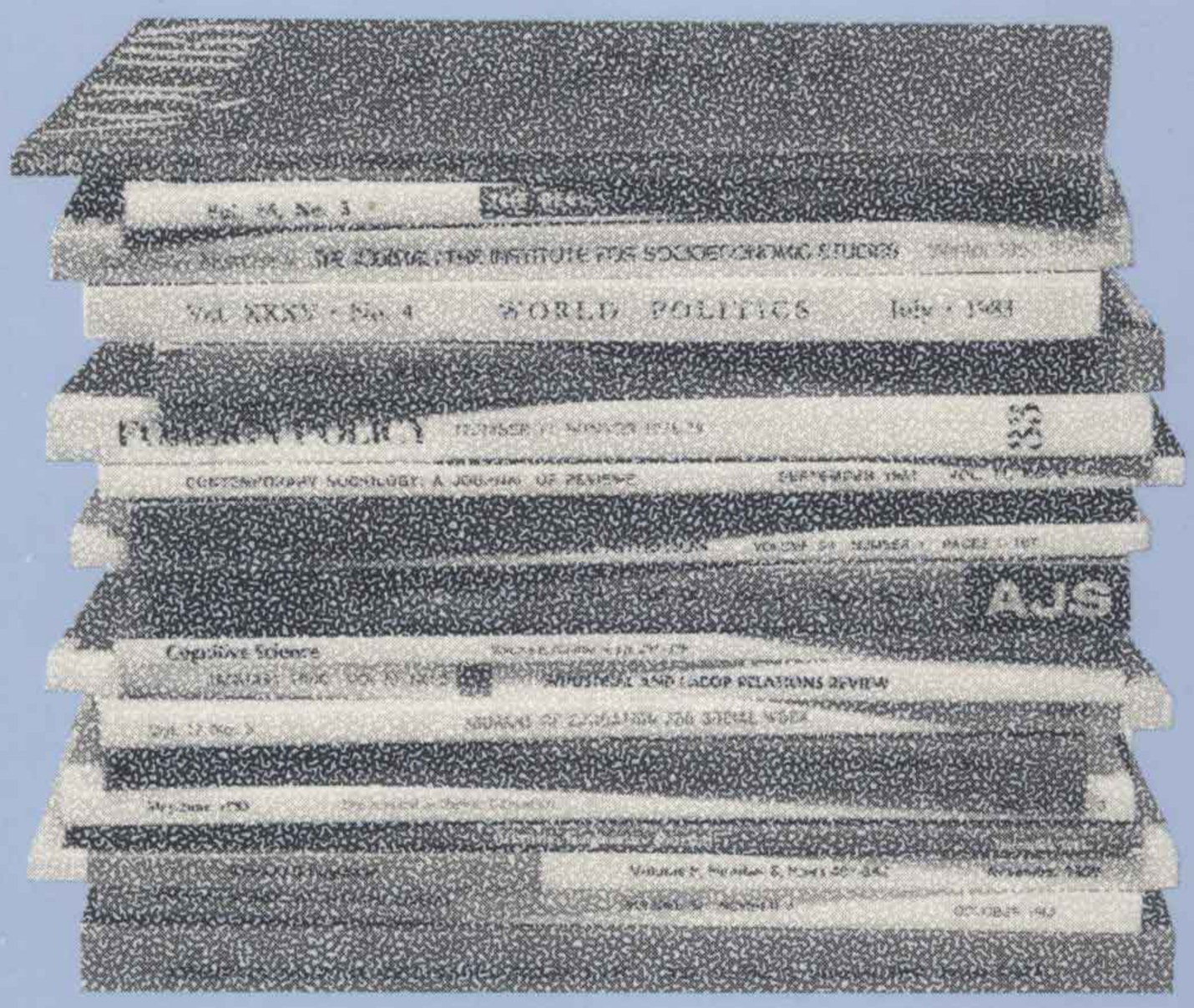
RICHARD J. LIGHT
AND
DAVID B. PILLEMER



SUMMING UP

THE SCIENCE OF REVIEWING RESEARCH

SUMMING UP

THE SCIENCE OF REVIEWING RESEARCH

RICHARD J. LIGHT
DAVID B. PILLEMER

Harvard University Press
Cambridge, Massachusetts, and London, England

Printed in the United States of America
10 9 8

LIBRARY OF CONGRESS CATALOGING IN PUBLICATION DATA

Light, Richard J.
Summing up.

Bibliography: p.
Includes index.
1. Social sciences—Research. 2. Social sciences—Bibliography—Methodology. I. Pillemer, David B., 1950– . II. Title.
H62.L46 1984 300'.72 84-4506
ISBN 0-674-85430-6 (alk. paper) (cloth)
ISBN 0-674-85431-4 (paper)

PREFACE

We decided to write this book because of persistent questions about how to strengthen new research. Each of us has been sought out by others for help in designing new investigations. These requests have come from public agencies and decision-makers as well as academic colleagues and students. And although the details of the questions change, the broad outlines are extraordinarily similar.

The dialogue begins with a request for help in designing a study. We then ask what has been learned in that particular subject area from earlier studies. After all, new investigations should build upon existing knowledge. The response, nearly always, is that the group of earlier studies is complex and even contradictory. Indeed, the contradictions are an important reason for conducting a new study.

Our questioners seldom consider investing time and re-

sources to synthesize the information that already exists. We wondered why not. This seems to be a sensible first step. Without a clear picture of where things now stand, simply adding one new study to the existing morass is unlikely to be very useful.

Scientists are not the only ones who ask questions about "what research adds up to." Policymakers must make practical decisions based on what is known *now*. In the early 1970s Walter Mondale, then a Senator, in an address to the American Educational Research Association, spoke about research on racial integration in America's public schools: "What I have not learned is what we should do about these problems. I had hoped to find research to support or to conclusively oppose my belief that quality integrated education is the most promising approach. But I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumptions or conclusions of the first. No one seems to agree with anyone else's approach. But more distressing: no one seems to know what works. As a result, I must confess, I stand with my colleagues confused and often disheartened."

The frustration Mr. Mondale expressed is both widespread and understandable. He wants some firm information, whether it comes down on one side or the other, and he cannot find it. His description of the lack of consistency in scientific findings unfortunately applies not only to research on racial integration but to many other issues as well.

Apart from the formulation of policy, difficulties in reconciling contradictory conclusions from similar studies cripple a fundamental component of the scientific process: the systematic accumulation of knowledge. Mark Twain said in his autobiography, "The thirteenth stroke of a clock is not only false of itself, but casts grave doubts on the credibility of the preceding twelve." This statement captures a critical part of the experience of doing applied research. It seems that for every twelve studies reaching any specific conclusion there is al-

ways a thirteenth that disagrees. Mark Twain's solution might well have been to put all thirteen behind him and light out for the Territories. The equivalent of this action in research would be to discard the conflicting evidence and initiate a new study. But such a step would incur three costs: a great deal of information, some potentially valuable, would be thrown away; a decision would be postponed for at least the length of time the new research takes; and, from the point of view of the next reviewer of the literature, the new research would simply be the fourteenth in the set of studies. Even with difficult problems, it is worth trying to combine and reconcile conflicting outcomes.

Clearly, society must improve its efforts to learn from existing findings, to "discover what is known." In this spirit we began a search for procedures, accessible to nonstatisticians, that would enhance the scientific quality of research summaries. We examined what reviewers currently do when they tackle a group of studies done by different people in different places at different times. Our search led ultimately to the writing of this book. In it we present circumstances under which it makes sense to use various statistical techniques. We suggest new ways of using simple graphical displays to examine patterns among findings. We emphasize conceptual issues throughout, because carefully planned reviews are nearly always stronger and more useful than atheoretical foraging. We also provide strategies for using different kinds of information from many studies. Some studies are primarily numerical; others are narrative or qualitative. Some have large sample sizes; others are tiny. Some have controlled research designs; othes do not. Our broad goal is to help readers organize existing evidence in a systematic way, whether a review is motivated by a scientific problem or the need for a policy decision. The book should also help readers examine and evaluate reviews prepared by others.

Our suggestions apply to many fields, including education, health, psychology, and policy sciences, and we include illustrations from each. We have tried to write at a technical level

accessible to a broad audience, including academic researchers, policy analysts, and students beginning their careers. We hope this book will help them to strengthen connections between current research and accumulated knowledge from the past.

ACKNOWLEDGMENTS

We both have many thanks to offer. A grant from the Spencer Foundation greatly facilitated our effort. H. Thomas James, President of the Spencer Foundation, has been supportive for years. We have benefited from conversations about research reviews with Robert F. Boruch, Eleanor Chelimsky, Thomas D. Cook, David S. Cordray, Gene V. Glass, Larry V. Hedges, Linda Morra, Robert Rosenthal, Michael A. Stoto, Herbert J. Walberg, and Sheldon H. White. We especially would like to acknowledge detailed comments on earlier drafts of this manuscript from Anthony S. Bryk, Robert E. Klitgaard, Frederick Mosteller, Stephen W. Raudenbush, Paul V. Smith, Terrence Tivnan, and Eric Wanner. Alicia Schrier helped with computer simulations and analyses, and William Minty prepared figures and charts. Camille Smith's superb editing improved the manuscript substantially.

ACKNOWLEDGMENTS

Each of us has additional acknowledgments. Thanks to the generosity of Wellesley College, David Pillemer received a year's leave from teaching to work on this book; he spent the year as a Visiting Scholar in the Department of Psychology and Social Relations, Harvard University. Richard Light has benefited from working with the staff of the Program Evaluation and Methodology Division of the General Accounting Office. As chairman of a National Academy of Sciences panel on evaluating children's programs, funded by the Carnegie Corporation, he learned a great deal about research reviews in education. His wife, Pat, and daughters, Jennifer and Sarah, provided the greatest support of all.

We are grateful for permission to use the following illustrative materials in this book. Box 1.2: editorial by permission of *The Washington Post,* copyright © 1982 by *The Washington Post;* letter from George D. Wilson by permission of George D. Wilson and the American Meat Institute. Table 2.1 by permission from J. M. Lachlin, N. Tygstrup, and E. Juhl, eds., *The Randomized Clinical Trial and Therapeutic Decisions* (New York: Marcel Dekker, 1982). Box 3.2: Figure A by permission from *Annual Review of Nuclear Science,* 25 (1975), copyright © 1975 by Annual Reviews Inc. Figure 3.4 by permission from C. C. Kulik and J. A. Kulik, "Effects of ability grouping on secondary school students: a meta-analysis of evaluation findings," *American Educational Research Journal,* 19 (1982), copyright © 1982 by the American Educational Research Association, Washington, D.C. Figure 3.14 by permission from G. V. Glass, "Integrating findings: a meta-analysis of research," *Review of Research in Education,* 5 (1977), copyright © 1977 by the American Educational Research Association. Figure 3.15 by permission from R. E. Klitgaard and G. Hall, *A Statistical Search for Unusually Effective Schools* (Santa Monica, Calif.: Rand, 1973), copyright © 1973 by the Rand Corporation. Figure 4.2 by permission from R. J. Light and David B. Pillemer, "Numbers and narrative: combining their strengths in research reviews," *Harvard Educa-*

tional Review, 52 (1982). Box 5.1: Figure B by permission from G. V. Glass and M. L. Smith, "Meta-analysis of research on class size achievement," *Educational Evaluation and Policy Analysis,* 1 (1979), copyright © 1979 by the American Educational Research Association.

CONTENTS

INTRODUCTION

1

A professor is called to testify before Congress as to whether a program offering nutritional supplements to low-income pregnant women should be expanded. Do the supplements improve maternal health? Do they improve child health?

A policymaker faces the challenge of restructuring the parole board system in a state. Are changes necessary? What changes would be most constructive?

An ambitious graduate student wants to try out an innovative housing program for elderly citizens. The plan is to have residents make decisions collectively. Are participants happier and healthier as a result?

Each of these people would benefit enormously by pausing before taking action and asking a few questions: What is known about the magnitude of the problem? What efforts have been made in the past to ameliorate it? Were they suc-

cessful? Does existing evidence suggest any promising new directions? These questions demand some way to formalize "what we already know."

Where can one turn for answers? Consider the graduate student and his housing innovation. Knowing that a good review of existing research should precede fieldwork, he approaches his faculty adviser for guidance. How does a scientist conduct a research review? What are the essential steps?

It is easy to imagine the student being slightly embarrassed to ask these questions, and the adviser feeling mild annoyance. Reviewing the literature is something a competent young scholar *should* know how to do. The professor's first reaction is likely to be that while the procedures are not carved in stone, some are quite standard. Go to the library. Use the social science abstracts. Thumb through current journals. Identify relevant articles. Briefly summarize them and draw some coherent overall conclusions.

Yet if the faculty member is pressed to give *explicit* guidelines, her annoyance may turn to frustration. How can relevant articles be identified? Which of tens or hundreds of studies of programs for the elderly should a summary present? How should conflicting findings from different studies be resolved? Trying to answer these questions may make it clear that the professor's "scientific" procedures are implicit rather than explicit, as much art as science.

Feeling this frustration, the faculty adviser takes the offensive. The absence of formal reviewing procedures is an inconvenience, but this does not undermine the research process. *New* research is the basis of scientific achievement. A research review is a chore to dispose of as quickly and painlessly as possible, usually by delegating it to subordinates. The student meekly replies that his *new* research will soon be somebody else's *old* data, receiving short shrift in a review article. But the lesson has been passed on to a new generation of scientists.

Why do scientists think that new research is better, or

more insightful, or more powerful? The underlying assumption must be that new studies will incorporate and improve upon lessons learned from earlier work. Novelty in and of itself is shallow without links to the past. It is possible to evaluate an innovation only by comparisons with its predecessors.

For science to be cumulative, an intermediate step between past and future research is necessary: synthesis of existing evidence. The casual attitude of some scientists toward this step undermines the value of many new research initiatives. With tens of studies examining questions such as the effectiveness of Head Start, the value of heart bypass graft surgery, or the impact of television advertising directed to young children, producing a useful summary requires systematic methods. Studies are done by different people, at different places, at different times. They may use different outcome measures, research designs, and analysis formats. The number and diversity of studies challenge even an expert's ability to "pull it all together" without formal tools.

Current Status of the Research Review

For many years, the "literature review" has been a routine step along the way to presenting a new study or laying the groundwork for an innovation. Journals such as *Psychological Bulletin, Review of Educational Research, American Public Health Journal,* and *New England Journal of Medicine* publish the best of such reviews. Traditionally, these efforts to accumulate information have been unsystematic. Studies are presented in serial fashion, with strengths and weaknesses discussed selectively and informally. These informal reviews often have several shortcomings:

1. The traditional review is *subjective.* Since the process has few formal rules, two well-intentioned reviewers can disagree about issues as basic as what studies to include and how to resolve conflicts between studies. The result is that rather than organizing diverse outcomes into a set of reason-

ably conclusive findings, the reviews themselves are open to attack for including inappropriate or poorly done studies or for drawing conclusions subjectively. Instead of resolving conflicts among the various studies, the review may only generate new conflicts.

2. The traditional review is *scientifically unsound.* Without formal guidelines, a reviewer may reach conclusions using methods inconsistent with good statistical practice. For example, when some studies show a positive program effect while others show no relationship or even a negative effect, a common way to summarize these findings is to use a "vote count." A reviewer counts up the number of studies supporting various sides of an issue and chooses the view receiving the most "votes." This procedure ignores sample size, effect size, and research design. Serious errors can result (Hedges and Olkin, 1980; Light and Smith, 1971).

3. The traditional review is an *inefficient way to extract useful information.* This is especially true when the number of studies is large, perhaps thirty or more. A reviewer unarmed with formal tools to extract and summarize findings must rely on an extraordinary ability to mentally juggle relationships among many variables. Systematic ways of exploring such relationships would make it far easier both to detect and to understand them. (Box 1.1 gives an illustration of the difficult issues facing narrative reviewers.)

One contemporary response to these shortcomings is to use statistical procedures for combining findings. Excellent books presenting quantitative methods include Glass, McGaw, and Smith (1981) and Hunter, Schmidt, and Jackson (1982). Quantitative procedures appeal to the wish for a sense of order that a complex body of findings can generate. We present some of these techniques in the chapters that follow. But our focus is on broader questions. How does one *structure* a research review? How does one even *think* about different ways of aggregating information? What *qualitative* sources of information are especially valuable?

BOX 1.1. CONFLICTS BETWEEN NARRATIVE REVIEWS

For years scientists have debated the extent to which schools and home environments influence children's IQ test scores. One way of assessing this impact is to examine the cognitive performance of adopted children. Munsinger (1974) examined a group of adoption studies and concluded that environmental effects are small: "Available data suggest that under existing circumstances heredity is much more important than environment in producing individual differences in IQ" (p. 623). Kamin (1978) later reviewed the *same* group of studies and reached the opposite conclusion.

That two distinguished scientists interpret a set of results so differently is only slightly surprising, since the personal beliefs of a reviewer can play a role in resolving disparate findings. This is especially true for a topic as controversial as nature-nurture. Far more striking are their different views on what constitutes acceptable review standards. According to Kamin, "Munsinger's review of the adoption literature is in general unreliable. Though any review must be selective in its presentation and analysis of data, Munsinger's is excessively so" (p. 194). Munsinger (1978) replies: "Kamin accuses me of errors and selective reporting of the adoption data, but in fact Kamin's comments are quite selective and often incorrect" (p. 202). These conflicting views about evidence are particularly apparent in comments on a study by Freeman, Holzinger, and Mitchell (1928): Kamin describes it as "large-scale and extraordinarily interesting" (p. 200); Munsinger argues that it is "replete with methodological and statistical difficulties" (1974, p. 635).

Kamin (1978) concludes: "perhaps the major point to be made is that readers interested in evaluating the evidence on heritability of IQ ought not to depend on published summaries. Those who wish to speak or to teach accurately about what is and is not known have no realistic alternative but to read the literature themselves" (p. 200). Taken literally, this statement eliminates the review as a scientific or practical tool. It is not practical to expect all people interested in a medical treatment, or a Head Start program, or even an issue as complicated as environmental impact on IQ, to read dozens of original scientific studies. Surely it is worth trying to develop systematic procedures for summarizing the literature. If two reviewers using explicit procedures reach different conclusions, at least readers can see why and then make an informed choice.

6
SUMMING UP

The science of preparing reviews has experienced a revolution of sorts in recent years. But the fruits of this work have not yet entered into the training of most social scientists, educators, and policymakers. For example, Jackson (1980) reports that *none* of a sample of 39 books on general methodology in social science devotes more than two pages to literature reviews. Jackson's investigation of the quality of social science reviews published in the period 1970–1976 turned up an almost complete lack of systematic procedures. Most contemporary reviews are still informal and discursive.

For social science to get the maximum benefit from prior research, sound reviewing strategies must become more accessible, more highly valued, and a routine part of advanced undergraduate and graduate training. We have designed this book as a small contribution toward these goals. (Box 1.2 presents a public debate about the value of synthesis.)

BOX 1.2. COMMISSIONING A NEW STUDY VERSUS SYNTHESIZING AVAILABLE EVIDENCE

In mid 1982 the National Academy of Sciences issued a long-awaited report on the link between diet and cancer. Part of this report described the research tying consumption of different kinds of meat to the likelihood of developing cancer. On June 19, 1982, the *Washington Post* published an editorial entitled "Food and Cancer," which said in part:

> If you are one of those people who have just about given up on making sense of the conflicting medical advice about what to eat, help—at least of a kind—is on the way. A striking convergence of expert opinion is coming about. More and more evidence shows that diet strongly influences the risk of coronary heart disease, cancer, hypertension and other major killers. And the recommended changes in diet for lowering the risk of each of these diseases reinforce, rather than contradict, each other.
>
> The newest evidence comes from a two-year study of the connections between diet and cancer, issued this week by the National Academy of Sciences. The group found first of all that