

STRENGTHENING THE ROLE OF RESEARCH IN POLICY DECISIONS:

The Campbell Collaboration and the Promise of
Systematic Research Reviews

Harris Cooper
Professor of Psychology
University of Missouri - Columbia

In the past three decades, there has been a dramatic increase in the amount of social research available to policy makers. From drug abuse prevention to school desegregation, it is not uncommon to find dozens, if not hundreds, of studies that examine the effectiveness of social policies and programs. Policy makers look to these bodies of evidence in the hope that research will assist in making sound decisions about which programs to continue, expand, or abandon. Practitioners look to research for prescriptions about how best to carry out their work. Participants want to know that programs will have the desired effect. The public seeks evidence that tax and philanthropic dollars are being spent wisely.

The promise of evidence-based decision making in the social policy arena remains largely unfulfilled. In fact, skepticism, if not outright cynicism, exists about the value of research in creating social policy. Some of the barriers to the effective use of research are endemic to the policy arena. Other barriers reside within the research community. Advocacy groups on opposite sides of an issue point to studies that support their position but conflict with one another. Researchers producing disparate results ignore flaws in their own work while questioning the trustworthiness of other's findings. Both behaviors lead to diminished credibility for all research.

These episodes, and the resulting perception of a diminished value for empirical evidence in setting public policy, can be traced to at least three characteristics of social research. First, broad-based policies and programs are carried out in real world contexts. The complexities of setting introduce factors that influence whether or not a policy or program will produce the desired results. The important nuances of setting are difficult to recognize and even more difficult to represent within the confines of a single study.

Second, for both ethical and practical reasons, social research frequently will include design flaws. The flaws mean that explanations for the outcome of a study other than the effectiveness of the policy or program itself will remain plausible. Most typical among these design flaws are that program participants often cannot be randomly assigned to receive or not receive a treatment. This leaves open the possibility that preexisting differences between the treated and untreated participants account for outcome differences.

Third, the outcomes of single studies are probabilistic in nature, based as they are on samples drawn from populations. Therefore, variation in outcomes when many studies on the same topic have been conducted, in direction as well as the magnitude of treatment effects, is not surprising. Indeed, it is even expected. Often, this variation due to sampling uncertainty is mistakenly called conflicting results.

A solution to all three of these problems can be found in how individual studies are carried out. Additionally, after decades of neglect, social scientists now agree that a solution can also be found in how bodies of evidence are treated after multiple studies have accumulated (see Appendix A for a brief history of these developments). The influence of context on policy and program evaluations can be examined in research synthesis by comparing the outcomes of groups of evaluations that include different types of participants, settings, and treatment characteristics, even though no single study contained all the variations. Multiple studies can also be grouped according to the characteristics of their research designs. If studies with different design strengths and weaknesses lead to similar results, greater confidence can be placed in a review's conclusion than in the results of any single evaluation. If results are different, rival hypotheses can be precisely identified for testing in future study. Finally, by combining the results of multiple studies the general effect of a policy or program can be pinpointed much more precisely than in a single investigation. The expected variation about this midpoint can also be estimated.

In each instance, the use of proper procedures for the synthesis of multiple studies does more than simply ameliorate the problems currently associated with the use of research in policy making. Systematic review procedures transform the difficulties into strengths. Variation in the context, design, and sampling characteristics of individual studies are the source of consternation when studies are examined individually, serially, and narratively. When multiple studies, each limited in their representation of context, design, and sample, are treated as data points in a second round of scientific investigation they contribute jointly to more confident, general, and properly contextualized guides to decision making.

Because of the potential value of systematic research reviews in the policy domain, both the producers and consumers of reviews now agree they must think about what distinguishes good from bad reviews. Further, they agree that without high-quality reviews, consumers will question the value of research for assisting the development of effective public policy. *The issues now facing social scientists concern how to define high-quality reviews, how to train producers to carry them out, and how to disseminate reviews to those who might formulate and implement policy and practice based on their result.*

Efforts are underway to “raise the bar” regarding how both primary research and systematic reviews are conducted in the policy arena. In health care, the Cochrane Collaboration has become a recognized vehicle for the production and dissemination of high-quality systematic reviews of research. In social policy, the recent emergence of a parallel organization, the Campbell Collaboration, promises to bring the same kind of rigorous treatment of literatures to research on education, crime and justice, and social welfare.

The Cochrane Collaboration on Health Care

In 1979, Archie Cochrane, a British epidemiologist, noted that a serious criticism of his field was that it had not organized critical summaries of relevant randomized controlled trials (RCTs). In 1987, Cochrane found an example in health care of the kind of review he was looking for. He called this systematic review of care during pregnancy and childbirth “a real milestone in the history of randomized trials and in the evaluation of care,” and suggested that other specialties should copy the methods (Cochrane, 1989). In the same year, the scientific quality of many published reviews in medicine was shown to leave much to be desired (Mulrow, 1987).

The Cochrane Collaboration was developed in response to the call for systematic, up-to-date reviews of RCTs of health care practices. Funds were provided by the United Kingdom’s National Health Service to establish the first Cochrane Center. When the Center opened at Oxford in 1992, those involved expressed the hope that there would be a collaborative international response to Cochrane's agenda. This idea was outlined at a meeting organized six months later by the New York Academy of Sciences. In October 1993, at what was to become the first in a series of annual Cochrane Colloquia, 77 people from eleven countries co-founded The Cochrane Collaboration.

The principles and products of the Cochrane Collaboration. The Cochrane Collaboration has evolved rapidly since the First Colloquium, but its basic objectives and principles have remained the same. It is an

international organization that aims to help people make well-informed decisions about health care by preparing, maintaining and ensuring the accessibility of systematic reviews of the effects of health care interventions. Detailed information on the Cochrane Collaboration can be found at <http://www.cochrane.org>

The Collaboration is built on the principles of joint effort, avoiding unnecessary duplication of effort, minimizing bias in review outcomes, ensuring relevance and access for people other than researchers, and continually updating and improving the quality of its work.

The core products of the Cochrane Collaboration are contained in the Cochrane Library, a set of electronic and web-based databases. The Cochrane Database of Systematic Reviews contains reviews that have been carried out by Collaboration review groups and that meet the standards set by the Collaboration's members. The Cochrane Controlled Trials Register, is an exhaustive reference database of randomized controlled trials of health care practices. The Database of Abstracts of Reviews of Effectiveness includes structured abstracts of systematic reviews completed outside the Collaboration that have gained approval after critical appraisal. The Cochrane Review Methodology Database is a bibliography of articles on the science of research synthesis. Also included in The Cochrane Library is a Reviewers' Handbook on the process of reviewing research.

There are several other unique aspects of the Cochrane Library. First, it contains comments and criticisms of its own work. Second, it remains a live document because review groups are constantly revising and updating their entries to reflect the results of new studies and improvements in review methodology. Thus, the quality of Cochrane reviews is enhanced by means of an iterative system through which successive versions of each review reflect not only the emergence of new data, but also valid criticisms, solicited or unsolicited, from whatever source. Successive versions of a particular review, together with any intervening criticisms, are archived electronically.

The organizational structure of the Cochrane Collaboration. Cochrane reviews are published electronically in quarterly issues of The Cochrane Database of Systematic Reviews. Preparation and maintenance of reviews is the responsibility of international collaborative Review Groups. Over 40 existing and planned review groups cover most of the important areas of health care. The members of these groups—researchers, health care professionals, consumers, and others—share an interest in generating reliable, up-to-date evidence relevant to the prevention, treatment and rehabilitation of particular health problems or groups of problems.

As they carry out their work, review groups employ a series of methods to assemble, appraise, and sometimes synthesize data from the trials that are relevant to their question. In doing so, they draw on the work of Methods Groups, which are created to organize and disseminate the work of methodologists who have come together to improve the validity and precision of systematic reviews. For example, collaborative review groups benefit from a Methods Group that developed high-quality, uniform methods for handsearching journals. Members from a number of Methods Groups have played major roles in the creation and maintenance of the Review Manager software that helps reviewers organize, prepare, analyze and present their systematic reviews.

The work of the Cochrane review groups also is facilitated in a variety of ways by the work of Cochrane Centers that advise on organizational policy and facilitate training and communication. Review groups are also assisted by field panels that monitor reviews to ensure that concerns of particular stakeholders are represented in reviews (e.g., child health). A consumer network also exists within the collaboration.

The Campbell Collaboration on Public Policy

The inaugural meeting of the Campbell Collaboration was held in Philadelphia, Pennsylvania, on February 24 and 25, 2000. Patterned after the Cochrane Collaboration, and championed by many of the same people, The Campbell Collaboration aims to bring the same quality of systematic evidence to issues of public policy as the Cochrane does to health care. It seeks to help policy makers, practitioners, consumers, and the general public make informed decisions by preparing, maintaining, and promoting access to systematic reviews of studies on the effects of public policies, social interventions, and educational practices.

The Campbell Collaboration was named after the American psychologist and methodologist, Donald Campbell, who drew attention to the need for society to assess more rigorously the effects of social and educational experiments. These experiments take place in education, delinquency and criminal justice, mental health, welfare, housing, and employment, among other areas.

Over 80 people from North America and Europe attended the inaugural meeting. In addition to general sessions, the meeting began the process of developing review groups. Attendees interested in education, crime and justice, and social welfare, met in breakout groups to define their scope and begin the process of building an organizational infrastructure. Similar breakout groups met to discuss organizational needs concerning primary research and systematic review methods, and

software and dissemination. Review groups in other areas are expected to emerge in coming years. The Campbell Collaboration web site is: <http://campbell.gse.upenn.edu>

Much time was spent examining ways in which the Cochrane and Campbell Collaborations could cooperate so as to share scarce resources and avoid duplication. This issue was especially salient to the incipient Methods Group because of the considerable overlap in methods used by medical and behavioral scientists. The Methods Group established a working committee of four members that will be joined by a similar group from the Cochrane Collaboration to look at ways to integrate activities, where appropriate. The author of this paper was appointed to convene the methods working committee and represent the methods groups on the Campbell Collaboration Steering Committee.

Implications for Policy

Currently, the use of research in the formation and evaluation of public policy can be described as marginal, at best. Causes for this lack of use include the public perception that research results are often equivocal. This inconsistency has its roots in complex settings, suboptimal research methodology, and misinterpretation of research findings on the part of researchers, policy makers, practitioners, and the public.

The Campbell Collaboration is an emerging international organization that aims to help make well-informed decisions by preparing, maintaining, and disseminating high-quality, systematic reviews of research on topics related to public policy, beginning with education, crime and justice, and social welfare.

By supporting the production of trustworthy reviews and by disseminating results in an accessible fashion, the Campbell Collaboration will play a crucial role in improving the quality of evidence-based decisions in the public policy arena.

References

Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.

Cochrane, A.L. (1989). Forward. In I. Chalmers, M. Enkin & Keirse, M. (Eds.). *Effective Care in Pregnancy and Childbirth*. Oxford, England: Oxford University Press.

Cook, T. D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37, 131-146.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52, 291-302.

Cooper, H. & Hedges, L.V. (1994). *Handbook of Research Synthesis*. New York: Russell Sage.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.

Feldman, K.A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, 4, 86-102.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8.

Glass, G.V. & Smith, M.L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1, 2-16.

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

Hunter, J.E., Schmidt, F.L. & Jackson, G.B. (1979). *Meta-Analysis: Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage.

Jackson, G.B. (1980). Methods for integrative reviews. *Review of Educational Research*, 50, 438-460.

Light, R.J. (Ed.). (1983) *Evaluation Studies Review Annual, Vol. 8*. Beverly Hills, CA: Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing Up: The Science of Research Reviewing*. Cambridge, MA: Harvard University Press.

Light, R.J. & Smith, P.V. (1971). Accumulating evidence: Procedures for resolving contradictions among research studies. *Harvard Educational Review*, 41, 429-471.

Mulrow, C.D. (1987). The medical review article: State of the Science. *Annals of Internal Medicine*, 106, 465-468.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.). *The Future of Meta-Analysis*. New York: Russell Sage.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3, 1243-1246.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 3, 377-415.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.

Appendix: A Brief History of Systematic Review Methodology

In 1904, Karl Pearson conducted what is believed to be the first statistical synthesis of research. Having been asked to review the evidence on a vaccine against typhoid, Pearson gathered data from eleven relevant studies and for each study he calculated a statistic called the correlation coefficient. He averaged these measures of the treatment's effect across two groups of studies distinguished by the nature of their outcome variable. Based on the average correlations, Pearson concluded that other vaccines were more effective (Pearson, 1904).

In 1932, Ronald Fisher, in his classic text *Statistical Methods for Research Workers*, noted that:

... although few or [no statistical tests] can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are lower than would have been obtained by chance. (Fisher, 1932, p.99).

Fisher then presented a technique for combining the p-values that came from independent tests of the same hypothesis. His work would be followed by more than a dozen papers published prior to 1960 on the same topic (cf., Olkin, 1990).

This early development of procedures for statistically combining results of independent studies largely went unused. However, beginning in the 1960s, social science research experienced a period of rapid growth. By the mid-1970s when Robert Rosenthal and Donald Rubin undertook a review of research studying the effects of interpersonal expectations on behavior they found 345 studies that pertained to their hypothesis (Rosenthal & Rubin, 1978). Almost simultaneously, Gene Glass and Mary Lee Smith were conducting a review of the relation between class size and academic achievement (Glass & Smith, 1979). They found 725 estimates of the relation, based on data from nearly 900,000 students. Smith and Glass also gathered assessments of the effectiveness of psychotherapy. This literature revealed 833 tests of the treatment (Smith & Glass, 1977). Likewise, John Hunter and Frank Schmidt uncovered 866 comparisons of the differential validity of employment tests for black and white workers (Hunter, Schmidt & Hunter, 1979).

Each of these research teams realized that for some topic areas, prodigious amounts of empirical evidence had been amassed on why people act and feel the way they do and on the effectiveness of psychological, social, educational, and health care interventions. These

researchers concluded that the traditional systematic review of research simply would not suffice. Largely independently, the three research teams rediscovered and reinvented Pearson's and Fisher's solutions to their problem.

In discussing his solution, Glass coined the term meta-analysis to stand for "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (Glass, 1976, p. 3). Shortly thereafter, other proponents of meta-analysis demonstrated that traditional review procedures led to inaccurate or imprecise characterizations of the literature, even when the size of the literature was relatively small (Cooper, 1979; Cooper & Rosenthal, 1980).

Rosenthal (1984) presented a compendium of meta-analytic methods covering, among other topics, the combining of significance levels, effect size estimation, and the analysis of variation in effect sizes based on a set of techniques involving assumptions tailored specifically to the analysis of study outcomes.

Another text that appeared in 1984 also helped elevate the research review to a more rigorous level. Light and Pillemer (1984) focused on the use of research synthesis to help decision-making in the social policy domain. Their approach placed special emphasis on the importance of meshing both numbers and narrative for the effective interpretation and communication of synthesis results.

In 1985 with the publication of *Statistical Procedures for Meta-Analysis*, Hedges and Olkin (1985) helped to elevate the quantitative synthesis of research to an independent specialty within the statistical sciences. This book, summarizing and expanding nearly a decade of programmatic developments by the authors, not only covered the widest array of meta-analytic procedures but also established their legitimacy by presenting rigorous statistical proofs.

Simultaneous with the development of meta-analysis procedures, several attempts were undertaken to frame the research review in the terms of a scientific process. In 1971, Feldman wrote, that systematically reviewing and integrating the literature of a field "may be considered a type of research in its own right—one using a characteristic set of research techniques and methods" (Feldman, 1971, p.86). In the same year, Light and Smith (1971) presented a "cluster approach" to research synthesis that was meant to redress some of the deficiencies in the existing strategies. They argued that if treated properly the variation in outcomes among related studies could be a valuable source of information, rather than a source of consternation as it appeared to be when treated with traditional reviewing methods.

Two papers that appeared in the *Review of Educational Research* in the early 1980s brought the meta-analytic and review-as-research perspectives together. First, Jackson (1980) proposed six reviewing tasks "analogous to those performed during primary research" (p. 441). His paper employed a sample of 36 review articles from prestigious social science periodicals to examine the methods used in syntheses of empirical research. His conclusion was that "relatively little thought has been given to the methods for doing integrative reviews" (p. 459).

Cooper (1982) drew the analogy between research synthesis and primary research to its logical conclusion. He presented a five stage model of the review that viewed research synthesis as a data gathering exercise and, as such, applied to it criteria similar to those employed to judge primary research. Cooper argued that, similar to primary research, a research review involves problem formulation, data collection (the literature search), data evaluation, data analysis and interpretation (the meta-analysis), and public presentation. For each stage, Cooper codified the research question, its primary function in the review, and the procedural differences that might cause variation in reviews' conclusions. Also, Cooper applied the notion of threats-to-inferential-validity introduced by Campbell and Stanley (1966; also see Cook & Campbell, 1979) for evaluating the utility of primary research designs to research synthesis. He identified numerous threats to validity associated with reviewing procedures that might undermine the trustworthiness of a research synthesis' findings.

During and after the years that the works mentioned above were appearing, the use of meta-analysis spread from psychology and education through many disciplines, especially social policy analysis (Light, 1983) and the medical sciences (see *Statistics in Medicine*, 1987, Volume 6, Number 3). In 1994, the first edition of *Handbook of Research Synthesis* was published (Cooper & Hedges, 1994).