A BRIEF HISTORY OF RESEARCH SYNTHESIS

IAIN CHALMERS
U.K. Cochrane Centre
LARRY V. HEDGES
University of Chicago
HARRIS COOPER
University of Missouri

Science is supposed to be cumulative, but scientists only rarely cumulate evidence scientifically. This means that users of research evidence have to cope with a plethora of reports of individual studies with no systematic attempt made to present new results in the context of similar studies. Although the need to synthesize research evidence has been recognized for well over two centuries, explicit methods for this form of research were not developed until the 20th century. The development of methods to reduce statistical imprecision using quantitative synthesis (meta-analysis) preceded the development of methods to reduce biases, the latter only beginning to receive proper attention during the last quarter of the 20th century. In this article, the authors identify some of the trends and highlights in this history, to which researchers in the physical, natural, and social sciences have all contributed, and speculate briefly about the "future history" of research synthesis.

AUTHORS' NOTE: This article is dedicated to Frederick Mosteller and Thomas C. Chalmers, whose work has played such a key role in the recent history of research synthesis. We are grateful to Doug Altman, Gerd Antes, Bob Boruch, Mike Clarke, Gene Glass, Tim Horder, Janneke Horn, Andrew Jull, Bruce Kupelnick, Steff Lewis, Alison Macfarlane, Harry Marks, Fred Mosteller, George Davey Smith, Mark Petticrew, Peter Sandercock, and the editors for helpful comments on earlier drafts of this article. Please address correspondence to Iain Chalmers, U.K. Cochrane Centre, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK; telephone: +44 1865 516300; fax: +44 1865 516311; e-mail: ichalmers@cochrane.co.uk.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 25 No. 1, March 2002 12-37

> If, as is sometimes supposed, science consisted in nothing but the laborious accumulation of facts, it would soon come to a standstill, crushed, as it were, under its own weight. The suggestion of a new idea, or the detection of a law, supersedes much that has previously been a burden on the memory, and by introducing order and coherence facilitates the retention of the remainder in an available form. . . . Two processes are thus at work side by side, the reception of new material and the digestion and assimilation of the old; and as both are essential we may spare ourselves the discussion of their relative importance. One remark, however, should be made. The work which deserves, but I am afraid does not always receive, the most credit is that in which discovery and explanation go hand in hand, in which not only are new facts presented, but their relation to old ones is pointed out. (Rayleigh, 1885, p. 20)

So said the professor of physics at Cambridge University in his presidential address to the 54th meeting of the British Association for the Advancement of Science held in Montreal in 1884. More than a century later, research funding agencies, research ethics committees, researchers, and journal editors in most fields of scientific investigation have not taken his injunction seriously. It is true that there have been some improvements recently in the scientific quality of "stand-alone" reviews. When assessing the relation between "new facts" and "old facts" in the Discussion sections of reports of new research, however, scientists very rarely use methods designed to reduce the likelihood that they and their readers will be misled by biases and the play of chance (Clarke & Chalmers, 1998).

SOME EARLY EXAMPLES OF RECOGNITION OF THE NEED FOR RESEARCH SYNTHESIS

Efforts to reduce the likelihood of being misled by biases and chance in research synthesis have quite a long history (Cooper & Hedges, 1994; Hedges, 1987a; Hunt, 1997). In the 18th century, for example, James Lind, a Scottish naval surgeon, was confronted with a plethora of reports about the prevention and treatment of scurvy. The title page of his famous treatise on the disease declares that it contains "An inquiry into the Nature, Causes, and Cure, of that Disease. *Together with a Critical and Chronological View of what has been*

published on the subject [italics added]." Lind (as cited in Hampton, 1998) observed in his text,

> As it is no easy matter to root out prejudices . . . it became requisite to exhibit a full and impartial view of what had hitherto been published on the scurvy, and that in a chronological order, by which the sources of these mistakes may be detected. Indeed, before the subject could be set in a clear and proper light, it was necessary to remove a great deal of rubbish. (p. x).

A couple of decades later, Arthur Young, a gentleman farmer who played a pioneering role in the development of sample surveys, noted that "it is impossible from single experiments, or from a great number, in different lands, separately considered, to deduce a satisfactory proof of the superiority of any method" (as cited in Brunt, 2001, p. 181).

In the early 19th century, the French statistician Legendre developed the method of least squares to solve the problem of combining data from different astronomical observatories where the errors were known to be different (Stigler, 1986), and by the end of the century, some impressive examples of application of the principles of research synthesis had begun to appear. In 1891, for instance, Herbert Nichols published a 76-page review of theories and experiments on the psychology of time.

It was not really until the 20th century, however, that the science of research synthesis as we know it today began to emerge. In 1904, Karl Pearson, director of the Biometric Laboratory at University College London, published a key paper in the *British Medical Journal*. Having been asked to review evidence on the effects of a vaccine against typhoid, Pearson gathered data from 11 relevant studies of immunity and mortality among soldiers serving in various parts of the British Empire. He calculated correlation coefficients for each of the 11 studies (noting that these were very variable and discussing how this variation might be explained) and then synthesized the coefficients within two subgroups, thus producing average correlations (Table 1).

Three years later, Joseph Goldberger (as cited in Winkelstein, 1998), who was working in the laboratory that later became the National Institutes of Health, published an analysis of statistics on bacteriuria in typhoid fever in the District of Columbia. Warren Winkelstein (1998) noted how Goldberger's analysis addressed many of the criteria that research syntheses are now expected to satisfy:

TABLE 1
Inoculation Against Enteric Fever

Correlation Between Immunity and Inoculation			
I.	Hospital staffs	+0.373	±0.021
II.	Ladysmith garrison	+0.445	±0.017
III.	Methuen's column	+0.191	±0.026
IV.	Single regiments	+0.021	±0.033
V.	Army in India	+0.100	±0.013
	Mean value	+0.226	
Correlation Between Mortality and Inoculation			
VI.	Hospital staffs	+0.307	±0.128
VII.	Ladysmith garrison	–0.010	±0.081
VIII.	Methuen's column	+0.300	±0.093
IX.	Single regiments	+0.119	±0.022
X.	Various military hospitals	+0.194	±0.022
XI.	Army in India	+0.248	±0.050
	Mean value	+0.226	

> First, a review of the literature identifies pertinent studies. Goldberger identified 44 studies and provided comprehensive references in a bibliography. Second, specific criteria are used to select studies for analysis. Goldberger used a newly developed serum agglutination test to separate reliable studies from those he considered unreliable. Third, data from the selected studies are abstracted. Goldberger tabulated the raw data from 26 selected studies. Fourth, statistical analysis of the abstracted data is implemented. Goldberger calculated the mean rate of bacteriuria from the pooled data. (p. 717)

Goldberger's attention to each of these steps is an early exemplar of the need to distinguish these two distinct methodological challenges in research synthesis—first, to take measures to reduce bias, then to consider whether meta-analysis can be used to reduce statistical imprecision.

There are other examples of approaches to research synthesis during the first half of the 20th century. In 1916, for example, Thorndike and Ruger derived average results from two experiments comparing the effects of outside air and recirculated air in classrooms on children's ability to add, check numbers and letters, and to find and copy addresses. In 1933, Peters presented a summary of more than 180 experiments on the effects of "character education" on schoolchildren in Pennsylvania. And during the 1930s, research synthesis also began in physics (Birge, 1932) and agriculture (Yates & Cochran, 1938).

A NOTE ON TERMINOLOGY

A variety of terms have been used to describe all or some of the processes to which we have alluded—particularly *research synthesis, systematic review*, and *meta-analysis*.

Our reason for using the term *research synthesis* is primarily because the term has been used extensively by the social scientists who led the development of the science and practice of this kind of research over the post–World War II period.

We might have chosen *systematic review* as an alternative term. There are certainly instances of use of the term *systematic review* earlier than *research synthesis* (Mandel, 1936), but it is uncertain whether use of the former during the pre–World War II period reflected the very structured process that we understand by the term today. Although it was used in the 1970s (Shaikh, Vayda, & Feldman, 1976), it was not until the late 1990s that the term *systematic review* became more widely used. This probably reflected two factors in particular. First, it was the term used by Cochrane (1989) in his foreword to a compilation of research syntheses relating to many aspects of care during pregnancy and childbirth published during the late 1980s (I. Chalmers, Enkin, & Keirse, 1989). The term was subsequently promoted by people concerned to draw a distinction between a process involving measures to control biases in research synthesis and the optional element of that process involving quantitative, statistical procedures, for which they suggested reserving the term *meta-analysis* (I. Chalmers & Altman, 1995; Egger, Smith, & Altman, 2001).

Glass introduced the term *meta-analysis* in 1976 in a presidential address stressing the need for better synthesis of research results. Those who liked neologisms adopted it rapidly, and it was used in the titles of some of the earliest substantive texts on statistical methods for quantitative synthesis (Hedges & Olkin, 1985). It became gradually clear, however, that the word was being used in a variety of ways and that it was intensely antigenic to some people, particularly those who challenged the use of quantitative synthesis to reduce statistical imprecision. Thus, Eysenck (1978) referred to "mega-silliness," Shapiro (1994) to "shmeta-analysis," and Feinstein (1995) to "statistical alchemy for the 21st century." These critics and others showed no appreciation of the need to adopt methods to reduce bias in reviews of research—regardless of whether statistical synthesis could be used to

reduce statistical imprecision. Restricting the term *meta-analysis* to the process of statistical synthesis seemed a way of helping people understand that the science of research synthesis comprises a variety of methods addressing a variety of challenges.

This convention has now been adopted in some quarters. For example, the second edition of the publication *Systematic Reviews* is subtitled *Meta-Analysis in Context* (Egger, Davey Smith, & Altman, 2001), and the fourth edition of Last's (2001) *Dictionary of Epidemiology* gives definitions as follows:

> SYSTEMATIC REVIEW The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. Meta-analysis may be, but is not necessarily, used as part of this process. (pp. 176-177)
>
> META-ANALYSIS The statistical synthesis of the data from separate but similar, i.e. comparable studies, leading to a quantitative summary of the pooled results. (p. 114)

A definition of our chosen term—*research synthesis*—will have to await publication of the fifth edition of the dictionary!

REDUCING STATISTICAL IMPRECISION IN RESEARCH SYNTHESIS (META-ANALYSIS)

The development of methods for reducing statistical imprecision in research synthesis (meta-analysis) antedated the development of methods for controlling biases. Most statistical techniques used today in meta-analysis have their origins in Gauss's and Laplace's work (Egger, Smith, & O'Rourke, 2001), which was disseminated in a "textbook" on "meta-analysis" for astronomers published in 1861 by the British Astronomer Royal (Airy, 1861). Karl Pearson's (1904) use of statistical methods for research synthesis (see earlier discussion) at the beginning of the following century is an early example of the use of these techniques in medical research. A statistical paper published a few years later by the physiologists Rietz and Mitchell (1910-1911) considered what kind of information a series of experiments can produce.

Several statisticians working in agricultural research in Britain in the 1930s developed and applied these approaches in that field

(Cochran, 1937; Fisher, 1932; Pearson, 1933; Tippett, 1931; Yates & Cochran, 1938). In particular, Ronald Fisher (1932), 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" (p. 99).

Fisher (1932) then presented a technique for combining the *p* values that came from independent tests of the same hypothesis. Interest in research synthesis among statisticians continued through the Second World War, and Fisher's work was followed by more than a dozen papers published on the topic prior to 1960 (see, e.g., Cochran, 1954; Jones & Fiske, 1953; Mosteller & Bush, 1954).

These statistical procedures for combining results of independent studies were not widely used until the 1960s, when social science research began to experience a period of rapid growth. By the mid-1970s, social scientist reviewers in the United States found themselves having to deal with, for example, 345 studies of the effects of interpersonal expectations on behavior (Rosenthal & Rubin, 1978), 725 estimates of the relation between class size and academic achievement (G. Glass & Smith, 1979), 833 tests of the effectiveness of psychotherapy (M. Smith & Glass, 1977), and 866 comparisons of the differential validity of employment tests for Black and White workers (Hunter, Schmidt, & Hunter, 1979). Largely independently, the research teams addressing these issues rediscovered and reinvented Pearson's and Fisher's solutions to the problem they faced. In discussing his solution, Gene Glass (1976) coined the term *meta-analysis* to refer to "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (p. 3). By the middle of the following decade, Rosenthal (1984) had presented a compendium of meta-analytic methods.

The publication of *Statistical Methods for Meta-Analysis* by Hedges and Olkin in 1985, a key methods paper by Richard Peto and his colleagues published the same year (Yusuf, Peto, Lewis, Collins, & Sleight, 1985), and the proceedings of a meeting convened by the U.S. National Heart, Lung and Blood Institute and the National Cancer Institute published as a special issue of *Statistics in Medicine* in 1987 all helped to secure recognition of the practice of quantitative synthesis of research among statisticians.

REDUCING BIASES IN RESEARCH SYNTHESIS

The development and adoption of methods to reduce biases in research synthesis has tended to lag behind the development of methods to reduce statistical imprecision. With the massive increase in the scale of scientific research after the Second World War, people working in a wide variety of fields began to recognize a need to organize and evaluate the accumulating bodies of research evidence (see e.g., Chase, Sutton, & First, 1959; Greenhouse, 1958; Herring, 1968; Lide, 1981; Lide & Rossmassler, 1973; Schoolman, 1982). It soon became clear that research synthesis threw up a far more complex range of methodological issues than simply the choice of methods for statistical synthesis. In many of the physical sciences, for example, research synthesis became referred to as "critical evaluation," with a substantial emphasis on discovering biases in the individual experiments themselves and developing sets of values of related physical properties that were as consistent and free from bias as possible (see Rosenfeld, 1975; Touloukian, 1975; Zwolinski & Chao, 1972).

The challenge was spelled out well by an American social scientist, David Pillemer (1984), who characterized the usual approach to reviews as

> subjective, relying on idiosyncratic judgments about such key issues as which studies to include and how to draw overall conclusions. Studies are considered one at a time, with strengths and weaknesses selectively identified and casually discussed. Since the process is informal, it is not surprising that different reviewers often draw very different conclusions from the same set of studies. (p. 28)

With a growth of acknowledgment that methodological rigor is needed to secure the validity of research reviews, just as it is for primary research (Cooper, 1982; Jackson, 1980), there was increased appreciation of the range of methods required to prepare unbiased syntheses of research. Social scientists in the United States led the way in this respect. They recognized, for example, that the methods used to select evidence for inclusion in reviews were potentially major sources of bias, particularly as methodological research began to reveal that researchers were more likely to report studies that had yielded "positive" (statistically significant) results. A study of reports published in a sample of psychology journals published in the late

1950s revealed that a very high proportion reported statistically significant results (Sterling, 1959). Investigations of the magnitude of the resulting publication biases made it clear that efforts to control biases in research synthesis would need to address these (Hedges, 1984; Rosenthal, 1979).

With some isolated exceptions (Beecher, 1955; Greenhouse, 1958), people working in health research were relative latecomers to research synthesis. In 1972, Cochrane drew attention to the adverse consequences for the British National Health Service of collective ignorance about the effects of many elements of health care, and in an essay published in 1979, he observed that "it is surely a great criticism of our profession that we have not organised a critical summary, by speciality or subspeciality, adapted periodically, of all relevant randomised controlled trials" (p. 8).

Cochrane's emphasis on randomized controlled trials was relevant to one element of an issue that had emerged among social scientists, namely, which criteria to use for judging when studies could be regarded as sufficiently unbiased for inclusion in research syntheses.

A few "critical summaries of randomized trials" in health care were done during the 1970s (Andrews, Guitar, & Howie, 1980; "Aspirin After Myocardial Infarction," 1980; I. Chalmers, 1979; T. Chalmers, Matta, Smith, & Kunzler, 1977; Stjernsward, Muenz, & von Essen, 1976), but it was not until the following decade that research syntheses of health research began to appear in any numbers and that the scientific issues that needed to be addressed were articulated clearly for people in the health professions. In Kenneth Warren's (1981) seminal book on coping with the biomedical literature, Edward Kass (1981) noted that "reviews will need to be evaluated as critically as are primary scientific papers" (p. 82). Cynthia Mulrow began that process in a seminal article published in the *Annals of Internal Medicine* in 1987 that concluded that review articles published in four major medical journals had not used scientific methods to identify, assess, and synthesize information. Other influential articles addressed to a medical readership were published the same year (L'Abbé, Detsky, & O'Rourke, 1987; Peto, 1987; Sacks, Berrier, Reitman, Ancona-Berk, & Chalmers, 1987).

During the late 1980s, global collaboration among investigators responsible for randomized trials in cancer and cardiovascular disease resulted in research syntheses based on collaborative reanalyses of

individual patient data derived from almost all the randomized trials of certain therapies (Advanced Ovarian Cancer Trialists' Group, 1991; Antiplatelet Trialists' Collaboration, 1988; Early Breast Cancer Trialists' Collaborative Group, 1988). These endeavors became yardsticks against which the scientific quality of other research syntheses in the field of health care would be judged. International collaboration during this time also led to the preparation of hundreds of systematic reviews of controlled trials relevant to the care of women during pregnancy and childbirth. These were published in a 1,500-page, two-volume book, *Effective Care in Pregnancy and Childbirth* (I. Chalmers et al., 1989), deemed an important landmark in the history of controlled trials and research synthesis (Cochrane, 1989; Mosteller, 1993). Three years later, the results were published of a similar project assessing the effects of care of newborn infants (Sinclair & Bracken, 1992).

Within the social sciences, the importance of this phase in the history of research synthesis was reflected in Lipsey and Wilson's (1993) assessment of more than 300 quantitative research syntheses of behavioral and educational intervention studies and Cooper and Hedges's (1994) 570-page *Handbook of Research Synthesis*.

Within health care, the practical importance of improving the scientific quality of reviews was given great impetus by an analysis conducted by a group of researchers led by Thomas Chalmers and Frederick Mosteller: A comparison of textbook advice on the treatment of people with myocardial infarction with the results of systematic syntheses of relevant randomized controlled trials showed that valid advice on some lifesaving treatments had been delayed for more than a decade, and other forms of care had been promoted long after they had been shown to be harmful (Antman, Lau, Kupelnick, Mosteller, & Chalmers, 1992). This report made it abundantly clear that the failure of researchers to prepare reviews of therapeutic research systematically could have very real human costs.

ACADEMIC RECOGNITION OF RESEARCH SYNTHESIS AS RESEARCH

Over recent decades, research synthesis has been widely seen within academia as second-class, scientifically derivative work, unworthy of mention in reports and documents intended to confirm the scientific credentials of individuals and institutions. Indeed,

systematic reviews are sometimes characterized as "parasitic recycling" of the work of those engaged in the real business of science—which is to add yet more data to the atomized output of the overall scientific enterprise.

As Bentley Glass (1976) noted more than a quarter of a century ago,

> The vastness of the scientific literature makes the search for general comprehension and perception of new relationships and possibilities every day more arduous. [Yet] the editor of the critical review journal finds each year a growing reluctance on the part of the best qualified scientists to devote the necessary time and energy to this task. (p. 417)

As Glass observed elsewhere in the article,

> The man who adds his bits of fact to the total of knowledge has a useful and necessary function. But who would deny that a role by far the greater is played by the original thinker and critic who discerns the broader outlines of the plan, who synthesises from existing knowledge through detection of the false and illumination of the true relationships of things a theory, a conceptual model, or a hypothesis capable of test. (p. 417)

Horder's (2001) recently published discussion of the relationship within developmental biological thinking between the organizer concept (articulated in the 1920s) and the concept of positional information (proposed in the 1970s) provides a compelling contemporary illustration of the kind of review for which B. Glass (1976) was calling. Horder concluded his review by noting that "'science' must be acknowledged as being a historical edifice: it not only consists of the latest results, but, more accurately, it is composed of the sum total of a massive accumulation of earlier-acquired data, interpretation and assumptions" (p. 124).

Most people within contemporary academia have not yet recognized (let alone come to grips with) the rationale for and methodological challenges presented by research synthesis. Neither have they grasped that the rationale applies in all spheres of research, not only in the areas of applied social and medical research in which it has begun to flourish. Researchers in applied medical research who have begun to apply the methods of rigorous research synthesis to animal experiments (Horn, de Haan, Vermeulen, Luiten, & Limburg, 2001; I.

Roberts, personal communication, July 2001), for example, have begun to uncover some unsettling findings. A systematic review of the effects of a calcium antagonist (nimodipine) in animal model experiments of focal cerebral ischaemia has raised questions about whether it was ever justified to proceed to controlled trials in humans involving nearly 7,000 patients. A systematic review of the studies in patients did not detect any evidence of beneficial effects of this drug (Horn & Limburg, 2001).

As early as 1971, Feldman wrote that systematically reviewing and integrating research evidence "may be considered a type of research in its own right—one using a characteristic set of research techniques and methods" (p. 86). In the same year, Light and Smith (1971) noted that it was impossible to address some hypotheses other than through analysis of variations among related studies and that valid information and insights could not be expected to result from this process if it depended on the usual, scientifically undisciplined approach to reviews.

In 1977, Eugene Garfield drew attention to the importance of scientific review articles to the advancement of original research: Review articles have high citation rates, and review journals have high impact factors. He proposed a new profession—"scientific reviewer" (Garfield, 1977)—and his Institute for Scientific Information went on to cosponsor (with Annual Reviews Inc.) an annual award for "Excellence in Scientific Reviewing" administered by the National Academy of Sciences (Garfield, 1979).

In the early 1980s, this reviews-as-research perspective was made explicit in two papers published in the *Review of Educational Research*. First, after examining the methods used in 36 review articles sampled from prestigious social science periodicals and concluding that "relatively little thought has been given to the methods for doing integrative reviews," Jackson (1980) proposed six reviewing tasks "analogous to those performed during primary research." A couple of years later, one of us (HC) drew the analogy between research synthesis and primary research and presented a five-stage model of research synthesis involving problem formulation, data collection (the search for potentially eligible studies), data evaluation (quality assessment), data analysis and interpretation (meta-analysis when appropriate), and public presentation (Cooper, 1982). The paper also applied to research synthesis the notion of threats to inferential

validity that had been introduced by Campbell and Stanley (1966) for evaluating the design of primary research (also see Cook & Campbell, 1979).

The promotion of this perspective was given impetus by the publication of two important books in the early 1980s. The more "scholarly" of these was a multiauthor issue of *Evaluation Studies Review Annual* edited by Richard Light (1983) that contained 15 contributions addressing methodological issues and procedures, followed by 20 separate articles illustrating how the methodologies had been applied in practice. In 1984, Richard Light and David Pillemer published their highly readable and influential book titled *Summing Up: The Science of Reviewing Research*. This became a key resource not only for their fellow social scientists but also for the people who were beginning to take this agenda seriously in health care. Building on the principles and resources developed by social scientists, Oxman and Guyatt (1988), for example, published guidelines for assessing the scientific quality of reviews in health care research.

Academic recognition of the science of research synthesis has been growing over recent years. There are examples of its wholehearted incorporation in the methods used in some areas of basic research (e.g., small particle physics and some areas of psychology) and in some areas of applied research (e.g., education and some aspects of health care). As Mark Petticrew (2001) noted in an article exposing some myths and misconceptions about research synthesis, there are research syntheses in such diverse topics as advertising, agriculture, archaeology, astronomy, biology, chemistry, criminology, ecology, education, entomology, law, manufacturing, parapsychology, psychology, public policy, zoology, and even eyewitness accounts of the Indian rope trick.

Even the graphical devices for presenting the results of research syntheses show similarities across widely different spheres of investigation. A form of presentation now often referred to as a "forest plot" (Lewis & Clarke, 2001) plots point estimates from different experiments along with their error bars. This form of presentation is now widely used by health researchers but has also been very commonly used by physicists. For example, Taylor, Parker, and Langenberg (1969) used this method to illustrate the empirical evidence from 12 experiments on an atomic constant called the fine structure constant (Hedges, 1987b).

Because the eye is drawn to the longer error bars in these forest plots, data from the less informative studies have a relatively greater visual effect. To compensate for this distorting feature, boxes with sizes reflecting the inverse of the variance of the estimate derived from each study have been used to mark the point estimates. This device was introduced during the 1980s, principally by medical researchers, and appears to have been inspired by a paper published in 1978 by McGill, Tukey, and Larsen (S. Lewis, personal communication, August 2001).

Even when no study within a group of related studies is sufficiently large to be informative, forest plots may help to reveal a discernable pattern. For example, to test the hypothesis that a widely used form of resuscitation used in critically ill patients—infusion of human albumin solution—reduces mortality, the Albumin Reviewers (2001) analyzed mortality data in 18 randomized trials. In 4 of these trials, none of the participants died, and the number dying in the remaining 14 trials ranged from only 1 to 12. Nevertheless, not only did the forest plot of estimates derived from the 64 deaths that did occur provide no evidence to support the use of a treatment that has been used widely for more than half a century, it actually suggested that human albumin solution increases the risk of death in critically ill patients.

Partly because research synthesis sometimes yields unwelcome results that challenge strongly held opinions and other vested interests, there is very variable acceptance of the scientific principles on which the process is founded. For example, although there is a strong tradition of research synthesis among American social scientists, only a tiny minority of British social scientists has any experience of this form of research, and many appear to be actively hostile to it. Within health research too, attitudes to research synthesis can vary dramatically. Thus, although the *New England Journal of Medicine* published some very important research syntheses during the 1980s, the journal has been overtly hostile to reports of such studies more recently.

As we discuss next, however, we believe that the future status of research synthesis as research is more likely to be shaped by forces outside academia than by those within it. Consumers of research have begun to point out more forcibly that "atomized," unsynthesized products of the research enterprise are of little help to people who wish to use research to inform their decisions.

THE USE OF RESEARCH SYNTHESES TO INFORM POLICY AND PRACTICE

One of the forces shaping perceptions of research synthesis is the growing appetite for research evidence among policy makers, practitioners, and the public more generally. This appetite started to become manifest during the last decade of the 20th century, but earlier examples exist. In a biographical article about the statistician Frank Yates, Michael Healy (1995) noted that

> as the war began and it became clear that phosphate and potash fertilizers were going to be extremely scarce, Yates with E. M. Crowther, the head of the Chemistry Department at Rothampsted, brought together and analyzed all the published experiments on fertilizer responses that they could lay their hands on (Yates & Crowther, 1941). . . . An example of its findings is the statement that the application of 1 cwt/acre of sulphate of ammonia at a cost of £4m would be expected to yield an extra crop to the value of £11m. As a result of this study, fertilizer rationing in the UK was placed on a rational basis and some of the survival of wartime Britain can be set to its credit. Other studies of a similar nature were undertaken at the same time, notably one on the feeding of dairy cows (Yates, Boyd, & Pettit, 1942). It was to be some twenty years before other fields of application began to realise that it was absurd not to look critically from time to time at the collected results of experimental work before deciding upon action, whether in the application of the research or in deciding upon a programme for further research. (p. 277)

It is indeed "absurd not to look critically from time to time at the collected results of experimental work before deciding upon action," but it was not really until the late 1980s that acceptance of the need for research synthesis among policy makers and practitioners emerged, if only because the volume of primary data they were having to cope with was becoming overwhelming. Eleanor Chelimsky (1994), formerly Assistant Comptroller General for Program Evaluation and Methodology at the U.S. General Accounting Office, described the situation that she and her colleagues faced at the beginning of the 1980s:

> I hoped that synthesis could dramatize, for our legislative users, not only what was, in fact, known, but also what was *not* known. In that way, I thought we could then focus attention on what needed to be learned (and how to learn it), in time to answer that policymaker's questions before, say, the next program reauthorization. Based on the

> legislative record for some programs, it seemed obvious that, on the one hand, the distinction between well-established knowledge and mere opinion was not always recognized, and on the other, that what needed to be research *as a next step* was sometimes not even glimpsed. . . . In short, it seemed reasonable to try to develop a systematic method for using synthesis as a way to channel relevant existing information to answer specific congressional questions. (pp. 3-4)

By 1994, 30 research syntheses had been prepared for Congress by the U.S. General Accounting Office on topics ranging from access to special education to the effectiveness of chemical weapons (Chelimsky, 1994).

Syntheses of the results of controlled trials in cancer, cardiovascular disease, and the various forms of care offered to women during pregnancy and childbirth became increasingly accepted during the 1990s as helpful by those wishing to make more informed decisions in health care. Research syntheses were identified for early support when a Research and Development Programme to support the U.K.'s National Health Service (NHS) was launched in 1991 (Peckham, 1991), and this was reflected in the creation of two centers—the NHS Centre for Reviews and Dissemination and the U.K. Cochrane Centre—to help tackle this agenda.

During the 1990s, the importance of research synthesis also became acknowledged among those considering proposals for new research. The NHS Health Technology Assessment Programme and the British and Dutch Medical Research Councils, for example, all began to require systematic reviews of existing research as a precondition for considering funding for proposed additional studies. In Denmark, the national research ethics committee system began to require applicants for ethical approval of proposed new research to show by reference to syntheses of existing evidence that proposed new studies were necessary and that they had been designed to take account of the lessons from previous research (I. Chalmers, 2001). These developments among organizations responsible for the funding and ethical approval of research began to force academia to take research synthesis more seriously. This trend is likely to be given further impetus by the widely publicized death of a young volunteer in a physiological experiment, the design of which had been inadequately informed by a systematic review of preexisting evidence about hazards (Clark, Clark, & Djulbegovic, 2001).

In a history of research synthesis published in 1997, Morton Hunt concluded that systematic reviews of research evidence appear to be having an influence on policies and practices in schools, hospitals, state welfare programs, mental health clinics, courts, prisons, and other institutions. Today's questions about the deployment of limited resources for the benefit of the public may not be those about phosphate and potash fertilizers to which answers were sought more than half a century ago, but the potential for research synthesis to inform decisions about policy and practice remains substantial and still inadequately exploited.

This is not to suggest that there have been no areas in which rigorously conducted systematic reviews have been uncontentious, even when the component studies of the review have been controlled experiments. Reactions to the Cochrane review of the effects human albumin solution in critically ill patients (Albumin Reviewers, 2000) provide a celebrated or notorious example, depending on one's point of view. Reviews of observational data can be relied on to generate even more heat, however, particularly if meta-analysis has been used to synthesize data from nonexperimental studies (Egger, Schneider, & Davey Smith, 1998).

USING ELECTRONIC MEDIA TO KEEP RESEARCH SYNTHESES UP TO DATE AND CORRECT

The growth in appetite for research syntheses among policy makers, practitioners, people using services, and others is a growth in appetite for information that is up to date and correct. This reasonable expectation has posed additional challenges to the research community. The potential for meeting these challenges increased dramatically with the evolution of electronic publishing. In the late 1980s, the international group that had prepared syntheses of research on the effects of forms of care offered during pregnancy and childbirth published their findings in various forms, one of which used electronic media (I. Chalmers, 1988). This meant that syntheses published on paper could be updated and corrected as new data or errors were identified.

At the end of 1992, the U.K. Cochrane Centre was established to draw on this experience and to facilitate the creation of an international network to prepare and maintain systematic reviews of the

effects of interventions across the whole of health care. At the end of the following year, an international network of individuals—the Cochrane Collaboration—emerged from this initiative (Antes & Oxman, 2001; Bero & Rennie, 1995; I. Chalmers, 1993; I. Chalmers, Sackett, & Silagy, 1997; Dickersin & Manheimer, 1998; Oxman, 2001). Since the launch of *The Cochrane Database of Systematic Reviews* in 1995, the research syntheses that have been published by this still young organization have been having an encouraging effect on the content of international guidelines and policies in health care.

Others have recognized that considerable scope exists for extending the collaborative, international arrangements developed by the Cochrane Collaboration for preparing, maintaining, and disseminating research syntheses. In his presidential address to the Royal Statistical Society in 1996, Adrian Smith, professor of statistics at Imperial College London, welcomed the creation of the Cochrane Collaboration and asked,

> But what is so special about medicine? We are, through the media, as ordinary citizens, confronted daily with controversy and debate across a whole spectrum of public policy issues. But typically, we have no access to any form of systematic "evidence base"—and therefore no means of participating in the debate in a mature and informed manner. Obvious topical examples include education—what *does* work in the classroom?—and penal policy—what *is* effective in preventing reoffending? Perhaps there is an opportunity here for the Society—together with appropriate allies in other learned societies and the media—to launch a campaign, directed at developing analogues to the Cochrane Collaboration, to provide suitable evidence bases in other areas besides medicine, with the aim of achieving a quantal shift in the quantitative maturity of public policy debates. (pp. 369-370)

The same principles that have led to the rapid evolution of the Cochrane Collaboration were adopted when the Campbell Collaboration was inaugurated at the beginning of the 21st century. This sibling organization, which draws particularly on the wealth of relevant experience among social scientists in the United States, is preparing, maintaining, and disseminating systematic reviews of the effects of social and educational policies and practices (Boruch, Petrosino, & Chalmers, 1999; Campbell Collaboration Steering Group, 2000). Importantly, the Cochrane and Campbell Collaborations will work

together to develop methods to improve the quality of research syntheses (Clarke & Cooper, 2000).

THE "FUTURE HISTORY" OF RESEARCH SYNTHESIS

> Upon this gifted age, in its darkest hour,
> Rains from the sky a meteoric shower of facts . . .
> They lie unquestioned, uncombined.
> Wisdom enough to leach us of our ill is daily spun;
> But there exists no loom to weave it into fabric . . .
>
> —Edna St. Vincent Millay (1892-1950)
> "Huntsman, What Quarry?"

> An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganised. We need a sort of clearing-house for the mind: a depot where knowledge and ideas are received, sorted, summarised, digested, clarified and compared.
>
> —H. G. Wells
> (quoted in *The Sunday Independent*, August 30, 1997)

Although it is widely agreed that science is cumulative, people have only very recently begun to acknowledge that scientists have a responsibility to cumulate scientifically. As this article has shown, there is scattered evidence that this has been acknowledged by some scientists for at least a century, but it was really only during the last quarter of the 20th century that the need to develop and apply methods to improve research synthesis became more widely recognized.

So far, most of the resulting activity has been directed at preparing stand-alone research syntheses. As Lord Rayleigh (1885) noted more than a century ago, however,

> The work which deserves, but I am afraid does not always receive, the most credit is that in which discovery and explanation go hand in hand, in which not only are new facts presented, but their relation to old ones is pointed out. (p. 20)

The digestion and assimilation of old material and the integration of new material with existing evidence are both essential elements of

TABLE 2
Classification of Discussion Sections in Randomized Controlled Trial Reports Published in May 1997 and May 2001 in Five Major General Medical Journals

Classification	*May 1997* (n = *26*)	*May 2001* (n = *33*)
First trial addressing the question	1	3
Contained an updated systematic review integrating the new results	2	0
Discussed a previous review but did not attempt to integrate the new results	4	3
No apparent systematic attempt to set the new results in the context of other trials	19	27

scientific endeavors, and this needs to be reflected in the methodological quality of the Discussion sections of reports of primary research. As the data in Table 2 show, even in papers published in five highly respected general medical journals, it remains very rare for the results of new controlled trials to be set in the context of systematic reviews of other, similar studies (Clarke, Alderson, & Chalmers, 2001; Clarke & Chalmers, 1998).

Some years ago, the editor of this journal suggested that a case could be made for calling for a moratorium on proposals for additional primary research until the results of existing research had been incorporated in scientifically defensible reviews (Bausell, 1993). Although he may have thought this a radical a proposition at the time, there is evidence that funders of research are beginning to take account of such views.

The future status of and investment in research synthesis thus seem more likely to be shaped by external pressures from the users of research information than by traditional attitudes within academia to this kind of work. Indeed, we predict that we are moving toward a time when the public will begin to ask increasingly penetrating questions about why it has taken academia so long to begin to practice the kind of scientific self-discipline for which Lord Rayleigh called in 1885.

More radically, the public may also begin to ask why researchers addressing similar or related questions do not collaborate effectively or make their raw data publicly available for others to exploit. The advantages of collaborative investigations using pooled raw data have been made abundantly clear by the global clinical trialists' collaborations in cancer and heart disease in particular (Advanced Ovarian

Cancer Trialists' Group, 1991; Antiplatelet Trialists' Collaboration, 1988; Early Breast Cancer Trialists' Collaborative Group, 1988). Physicists have led the way in making raw data publicly available in electronic form (Ginsparg, 1998). As Gene Glass (2001) noted, "Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in the abbreviated reports of research in journals and other printed sources" (p. 12). We agree with him that the future history of research synthesis should be based increasingly on the creation of publicly accessible archives of raw data.

REFERENCES

Advanced Ovarian Cancer Trialists' Group. (1991). Chemotherapy in advanced ovarian cancer: An overview of randomised clinical trials. *British Medical Journal*, *303*, 884-893.

Airy, G. B. (1861). *On the algebraical and numerical theory of errors of observations and the combination of observations*. London: Macmillan.

Albumin Reviewers (Alderson, P., Bunn, F., Lefebvre, C., Li Wan Po, A., Li, L., Roberts, I., et al.). (2000). *Human albumin solution for resuscitation and volume expansion in critically ill patients* (Cochrane Review). Retrieved June 2000 from *The Cochrane Library* (Issue 2) database.

Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech and Hearing Disorders*, *45*, 287-307.

Antes, G., & Oxman, A. D. (for the Cochrane Collaboration). (2001). The Cochrane Collaboration in the 20th century. In M. Egger, G. Davey Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 447-458). London: BMJ Books.

Antiplatelet Trialists' Collaboration. (1988). Secondary prevention of vascular disease by prolonged anti-platelet treatment. *British Medical Journal*, *296*, 320-331.

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. *Journal of the American Medical Association*, *268*, 240-248.

Aspirin after myocardial infarction [Editorial]. (1980). *Lancet*, *1*, 1172-1173.

Bausell, B. B. (1993). After the meta-analytic revolution. *Evaluation and the Health Professions*, *16*, 3-12.

Beecher, H. K. (1955). The powerful placebo. *Journal of the American Medical Association*, *159*, 1602-1606.

Bero, L., & Rennie, D. (1995). The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Journal of the American Medical Association*, *274*, 1935-1938.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review*, *40*, 207-227.

Boruch, R., Petrosino, A., & Chalmers, I. (1999). The Campbell Collaboration: A proposal for systematic, multinational, and continuous reviews of evidence. In P. Davies, A. Petrosino, & I. Chalmers (Eds.), *The effects of social and educational interventions: Developing an*

infrastructure for international collaboration to prepare, maintain and promote the accessibility of systematic reviews of relevant research (pp. 1-22). London: University College London School of Public Policy.

Brunt, L. (2001). The advent of the sample survey in the social sciences. *The Statistician*, *50*, 179-189.

Campbell Collaboration Steering Group. (2000). *Decisions and action plans made at the working inaugural meeting of the Campbell Collaboration*. Retrieved October 2001 from http://campbell.gse.upenn.edu/papers/2_decisions.html

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chalmers, I. (1979). Randomized controlled trials of fetal monitoring 1973-1977. In O. Thalhammer, K. Baumgarten, & A. Pollak (Eds.), *Perinatal medicine* (pp. 260-265). Stuttgart, Germany: Georg Thieme.

Chalmers, I. (Ed.). (1988). *The Oxford database of perinatal trials*. Oxford: Oxford University Press.

Chalmers, I. (1993). The Cochrane Collaboration: Preparing, maintaining and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences*, *703*, 156-163.

Chalmers, I. (2001). Using systematic reviews and registers of ongoing trials for scientific and ethical trial design. In M. Egger, G. Davey Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 429-443). London: BMJ Books.

Chalmers, I., & Altman, D. G. (Eds.). (1995). *Systematic reviews*. London: BMJ Books.

Chalmers, I., Enkin, M., & Keirse, M.J.N.C. (Eds.). (1989). *Effective care in pregnancy and childbirth*. Oxford: Oxford University Press.

Chalmers, I., Sackett, D., & Silagy, C. (1997). The Cochrane Collaboration. In A. Maynard & I. Chalmers (Eds.), *Non-random reflections on health services research: On the 25th anniversary of Archie Cochrane's* effectiveness and efficiency (pp. 231-249). London: BMJ Books.

Chalmers, T. C., Matta, R. J., Smith, H., & Kunzler, A.-M. (1977). Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine*, *297*, 1091-1096.

Chase, R. A., Sutton, S., & First, D. (1959). Bibliography: Delayed auditory feedback. *Journal of Speech and Hearing Research*, *2*, 193-200.

Chelimsky, E. (1994, June). *Politics, policy, and research synthesis*. Keynote address before the National Conference on Research Synthesis, sponsored by the Russell Sage Foundation, Washington, DC.

Clark, O., Clark, L., & Djulbegovic, B. (2001). Is clinical research still too haphazard? *Lancet*, *358*, 1648.

Clarke, M., Alderson, P., & Chalmers, I. (2001). *Discussion sections in reports of controlled trials published in general medical journals: No evidence of progress between Prague and Barcelona*. Paper presented at the 4th International Congress on Peer Review in Biomedical Publication. Barcelona, Spain, 14-16, September. Manuscript submitted for publication.

Clarke, M., & Chalmers, I. (1998). Discussion sections in reports of controlled trials published in general medical journals: Islands in search of continents? *Journal of the American Medical Association*, *280*, 280-282.

Clarke, M., & Cooper, H. (2000). *Discussion paper on Cochrane and Campbell methods groups*. Retrieved October 2001, from http://campbell.gse.upenn.edu/contents.html

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, *4*(Suppl.), 102-118.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101-129.

Cochrane, A. L. (1972). *Effectiveness and efficiency. Random reflections on health services*. London: Nuffield Provincial Hospitals Trust.

Cochrane, A. L. (1979). 1931-1971: A critical review, with particular reference to the medical profession. In *Medicines for the year 2000* (pp. 1-11). London: Office of Health Economics.

Cochrane, A. L. (1989). Foreword. In I. Chalmers, M. Enkin, & M.J.N.C. Keirse (Eds.), *Effective care in pregnancy and childbirth* (pp. vii). Oxford, UK: Oxford University Press.

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

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

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

Dickersin, K., & Manheimer, E. (1998). The Cochrane Collaboration: Evaluation of health care and services using systematic reviews of the results of randomized controlled trials. *Clinical Obstetrics and Gynecology*, *41*, 315-331.

Early Breast Cancer Trialists' Collaborative Group. (1988). Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. *New England Journal of Medicine*, *319*, 1681-1692.

Egger, M., Davey Smith, G., & Altman, D. (Eds.). (2001). *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). London: BMJ Books.

Egger, M., Davey Smith, G., & O'Rourke, K. (2001). Rationale, potentials, and promise of systematic reviews. In M. Egger, G. Davey Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 3-19). London: BMJ Books.

Egger, M., Schneider, M., & Davey Smith, G. (1998). Spurious precision? Meta-analysis of observational studies. *British Medical Journal*, *316*, 140-144.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychology*, *33*, 517.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71-79.

Feldman, K. A. (1971). Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, *44*, 86-102.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver and Boyd.

Garfield, E. (1977). Proposal for a new profession: Scientific reviewer. *Essays of an Information Scientist*, *3*, 84-87.

Garfield, E. (1979). The NAS James Murray Luck Award for excellence in scientific reviewing. *Essays of an Information Scientist*, *4*, 127-131.

Ginsparg, P. (1998). *Electronic research archives for physics*. Retrieved December 2001 from http://tiepac.portlandpress.co.uk/books/online/tiepac/session1/ch7.htm

Glass, B. (1976). The critical state of the critical review article. *Quarterly Review of Biology*, 50th Anniversary Special Issue (1926-76), 415-418.

Glass, G. V (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, *10*, 3-8.

Glass, G. V (2001). *Meta-analysis at 25*. Retrieved December 2001 from http://glass.ed.asu.edu/gene/papers/meta25.html

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

Greenhouse, S. W. (1958). Some statistical and methodological aspects in the clinical evaluation of the tranquilizers in mental illness. *Biometrics*, *14*, 135.

Hampton, J. R. (1998). The end of medical history? *Journal of the Royal College of Physicians of London*, *32*, 366-375.

Healy, M.J.R. (1995). Frank Yates, 1902-1994—The work of a statistician. *International Statistical Review*, *63*, 271-288.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61-85.

Hedges, L. V. (1987a). Commentary. *Statistics in Medicine*, *6*, 381-385.

Hedges, L. V. (1987b). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist*, *42*, 443-455.

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

Herring, C. (1968). Distil or drown: The need for reviews. *Physics Today*, *21*, 27-33.

Horder, T. J. (2001). The organizer concept and modern embryology: Anglo-American perspectives. *International Journal of Developmental Biology*, *45*, 97-132.

Horn, J., & Limburg, M. (2001). Calcium antagonists for acute ischemic stroke (Cochrane Review). Retrieved December 2001 from the Cochrane Library (Issue 3) database.

Horn, J., de Haan, R. J., Vermeulen, M., Luiten, P.G.M., & Limburg, M. (2001). Nimodipine in animal model experiments of focal cerebral ischaemia. *Stroke*, *32*, 2433-2438.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russell Sage.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, *86*, 721-735.

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

Jones, L. V., & Fiske, D. (1953). Models for testing the significance of combined results. *Psychological Bulletin*, *50*, 375-382.

Kass, E. H. (1981). Reviewing reviews. In K. S. Warren (Ed.), *Coping with the biomedical literature* (pp. 79-91). New York: Praeger.

L'Abbé, K. A., Detsky, A. S., & O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine*, *107*, 224-232.

Last, J. M. (2001). *A dictionary of epidemiology.* Oxford: Oxford University Press.

Lewis, S., & Clarke, M. (2001). Forest plots—Trying to see the wood and the trees. *British Medical Journal*, *322*, 1479-1480.

Lide, D. R. (1981). Critical data for critical needs. *Science*, *212*, 1343-1349.

Lide, D. R., & Rossmassler, S. A. (1973). Status report on critical compilation of physical chemical data. *Annual Review of Physical Chemistry*, *29*, 135-158.

Light, R. J. (Ed.). (1983). *Evaluation studies review annual*. Beverly Hills, CA: Sage.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. 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.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational and behavioral treatment. *American Psychologist*, *48*, 1181-1209.

Mandel, H. (1936). *Racial psychic history: A detailed introduction and a systematic review of investigations*. Leipzig, Germany: Heims.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *American Statistician*, *32*, 12-16.

Mosteller, F. (1993). The prospect of data-based medicine in the light of ECPC. *Milbank Quarterly*, *71*, 523-532.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindsay (Ed.), *Handbook of social psychology: Vol. 1. Theory and method* (pp. 289-334). Reading, MA: Addison-Wesley.

Mulrow, C. D. (1987). The medical review article: State of the science. *Annals of Internal Medicine*, *106*, 485-488.

Nichols, H. (1891). The psychology of time. *American Journal of Psychology*, *3*, 453-529.

Oxman, A. D. (2001). The Cochrane Collaboration in the 21st century: Ten challenges and one reason why they must be met. In M. Egger, G. Davey Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 459-473). London: BMJ Books.

Oxman, A. D., & Guyatt, G. H. (1988). Guidelines for reading literature reviews. *Canadian Medical Association Journal*, *138*, 697-703.

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

Pearson, K. (1933). On a method of determining whether a sample of given size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, *25*, 370-410.

Peckham, M. (1991). Research and development in the National Health Service. *Lancet*, *338*, 367-371.

Peters, C. C. (1933). Summary of the Penn State experiments on the influence of instruction in character education. *Journal of Educational Psychology*, *7*, 269-272.

Peto, R. (1987). Why do we need systematic overviews of randomized trials? *Statistics in Medicine*, *6*, 233-240.

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: Myths and misconceptions. *British Medical Journal*, *322*, 98-101.

Pillemer, D. B. (1984). Conceptual issues in research synthesis. *Journal of Special Education*, *18*, 27-40.

Rayleigh, The Right Honorable Lord. (1885). *Presidential address at the 54th meeting of the British Association for the Advancement of Science, Montreal, August/September 1884*. London: John Murray.

Rietz, H. L., & Mitchell, H. H. (1910-1911). On the metabolism experiment as a statistical problem. *Journal of Biological Chemistry*, *8*, 297-326.

Rosenfeld, A. H. (1975). The particle data group: Growth and operations. *Annual Review of Nuclear Science*, *25*, 555-599.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

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.

Sacks, H. S., Berrier, J., Reitman, D., Ancona-Berk, V. A., & Chalmers, T. C. (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine*, *316*, 450-455.

Schoolman, H. M. (1982). Anatomy, physiology and pathology of biomedical information. *Western Medical Journal*, *137*, 460-466.

Shaikh, W., Vayda, E., & Feldman, W. (1976). A systematic review of the literature on evaluative studies of tonsillectomy and adenoidectomy. *Pediatrics*, *57*, 401-407.

Shapiro, S. (1994). Meta-analysis/shmeta-analysis. *American Journal of Epidemiology*, *140*, 771-778.

Sinclair, J. C., & Bracken, M. B. (Eds.). (1992). *Effective care of the newborn infant*. Oxford: Oxford University Press.

Smith, A.F.M. (1996). Mad cows and ecstasy: Chance and choice in an evidence-based society. *Journal of the Royal Statistical Society*, *159*, 367-383.

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*, 30-34.

Stigler, S. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stjernsward, J., Muenz, L. R., & von Essen, C. F. (1976). Postoperative radiotherapy and breast cancer. *Lancet*, *1*, 749.

Taylor, B. N., Parker, W. H., & Langenberg, D. N. (1969). Determination of e/h, using macroscopic quantum phase coherence in superconductors: Implications for quantum electrodynamics and the fundamental physical constants. *Reviews of Modern Physics*, *41*, 375-496.

Thorndike, E. L., & Ruger, G. J. (1916). The effects of outside air and recirculated air upon the intellectual achievement and improvement of school pupils: A second experiment. *School and Society*, *4*, 261-264.

Tippett, L.H.C. (1931). *The method of statistics*. London: Williams and Norgate.

Touloukian, Y. S. (1975). Reference data on thermophysics. In H. A. Skinner (Ed.), *International review of physical chemistry: Vol. 10. Thermochemistry and thermodynamics* (pp. 119-146). Newton, MA: Butterworth-Heinemann.

Warren, K. S. (Ed.). (1981). *Coping with the biomedical literature*. New York: Praeger.

Winkelstein, W. (1998). The first use of meta-analysis? *American Journal of Epidemiology*, *147*, 717.

Yates, F., Boyd, D. A., & Pettit, G.H.N. (1942). Influence of changes in levels of feeding on milk production. *Journal of Agricultural Science*, *32*, 428-456.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, *28*, 556-580.

Yates, F., & Crowther, E. M. (1941). Fertilizer policy in wartime: The fertilizer requirements of arable crops. *Empire Journal of Experimental Agriculture*, *9*, 77-97.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomised trials. *Progress in Cardiovascular Research*, *27*, 336-371.

Yusuf, S., Simon, R., & Ellenberg, S. S. (Guest eds). (1987). Meta-analysis of controlled trials [Special issue]. *Statistics in Medicine*, *6*(3).

Zwolinski, B. J., & Chao, J. (1972). Critically evaluated tables of thermodynamic data. In H. A. Skinner (Ed.), *International review of physical chemistry: Vol. 10. Thermochemistry and thermodynamics* (pp. 93-120). Newton, MA: Butterworth-Heinemann.