The Campbell Collaboration: A Reliable Source of Evidence for Practice

Julia H. Littell, PhD

Where can professionals find the current best evidence for practice with vulnerable children and families? Many look to published research reviews or the many summaries of "evidence-based practices" (EBP) produced by scholars, government agencies, and professional organizations. But, as I explain, the accuracy of these sources varies. How can we tell whether a research summary is accurate or whether it is an unbiased assessment of the evidence?

This article considers what constitutes "good" evidence and introduces helping professionals to The Campbell Collaboration, a new international organization devoted to producing the highest-quality systematic reviews of evidence on "what works" and what doesn't in the fields of social care. By way of introduction, let us examine the contexts in which the need for more reliable information on intervention effects emerged.

Evidence-Based Practice

For more than a century, there have been movements to use "scientific" evidence to inform practice and policy and to improve the health and well-being of vulnerable children and families. In the scientific charity movement of the late nineteenth century, practitioners looked to science to solve social problems. As the medical and social sciences evolved, attempts to link science and practice also emerged. In the scientist-practitioner movement in social work in the 1970s, practitioners were expected to generate scientific evidence and use it in practice. Many experienced a fundamental conflict between the faith required for practice and the skepticism required for research. The current movement toward evidence-based practice takes into account a division of labor in the helping professions (i.e., most practitioners are not researchers and vice versa), and it links practice and research in other ways.

Evidence-based practice derives from evidence-based medicine, which was developed in the early 1980s in Canada and has had its greatest impact in the United Kingdom. David Sackett, one of the founders of evidence-based medicine, defined it as "the conscientious, explicit, and judicious use of current best evidence in making decisions" about individual cases (Sackett et al., 1996, p. 71). EBP is *a process performed by clinicians*. The process of EBP involves formulating answerable questions, seeking answers, appraising the evidence, applying the results, and assessing the outcomes.

Although interest in EBP is growing (certainly much lip service is paid to it), current discourse includes considerable confusion about what EBP is and isn't. Some experts appear to favor a cookbook approach to EBP, linking certain diagnoses or conditions to practices that have been "proven" effective in similar cases and then promoting or even legislating the use of these treatments. That is certainly *not* what Sackett had in mind. Other oversimplifications include the classification of interventions as "effective," "promising," or "not effective," based on criteria that vary (considerably) from one classification schema to the next. Widespread misconceptions include the notions that the only evidence that matters in EBP is evidence about outcomes, and that the only credible evidence on outcomes comes from randomized controlled trials.

Evidence for Practice

Many sources and types of evidence are relevant for practice. Current models of evidence-based practice and policy (EBP) encourage professionals to seek and carefully consider credible information about clients' needs, values and preferences, contexts and constraints, and interventions effects. Practitioners may come up empty-handed, unable to find credible evidence on one or more of these topics; nevertheless, they must act. In these circumstances, EBP helps practitioners clarify what we *don't* know.

Scientific evidence is always tentative, constantly evolving, and incomplete. EBP practitioners can avoid ruts and fads by recognizing that the current best evidence is not the last word on the subject.

Next, let us focus on *one* type of evidence for practice—results of empirical research on intervention effects. This is not necessarily any better or more important than other types of evidence. However, if we are concerned about intervention effects, we ought to examine and synthesize this kind of evidence carefully.

How Do We Know What Works?

A range of evaluation methods is used to identify effects of interventions, and diverse review methods are used to synthesize results of multiple studies of intervention effects.

Effect Studies

Randomized controlled trials (RCTs) are generally considered the gold standard in evaluation research because RCTs are most likely to provide unbiased estimates of intervention effects. However, RCTs are not always appropriate nor are they foolproof. There are many alternatives to the RCT, some more reliable than others.

RCTs conducted under carefully controlled conditions, often in university clinics and in close collaboration with program developers, are sometimes called "efficacy studies." RCTs conducted under real world conditions are called "effectiveness" studies. In general, efficacy studies inform us about the *potential* impact of an intervention under ideal conditions, while effectiveness studies show the *likely* impact in practice settings.

An RCT can yield two different kinds of information. First, researchers can assess outcomes for all cases in the original treatment and comparison groups to provide an unbiased estimate of the overall impact of the treatment as it was implemented. This is called an "intent-to-treat" (ITT) analysis. This information is often what policy makers and agency administrators want. It tells us what effects we are likely to obtain if an intervention is implemented, even though participation levels vary and some people drop out.

Second, RCTs may yield an analysis of "treatment on the treated" (TOT). This includes only those cases that received the "full dose" of treatment. Because drop-outs are excluded (and attrition is often not random, i.e., people drop-out for various reasons) the TOT analysis does not follow the original experimental design. Hence, these results are essentially quasi-experimental.

cont'd on page 8

The validity (accuracy or credibility) of research results is not simply a function of research design or methods. Validity is a property of the inferences one can draw from a study. For example, characteristics of the sample and setting affect our ability to extrapolate results to other people and places. Characteristics of measurement instruments affect our ability to examine central ideas and associations between constructs. Competing explanations for differences between experimental and comparison groups affect our ability to identify intervention effects. Small studies often lack the statistical power necessary to detect clinically meaningful effects, and some studies are unduly influenced by extreme cases (outliers).

There is no such thing as a perfect study. Single studies, no matter how rigorous, have limited generalizability. Multiple studies are needed to enhance confidence in results, or modify or refute previous findings. Independent replications are necessary to counter "allegiance effects" that may appear when interventions are studied by their developers.

Synthesizing Results of Multiple Studies: Traditional Reviews

The most common method of synthesizing results of multiple studies is the traditional literature review. The traditional method involves finding relevant studies, describing them, and generating conclusions about what the weight of the evidence suggests. This approach is vulnerable to several types of bias.

Sampling bias. Reviewers tend to obtain a convenience sample of published studies. If I do a keyword search of the electronic bibliographic databases that happen to be available to me, my results will differ from those obtained by a colleague who uses the same keywords in a different location with access to different databases. Further, because journal indexing is fallible, relevant studies may be missed by electronic keyword searches.

Publication bias. Reports with positive, statistically significant findings are more likely to be submitted for publication and more likely to be published than those with negative or null results. This publication bias may suppress reporting of nonsignificant findings in studies that have mixed results. If a research review considers only published reports and there are many, relevant unpublished studies on the topic, positive effects will be overestimated.

Because journals limit the length of published articles, descriptions of intervention and research methods are often incomplete. Some are more tidy than accurate. If reviewers rely only on published accounts of studies, they may miss important implementation issues that affect the interpretation of results.

Confirmation bias. Researchers and reviewers who expect certain results are likely to find them if they use evidence selectively. People tend to accept evidence that confirms their expectations and dismiss that which does not.

The task of combining results of multiple studies is quite complex. No two studies are identical. Studies on the same topic may have different sample characteristics, designs, outcome measures, and results. Traditionally, reviewers have used their best cognitive algebra to sort out differences among studies and sum up results. This mental math is not very accurate. Studies have shown that reviewers' conclusions may be affected by irrelevant information (e.g., the wording of titles of research articles and authors' reputations or affiliations).

Further, what is being "counted" in the traditional literature review depends entirely on original study report. Since many small studies lack power to detect effects, adding them up can lead reviewers to underestimate intervention effects.

The Science of Research Synthesis

For the past century, statisticians and scholars have worked to develop methods to combat the biases inherent in traditional narrative reviews. Beginning with Pearson's work in medicine in 1904, researchers have created systematic approaches to synthesizing results of multiple studies. This includes meta-analysis and "systematic reviews."

Meta-analysis

Meta-analysis refers to the quantitative synthesis of findings from two or more studies. In a meta-analysis, original data from each study are converted to a common metric called an "effect size" (ES). There are several ES metrics, but the most common is the *d* index (also called the standardized mean difference), which expresses differences between treatment and control groups in standard deviation units.

Meta-analysis can be used to answer a number of different questions about a body of research. Usually we want to know whether all of the studies in our sample point to the same conclusion (heterogeneity tests address this issue). We may also want to estimate an average or overall effect across studies. Average ES are created by weighting study ES by their precision (inverse variance) and then averaging results across studies. This means that more precise studies (usually larger studies and those with more consistent results) "count" more in the overall average. This is as it should be. (We wouldn't want a study with 10 cases to count as much as one with 1,000; nor would we want a study with widely varying results to count as much as one with consistent results.)

Because meta-analysis includes data from all subjects in the original studies, many underpowered studies with statistically nonsignificant results can add up to an average ES that is statistically significant and clinically meaningful.

If many effect studies are available, meta-analysts can examine differences among them to address other questions relevant to policy and practice. For example, we might want to know whether a program tends to be more effective with younger children or older ones, whether low-income families benefit as much as wealthier families, whether more intensive programs have stronger effects, and so on.

Meta-analysis is not always possible or desirable. It does not make sense to combine studies that address different questions, and the quantitative synthesis of results of many weak studies is still weak. Further, meta-analysis attends to only one phase of the review process—the analysis. The science of research synthesis can be brought to bear on other aspects of the review process.

Systematic Reviews

Systematic reviews focus on the scientific aspects of *all* phases of research synthesis. Unfortunately, the term has been widely misused, and many so-called systematic reviews aren't.

A systematic review has explicit objectives, uses transparent procedures, and attempts to minimize bias in the identification, assessment, and synthesis of research results. Procedures are spelled out in advance and are documented so that others can critically appraise or replicate the review, or both. A systematic review follows the basic steps in the research process (further, while most studies sample people, families, organizations, and the like, a systematic review samples and analyzes prior studies). The steps are as follows:

First, explicit objectives are stated and eligibility criteria (inclusion and exclusion criteria) are formulated to specify the types of study designs, interventions, populations, and outcomes that will be included in the review.

Second, a systematic search strategy is designed to reduce bias in the identification of eligible studies. The search strategy specifies keyword strings and sources that will be used to find relevant studies in bibliographic databases and other electronic media. The search may be bounded by dates, journals, databases, and so forth, as long as the search procedures are transparent and replicable. The search strategy includes efforts to find "gray" (unpublished) literature; this usually involves contacts with experts in the field and careful scanning of relevant bibliographies. Many reviewers use hand searches of key journals to identify relevant studies that are not fully indexed.

Next, reviewers conduct the search and document results. Decisions about full text retrieval and study eligibility are usually made by two independent raters to increase reliability. Specific reasons are given for each study exclusion.

The data from eligible outcome studies are extracted by independent raters onto uniform paper or electronic forms. These data include characteristics of the study (e.g., design, attrition), interventions, samples, outcome measures, and findings. Again, coding is conducted by multiple reviewers to increase reliability; differences among coders are discussed and resolved (sometimes by a third person). Reviewers assess many qualities of eligible studies and seek additional information from primary authors as needed.

A systematic review *may* include meta-analysis if there are two or more similar studies that meet the eligibility criteria.

Finally, the review process and results are reported in a structured and detailed document.

Potential problems with systematic reviews. A systematic approach does not guarantee that a review will be free of bias, although transparent methods facilitate commentary and debate about the integrity of a review.

Systematic reviews are very labor intensive and, therefore, expensive. Costs depend on the duration and complexity of the review and range from $40K to $200K per review.

Once completed, systematic reviews may become outdated as results of new studies become available. To remain relevant, reviews must be updated every two or three years.

What do these problems mean for practitioners and policy makers? Many people underestimate the complexities of finding, assessing, and synthesizing evidence scientifically. Will practitioners and policy makers be able to do this and keep their day jobs? In the EBP framework, ultimate responsibility for the assessment and use of evidence lies in the hands of the practitioner and policy maker. Realistically, decision makers need help with this process. Reliable sources of evidence—a body of systematic reviews on topics relevant to practice and policy—will be enormously useful for policy and practice. Again, this will not be the last word on the issues, nor should it obscure other types and sources of evidence that practitioners need to consider in making decisions (i.e., evidence about the needs, preferences, contexts, and constraints present in individual cases).

Some systematic reviews find that there is no credible evidence on a topic. When this occurs, decisions must be made on other grounds. However, arming practitioners and policy makers with the knowledge that there is no good information on a topic helps them fulfill the EBP dictum to consider the current best evidence. In short, systematic reviews are merely a way of carefully compiling and making available one type of information for practice.

Where Is the Evidence That Is Needed?

During the past decade, important advances in the science and practice of research synthesis included improved methods of information retrieval, better understanding of relationships between research design and outcomes, and development of statistical techniques and software for meta-analysis. At the same time, many organizations and individuals made extensive efforts to compile and synthesize empirical evidence on intervention effects for specific conditions and problems. Practitioners and policy makers who want to know "what works" and "what works best for whom" can find many lists of empirically-supported programs on Web sites sponsored by government agencies, foundations, and professional organizations. More thorough treatments of these topics are available in government reports and peer-reviewed publications.

This said, with a few exceptions, advances in the science and practice of research synthesis have not been connected. As a result, putatively authoritative reviews and lists of effective practices have proliferated with little attention to the science of research synthesis. Ironically, while these lists and reviews are aimed at providing evidence for practice and policy, they are not themselves based on evidence about how to find, summarize, and synthesize research findings.

The Campbell Collaboration

Building on the successful, collaborative model of rigorous research synthesis pioneered by the Cochrane Collaboration in medicine (see www.cochrane.org), the Campbell Collaboration was created in 1999 to bridge the science and practice of research synthesis and produce the highest-quality systematic reviews of research on intervention effects in the fields of social care.

The Campbell Collaboration is an independent, nonprofit organization devoted to producing reliable information on effects of behavioral, social, and psychological interventions (see www.campbellcollaboration.org). Named for Donald T. Campbell, the Collaboration (fondly known as C2) strives to minimize bias and maximize the quality, relevance, timeliness, and accessibility of information for policy and practice.

C2 develops standards for systematic reviews, offers training in systematic review methods, provides technical assistance to review

cont'd on page 10

teams, ensures that C2 reviews meet C2 standards through a peer-review process, and provides Web-based access to C2 systematic reviews. C2 hosts annual colloquia and fosters international, interdisciplinary perspectives on social problems. In addition, C2 maintains a unique, electronic register of studies of the effects of psychosocial, behavioral, and educational interventions.

C2 is organized by a corporate Board, an international Steering Group, a Secretariat's office, and six Coordinating Groups. The Coordinating Groups cover three substantive areas (education, social welfare, and crime and justice) and three cross-cutting topics (methods, communication, and users). Some of the Coordinating Groups have subgroups (e.g., the Methods Group has subgroups on topics such as training, information retrieval, research design, and statistics). The Cochrane Collaboration and C2 relate to each other through overlapping steering groups and subgroups.

C2 has been supported by largely volunteer efforts from an international, interdisciplinary network of scholars, practitioners, and policy makers and by the work of other nonprofit organizations, government agencies, and foundations (particularly in the United Kingdom and Nordic countries). Recent support from the American Institutes for Research has allowed C2 to begin to build a more permanent infrastructure.

C2 Reviews

Like Cochrane reviews, C2 reviews are produced by independent review teams that follow certain policies and procedures (policy statements on information retrieval, research design, and statistics are available on the C2 Web site). C2 reviews are *not* limited to RCTs, but evidence from RCTs is analyzed separately from evidence from other kinds of studies.

The process begins when a review team registers a "title" for the review with a C2 Coordinating Group. The title declares the review team's objectives and outlines a preliminary approach to the review topic. Next, the team develops a detailed "protocol" or plan for the review. The protocol is vetted by substantive and methodological experts within and outside of C2, who comment on the relevance of the proposed review for practice and policy in a particular field, along with its methodological rigor. The protocol includes an explicit statement about any potential conflicts of interest. Completed reviews are also vetted by substantive and methodological experts within and outside C2. C2 reviews are expected to be updated every two or three years.

Once accepted by a C2 Coordinating Group, all products (titles, protocols, and completed reviews) are posted on the C2 Web site. Commentaries may be posted as well.

The C2 Social Welfare Coordinating Group

The C2 Social Welfare Group may be of interest to APSAC members because this group covers topics related to child abuse. Within the Social Welfare Group, the Developmental, Psychosocial, and Learning Problems (DPLP) Subgroup, which is coregistered with the Cochrane Collaboration, has produced systematic reviews on topics such as cognitive-behavioral interventions for sexually abused children and school-based programs for prevention of child sexual abuse.

To date, most of the interest, effort, and funding for systematic reviews in social welfare has been located in Europe; hence, the Social Welfare Group is only beginning to organize networks of scholars, practitioners, policy makers, and funders in North America. The group covers a wide range of topics, including child welfare, mental health, substance abuse, public health, aging, poverty, housing, welfare, work, and family life.

Using the Science of Research Synthesis

Practitioners and policymakers can use valid, up-to-date research syntheses to make informed decisions about the likely impacts of social and behavioral interventions. However, research reviews might be biased, particularly when they are not based on understanding of common problems and methods of research synthesis. Consumers should be wary of traditional research reviews that rely on narrative summaries of convenience samples of published studies. Valid summaries of available evidence are more likely to come from systematic reviews that use transparent methods and attempt to minimize bias at every step in the review process.

To illustrate the differences between traditional and systematic reviews, my colleagues and I recently completed a jointly-registered Cochrane/C2 review on effects of multisystemic therapy (Littell, Popa, & Forsythe, 2005). Results of this systematic review (available in the *Cochrane Library, Issue 4*, and on the C2 Web site) are not consistent with those of traditional, narrative reviews or partially-systematic reviews of the same body of evidence.

This suggests a need to reassess and update empirical evidence that has been reviewed by traditional methods. The Campbell Collaboration provides a platform for this purpose by bridging the science and practice of research synthesis, developing reliable syntheses of evidence on intervention effects in the fields of social care, and promoting open debate about the evolving evidentiary status of interventions. Practitioners and policy makers are welcome to join the Campbell Collaboration, suggest topics for systematic reviews, participate in review teams, and critique C2 products (contact: jlittell@brynmawr.edu).

About the Author

Julia H. Littell, PhD, is Associate Professor in the Graduate School of Social Work and Social Research at Bryn Mawr College, specializing in children and family issues and child welfare. She is a member of the editorial board of Children and Youth Services Review and a member of the Steering Committee of the Campbell Collaboration.

References

Littell, J. H., Popa, M., & Forsythe, B. (2005). Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10-17 (Cochrane Review). In *The Cochrane Library, Issue 4, 2005*. Chichester, UK: Wiley.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*, 71-12.