CHAPTER TWENTY-FIVE

META-ANALYSES, SYSTEMATIC REVIEWS, AND EVALUATION SYNTHESES*

Robert Boruch, Anthony Petrosino, Claire Morgan

What is a *meta-analysis*? A *systematic review*? An *evaluation synthesis*? A variety of phrases are used to describe scientifically disciplined approaches to searching literatures, assembling studies for review, and analyzing, interpreting, and reporting the results. Here, we adopt the definitions given by Chalmers, Hedges, and Cooper (2002). A *systematic review* involves the application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic. *Meta-analysis* is the statistical synthesis of data from separate but similar (that is, comparable) studies, leading to a quantitative summary of the pooled results. *Evaluation synthesis* is an attempt to "integrate empirical evaluations for the purpose of creating generalizations . . . [in a way that] is initially nonjudgmental vis-a-vis the outcomes of the synthesis and intends to be exhaustive in the coverage of the database" (Cooper, Hedges, and Valentine, 2009, pp. 5, 19).

The word *bias*, as seen in the definition of systematic review and implied in the other definitions, has a basic meaning of "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer

*Work on the topic discussed in this chapter has been funded by the Institute for Education Sciences, the research arm of the U.S. Department of Education. The views expressed here do not necessarily reflect the views of funding organizations.

over others" (as stated in *Merriam-Webster's Collegiate Dictionary*) but exists in many forms. Identifying and depending only on reports that suit the reviewer's ideological or theoretical preference is an obvious source of bias, for example. The tactic has been exploited shamelessly in political, professional, and even ostensibly dispassionate arenas, such as the university. Paying attention only to reports that are published in refereed academic journals also implies a biased sample of pertinent reports: those not published in such journals are ignored or not identified. Bias also refers to the study design for each study in an assembly of studies and, in particular, bias in the statistical estimates of an intervention's effect that is produced by each design. Randomized trials, for instance, when they are carried out well, produce statistically unbiased estimates of the relative effect of an intervention. The statistical bias in estimates of effect produced by alternative approaches, such as a before-after evaluation, cannot always be identified, much less estimated.

Simple definitions are necessary but not sufficient. There is a science to reviewing research, including meta-analyses. The rationales, principles, and procedures used and the scientific standards of evidence employed have to be made clear.

Why Be Conscientious in Reviewing Studies of Intervention Effects?

Any college student or professor, legislative staffer or public lobbyist, journalist, or thoughtful citizen can do a Google, Bing, or other Internet search on phrases such as "what works." Our rudimentary search on just these terms in July 2014 yielded a staggering two billion hits in .05 seconds. Obviously, more careful and systematic procedures are necessary to reduce this volume and focus exclusively on those studies that directly bear on the effectiveness of an intervention. The following gives some other reasons to justify conscientious review procedures in synthesizing evaluations.

Multiple Evaluations Versus a Single Evaluation

Other things being equal, examining multiple, independent, and high-quality evaluations of an intervention or a class of interventions is a better way to understand the intervention's effects than examining one evaluation. Findings from a single study done in one place, by one team, and with one actualization

of the intervention, for instance, usually cannot easily be generalized to other settings, other teams, or other actualizations. Replication or a near-replication is important for supporting statements about how often, to what degree, and in what circumstances the intervention works. Meta-analyses, systematic reviews, and evaluation syntheses try to get beyond the single study, if indeed there are more studies to examine.

For instance, Petrosino, Turpin-Petrosino, and Guckenburg (2010) examined the effects of juvenile system processing on delinquency. Less serious juvenile offenders can be handled with considerable discretion. Juvenile system practitioners can opt to bring the child formally through the juvenile justice system (official processing), divert the child out of the system to a program or service, or release the child to parents or guardians with no further action. To some observers' surprise, at least twenty-nine randomized trials have been mounted since 1972 that have compared assignment of juveniles to an "official" system processing condition (that is, petitioned before the court, appearance before a judge, case moving forward in the system) with at least one release or diversion program condition.

Across these twenty-nine experiments, there is considerable variation. The selective reader could cite any single study—or selective number of studies—as "evidence" for a position that processing has a "deterrent" effect and reduces subsequent delinquency. Indeed, about ten studies show positive results for processing. Relying on a selective gathering of evidence might lead decision makers to opt for processing juvenile offenders formally through the court system as a deterrent measure.

The totality of the evidence reviewed by Petrosino and his colleagues, however, paints a different picture. Figure 25.1 presents the effect sizes for juvenile system processing versus a diversion program or release condition. (In this review, standardized difference [Std. Mean Diff.] refers to the difference between the outcomes for the treatment and control group divided by the standard deviation.) In this instance the assembly of evidence suggests that across all twenty-nine studies, the effect size was –.11. Although this is a negative effect, indicating that processing led to an increase in delinquency, this would be considered by most readers to be a small effect size. But keep in mind that juvenile system processing is a more expensive option for most jurisdictions than simple release and likely more expensive than almost all but the most intensive diversion programs. If there is no deterrent impact of official judicial processing but in fact a small negative effect, and if it is a more expensive option, a judge, citizen, or policymaker could clearly ask if it would be better to divert or release less serious juvenile offenders.

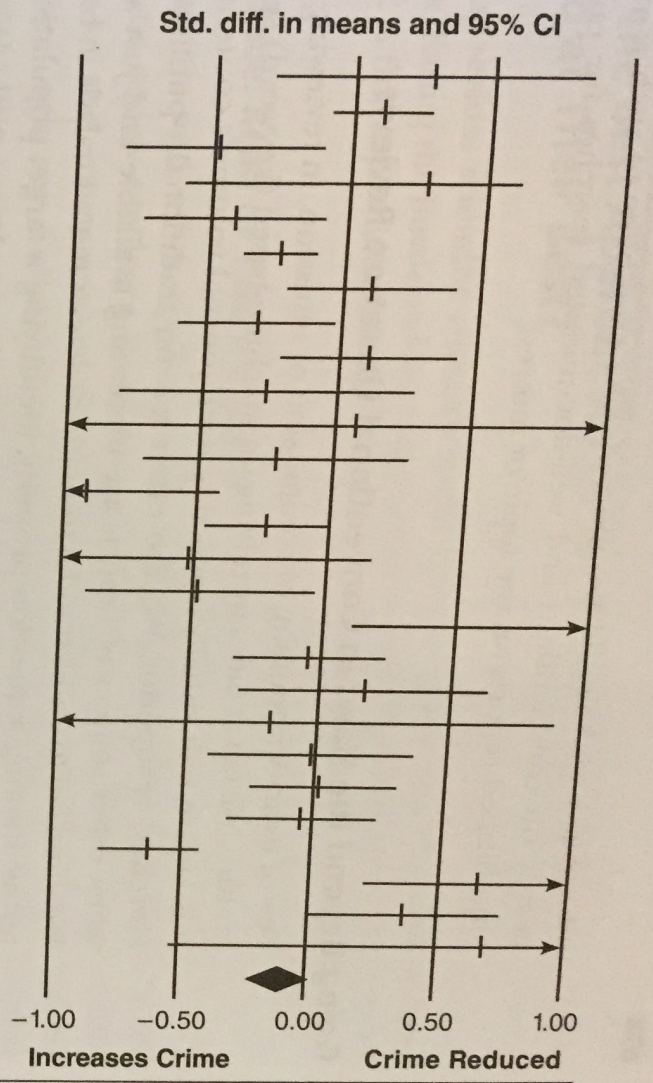
Identifying High-Quality Evidence

The best high-quality systematic reviews, meta-analyses, and evaluation syntheses identify high-quality evidence that has been produced on the effects of interventions and where such evidence is unavailable. The review discussed in the previous section is an illustration of this. It identifies the dependable grounds on which decisions can be made to adopt, avoid, or improve the intervention. As will be discussed below, to be determined of high quality and worthy of inclusion in a systematic review, the evidence must meet transparent inclusion criteria identified by the researchers. An example of how important it can be to establish the absence of high-quality evidence can be found in a review conducted by Fisher, Montgomery, and Gardner (2008). These investigators conducted a systematic review of research on the effects of providing employment or educational opportunities (that is of "opportunities provision") to prevent gang involvement. They searched widely for dependable evidence from experimental and quasi-experimental studies that tested the provision of opportunities to actual or prospective gang members in the interest of preventing or reducing participation in gangs. They did not find a single study meeting their eligibility criteria. The value of such a systematic review lies in establishing that *no* high-quality evaluations have been carried out on a particular topic. Such a review establishes the need for funding primary evaluation studies that test promising interventions.

Governments and non-governmental aid organizations often want evidence to show where their funding can be most effectively allocated. Recently, for example, some multinational organizations that are concerned about the availability of quality evidence for low-income countries have begun to fund production of such evidence in the interest of better interventions and decision making. The International Initiative for Impact Evaluation (3ie), for instance, promotes rigorous studies, particularly randomized controlled trials, in the developing nation context. This organization also sponsors systematic reviews—*synthetic reviews* in the organization's vernacular. With 3ie's support, Petrosino, Morgan, Fronius, Tanner-Smith, and Boruch (2014) conducted systematic searches to identify experimental and quasi-experimental studies that tested the impact of an intervention on school enrollment, attendance, dropout rates, and the like, in developing countries. The project initially identified some high-quality evidence (seventy-six eligible studies), and a large number of studies that either did not use an experimental or quasi-experimental design or did not include an outcome measure of school enrollment. This kind of mapping has benefits for decision making not only about what programming to implement but also where future studies are likely to be

FIGURE 25.1. PROCESSING EFFECTS ON PREVALENCE: FIRST EFFECTS

Study Name	Outcome	Statistics for Each Study: Std. Mean Diff.	Lower Limit	Upper Limit	*p*-Value
Patrick & Marsh (2005)	First Effect-P	0.28	−0.30	0.86	0.34
Severy & Whitaker (1982)	First Effect-P	0.10	−0.08	0.27	0.30
Klein (1986)	First Effect-P	−0.48	−0.84	−0.12	0.01
Smith et al. (1979)	First Effect-P	0.00	−0.61	0.61	1.00
Baron & Feeney (1976) 602	First Effect-P	−0.43	−0.76	−0.10	0.01
Baron and Feeney (1976) 601	First Effect-P	−0.25	−0.38	−0.12	0.00
Dunford et al. (1982) KC	First Effect-P	0.09	−0.22	0.40	0.55
Dunford et al. (1982) NY	First Effect-P	−0.32	−0.61	−0.03	0.03
Dunford et al. (1982) FL	First Effect-P	0.10	−0.22	0.42	0.56
Koch (1985)	First Effect-P	−0.2	−0.82	0.27	0.32
Blakely (1981)	First Effect-P	0.07	−1.03	1.16	0.91
Davidson II et al. (1987)	First Effect-P	−0.23	−0.72	0.27	0.37
Davidson II et al. (1980)	First Effect-P	−0.94	−1.44	−0.43	0.00
Quay and Love (1977)	First Effect-P	−0.24	−0.47	−0.02	0.03
Bauer et al. (1980)	First Effect-P	−0.51	−1.18	0.16	0.13
Quincy (1981)	First Effect-P	−0.47	−0.90	−0.04	0.03
Hintzen et al. (1979)	First Effect-P	1.00	0.11	1.99	0.03
Smith et al. (2004)	First Effect-P	−0.05	−0.34	0.24	0.73
Povitsky Stickle et al. (2008)	First Effect-P	0.16	−0.31	0.63	0.50
University Associates (1986) OTSEGO	First Effect-P	−0.19	−1.28	0.90	0.73
University Associates (1986) BAY	First Effect-P	−0.03	−0.42	0.37	0.89
University Associates (1986) KALAMAZOO	First Effect-P	0.03	−0.25	0.31	0.84
University Associates (1986) DETROIT	First Effect-P	−0.05	−0.34	0.24	0.73
Curran et al. (1977)	First Effect-P	−0.64	−0.82	−0.45	0.00
Sherman et al. (2000 JPP	First Effect-P	0.65	0.22	1.08	0.00
McCold and Wachtel (1998)	First Effect-P	0.368	−0.01	0.74	0.06
True (1973)	First Effect-P	0.68	−0.54	1.91	0.28
		−0.11	−0.24	0.02	0.10



Source: Petrosino, Turpin-Petrosino, and Guckenburg, 2010.

informative and should be initiated. The credibility of such a map depends heavily on conscientious and well-documented searches for published and unpublished reports on the topic of interest.

Going Beyond the Flaws in Conventional Literature Reviews

Few of us are without sin, of commission or omission, in reviewing a body of literature. We fail at times by relying on machine-based (keyword) searches when it is known that visually inspecting each journal volume's contents (i.e., hand searches) is superior. We also often rely on traditionally-published literature when other sources of reports are increasing available and just as important. We often fail to understand systematic review or meta-analysis in basic scientific terms: framing a question properly, identifying a target population of studies, sampling the studies well, and analyzing the results properly. When we do literature reviews, we may fail to make our standards of evidence and procedures explicit. The modern approaches to reviews assist us in being scientifically virtuous, or at least in understanding what virtue is.

Farrington and Petrosino (2000) have contrasted the imperfections of "common reviews of the literature" with the quality of the reviews produced by organizations such as the international Cochrane and Campbell Collaborations. They point out that on the one hand, common literature reviews are usually one-off exercises that fail to be updated or to exploit new technologies of searching, reviewing, and summarizing studies. The Cochrane and Campbell Collaborations, on the other hand, capitalize on contemporary technical methods and attempts to periodically update reviews. Farrington and Petrosino remind us that conventional reviews are usually based on one country's research and on English language publications, whereas organizations such as Cochrane and Campbell are international. Common reviews often do not present explicit details on such important components as what literature will be included, how it will be assessed, and criteria for determining success of an intervention. Cochrane and Campbell Collaboration reviews stress explicit and transparent methods, including an externally peer-reviewed and electronically published protocol or plan for the review. Finally, these authors point out that conventional reviews are published in a variety of outlets that each have their own jargon and standards of evidence, which presents substantial difficulties for policy people, practitioners, and researchers who work across disciplines. The purpose of organizations such as Cochrane and Campbell Collaboration is to provide an electronically accessible data base of high-quality, uniformly structured, systematic reviews and evaluation syntheses.

How Are the Best Approaches to Systematic Reviews Employed at Their Best?

Doing serious scientific research in the context of systematic reviews is hard work. Easing the burden without degrading the quality of the product is a good idea. A fine aspiration at least.

Practical Advice: Read or Take a Course

Evaluators and other applied researchers who know nothing about a systematic review can learn by reading a good one. Since 1993, the most uniform of and transparent of the genre in health care have been produced by the Cochrane Collaboration (www.cochrane.org). More recent parallel efforts in the social sector are being produced by the Campbell Collaboration (www.campbellcollaboration.org). Both rely heavily on voluntary efforts. In education in the United States, the What Works Clearinghouse (WWC) has been well funded to produce remarkably detailed reviews of particular curriculum packages that can withstand the legal threats of commercial publishers and package developers. Smaller but equally noble efforts have been mounted by the Coalition for Evidence-Based Policy, and by Robert Slavin in his "Best Evidence for Education (BEE)."

These are among the best partly because they get well beyond the flaws of run-of-the-mill literature reviews. They depend on organizational innovations and technology, including transparent standards of evidence on effects of interventions.

Perish the thought of reading or taking a course for those who prefer only Google. But people who are serious in their interest might read a book. A comprehensive handling of advances in the area is edited by Cooper, Hedges, and Valentine (2009). Their book requires stamina but is mighty thorough.

Short courses on systematic reviews, meta-analysis, and the activities they require, such as hand searches of journals and adherence to explicit standards, are valuable. The Cochrane Collaboration and Campbell Collaboration offer these at annual meetings and at other times. The WWC has developed training courses for reviewing education evaluations (http://whatworks.ed.gov). Among other organizations, the Society for Prevention Research has initiated presentation on the topic at its annual meeting. Academic institutions, such as the Evidence for Policy and Practice Information and Coordinating Centre (EPPI Centre) at the University of London, now offer programs and courses in research synthesis.

Practical Advice: Contribute to a Meta-Analysis, Systematic Review, or Evaluation Synthesis

Conducting a meta-analysis, systematic review, or evaluation synthesis that is governed by high standards can be demanding. The opportunities for voluntary contributions are ample; for example, the Cochrane and Campbell Collaborations seek such voluntary efforts (e.g., as authors, reviewers, passing along eligible studies). For the opportunities in the international Campbell Collaboration more generally, see http://campbellcolaboration.org. In Copenhagen, SFI Campbell (formerly the Nordic Campbell Centre), which "works with evidence and measuring of effects of social welfare interventions," (www.sfi.dk/Default.aspx?ID=432), provides seed money to talented people who want to contribute to systematic reviews that are far better, and more demanding, than the more common reviews of literature. Nowadays, substantial numbers of good reviews are produced through government agencies, such as the WWC, and through contracts with or grants to organizations whose staff and consultants produce the reviews. These arrangements typically depend on salaried professional staff rather than volunteers.

Producing a Meta-Analysis, Systematic Review, Evaluation Synthesis

The major steps in a systematic review, meta-analysis, or evaluation synthesis are easy to lay out. However, they are not easy to take, just as the analogous steps in field evaluations and other applied research are not easy. The simplified list in the following section capitalizes on the guidelines of the Quality of Reporting of Meta-analyses (QUOROM) group (Moher and others, 1999) and on Cooper, Hedges, and Valentine (2009), the Campbell and Cochrane Collaborations, the WWC, and other sources.

Specify the Topic Area. In the WWC, for instance, specifying the topic means identifying (Gersten and Hitchcock, 2009):

1. A rationale for addressing the problem
2. The specific question(s) that will be addressed
3. The relevant outcome variables
4. The relevant target populations and subpopulations of interest
5. The relevant class of interventions that address the problem

Reviewers proposing new topics for review in the Campbell Collaboration must fill out a title registration page containing such information. This

is done to enhance transparency and uniformity as well as to avoid duplication of effort.

As important, authors of proposed reviews must indicate what types of studies are going to be reviewed and the relevant outcomes and the targeted populations of interest. For example, in the Petrosino, Morgan, Fronius, Tanner-Smith, and Boruch (2014) review on the effects of school enrollment strategies in developing nations, the research team specified that it would examine evaluations of school enrollment policies and practices based on randomized trials or rigorous quasi-experimental designs. The team also required that eligible studies report at least one outcome of enrollment, attendance, or dropout and that these studies be conducted in developing nations with primary and secondary school students.

Develop a Management Strategy and Procedures. Managing a single systematic review, meta-analysis, or evaluation synthesis requires a strategy that does not differ *in principle* from the management requirements of a field study. This includes identifying who will do what tasks, when, with what resources, and under what ground rules. A plan for conducting the review is required by the Campbell and Cochrane Collaborations and by funding agencies, such as the Institute for Education Sciences and 3ie, that support such syntheses. This protocol lays out the plan for the review and indicates the timeline for completing the review and submitting deliverables such as the final review draft. Such protocols, especially when published electronically by organizations like the Campbell and Cochrane Collaborations, also provide a level of transparency in that one can determine if and how review teams deviated from the plan.

The time required and difficulty encountered in doing a review, and the funding and other resources needed to complete one, are influenced heavily by the size and complexity of the studies that will be included. A review that does not find any eligible studies will of course be substantially cheaper and quicker than a review including hundreds of studies.

Specify the Search Strategy. Specifying what literatures will be searched, how, and with what resources is crucial. The best reviews are exhaustive, and usually exhausting, in searching for reports published in peer-reviewed social science journals or issued by organizations with high-quality editorial screening, or both. Doing both is better, at least in the United States, where some evaluation organizations have external peer review systems with standards that get beyond those of some professional journals. Will evaluations that are relevant but not reported widely also be included in the systematic review? Many organizations, for-profit and otherwise, for instance, do not publish articles in peer-reviewed

TABLE 25.1. SYSTEMATIC REVIEW SEARCH STRATEGIES.

Conduct electronic searches of bibliographic databases using specified keywords and strings
Conduct online "hand searches" of relevant journals
Examine online holdings of relevant organizations and research firms
Scan the references of each retrieved report
Contact researchers working in the topic area

journals. Unless they put a report on an easily accessed website, that report might not be uncovered. Evaluation reports by school district research offices, and by vendors of educational software and curriculum packages are not circulated widely, if at all. The systematic review team has to decide whether to survey these and how to do so. The Institute for Education Science's WWC, for example, posts the topical protocol for each review that is planned on its public website. The WWC tells the formal WWC Network about each, so as to invite people to submit studies that seem pertinent for inclusion in a particular review. Surveying researchers in the field, as Waddington, Snilstveit, White, and Fewtrell (2009) did, is one approach reviewers have used in an attempt to identify what is referred to as *grey* or *fugitive* literature that might otherwise remain stuck in file drawers.

Researchers may undertake online "hand searches" of certain peer-reviewed journals, knowing that such a search yields a far more reliable and complete assembly of relevant studies than a search engine. The best systematic reviews undertaken under the guidelines of the Campbell and Cochrane Collaborations, the WWC, and others make plain what literatures have been covered in the search. For example, a review of studies of the effect of water, sanitation, and hygiene practices intended to combat diarrhea in developing nations (Waddington, Snilstveit, White, and Fewtrell, 2009) searched ten electronic bibliographical databases, contacted key scholars working in the area, and conducted specialized searches of the Web sites of approximately twenty-five leading international organizations, such as the International Federation of the Red Cross and Red Crescent Societies. The searches yielded 76 experimental and quasi-experimental impact studies that appeared dependable for estimating the effects of the interventions.

Table 25.1 outlines search strategies commonly undertaken in a thorough systematic review. Beyond identifying the target for the literature search, the way the search is conducted has to be specified. What keywords, constructed how and why, will be used with what electronic search engine and with what electronic databases? Randomized trials, for instance, are sometimes hard to locate given that relevant keywords often do not appear in a journal article's

TABLE 25.2. EXAMPLES OF INCLUSION CRITERIA FOR SYSTEMATIC REVIEWS.

Was the evaluation conducted in the region or with the population of interest?
Does the evaluation include the outcome measure(s) of interest?
Was the evaluation conducted during the timeframe of interest?
Does the evaluation report on construct validity that ties the outcome variable to interventions?
Does the evaluation employ a design that permits unbiased and relatively unequivocal estimates of the intervention's effects
Does the evaluation report sufficient information to estimate effect sizes?
Does the evaluation meet methodological quality criteria, e.g.: Was the intervention implemented with fidelity? Were there selection bias or attrition issues?

abstract or title. Consequently, trying out different words in each database may be warranted. In searching for study trials in the crime and justice arena, Petrosino's (1995) search suggested that the following keywords had a high yield: *random, experiment, controlled, evaluation, impact, effect,* and *outcome.* Depending on the vernacular employed in the discipline, databases, search engines, and so on, another researcher's list could be appreciably different from this. The aforementioned water sanitation review (Waddington, Snilstveit, White, and Fewtrell, 2009) reported that good success arose in part from pairing terms such as *sanitation, water quality, water quantity,* and *hygiene* with *diarrhea.*

Develop Inclusion and Exclusion Criteria for Studies in the Review. This step focuses on identifying the studies that will be regarded as potentially legitimate data for a systematic review. Efforts to make inclusion standards uniform, explicit, and scientific in orientation have been made by the Cochrane Collaboration, the Campbell Collaboration, the Coalition for Evidence-Based Policy, and the WWC, among others. Table 25.2 shows some examples of questions to ask when evaluating a study for inclusion in a systematic review.

Once a study is tentatively included, more detailed questions on implementation fidelity, rates of missing data and loss of participants or attrition from study samples quality of measurement, and so on are posed. In the WWC, for example, data drawn from study reports are coded, preferably by two independent coders, so as to permit further determinations about how much one can depend on the study at hand. For instance, a randomized trial or quasi-experiment with a 30 percent difference in the attrition rates for the intervention and the control groups would be ruled out as a dependable resource by reviewers who understand how vulnerable this difference in attrition rate renders the study's results. An exception may be made if evidence can be produced to argue that plausible biases are negligible.

Under the WWC standards, a study is rejected from a systematic review if it (1) fails to report on construct validity that ties the outcome variable to interventions, (2) fails to employ an evaluation design that permits unbiased and relatively unequivocal estimates of the intervention's effects, (3) does not test the intervention on appropriate target populations, or (4) fails to report information sufficient to estimate effect sizes. Studies that do report information in all these areas are tentatively included in the review.

As the definitions given earlier suggest, inclusion criteria in systematic reviews focus on eliminating biased estimates of the effects of interventions. Generalizing from the studies at hand is often subordinate to the aim to eliminate biases from the studies being examined. Nonetheless, when a systematic review includes a number of studies conducted in a wide range of jurisdictions (including multinational settings, on occasion), conducted over a long time period, and using various measurement outcomes of a construct, the findings can be construed as having higher external validity than findings from single studies. Various advanced statistical methods can help one understand the assumptions underlying such generalizations.

Reviewers are often surprised, despite the number of publications on studies conducted in a field, at the number of studies that do not meet the eligibility criteria for dependability of the evidence. For example, a U.S. Government Accountability Office (GAO) review of sixty-one studies of interventions for the low-income participants in the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) depended heavily on only 37, which were declared "relatively credible" (Hunt, 1997, p. 41). Mark Lipsey's review of studies on juvenile delinquency prevention and treatment programs initially amassed more than 8,000 citations, and after screening depended on 443 that met the researchers' standard for good design and execution (Hunt, 1997, p. 129). In a review of effects of a marital and family therapy, "a year and a half of such efforts netted [William] Shadish a haul of roughly two thousand references" (Hunt, 1997, p. 45). About 160 of these met high standards of evidence and were included in the review. Gersten and Hitchcock (2009) identified 700 publications related to English-language learners and interventions for them, and covered just two dependable evaluations in their WWC review.

Develop a Scheme for Coding Studies and Their Properties. Evaluation syntheses, systematic reviews, and meta-analyses direct one's attention to an assembly of studies. The assembly is often a mob. The implication is that reports on evaluations of the effects of interventions, when included in a disciplined review, need to be construed as objects for interrogation and categorized in a variety of ways. As David Wilson (2009) put it, coding for a systematic review

TABLE 25.3. EXAMPLES OF VARIABLES TO CODE IN INCLUDED EVALUATIONS.

Characteristics of Intervention	Characteristics of Study Population	Characteristics of Study Methods
Description of intervention		
Detailed descriptions of the intervention and control condition, including the "dosage" of the treatment being implemented, and the number of participants assigned to each group	Detail about the type of participants in the trials	
Whether the program experienced significant implementation and fidelity problems	Setting and context in which trial was conducted	Information about randomization or quasi-experimental assignment
Issues with crossovers (persons receiving a condition they were not assigned to)		Level of assignment and whether the study included multiple analyses at different levels
		How the groups were equated and whether any problems with equating were reported
		Loss of participants due to attrition or database matching issues and whether the attrition differentially affected the groups
		Selection bias (e.g., breakdowns in randomization or unusual unequal distributions in groups)

is akin to "interviewing the studies." In best practice, coding and abstraction of each study considered for a systematic review involves development of coding schema, training of coders, and the use of at least two independent coders (double coding) so as to provide reliability checks. The codes address details of the intervention, characteristics of the samples used in the study, definitions of specific outcomes, dose levels, and so on. Table 25.3 lists examples of variables that could be included in a coding instrument.

Consider an example. Wilson, Lipsey, and Soydan's (2003) award-winning review of the effect of mainstream delinquency programs on minority youths is based on the double coding of about 150 features of each study in the review. The authors' early attention to detail in coding permitted their later research

on subsamples of minority youths in evaluations that included small to moderately sized subsamples. Coding categories in this review were similar to those used in Cochrane, Campbell, and WWC reviews, at least with respect to the evaluation's design; for example, randomized trials are distinguished routinely from nonrandomized trials. Codes identify detailed features of the interventions, such as the kinds of staff delivering the treatment, the format (group versus individual), the site, and so on.

A review may have to discard studies following the detailed coding of reports. Upon closer inspection, for instance, some studies originally thought to be eligible may be put aside because they do not provide the necessary data for meta-analysis. To judge from Gersten and Hitchcock's (2009) examination of the flaws in reporting, common problems are that the published reports do not provide any quantitative data to permit the computation of an effect size or do not analyze data correctly and do not provide enough information to correct the original analysis.

Compute Effect Size Estimates, Code Them, and Estimate Their Variances. An *effect size* in any science is estimated relative to some basis for comparison, reference, or benchmark. In a two-arm, randomized controlled trial, for instance, the common estimate of effect size involves computing the difference between mean outcomes for the two interventions being compared, and then dividing this difference by the square root of a pooled estimate of variance within the intervention groups. Odds ratios are common in the health sector and are being used more often in meta-analyses of social interventions (Cooper, Hedges, and Valentine, 2009). Neither of these statistical indicators of effect size or odds ratios is easily understandable to many people. Consequently, graphic portrayals that meet good statistical standards, such as the example in Figure 25.1, are now common. The technology and the art of portraying results in numbers, prose, and charts are still developing and deserve serious research on how people understand and value the portrayals (Boruch and Rui, 2008).

Impact evaluation reports do not always contain sufficient information for the reviewer to estimate effect size. This may lead to a study being eliminated from a review. However, many procedures have been developed that permit estimates of effect size to be computed from minimal data, such as the actual statistical test value (t, F, or chi-square distributions) or the statistical probability that the observed result occurred by chance. Lipsey and Wilson (2001), among others, make such conversion procedures readily available in their texts. Helpful software programs have been developed to assist researchers in computation of effect sizes and analysis of samples; one such is

Comprehensive Meta-Analysis Version 2.0 (see Borenstein, Hedges, Higgins, and Rothstein, 2005).

Develop an Analysis Strategy. The purpose of systematic review, meta-analysis, and evaluation synthesis is to reach conclusions based on a summary of results from an assembly of studies. Analysis steps are put simply in the following paragraphs:

First, arrange your thinking about the data at hand (studies of interventions) in terms of the studies' target populations, samples observed and samples not observed, and the effect sizes produced. Ensure that these effect sizes are constructed so as to make their interpretation plain. And ensure that outliers and artifacts of particular studies are identified and taken into account.

Second, focus attention on the distributions of the effect sizes. For instance, any given randomized trial on an intervention produces an effect size for which a confidence interval can be constructed. Other studies you have included will also produce effect sizes, each of which associated with a confidence interval. All these effects can be plotted out in a chart of the distribution of effect sizes. Systematic reviews under the definition given earlier typically include such a chart. A meta-analysis involves combination of effect sizes, and (often) the analysis of effect sizes as a function of the coded characteristics of the studies that are included in the review.

Describing the effects sizes and their distribution for an assembly of interventions in a class is essential for a high-quality review. Petrosino, Turpin-Petrosino, and Guckenburg (2010) did so in their review of juvenile system processing (Figure 25.1). This satisfies the interest of some readers who want to know whether an intervention resulted in doing some good, relative to high standards of evidence, and whether it did no good, relative to the same standards.

Beyond this, sophisticated statistical machinery and substantive understanding might be brought to bear on the question: What seems to "explain" the variation in effect sizes among studies that were reviewed? For instance, one may examine effect sizes for the studies as a statistical function of characteristics of study design, such as whether the design is a randomized trial or not, sample size, and so on. One may examine effect sizes as a function of coded characteristics of the intervention. Lösel and Beelman (2003), for instance, undertook a meta-analysis of eighty-four reports on randomized trials that were designed to estimate the effect of child skills training on antisocial behavior. They depended on different kinds of statistical models to understand the relationship between effect sizes (dependent variable) and characteristics of each study, the characteristics of the interventions, and the characteristics of

the children in each study sample. For example, studies with smaller samples tended to be associated with larger effect sizes. Treatment dosage appeared not to be related to effect size. Interventions administered by study authors or research staff or supervised students were associated with larger effect sizes.

As Berk (2007) and others point out, statistical modeling in this meta-analytic context has the same merits and shortcomings as those of model-based analyses of data from passive observational studies. That is, the studies in a systematic review are units of observation; they are observed passively by the reviewer. The observations are the results of a kind of survey. Conventional regression analyses of effect size then can help to illustrate relationships. But misspecification of the regression model, unobserved variables that are related to variables in the model (i.e., *confounders*), and relations among the independent variables usually do not permit unequivocal statements about what *causes* the effect size to vary.

Interpret and Report the Results. In the best systematic reviews, reports of at least two kinds are produced. The first is exquisitely detailed and contains all the scientific information sufficient for an independent analyst or scientist to conduct an identical review, that is, to replicate. As a practical matter, such detailed reviews are published in electronic libraries, and unlike hard-copy reports and research journals, have no page limitations. In the best, the topical coverage is uniform and standards are uniformly transparent, to make it easy for readers to move from one systematic review to the next. The Cochrane Collaboration, Campbell Collaboration, and WWC products have this character.

A second kind of report, a summary in hard-copy or electronic form, is crucial to users of evaluations who are not themselves researchers. Users such as policy decision makers and other practitioners typically value a summary that is uniform from one review to another and in language that is as plain as possible. The Cochrane Collaboration's reviews in recent years have included such summaries. The WWC produces these routinely and not without serious effort.

In the most sophisticated production of systematic reviews, reporting may involve the engagement of networks of users who were parties to a review's production, networks of potential users who might repackage and distribute the results, information brokers, and so on. The hard problem is developing networks of users and information brokers. The Institute of Education Sciences has invested resources in developing a network to ensure that products of the WWC are understood and influenced by a network of potential users. The practical advice on this is to engage potential users at the front end.

There is a third kind of report that is not yet common. It involves publication of all micro-records from all studies that are covered in a systematic review. Such a report, compiled with good definitions and numbers, would permit secondary analysis of micro-records by anyone with access to a spreadsheet and a way of importing files. This opportunity for transparency is part of the future.

What Resources Can Be Employed to Do the Job Well?

There are now many organizations that are conducting systematic reviews, and several, such as the Cochrane and Campbell Collaborations and the WWC, that are producing them on a grand scale. Besides these large-scale efforts, technological advances are improving the ability of researchers to identify, catalog, and analyze the results of separate but similar evaluation studies.

Independent International and Domestic Resources

The international Cochrane Collaboration was formed in 1993 to prepare, maintain, and make accessible systematic reviews of evaluations of the effects of health-related interventions. As of June 2014, the Cochrane Collaboration had produced over 6,000 completed systematic reviews based on explicit and uniform operating principles and transparent standards of evidence, with over 2,300 published protocols indicating current reviews in progress. The international Campbell Collaboration is the Cochrane Collaboration's young sibling. Created in 2000, its aims in its area of interest are identical to Cochrane's: to prepare, maintain, and make accessible systematic reviews of studies of the effects of interventions. This is to inform people about what works in the arenas of crime and justice, education, social welfare and international development. The *Cochrane Handbook for Systematic Reviews of Interventions* is used by both organizations to meet technical, quality control, and uniformity standards.

The Coalition for Evidence-Based Policy (www.coalition4evidence.org) has been remarkably influential, partly on account of its informed advocacy of randomized trials in the United States and partly on account of its efforts to identify top-tier programs in the United States that depend on basic standards of evidence (www.toptierevidence.org). The Best Evidence Encyclopedia (www.bestevidence.org) is a U.S.- and U.K.-based effort that uses some of the basic evidence standards for identifying dependable studies.

Among U.S. states, California's Evidence-Based Clearinghouse for Child Welfare (www.cachildwelfareclearinghouse.org) is a precedent. The Washington State Institute for Public Policy (www.wsipp.wa.gov) has been remarkable in uncovering and using systematic reviews of high-quality evidence and ensuring that such evidence gets to state legislators; its efforts get to a macro-level that involves reviews of many reviews. WSIPP also uses systematic reviews to estimate anticipated benefits from investing in particular programs, in their economic analyses (i.e., cost-benefit studies).

Government Organizations and Government-Sponsored Entities

In the United States, a variety of government organizations have undertaken systematic reviews of the applied research and evaluation literature or have provided funds to others to do so. Some of these organizations, such as the GAO, have helped to advance the state of the art since the 1980s (Cordray and Morphy, 2009). The U.S. Department of Education's What Works Clearinghouse has developed technical resources, such as uniform standards and procedures for determining whether each evaluation study in an assembly of studies can be used as a basis for a causal inference about an intervention's effect. It is moreover a remarkable source of guidance on technical issues in analysis of cluster randomized trials, statistical power analysis, and missing data analysis.

An initiative known as the Community Guide has been undertaken by the U.S. Centers for Disease Control and Prevention (www.thecommunityguide.org). The initiative's Task Force on Community Preventive Services conducts systematic reviews of research on the effects of interventions relevant to preventing health problems, including violence and injuries. For example, recent reviews focus on firearm laws, early childhood visitation programs, school-based violence prevention, reducing exposure to environmental tobacco smoke, and worksite obesity prevention programs.

Police agencies in the United Kingdom and private foundations such as the Jerry Lee Foundation in the United States have supported systematic reviews of evidence in crime prevention approaches such as closed circuit television. Farrington and his colleagues (2011) reports on sponsored projects in the context of the Campbell Collaboration's Crime and Justice Group. The Norwegian, Danish, and Swedish governments have sponsored systematic reviews in the education and social services sectors under the auspices of the Campbell Collaboration. Canada has supported systematic reviews under the same auspices in social welfare and under the auspices of the Cochrane Collaboration in health.

Technical Resources

Technical resources include the monographs, books, and software identified earlier. They include the technical guidance documents being produced by the WWC and at times by voluntary organizations such as the Cochrane and Campbell Collaborations. Because technology in design, execution, and analysis of studies changes with time and because there are changes in procedures used in identifying, assembling, and screening studies, the aspiring systematic reviewer has to pay attention to new developments. Younger people have stamina, and we wish them good luck on this account.

Web-oriented databases and search engines that furnish the ingredients for an evaluation synthesis are low cost and access to them is easy. PsychInfo and ERIC, for instance, are databases that are accessible in most research universities and many research and evaluation organizations. Each database is accessed by different vendors' search engines, however, and costs and benefits of these may differ appreciably.

The electronic search engines are sometimes less helpful than one might expect. For instance, they often do not search the full text of the evaluation report for the keywords. As a consequence, studies are missed. For instance, a PsychInfo search of abstracts from the *Journal of Educational Psychology* (1997–2000) for randomized trials yielded about thirty reports on trials. A search of the full text of the journal's contents for the same years yielded 100 trials (Turner and others, 2003). Machine-based searches of *American Education Research Journal* (1963–2000) yielded less than a third of the evaluations based on randomized trials in math and science education. To complicate matters, abstracts of articles in refereed journals on evaluation and applied research are not uniform. One technological advance that constitutes a resource in hand searches is the electronic publication of full texts of journal articles and books. This greatly facilitates full-text searches of course, including immediate demarcation and reproduction of pertinent reports or portions of them.

Resources and Issues for the Future: Scenarios

Part of the future lies in the reviewer's access to micro-records from each study that is used in a review. During the 1970s, for instance, evaluation studies of programs began to yield micro-record data that were made available at times for independent secondary analysis. Micro-records from evaluations of the effects of capital punishment on crime in the United States, from randomized trials on the effects of cultural enrichment programs on children in Colombian *barrios*, and from randomized trials on graduated taxation plans

are among those that have been made accessible. These data have been reanalyzed to confirm earlier analyses, test new hypotheses, and for other reasons (Boruch, Wortman, and Cordray, 1981). In milestone studies, Mosteller (1995) and Krueger (1999) reanalyzed micro-records from the Tennessee Class Size randomized trial to verify earlier analyses by Finn and Achilles (1990) that had found that reducing class size had substantial effects on children's achievement. As a practical matter, the internet makes access to machine-readable micro-records on impact evaluations far more feasible than it has been. This in turn means that people who undertake systematic reviews, meta-analyses, and evaluation syntheses will be able to undertake deeper reviews that capitalize on micro-records rather than only on evaluation reports. The research literature on systematic reviews, meta-analysis, and research synthesis, however, is disconnected from the research literature on data sharing and secondary analysis of micro-records. Still to be worked out are ethical issues generated by using data collected from individuals for an earlier study they had consented to, for a different project.

To What End? Value Added and Usefulness

Systematic reviews have not only generated surprising results that countered widely believed notions, but they have also led to some important by- products.

Value Added: Surprises

What can evaluators and users of evaluations learn from a disciplined meta-analysis or systematic review? Surprises are important, as are independent confirmation of results of an earlier review.

Roberts and Kwan (2002) reviewed randomized trials on driver education programs to understand whether they worked. Given substantial investments in such programs in the United Kingdom, United States, and elsewhere, the public would expect that the programs would be found effective. Using Cochrane Collaboration standards and procedures, Roberts and Kwan found that the programs did not lead to lower accident rates among graduates of driver education programs. Because students got their licenses earlier than non-students as a consequence of graduating from these programs, their exposure risk was higher. This led to more accidents.

Shadish and others (1993) produced an award-winning systematic review showing that marital and family therapy, on average, placed about 70 percent of participants above the mean of control group members (50 percent base).

The origin of this review lay in serious doubts about the effectiveness of such therapy, including criticism of it by therapists whose work focused on individuals rather than couples or families. The doubts were put to rest, for a while at least, on scientific grounds.

Cooper, Robinson, and Patall (2006) examined a topic that brings anxiety, if not fear and loathing, to many parents, not to speak of children or teachers: homework. Their systematic review of studies of the effects of homework covered elementary, middle, and high school. It led to recommendations that in elementary school grades, one ought not to expect the homework assignments to yield better test scores. Rather, one should expect better study habits. It led to recommendations, based on reliable studies, that assignments for elementary school students ought to be short, and engage materials found at home. The academic benefits of homework kicked in at middle school and could be regarded as an extension to classroom and curriculum in high school. This review and the recommendations based on it have been featured in contemporary media, such as *The Wall Street Journal* and *The New York Times*, on TV shows, and in forums at the local school-level and national levels.

In the medical sector, Chalmers and others recognized that over a twenty-year period, over fifteen different approaches to handling acute myocardial infarction had been tested in randomized trials. Results varied. The main message, roughly speaking, was this: meta-analyses of diverse evaluative studies showed that anti-clotting drugs "almost certainly" reduced the risk of dying by 10 to 20 percent. Further, streptokinase is among these drugs, tested in over thirty trials. Over reported trials, cumulative odds ratios favor the interventions. Part of the surprise in this is that many physicians had paid no attention to the earlier evidence (Hunt, 1997).

Academic Disciplines, the Policy Sector, and Dependence on Systematic Reviews

An indicator of value added to the sciences is that meta-analyses and systematic reviews are undertaken in many disciplines, including agricultural sciences, physiological research, psychology, education, health research, and the physical sciences (see Chalmers, Hedges, and Cooper, 2002, and Cooper, Hedges, and Valentine, 2009, for specific references in each area). Recent workshops undertaken by the National Academy of Sciences (NAS), on field evaluation of methods and tools for intelligence and counterintelligence, made use of a Campbell Collaboration Crime and Justice Group review by Lum, Kennedy, and Sherley (2006), for instance.

To judge from Cordray and Morphy's (2009) empirical study and from deliberations of the Milbank Fund in the health sector, systematic reviews of the kinds discussed here are not well recognized and are infrequently used by policy-makers. Examples given earlier from the Washington State Institute for Public Policy and the U.S. Government Accountability Office are exceptional. Understanding how to ensure that policy people know about the evidence, understand it, have the capacity to use it, and are willing to use it in this context is as important as the challenge of encouraging use of dependable evidence in the policy sector more generally. John Graunt discussed the matter in the seventeenth century. (We will not tease the reader with a reference for this history. Unless asked.)

By-Products

Some by-products of organized efforts to produce systematic reviews are important. These include uniform transparent guidelines on classifying the quality of evaluations on the basis of their design and execution. Higher order guidelines make explicit the standards used in deciding whether an assembly of evaluations justifies a systematic review or meta-analysis. To take a simple example, the Campbell Collaboration and the Cochrane Collaboration require that each review make the standards explicit and, moreover, abide by Collaboration guidelines in doing so. Randomized trials are put high in the priority of designs that justify a causal inference. Simple before-after studies are low in priority unless some remarkable evidence or theory can be invoked to justify causal claims based on the results. To the extent that reviews and organizational efforts make standards of evidence explicit, we expect that the number of new studies that can sustain causal inferences will increase.

Another by-product is the development of better databases that can serve as the reservoir from which studies are drawn for systematic reviews. For instance, Medline searches routinely failed to identify *randomized trial* in that database until the 1990s. The Cochrane Collaboration's hand searches of journals revealed that these searches had a far higher yield of trials than Medline-based searches. Medline changed its database policy to ensure that randomized trials are more easily detectable to anyone, including Cochrane people who do reviews, trialists who are designing a study, and so on. For example, this effort resulted in adding, when applicable, the words "Randomized controlled trial" to the "publication type" heading for abstracts (Willis, 1995).

Organized networks to generate systematic reviews, supported by individual pro bono efforts, can be construed as another kind of product, notably social and intellectual capital. The Cochrane Collaboration has developed a

network of over 10,000 people involved in health-related reviews in nearly thirty countries, for instance. Cochrane's sibling, the Campbell Collaboration, has involved people from ten to fifteen countries in annual meetings since 1999. The people in these networks include evaluators and other applied researchers, policymakers, and practitioners of other kinds.

Conclusion

The title of this chapter could easily have been "Try All Things and Hold Fast to That Which Is Good," exploiting one of St. Paul's letters to the Thessalonians. We can find similar ideas in medieval Arabic literature, notably Ibn Khaldun's *al Muqaddimah*, in the writings of nineteenth-century scientists and practitioners such as Florence Nightingale, and elsewhere.

People who today do systematic reviews stand on the shoulders of such colleagues in at least two respects. First, they, as their departed colleagues did, try to understand what is good. That is, they take seriously the question of what evidence justifies the claim that the intervention, program, or policy worked better than an alternative in a fair comparison. Second, contemporary systematic reviewers also try to bring order out of the chaos of publications in academic journals, the ill-disciplined issuances on websites, the declarations on television and in blogs, tweets, and technological whatnot. They do so in ways that make the processes and standards of evidence plain. Ibn Khaldun would have admired. Ditto for Florence. Maybe even Paul.

References

Berk, R. A. "Statistical Inference and Meta-analysis." *Journal of Experimental Criminology*, 2007, *3*(3), 247–270.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. *Comprehensive Meta-Analysis Version 2.* Englewood Cliffs, NJ: Biostat, 2005.

Boruch, R. F., and Rui, N. "From Randomized Controlled Trials to Evidence Grading Schemes." *Journal of Evidence Based Medicine*, 2008, *1*, 4–49.

Boruch, R. F., Wortman, P. M., and Cordray, D. S. (1981). *Reanalyzing Program Evaluations: Policies and Practices for Secondary Analysis of Social and Education Programs.* San Francisco, CA: Jossey-Bass.

Chalmers, I., Hedges, L. V., and Cooper, H. "A Brief History of Research Synthesis." *Education and the Health Professions*, 2002, *25*(1), 12–37.

Cooper, H., Robinson, J. C. and Patall, E. A. (2006). "Does Homework Improve Academic Achievement? A synthesis of Research, 1987–2003." *Review of Educational Research*, *76*, 1–62.