American Psychologist

The Overview of Reviews: Unique Challenges and Opportunities When Research Syntheses Are the Principal Elements of New Integrative Scholarship

Harris Cooper and Alison C. Koenka

Online First Publication, February 20, 2012. doi: 10.1037/a0027119

The Overview of Reviews

Unique Challenges and Opportunities When Research Syntheses Are the Principal Elements of New Integrative Scholarship

Harris Cooper and Alison C. Koenka *Duke University*

In the past two decades, a new form of scholarship has appeared in which researchers present an overview of previously conducted research syntheses on the same topic. In these efforts, research syntheses are the principal units of evidence. Overviews of reviews introduce unique problems that require unique solutions. This article describes what methods overviewers have developed or have adopted from other forms of scholarship. These methods concern how to (a) define the broader problem space of an overview, (b) conduct literature searches that specifically look for research syntheses, (c) address the overlap in evidence in related reviews, (d) evaluate the quality of both primary research and research syntheses, (e) integrate the outcomes of research syntheses, especially when they produce discordant results, (f) conduct a second-order meta-analysis, and (g) present findings. The limitations of overviews are also discussed, especially with regard to the age of the included evidence.

Keywords: overviews of reviews, research synthesis, systematic reviews, meta-analysis

At first, the notion of an overview of systematic reviews—especially a quantitative integration of meta-analyses—conjures up an ad infinitum coupling of scientific evidence that would resolve in absurdity. In fact, not long after the introduction of meta-analysis, Kazrin, Durac, and Agteros (1979) lampooned the notion of a "meta-meta-analysis" by suggesting that such efforts could lead to meaningless levels of generalization, to wit: "Independent variables do work. Unfortunately, there were some unforeseen interactions. Apparently, not all independent variables work on all dependent measures, especially at the same time" (p. 398).

Despite Kazrin et al.'s (1979) concerns, overviews of reviews provide unique and important contributions to various bodies of literature and are appearing with greater frequency. Overviews emerged out of a need identified by numerous independent researchers. In the pursuit of particular applications, individual researchers created their own methodological templates. These efforts appeared under a variety of labels, including "overview of reviews," "review of reviews," "umbrella reviews," "meta-reviews," and even Kazrin et al.'s (1979) "meta-meta-analysis."

Thomson, Russell, Becker, Klassen, and Hartling (2010) searched health-related reference databases from 2000 to 2009 and found 139 overviews. Twenty-three appeared in 2009. So, it is time to take a look at how overviews are carried out, with an eye toward improving the process.

First Codification

The Cochrane Collaboration is an international consortium of medical researchers who produce systematic reviews[1] on health topics (http://www.cochrane.org). The *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2008) presents the reporting requirements for reviews in the Cochrane Library, which is perhaps the most respected source of scientific evidence on health care interventions. The most recent *Cochrane Handbook* concludes with a chapter not found in previous editions. The chapter is titled "Overviews of Reviews" (Becker & Oxman, 2008). In it, the authors codified and extended procedures for conducting this new form of scholarship. They defined a Cochrane overview as a review "designed to compile evidence from multiple systematic reviews of interventions into one accessible and usable document" (Becker & Oxman, 2008, p. 607). To our knowledge, this chapter represents the first attempt at a comprehensive treatment of the overview of reviews.

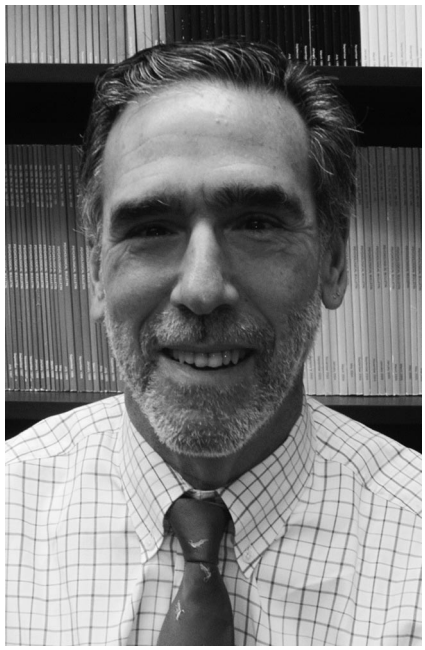
The purpose of the present article is to do for overviews in the behavioral sciences what Becker and Oxman (2008) did for overviews in the health sciences. It is time to begin developing standards of transparency and rigor and rules of best practice for overviews. We start with Becker and Oxman's definition but adjust its emphasis to reflect the predominant priorities of behavioral scientists. We then examine methods that behavioral scientists have developed and borrowed to help them carry out fair and open over-

Harris Cooper and Alison C. Koenka, Department of Psychology and Neuroscience, Duke University.

Correspondence concerning this article should be addressed to Harris Cooper, Department of Psychology and Neuroscience, Box 90086, Duke University, Durham, NC 27707. E-mail: cooperh@duke.edu

[1] Throughout the article, we use the terms *systematic review* and *research synthesis* interchangeably. The term *primary research* is used to refer to the individual studies that go into a research synthesis or systematic review. The term *meta-analysis* is used to refer to the quantitative integration of the results of primary research.

 0003-066X/12/$12.00
DOI: 10.1037/a0027119



Harris Cooper

views. We focus on aspects of overviews that are unique to this emerging approach to knowledge integration. Throughout, we use examples drawn from psychology, education, and behavioral medicine.

The Purposes and Tasks of Overviews

Becker and Oxman (2008) suggested several reasons why one might undertake an overview of reviews in the medical sciences. Overviews are most appropriate when systematic reviews focus on (a) the same problem using different interventions or (b) the same intervention but different outcomes, populations, or conditions. Also, Becker and Oxman (2008) suggested that overviews can be used to summarize evidence about adverse effects of an intervention and to provide a comprehensive overview of an area before conducting an updated systematic review.

This list of reasons for undertaking an overview of reviews is easily transferable to the behavioral sciences (see also Thomson et al., 2010), with just a few alterations primarily involving shifts in emphasis necessitated by a greater focus on basic processes and theory testing. Becker and Oxman (2008) were focusing their efforts on overviews meant for health care professionals. We would suggest this list is especially well-suited for psychologists interested in evaluating interventions in health and clinical psychology; indeed, most of the examples we describe below are of this nature. However, some behavioral scientists might undertake overviews for other reasons. First, overviews may be written in the service of summarizing research syntheses on correlational research. For example, Peterson (2001) conducted an overview of 30 meta-analyses that looked at how research results differed depending on whether college or noncollege samples were used as participants in studies. He found greater homogeneity in reported effect sizes for studies using college students than for those not using college students and even some differences in the direction of effects. The overview results were used to caution against overgeneralization of results from college samples.

Second, behavioral scientists conducting overviews of interventions tend to focus considerable attention on the mediating mechanisms underlying responses to interventions and on testing theoretical formulations about these mechanisms. We would suggest that with an eye toward process and theory, behavioral scientists might place more emphasis on overviews that catalog the mediator and moderator variables tested in past syntheses. Particular mediators and moderators might be the focus of attention in one or a few systematic reviews but not in others. Providing comprehensive lists of these variables would take high priority in theoretical domains and, indeed, might be where the test of a theory's viability resides.

Finally, the audience for overviews in behavioral science might not be primarily care providers (the principle audience for Cochrane overviews) but rather other researchers. Perhaps, then, a greater emphasis in behavioral science overviews would be on resolving discrepancies in the outcomes of the covered syntheses and on pointing the direction for future research.

With these caveats in mind, we offer the list of objectives for overviews presented in Table 1, for which we are beholden to Becker and Oxman (2008) but to which we have added modifications meant to be more inclusive of purposes pertinent to the more conceptual domains of behavioral inquiry.

The Structure and Sequence of Steps in an Overview

The structure and sequence of tasks confronting a scholar starting on an overview of reviews vary little from those widely adopted for use in conducting a research synthesis. Cooper (2010) parsed the activities involved in conducting a research synthesis into seven categories: (a) formulate the problem, (b) search the literature, (c) gather information from research reports, (d) evaluate the quality of the evidence, (e) analyze and integrate the outcomes of research, (f) interpret the evidence, and (g) present the results. These steps are easily converted into the steps of conducting an overview. Table 2, adapted from Cooper (2010), briefly presents each step in the overview process along with the question asked at that step, the step's primary function in the overview, and how variations in the performance of a step might lead to differences in the conclusions drawn in the overview.

Unique Characteristics of Overviews of Reviews

While the structure and sequence of tasks are similar for research syntheses and overviews, the techniques used to carry out the tasks can be similar or quite different. We now turn to an examination of these tasks. We begin with similarities but provide detailed descriptions of some of the tech-

Alison C. Koenka

niques used by overviewers to meet the unique challenges they face.

The Breadth of the Questions Asked

One distinction between primary research, research synthesis, and an overview of several related syntheses is in the breadth of the constructs of interest. In the narrowest case, the focus of the overview and the focus of *any one* of its constituent syntheses can occupy very similar problem spaces. For example, Wigal et al. (1999) presented an overview of research syntheses on the effectiveness of stimulants used to treat children with attention-deficit/hyperactivity disorder (ADHD). These overviewers found systematic reviews that had examined research on different stimulant drugs, looked at different outcomes, and did so in varying contexts. However, because their problem was narrowly and largely operationally defined, there was much overlap in the content and focus of the syntheses in the overview.

On the other hand, an overview can define a problem space much broader than that occupied by any given synthesis it contains. And, the foci of syntheses contained in the overview can have little resemblance to one another. For example, Lipsey and Wilson (1993) presented an overview on the very broad topic of the effectiveness of psychological, behavioral, and educational interventions. Their overview contained 302 independent meta-analyses on a wide variety of interventions. Further, these researchers *excluded* meta-analyses with overlapping content. Needless to say, they found not one meta-analysis with a problem space defined similarly to that of their overview, though every one of the contained meta-analyses examined an intervention falling within it. Thus, overviews often seek to answer questions that are much broader in scope than questions that are typically asked by a single research synthesis.

Table 1
Purposes of Overviews of Reviews in the Behavioral Sciences

- To summarize evidence from more than one research synthesis focused on the same or overlapping research problems or hypotheses
- To compare findings and resolve discrepancies in the conclusions drawn in more than one research synthesis focused on the same research problem or hypothesis
- To catalog the mediators and moderators of a revealed effect or relationship tested in research syntheses focused on the same research problem or hypothesis
- To identify gaps in the literature where multiple studies may exist on the same problem or hypothesis but a research synthesis has not been performed
- To supplement existing research syntheses by including studies they did not include, either because the studies were omitted or appeared after the syntheses were conducted

Searching for Research Syntheses Only

Another feature that is unique to conducting an overview arises in the search of the literature. When searching the literature, overviewers can greatly increase the precision of their search by including terms related to the document type—a research synthesis—that interests them. Montori, Wilczynski, Morgan, and Haynes (2005) ran an empirical test of different search strategies meant to retrieve systematic reviews using Medline as the reference database (see also Wong, Wilczynski, & Haynes, 2006). First, looking for instances of systematic reviews, they hand searched the issues of 161 journals covered in Medline and published in the year 2000. Then they calculated the sensitivity and specificity of the results of all possible combinations of their search terms related to the type of document.[2] Using document types—such as *meta-analysis* and *review, academic*—both singly and in combination, the authors were able to construct searches with near 100% sensitivity and around 50% specificity.[3]

What document types are best to search for in behavioral science? We conducted a search using PsycINFO in which we employed the terms *literature review*, *meta-analysis*, *systematic review,* and *research synthesis*. Table 3 shows the number of retrievals from a search of abstracts in the entire PsycINFO database when the abstract (a)

[2] Sensitivity was measured by the proportion of relevant documents that were retrieved, using the hand search as the gold standard, with higher proportions meaning more complete coverage. Specificity was the proportion of relevant documents retrieved, with higher proportions meaning searchers needed to examine fewer documents falling outside their problem space.

[3] It can be argued that for the early stages of a literature search, sensitivity is more important than specificity; looking at document records related to but just outside the boundaries of a problem space can help the searchers refine the placement of their conceptual boundaries.

Table 2
Steps in Conducting an Overview of Reviews

Step	Research question asked at this stage of the overview	Primary function served in the overview	Procedural variation that might produce differences in conclusions
Formulating the problem	What research syntheses will be relevant to the problem or hypothesis of interest?	Define (a) the variables and (b) the relationships of interest so that relevant and irrelevant syntheses can be distinguished.	Variation in the conceptual breadth and detail of definitions might lead to differences in the research syntheses included in the overview.
Searching the literature	What procedures should be used to find relevant syntheses?	Identify (a) sources and (b) index terms used to search for relevant syntheses.	Variation in searched sources and index terms might lead to different syntheses being retrieved.
Gathering information from syntheses	What information about each synthesis is relevant to the problem or hypothesis of interest?	Collect relevant information about syntheses in a reliable manner that can be verified by others.	Variation (a) in information gathered from syntheses might lead to different overview conclusions, (b) in coder training might lead to differences in entries on coding sheets and/or, (c) in rules for deciding what syntheses are independent might lead to differences in the amount and specificity of data used to draw cumulative conclusions.
Evaluating the quality of evidence	What syntheses should be included or excluded from the overview based on (a) the methods used in the studies contained in the research synthesis and/or (b) the methods used to conduct the research synthesis itself?	Identify and apply criteria that separate syntheses conducted in ways that correspond with the research question from those that do not.	Variation in criteria for decisions about inclusion of syntheses might lead to systematic differences in which syntheses are contained in the overview.
Analyzing and integrating the outcomes of syntheses	What procedures should be used to condense and combine the syntheses' results?	Identify and apply procedures for (a) combining results across syntheses and (b) testing for differences in results between syntheses.	Variation in procedures used to integrate results of individual syntheses can lead to differences in cumulative results.
Interpreting the evidence	What conclusions can be drawn about the cumulative state of the research evidence?	Summarize the cumulative evidence with regard to its strength, limitations, and generality.	Variation (a) in criteria for labeling results as "important" and (b) in attention to details of studies and syntheses might lead to differences in interpretation of findings.
Presenting the results	What information should be included in the report of the overview?	Identify and apply editorial guidelines and judgment to determine aspects of methods and results readers will need to know.	Variation in reporting might (a) lead readers to place more or less trust in overview outcomes and (b) influence others' ability to replicate results.

Note. From *Research Synthesis and Meta-Analysis: A Step-by-Step Approach* (pp. 14–15), by H. Cooper, 2010, Thousand Oaks, CA: Sage. Copyright 2010 by Sage Publications, Inc. Adapted with permission.

Table 3
Entries in PsycINFO Using Terms for the Elements of an Overview of Reviews (Searching in the Abstract Only)

Term	Term appears in abstract	Term appears in abstract and NOT the other terms
Literature review	58,235	53,724
Meta-analysis	9,748	7,912
Systematic review	9,432	5,476
Research synthesis	3,894	2,957

mentioned the terms regardless of whether another of the terms was also mentioned and (b) mentioned only that term (i.e., using the NOT function to exclude documents that contained more than one of the terms).

A few things are instructive to note from Table 3. First, the oldest and most expansive of the terms, *literature review,* returned by far the most abstracts. However, including this term will result in low specificity for an overviewer. The term is applied to forms of scholarship that do not necessarily review the results of research (see Cooper, 1988, for distinctions in literature reviews). If the search is amended to include *literature review* AND *research,* the retrieved abstracts are more than cut in half. It is also clear from Table 3 that a sensitive search of the literature for an overview would have to include all four terms. There are literally thousands of documents in PsycINFO that refer to one of these forms of scholarship but to no other.[4]

Typically, when an overview is written, the search description will appear in the text. However, because these searches are often quite broad and complex, overviewers have found that creating a table about the search procedures can present a more comprehensible summary. Table 4 presents an example taken from an overview examining school-based approaches to health promotion by Peters, Kok, Ten Dam, Buijs, and Paulussen (2009). These overviewers did not use keywords to screen for systematic review only.[5]

Screening Syntheses on Criteria Other Than Substantive Relevance

It goes without saying that the first criterion for screening syntheses for inclusion in an overview will be whether or not the synthesis deals with a topic that falls within the overviewers' problem space. Because the topics of overviews can be so broad, overviewers have occasionally relied on expert panels to help them set the boundaries of the problem space. For example, Wigal et al. (1999) used panels of clinicians, educators, and parents to help them identify 10 critical topics for their overview of stimulant effects on ADHD.

Once a set of substantively relevant syntheses has been amassed, overviewers will employ additional screens that further limit the documents they will include. The most prominent of these screens concerns the different criteria that can be used to address the problem of overlap in the content of syntheses.

Syntheses with overlapping evidence. In conducting a research synthesis, researchers must decide, for example, whether two studies reporting the effects of an intervention using the same participants at two different points in time are independent tests of effectiveness. At the other extreme—are two studies with different participants but conducted by the same investigators independent tests? Different decision rules can lead to markedly different data sets and conclusions.

Overviewers are faced with the same issue. However, overviewers may have the relative luxury of being able to assess quite precisely the amount of overlap among the syntheses. The content covered in most research syntheses, and especially meta-analyses, will allow the overviewers to list the primary studies included by the synthesists. The overviewers can then compare these studies with one another to assess the overlap among the syntheses, determine whether it is problematic, and if so, decide precisely which reviews interest them most. For example, Cooper (1989) presented a matrix using rows and columns to display the overlap in nine syntheses of the research on homework. He used this table to estimate the average overlap in evidence and suggested that this lack of overlap might be one reason for discrepant conclusions. He also used it to identify the most exhaustive synthesis.

Overviewers have adopted numerous strategies for handling overlap in their constituent research syntheses. To a degree, the choices will be dictated by the nature of the problem space. For example, Lipsey and Wilson's (1993) overview of meta-analyses examining the effectiveness of psychological, behavioral, and educational interventions was so broad in scope that they could eliminate all overlapping research syntheses when they calculated cumulative effect sizes. The criteria they used were to choose (a) the meta-analyses with the most complete information, and if these were equivalent, (b) the meta-analysis with the largest number of primary studies.

Butler, Chapman, Forman, and Beck (2006), in an overview of meta-analyses on the effectiveness of cognitive behavioral therapies, used criteria that favored the most extensive and methodologically rigorous meta-analyses. Typically, these were also the most recent meta-analyses:

> Methodological strengths we looked for were (1) inclusion of only randomized clinical trials, (2) sample-size weighting of effect sizes, (3) analysis of heterogeneity of effect sizes and outliers, and (4) inclusion of moderator variables in the analyses. These factors took precedence over the mere number of studies contained in the meta-analysis. (Butler et al., 2006, pp. 18–19)

Jepson, Harris, Platt, and Tannahill (2010) also used methodological quality as a rule to eliminate overlapping syntheses but listed the recentness of the syntheses as a final

[4] Using the OR function (meaning the abstract had to contain at least one of the four terms) resulted in over 75,000 retrievals. Remember, however, that this set of documents would then be crossed with (using the AND function) the terms specific to the problem of interest.

[5] Sampson et al. (2009) provided an excellent set of guidelines for judging the likelihood that a literature search has uncovered the relevant literature.

Table 4
Example of a Tabular Summary of the Databases and Keywords Used in an Overview Search

Databases			
PubMed keywords	PsycINFO keywords	ERIC keywords	Review initiatives websites
School health promotion: Curriculum Health-education Health-promotion School-health-services Health-plan-implementation Effectiveness: Program-evaluation Evaluation-studies Risk-reduction-behavior Behavior focus: Smoking Alcohol-drinking Sex-education Diet Food-habits	School health promotion: Curriculum Curriculum-development Educational-programs Schools School-environment Health-education Health-promotion Effectiveness: Effectiveness Educational-program-evaluation Treatment-effectiveness-evaluation Health-attitudes Health-behavior Health-knowledge Behavior focus: Tobacco-smoking Alcohol-abuse Safe-sex Sex-education Sexuality Sexually-transmitted-diseases Food-intake Nutrition Health-behavior Lifestyle	School health promotion: Curriculum School-health-services Health-programs Health-education Comprehensive-school-health-education Intervention Instruction Effectiveness: Program-effectiveness Program-evaluation Program-implementation Outcomes-of-education Knowledge-level Feedback Learning Behavior focus: Tobacco Smoking Alcohol-education Drinking Substance-abuse Sex-education Sexuality Nutrition Nutrition-instruction Eating-habits	Campbell Collaboration Centre for Reviews and Dissemination, York, UK Cochrane Collaboration Effective Public Health Practice Project, Hamilton, Canada EPPI-Centre, London, UK Guide to Community Preventive Services

Note. Publication year: January 1995 – October 2006. Language: English. The keywords within one group of keywords (e.g., school health promotion) were combined with "OR," the groups were combined with "AND." Reprinted from "Effective Elements of School Health Promotion Across Behavioral Domains: A Systematic Review of Reviews" by L. W. H. Peters, G. Kok, G. T. M. Ten Dam, G. J. Buijs, and T. G. W. M. Paulussen, 2009, *BMC Public Health, 9*(182), p. 3. Copyright 2009 by Peters et al.; licensee BioMed Central, Ltd.

criterion. Later we return to the issue of methodological rigor of syntheses, not just in the context of a decision rule to deal with overlap but also as a criterion with which to judge the quality of evidence presented by the syntheses.

Another strategy was used by Nation et al. (2003), who conducted an overview in search of the principles of effective programs meant to prevent drug abuse, delinquency, and violence among youth. They limited overlap by including only one review for each first author "unless it was clear the reviews used different data" (Nation et al., 2003, p. 450).

Finally, Green, Howe, Waters, Maher, and Oberklaid (2005) chose to ignore the issue of overlapping evidence in an overview examining school-based interventions to promote the social and emotional health of primary-school-aged children: "While several studies were included in more than one review, the precise degree of overlap of studies between reviews was difficult to ascertain and was not central to the objectives of this review" (p. 32).

In sum, then, strategies for handling overlap in the evidence base of research syntheses include (1) using the synthesis that (a) provides the most complete description, (b) is most recent, (c) contains the most evidence, (d) is methodologically most rigorous, or (e) is published (passed peer review or is in wide circulation) or (2) simply including all syntheses regardless of overlap. It is possible to argue that each of these strategies is most appropriate, depending on the nature of the problem under consideration.[6]

Publication status. Publication status might also be used to screen research syntheses for an overview. For example, this criterion was used by Mikton and Butchart (2009) in an overview on prevention of child maltreatment. These overviewers included only published peer-reviewed syntheses "since the aim was to focus on reviews with a wide influence on policy and practice" (Mikton & Butchart, 2009, pp. 353–354).

Accessibility and the fact that peer review means published syntheses have been judged on methodological rigor are frequent arguments for including only published work in a synthesis or overview. However, through the years, the arguments for excluding unpublished work have grown less persuasive. Most synthesists today apply their own methodological screens, and these are often more transparent (and many demand more rigor) than those used for publication. Many overviewers do the same. In addition, with the rise of the Internet, researchers (and policymakers and practitioners) have access to many well-conducted syntheses that have never been submitted for formal publication. For example, the systematic reviews contained in the library of the Campbell Collaboration (www.campbellcollaboration.org) and the National Institute for Health and Clinical Excellence (NICE; part of the National Health Service in the United Kingdom; www.evidence.nhs.uk) may or may not have been published anywhere else but on their websites.

Other screening criteria. In addition to overlap and publication status, many other criteria are used to screen syntheses from overviews. For example, it is not unusual for overviewers to restrict syntheses on the basis of their year of appearance, eliminating older syntheses.

A detailed approach for summarizing the search and screening process is provided in Figure 1, which is taken from an overview by Maniglio (2010) of research on sexual abuse as a cause of depression. This overviewer provided a flow chart describing the entire search and screening strategy. This approach has much to recommend it, and it parallels the approach to reporting the flow of data that is recommended by both the Consolidated Standards of Reporting Trials (2007) and the American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). Maniglio (2010) cast a wide net in the search, using the terms *child(hood) sexual abuse* and *child(hood) sexual maltreatment* but no set of terms to delimit the document type. When the search terms are more extensive, perhaps the optimum approach for overviewers would be to include both a table like the example in Table 3 and a flowchart like that in Figure 1.

Whatever approach to reporting is taken, overviewers must describe the screens they used to decide which systematic reviews to include and exclude. Along with the search strategy, this information is critical to any appraisal of the trustworthiness of the overview.

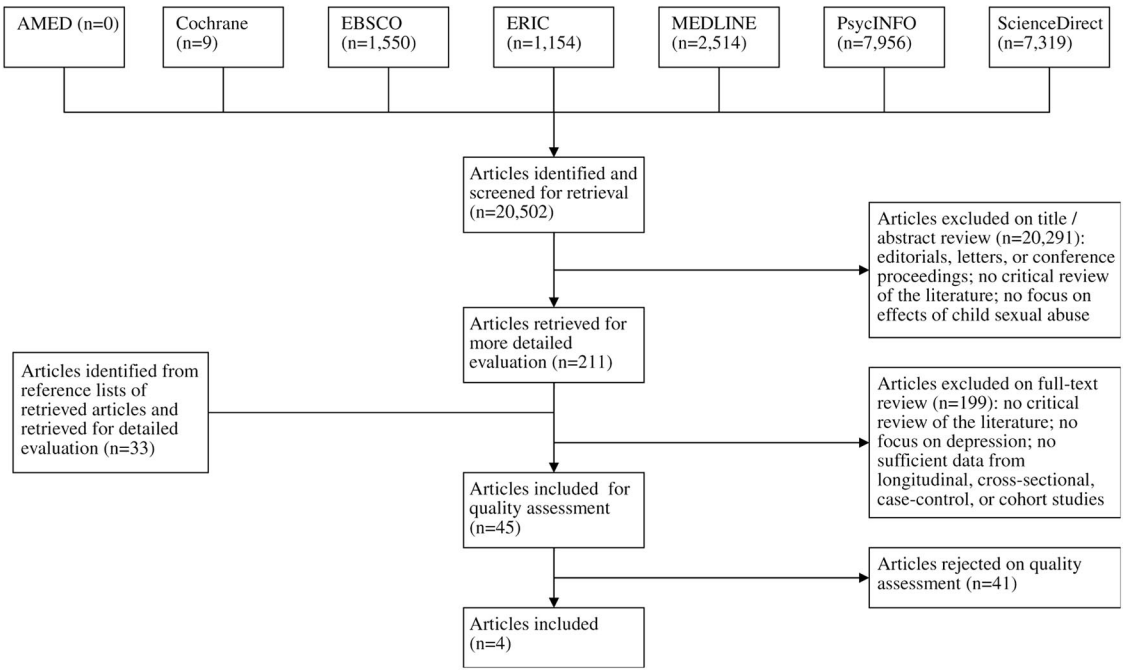
Coding of Syntheses

The methods for extracting and coding the information from the syntheses included in an overview parallel the methods used to code studies that are the elements of the syntheses themselves. Overviewers must (a) create a coding frame, (b) train coders, and (c) assess the accuracy of the coded information. Guidance for these tasks can be found elsewhere (e.g., Cooper, 2010; Lipsey, 2009; Orwin & Vevea, 2009; Wilson, 2009).

Extracting information from a research synthesis presents at least one challenge that needs a unique solution. This concerns what to do when the synthesis report is missing information. Overviewers might use a strategy developed by Valentine, Cooper, Patall, Tyson, and Robinson (2010) as part of their overview concerning the effects of afterschool programs on children's achievement and related outcomes. Their scheme focused on nine features of afterschool programs that synthesists might or might not have used to decide whether a study should be included in the synthesis. For example, the synthesis preparers might have wanted their review to be restricted to afterschool programs that served students at particular grade levels. For each such feature, coders had five response options: The reviewers *explicitly* stated that afterschool programs with that characteristic were (a) included in or (b) excluded from the synthesis; the reviewers did not explicitly state that the program characteristic was included in or excluded from the synthesis but it could be *inferred* to be (c) included or (d) excluded on the basis of other criteria; or (e) no explicit statement was made and no inference was possible. Valentine and colleagues had each of the syntheses

[6] Also, it is important to note that systematic reviews with a high degree of overlap still may disagree with one another or examine different moderator variables or outcomes. Thus, completely excluding overlapping reviews may be suboptimal.

Figure 1
Example of a Flowchart Describing the Search Process for an Overview



Note. Reprinted from "Child Sexual Abuse in the Etiology of Depression: A Systematic Review of Reviews" by R. Maniglio, 2010, *Depression and Anxiety, 27,* p. 633. Copyright 2010 by Wiley-Liss, Inc.

in their overview coded by two coders, and discrepancies were resolved in conference.[7]

Evaluating the Quality of the Evidence

Systems for judging the quality of evidence help scholars to determine the level of confidence they should place in research conclusions. Research synthesists typically apply such systems to primary studies to weigh the rigor of the studies' methodology. Sometimes this is done to eliminate studies whose methods do not permit the kinds of inferences the synthesists wish to make. Other times, synthesists will use the quality-judging system to (a) group studies on the basis of their trustworthiness and/or (b) explore methodology as a moderator of outcomes to see, for example, whether studies with more and less rigorous methods led to different outcomes. Considering the quality-judging system used in the relevant syntheses and assessing whether it was appropriate is an important step in conducting an overview.

Also, overviewers have the unique task of gauging the methodological rigor not of primary studies but of the research syntheses. As noted above, sometimes these judgments are used to decide which of overlapping syntheses should be included in the overview. Other times, overviewers use methodological rigor to exclude poorly conducted syntheses or to decide which of the included syntheses are most trustworthy. This can be especially important if there are differences in the conclusions drawn by different synthesists.

Systems for coding the methodology of primary research. Numerous systems have been proposed for judging the quality of primary research. Table 5 provides a list of several systems currently in use by behavioral scientists for coding the methodological rigor of primary research. Readers should also consult West et al. (2002) for a comprehensive examination of schemes available in health science prior to the early 2000s.

Over the past decade, systems for evaluating the methodology used in behavioral research have grown in complexity. Judgments of methodological rigor have come to be viewed as multidimensional and at least partly a function of the nature of the problem. For example, a synthesist might exclude studies because they do not employ random assignment of participants to conditions. But this single criterion ignores other aspects of design that might threaten a randomized study's legitimacy for drawing causal inferences. These include (a) the differential loss of participants from the treatment and control conditions after random assignment and (b) the contamination of treatments, to name just two. The single criterion also ignores the importance of other design features, such as the validity of the measures, whether the sampled

[7] These overviewers took the additional step of contacting the synthesis authors and asking them to verify the codes.

Table 5
Examples of Different Systems Used to Evaluate the Quality of Evidence in Primary Research

Author (Year)	Field
National Health Service (2006): Critical Appraisal Skills Program	Prevention
Higgins, Altman, & Sterne (2008)	Health care
Valentine & Cooper (2008)	Psychology, education, or other social sciences
National Institute for Health and Clinical Excellence (2009)	Public health technologies, interventions, and practice
Heck & Minner (2010)	Mathematics and science education interventions
Blueprints for Violence Prevention (2004)	Violence and drug use
National Registry of Evidence-Based Programs and Practices (2007)	Mental health and substance abuse
U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse (2008)	Education

participants represent the target population, and whether the data were analyzed correctly.

Also, including only studies using random assignment in a synthesis may not be the best strategy because evaluators of interventions might use other designs that can also lead to strong inferences about an intervention's effectiveness (Shadish, Galindo, Wong, Steiner, & Cook, 2011). These include regression discontinuity designs, interrupted time series or single-case designs, and in some cases, well-conducted quasi-experimental designs (see Cooper et al., 2012, for descriptions of these designs). These designs are likely to appear in topic areas where random assignment is not possible, perhaps those related to public policy and phenomena of rare occurrence.

As an example, Valentine and Cooper (2008) presented a multidimensional system for evaluating primary research called the Study Design and Implementation Assessment Device, or the Study DIAD. This system provides for the needs of numerous disciplines (i.e., it allows synthesists to choose different levels of abstractness for their problem space), and it underwent an extensive validation process. The Study DIAD requires users to (a) be detailed, operational, and transparent about their criteria, (b) define their criteria prior to beginning the evaluation of studies, and (c) apply the criteria consistently.[8] At the most abstract level, the Study DIAD answers four questions, one each about the construct validity, internal validity, external validity, and statistical conclusion validity of a study. Each of these four questions is subdivided into two more specific questions. At the most operational level, 32–34 questions are asked about specific aspects of research design and implementation (the exact number of questions depends on the design of the study). Finally, users of the Study DIAD apply algorithms to the answers to the operational questions that generate the answers to the more global questions.

In sum, it is important that overviewers pay careful attention to the systems that were used in the syntheses to judge the quality of primary studies that fall within the problem space. Overviewers need to state explicitly what systems for judging the quality of primary research were used in the constituent syntheses and, if syntheses were excluded on this account, what the basis was for the decision.

Systems for coding the methodology of research syntheses. The quality system applied by the overviewers themselves involves judging the methodological rigor of the research syntheses. Systems for coding the methodology of research syntheses are not as plentiful as systems for coding the methodology of primary research, although a handful of such systems have appeared, primarily in the health sciences. West et al. (2002) cataloged 11 independent efforts to develop such systems, and they rated each system on how well it covered 11 domains of synthesis methodology.

Since the early 2000s, a handful of systems for coding the methodology of research syntheses have been used by behavioral scientists conducting overviews, again mostly in the context of the health sciences. One such system comes from the Scottish Intercollegiate Guidelines Network (SIGN, 2011; http://www.sign.ac.uk/index.html). The levels and grades of evidence used in the SIGN system are presented in Table 6. SIGN uses separate systems for grading the quality of (a) the evidence, which includes an intertwined assessment of the methods of primary studies as well as the methods of synthesis, and (b) the strength of the recommendations based on that evidence. While SIGN conflates multiple judgments into a single system and leaves some judgments noticeably vague (i.e., it is difficult to define exactly what is meant by "high quality" and "low risk of bias"), the system is accompanied by a thorough process for operationalizing these judgments. Therefore, if SIGN is used by a small team, provision of details about the definition of terms in the particular problem space is essential, as is evidence of interrater agreement. For example, Newbury-Birch et al. (2009) used the SIGN system in their overview of reviews on the impact of alcohol consumption on young people.

Another system was developed by the National Institute for Health and Clinical Excellence (NICE; http://

[8] One drawback of the Study DIAD is that it currently focuses only on randomized and nonrandomized group studies. Modules are yet to be developed to cover other types of research designs, such as regression discontinuity and time series designs. However, the Study DIAD sections on construct, external, and statistical validity are applicable to all designs.

Table 6
Scottish Intercollegiate Guidelines Network (SIGN) System for Coding the Quality of Systematic Reviews

	Levels of evidence:
1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well-conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
1−	Meta-analyses, systematic reviews, or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well-conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2−	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Non-analytic studies, e.g., case reports, case series
4	Expert opinion
	Grades of recommendations:
A	At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population; *or* A body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating overall consistency of results
B	A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results; *or* Extrapolated evidence from studies rated as 1++ or 1+
C	A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; *or* Extrapolated evidence from studies rated as 2++
D	Evidence level 3 or 4; *or* Extrapolated evidence from studies rated as 2+

Note. From *SIGN 50: A Guideline Developer's Handbook* (p. 51), by Scottish Intercollegiate Guidelines Network, 2011, Edinburgh, Scotland: Author. Copyright 2011 by Scottish Intercollegiate Guidelines Network. Reprinted with permission.

www.nice.org.uk/aboutnice/), part of the National Health Service of Great Britain. This system asks five questions, one each about the clarity of the problem definition, the fit of studies to the problem, the adequacy of the literature search, the quality of the included studies, and the quality of the synthesis methods. For example, Jepson et al. (2010) conducted an overview on behavioral interventions to promote healthy behavior, including reduced tobacco, alcohol, and drug use, increased exercise and healthy eating, and reduced sexual risk-taking in young people. Jepson and colleagues' final screening device supplemented the NICE criteria by asking 12 additional questions, many of which focused on the literature searching strategy and the criteria for including and excluding syntheses.[9]

AMSTAR (for Assessment of Multiple Systematic Reviews) was developed by Shea et al. (2007). These researchers began with a pool of 37 items. The items were used by two judges to assess the quality of 151 systematic reviews. An exploratory factor analysis reduced the number of items to 29 that measured 11 different components of synthesis methodology. An international panel of clinicians, methodologists, epidemiologists, and reviewers new

[9] NICE was preceded by the National Health Service's Health Development Agency, which also developed screening devices and commissioned overviews (see, e.g., Ellis et al., 2003, on HIV prevention and Naidoo, Warm, Quigley, and Taylor, 2004, on smoking cessation).

to the field then critically examined the items and decided which should be on the final instrument. The AMSTAR score for any given systematic review involves awarding one point for each question answered "yes," so scores can range from 0 to 11. Shea et al. (2009) reported substantial interrater agreement, construct validity, and ease of use when applying AMSTAR to a set of randomly chosen systematic reviews.

Finally, the only system we could find that was developed with behavioral science research syntheses in mind was originally presented by Cooper (2007) and then slightly revised (Cooper, 2010). Table 7 presents Cooper's (2010) checklist. Similarly to the AMSTAR system, this checklist presents questions that are answered "yes" or "no," with "yes" suggesting that more rigorous methods were used in the conduct of the synthesis. While there are no inconsistencies between the two systems, there are differences in emphasis, as again might be expected given differences between the foci of medical and behavioral science research. For example, Cooper's system asks four questions about the definition of the problem space (clarity of concepts, correspondence of concepts and operations, specification of appropriate research designs, and placement of the problem in a broader context), whereas AMSTAR asks only one.[10]

The four systems for judging the quality of research syntheses described above differ in their details and emphasis. However, it is heartening that from a bird's eye view there seems to be much agreement about the categories of importance. Which of the systems would be most appropriate in a particular overview will depend on the problem space under consideration; the problem should suggest the relative importance among the different steps in the synthesis process. Also, the resources available to the overviewer may play a role in this decision. Regardless of which system is used, it is always good practice to have more than one person apply the system to each relevant synthesis and to have discrepancies resolved in conference or by a third rater (this is good practice for all coding tasks). If resources are limited, a check on individual raters' reliability should be undertaken.

It is also the case that only the SIGN system is a true grading system; the others are checklists.[11] While it is known that the different dimensions of quality are largely independent (Shea et al., 2007), at present the only advice developers have for users is to sum across the dimensions (although the Cooper, 2010, checklist can easily result in subscores). This is a practice that has been criticized in the context of judging the quality of primary research (Valentine & Cooper, 2008). It is no less true here. First, single scores derived from different systems may diverge because the systems emphasize different aspects of synthesis methodology rather than because the systems are unreliable. Conversely, syntheses with different strengths and weaknesses can end up with the same score. For example, using Cooper's (2010) system, a review that was based on a careful and thorough literature search but ignored the quality of the primary studies would receive the same score (by summing the "yes" responses) as a review based on a haphazard search but including a careful consideration of primary research methodology (all else being equal). While the systems used to judge the quality of primary research have evolved beyond the "single score" approach, the systems for research syntheses have not. This is an important next step for the developers of these systems.

Summing the Results of Syntheses

In the past quarter century, the methods for integrating primary research results have changed dramatically as the techniques of meta-analysis have developed rapidly and grown in acceptance (see Cooper, Hedges, & Valentine, 2009). Needless to say, methods for integrating the results of research syntheses are still in their infancy. We were able to identify three approaches that have garnered some attention in the literature.

Looking for sources of discordance in outcomes. One approach to integrating the results of systematic reviews begins by examining whether different reviews resulted in different conclusions about the literature. To carry out this analysis, overviewers first ask whether the constituent reviews come to the same or very similar conclusions. If so, then the results of any one review might stand for them all. If not, then the overviewers must examine the reviews in more depth and determine which come closest to giving a trustworthy answer to the question at hand.

Jadad, Cook, and Browman (1997) presented a decision tool for choosing between discordant systematic reviews. First, they pointed out that systematic reviews can differ (a) because their *results* are different or (b) because the reviewers' *interpretations of the results* are different. The former case is of greater concern than the latter; if interpretations differ, the overviewers can provide an interpretation of their own, perhaps parsing through the discordant interpretations using substantive, logical, or other criteria.

If the results of the syntheses are different, the overviewer must determine what conceptual or methodological differences are the sources of the discrepancies. The obvious first place to look would be in the research covered by the reviews. We discussed earlier some methods for examining overlap. In the case where the goal is to uncover why separate syntheses reached different conclusions, the overviewers first would focus on a *lack* of overlap. If a lack of overlap is discovered, the overviewers would examine the descriptions of the covered research and determine whether the discordant reviews defined the problem space in a similar or different manner. If different, the review that was more consistent with

[10] The Cochrane Collaboration also has a system, called Grades of Recommendation, Assessment, Development, and Evaluation (GRADE; Schünemann et al., 2008), that is applied to a body of evidence once a systematic review is complete. This system relates not to any individual primary study or to the methods of the research synthesis but to the quality of the evidence as a whole. It has been widely adopted by organizations interested in improving health care and medicine.

[11] Lepore and Coyne (2006) used a grading system they developed for categorizing reviews of the literature on psychological interventions to relieve distress in cancer patients. They categorized reviews by whether they included only quality randomized trials, only randomized trials, or both randomized trials and quasi-experiments. This criterion was crossed with whether only published research (a negative feature) or both published and unpublished research were included.

Table 7
A Checklist of Questions Concerning the Validity of Research Synthesis Conclusions

Step 1: Formulating the problem
1. Are the variables of interest given clear conceptual definitions?
2. Do the operations that empirically define each variable of interest correspond to the variable's conceptual definition?
3. Is the problem stated so that the research designs and evidence needed to address it can be specified clearly?
4. Is the problem placed in a meaningful theoretical, historical, and/or practical context?

Step 2: Searching the literature
5. Were proper and exhaustive terms used in searches and queries of reference databases and research registries?
6. Were complementary searching strategies used to find relevant studies?

Step 3: Gathering information from studies
7. Were procedures employed to assure the unbiased and reliable (a) application of criteria to determine the substantive relevance of studies and (b) retrieval of information from study reports?

Step 4: Evaluating the quality of studies
8. If studies were excluded from the synthesis because of design and implementation considerations, were these considerations (a) explicitly and operationally defined and (b) consistently applied to all studies?
9. Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?

Step 5: Analyzing and integrating the outcomes of studies
10. Was an appropriate method used to combine and compare results across studies?
11. If a meta-analysis was performed, was an appropriate effect size metric used?
12. If a meta-analysis was performed (a) were average effect sizes and confidence intervals reported and (b) was an appropriate model used to estimate the independent effects and the error in effect sizes?
13. If a meta-analysis was performed, was the homogeneity of effect sizes tested?
14. Were (a) study design and implementation features (as suggested by Question 8 above) along with (b) other critical features of studies, including historical, theoretical, and practical variables (as suggested by Question 4 above) tested as potential moderators of study outcomes?

Step 6: Interpreting the evidence
15. Were analyses carried out that tested whether results were sensitive to statistical assumptions, and if so, were these analyses used to help interpret the evidence?
16. Did the research synthesists (a) discuss the extent of missing data in the evidence base and (b) examine its potential impact on the synthesis findings?
17. Did the research synthesists discuss the generality and limitations of the synthesis findings?
18. Did the synthesists make the appropriate distinction between study-generated and review-generated evidence when interpreting the synthesis results?
19. If a meta-analysis was performed, did the synthesists (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?

Step 7: Presenting the results
20. Were the procedures and results of the research synthesis clearly and completely documented?

Note. From *Research Synthesis and Meta-Analysis: A Step-by-Step Approach* (pp. 18–19), by H. Cooper, 2010, Thousand Oaks, CA: Sage. Copyright 2010 by Sage Publications, Inc. Reprinted with permission.

the overviewers' definition would be given greater weight in their interpretation.

If the covered literatures overlap but one systematic review contains all or most of the research contained in the others and more, then again the choice is clear. The more comprehensive review will lead to more trustworthy conclusions, all else being equal. This is one criterion that Lipsey and Wilson (1993) used to cull through overlapping reviews in their overview of the effectiveness of psychological, behavioral, and educational interventions.

If the research in the reviews is roughly similar, then the overviewer would examine the methods of data integration used in the syntheses—the rigor of data extraction techniques, the validity of assumptions for meta-analysis, other potential sources of bias—and choose the review with the strongest methods.[12]

As an example of this approach, Swanson et al. (1993) examined narrative reviews of research on the effects of stimulant drugs on children with attention deficit disorder. These authors listed four areas of agreement between the reviews (that immediate effects were dramatic, that expectancy and placebo effects contributed to the finding, that

[12] It is also possible that reviews may agree about an effect size but disagree in their conclusions if one review attained statistical significance and the other did not. This may happen if one review covered more studies (and hence had greater precision) or if one used a fixed effects model and the other a random effect model. In these cases, the discordance is easily explained.

pretreatment differences among children did not predict effects, and that documented long-term effects were negligible), but they also noted that the syntheses differed in their overall interpretation of the literature. They then traced the disagreements to differences in literature coverage, in outcome variables, and in the emphasis on different subpopulations.

Also of interest to the overviewer will be whether the discrepancy in the results of reviews involves (a) different conclusions about the effectiveness of an intervention—whether it was or was not effective—or (b) disagreements about the magnitude of an effect. In the latter case, Jadad et al. (1997) pointed out that differences in the magnitudes of effects may or may not be important: "A decision-maker may consider differences between 2 reviews to be unimportant if the estimated treatment effects are of different magnitude but in the same direction, and are statistically significant and clinically important" (p. 1412). This would be a matter of interpretation for the overviewer to undertake. For example, returning to Swanson et al. (1993), these authors examined three meta-analyses and found a relatively small range of estimated effect sizes for the impact of stimulants on behavior and attention (ranging from .75 to .90) but a larger range for the impact on IQ and achievement (from .19 to .47). They also noted a discrepancy in the estimates of the placebo effect in the meta-analyses (.32 vs. .07) that they felt deserved reconciliation. With regard to placebo effects, the overviewers chose to accept the larger estimate because (a) it was a principal focus of that review and (b) it was consistent with placebo effects found in other research domains. Seeking guidance in related literatures is yet another potential criterion for resolving discordant results.

Second-order meta-analysis. One problem with using the analyses of discordance suggested above is that, like the results of a group of primary studies, the variation in the results of meta-analyses might best be explained by sampling error alone. What appear to be discordant results may simply be the tails of a single underlying distribution of estimated effects. A second strategy for integrating the results of research syntheses is to perform a second-order meta-analysis (Hunter & Schmidt, 2004). Here, the average effects found in meta-analyses conducted within the same problem space are themselves combined.

In conducting a second-order meta-analysis, overviewers are again confronted with the issue of what to do about meta-analyses with overlapping evidence. Several approaches have been taken. For example, Swanson et al. (1993) averaged the effect of stimulant drugs on outcomes for children with attention deficit disorder. These authors paid no attention to the independence, or lack thereof, of the three meta-analytic estimates. On the other hand, Wilson and Lipsey (2001; using the meta-analyses collected for Lipsey & Wilson, 1993) studied the impact of methods on the outcomes of treatment effectiveness research. To address overlap, they identified meta-analyses with 25% or more of their research in common and eliminated the one with the fewer studies in each comparison, except when multiple smaller meta-analyses (with little overlap) would include more studies if the largest meta-analysis was eliminated. Wilson and Lipsey reported that this strategy resulted in few meta-analyses with any overlap and that when overlap did occur, it was generally less than 10%. And finally, Barrick, Mount, and Judge (2001) conducted two separate analyses, one with and one without overlapping meta-analyses (15 in total), in their second-order meta-analysis of personality as a predictor of job performance.

Finally, second-order meta-analysts must decide how to calculate the second-order average effect and estimate its sampling error. Wilson and Lipsey (2001) estimated the central tendency using bootstrap means weighted by the harmonic means of the number of effect sizes contributing to the meta-analysis. They chose this approach "because it provided a method for estimating confidence intervals around the mean for each index without requiring assumptions about its underlying distributional properties" (Wilson & Lipsey, 2001, p. 416). Rothstein, Schmidt, Erwin, Owens, and Sparks (1990), in a study of biomarkers as predictors of job performance, presented another method for estimating average effect sizes and confidence intervals in second-order meta-analysis that was based on parametric distribution assumptions (see also Hunter & Schmidt, 2004).

Conducting analyses looking at moderators and mediators of effects in a second-order meta-analysis can be especially challenging. This is because the moderating and mediating variables must exist at the level of the research syntheses that are the constituent elements, not at the level of the individual studies. Despite the unique challenges second-order meta-analysts face when examining moderators and mediators, such an examination can be done. For example, in an overview of meta-analyses on the impact of technology on learning, Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) extracted the effect sizes and their standard errors from 25 meta-analyses. They were able to categorize the meta-analyses by their quality, by the primary use of the instructional technology (direct instruction or instructional support), and by the grade level of students (K–12, postsecondary). They found that higher quality meta-analyses, meta-analyses of support instruction, and meta-analyses of studies with younger students reported larger effect sizes.

First-order meta-analysis. Finally, overviewers can perform a new research synthesis or meta-analysis by identifying the individual, independent primary studies included in all of the covered research syntheses and reintegrating them. This approach might not be called an overview, since the overviewers might simply be using the research syntheses as a source of references for their new effort. The line becomes less clear if the overviewers use the effect sizes from primary research reported in the meta-analyses. This approach would clearly be cost-effective in that the overviewers would not have to recode the effect sizes from the primary studies. However, it requires that the reports of the meta-analyses present tables with the constituent effect sizes and that these be calculated in a similar fashion from one meta-analysis to the next. Also, the types of moderator analyses the overviewers can perform would be constrained by what other characteristics of

primary studies might have been tabled along with the effect sizes and whether these are repeated across meta-analyses. Of course, in addition to conducting their own moderator analyses, the overviewers can report the results of the moderator analyses conducted in the separate meta-analyses when they discuss the moderators of their main effect.

Cataloging the Moderators, Mediators, and Outcomes Addressed in Syntheses

One of the benefits of conducting systematic reviews of behavioral research is the expanded variety of people, places, and outcomes the synthesist discovers, relative to any single study. This variety can provide important insights into the generalizability and limitations of an intervention's effectiveness. Overviewers can cast a yet broader net, capturing even more variation. These variations can relate to (a) how and to whom the treatment was delivered, (b) what variables moderated and mediated the effectiveness of an intervention, and (c) how the outcome was defined and measured. For example, Nation et al. (2003) examined 35 systematic reviews in a search for principles of effective programs meant to prevent deleterious behaviors by young people. On the basis of the percentage of reviews endorsing a principle, they identified nine program characteristics related to effectiveness.

It is especially critical for overviewers to carefully catalog and distinguish between (a) the moderating, mediating, and outcome variables that have been tested in the empirical literature and the systematic reviews[13] and (b) those that have been mentioned often in the literature but not tested empirically. The latter list may be considerably longer than the former. These catalogs not only help readers understand what the evidence says about mechanisms that have been previously tested, but they also alert researchers to where evidence is needed. They are also important for helping readers understand the distinction between *no evidence about effects* and *evidence of no effect.*

Presentation of Findings

Throughout this article we have presented tables and figures that illustrate effective approaches for presenting the methods and results of overviews. Here, we highlight one innovative technique proposed by Ioannidis (2009).

It is often the case that a single intervention study will contain only one comparison between a new treatment and a no-treatment control, a placebo treatment, or treatment as usual. However, in conducting research syntheses, researchers may find that the treatment of interest has been compared with different types of controls but in different studies. Overviewers may then face even more complexity in that they might be interested in numerous approaches to treatment and how these treatments compare with one another. Ioannidis (2009) presented a graphical technique for displaying the different treatment comparisons. This is presented in Figure 2. Each treatment that has been tested is given a node in the figure and is linked with a line to each other treatment (or control) with which it has been compared.

Note that Ioannidis (2009) included only a placebo comparison group. In many behavioral science applications, there would likely be other types of controls that may or may not have links connecting them. For example, if Swanson et al. (1993), in their overview of stimulant drug effects on attention deficit disorder, had this technique available, they could have linked placebo and other types of controls because two of the meta-analyses they covered included data on this comparison. Ioannidis (2009) pointed out that the links between nodes can be accompanied by the meta-analytic estimates of the differences in treatment effectiveness.

Standards for Reporting Overviews

The *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins & Green, 2008) contains guidelines for the content and formatting of overviews. Not surprisingly, these guidelines suggest a format that differs little from that used for reporting systematic reviews.[14] Overviewers in the behavioral sciences might therefore look at the Meta-Analysis Reporting Standards (MARS) adopted by the American Psychological Association (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The content of the MARS was developed with input from many sources, including several related efforts in the health sciences. All of the content called for in the MARS (except perhaps the reporting of the conduct and results of the statistical analyses, which would only be relevant if a second-order meta-analysis was conducted) would be no different for an overview.

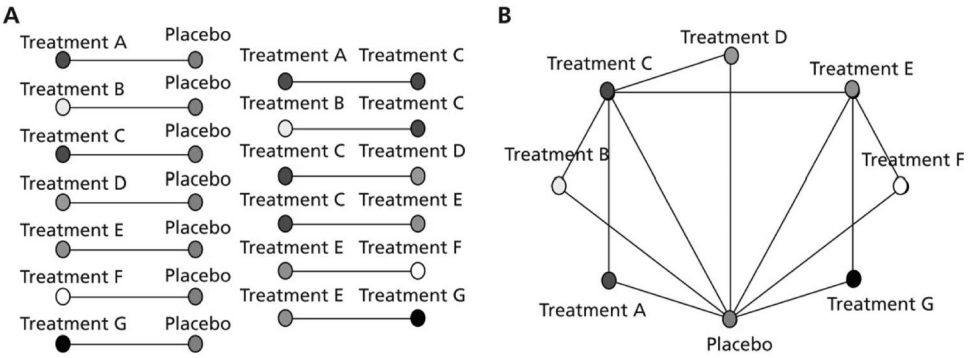
Limitations and Needed Future Developments

Overviews have great potential for providing "big picture" summaries of empirical research. They can be of great service to readers wanting to familiarize themselves with an area or looking for areas in which new research is needed. If written in nontechnical language, they can provide policymakers and practitioners with summaries of what is known about interventions and even the relative effectiveness of related interventions. But overviews are not without their limitations. Earlier, we noted several of these. One problem that reoccurred in our discussion concerns the overlap in the content of the constituent research syntheses and how the overviewers should choose among or weight syntheses so as to maximize the utility of the information they include. We cataloged several approaches to addressing overlap, but none of the currently available approaches is com-

[13] And it is important for overviewers to distinguish between *study-generated evidence* and *synthesis-generated evidence* (Cooper, 2010). Study-generated evidence, if based on strong experimental designs, permits the synthesist to make claims about causal relationships. Synthesis-generated evidence will always be about associations.

[14] Smith, Devane, Begley, and Clarke (2011) presented a brief discussion of criteria for evaluating the methods used in conducting overviews, focusing on those intended for inclusion in the Cochrane Collaboration database.

Figure 2
A Graphic Display of Treatment Comparisons



Note. Schematic representation of (A) an umbrella review encompassing 13 comparisons involving 8 treatment options (7 active treatments and a placebo) and (B) a network with the same data. Each treatment is shown by a node, and comparisons between treatments are shown with links between the nodes. Each comparison may have data from several studies that may be combined in a traditional meta-analysis. Reprinted from "Integration of Evidence From Multiple Meta-Analyses: A Primer on Umbrella Reviews, Treatment Networks and Multiple Treatments Meta-Analyses," by J. P. A.Ioannidis, 2009, *Canadian Medical Association Journal, 181,* p. 489. Copyright 2009 by the Canadian Medical Association.

pletely satisfactory, and choosing among them necessarily involves consideration of the nature of the problem under study.

Another related problem concerns the methods used to synthesize the syntheses. We described three approaches, one that was narrative and two that involved meta-analysis. The narrative approach is imprecise, can lack transparency, and therefore can be open to bias. The meta-analytic solutions are more precise but will be hard to apply until the standards of reporting systematic reviews are more widely used. Synthesists can use different metrics and different formulas to calculate effect sizes, and given the complexity of this task, it would not be surprising to find errors in these calculations. Finally, the search for moderating influences on meta-analytic results is limited. Most overviewers will need to rely on cataloging moderators tested in the individual reviews, and this is clearly an area where methodological work is needed.

Another problem overviewers will have involves the time lag between when their work is ready for public distribution and when the studies that formed the basis of the systematic reviews they cover were conducted. For example, Tamim et al.(2011) used a cutoff date of 1985 for inclusion of meta-analyses in an overview of computer-based instruction. But some of the covered syntheses included studies reported as far back as 1967, over 40 years before the overview appeared. In some problem areas the age of the research may not be a concern, but in others overviewers would need to consider whether the covered research is still relevant to current practice. Earlier, we noted that the age of the synthesis might be a criterion for inclusion in an overview; attention also needs to be paid to the age of the research within the syntheses.

Related to the problem of coverage of obsolete research is the problem of coverage of the most recent research. De Solla Price (1965) defined a literature review as a document meant to "replace those papers that have been lost from sight behind the research front" (p. 513). If an overview is meant to replace the syntheses it covers, then it is even further removed from the research front. Will overviewers of intervention evaluations ignore those that are so new that they have not generated enough research to warrant a synthesis? If so, they will be mute regarding the most recent advances on the problem, perhaps missing those innovations that have appeared in the past decade or so. To address this problem, we suggest that overviewers would be well served to include a section in their report in which they at least reference and discuss what these recent advances are and how they relate to the work covered in the constituent research syntheses (see also Thomson et al., 2010).

Conclusion

In conclusion, overviews of reviews have arisen organically in the psychological and allied research literatures. The impetus for their appearance has been (a) the growing number of research syntheses that summarize and integrate similar or closely related research questions and (b) the desire of scholars (and the audiences they write for) to tackle ever broader research questions. Both of these developments can be taken as indices of a maturing science.

While most efforts to date have focused on evaluating psychological and health interventions, overviews have also been used to help make sense of other literatures. There is little doubt that the future will see an increase in the number of overviews and that they will appear in more varied problem areas. To date, scholars wishing to conduct rigorous and systematic overviews have relied heavily on methodologies used in research syntheses. Generally speaking, their methodological choices have been appro-

priate. Sometimes, however, at some decision points the variety of approaches used suggests a need for clearer guidelines on what methods should be used under what circumstances. At other times, the jerry-rigging of methods suggests that there are unique aspects to conducting sound overviews that require new and unique solutions. It is our hope that this survey of issues and methods for this new form of scholarship (a) will assist the producers of overviews in knowing what their options are and which are the best practice, (b) will help consumers gauge the soundness of the overviews they encounter, and (c) will alert methodologists to the steps in the process of conducting an overview that need their attention.

REFERENCES

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839–851. doi:10.1037/0003-066X.63.9.839

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1-2), 9–30. doi:10.1111/1468-2389.00160

Becker, L. A., & Oxman, A. D. (2008). Overviews of reviews. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 607–631). Chichester, West Sussex, England: Wiley. doi:10.1002/9780470712184.ch22

Blueprints for Violence Prevention. (2004). *Blueprints model programs selection criteria.* Retrieved from http://www.colorado.edu/cspv/blueprints/criteria.html

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26,* 17–31. doi:10.1016/j.cpr.2005.07.003

Consolidated Standards of Reporting Trials. (2007). *CONSORT: Strength in science, sound ethics.* Retrieved from http://www.consort-statement.org/

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1,* 104–126. doi:10.1007/BF03177550

Cooper, H. (1989). *Homework.* New York, NY: Longman. doi:10.1037/11578-000

Cooper, H. (2007). *Evaluating and interpreting research syntheses in adult learning and literacy.* Cambridge, MA: National Center for the Study of Adult learning and Literacy.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cooper, H., Camic, P., Long, D., Panter, A., Rindskopf, D., & Sher, K. J. (Eds.). (2012). *APA handbook of research methods in psychology: Vol. 3. Data analysis and research publication.* Washington, DC: American Psychological Association.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149,* 510–515. doi:10.1126/science.149.3683.510

Ellis, S., Barnett-Page, E. Morgan, A., Taylor, L., Walters, R., & Goodrich, J. (2003). *HIV prevention: A review of reviews assessing the effectiveness of interventions to reduce the risk of sexual transmission*. London, England: National Health Service, Health Development Agency.

Green, J., Howe, F., Waters, E., Maher, E., & Oberklaid, F. (2005). Promoting the social and emotional health of primary school-aged children: Reviewing the evidence base for school-based interventions. *International Journal of Mental Health Promotion, 7,* 30–36.

Heck, D. J., & Minner, D. D. (2010). *Technical report: Standards of evidence for empirical research, Math and Science Partnership Knowledge Management and Dissemination.* Retrieved from Math and Science Partnership Knowledge Management and Dissemination website: http://www.mspkmd.net/papers/heck_minner_oct2010.pdf

Higgins, J. P. T., Altman, D. G., & Sterne, J. A. C. (2008). Assessing the risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 187–242). Chichester, West Sussex, England: Wiley.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions.* Chichester, West Sussex, England: Wiley. doi:10.1002/9780470712184

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.

Ioannidis, J. P. A. (2009). Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *Canadian Medical Association Journal, 181,* 488–493. doi:10.1503/cmaj.081086

Jadad, A. R., Cook, D. J., & Browman, D. P. (1997). A guide to interpreting discordant systematic reviews. *Canadian Medical Association Journal, 156,* 1411–1416.

Jepson, R. G., Harris, F. M., Platt, S., & Tannahill, C. (2010). The effectiveness of interventions to change six health behaviours: A review of reviews. *BMC Public Health, 10*(538). Retrieved from http://www.biomedcentral.com/1471–2458/10/538. doi:10.1186/1471-2458-10-538

Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta-analysis: A new method for evaluating therapy outcomes. *Behaviour Research and Therapy, 17,* 397–399. doi:10.1016/0005-7967(79)90011-1

Lepore, S. J., & Coyne, J. C. (2006). Psychological interventions for distress in cancer patients: A review of reviews. *Annals of Behavioral Medicine, 32,* 85–92. doi:10.1207/s15324796abm3202_2

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 147–158). New York, NY: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209. doi:10.1037/0003-066X.48.12.1181

Maniglio, R. (2010). Child sexual abuse in the etiology of depression: A systematic review of reviews. *Depression and Anxiety, 27,* 631–642. doi:10.1002/da.20687

Mikton, C., & Butchart, A. (2009). Child maltreatment prevention: A systematic review of reviews. *Bulletin of the World Health Organization, 87,* 353–361. doi:10.2471/BLT.08.057075

Montori, V. M., Wilczynski, N. L., Morgan, D., & Haynes, R. B. for the Hedges Team. (2005). Optimal search strategies for retrieving systematic reviews from Medline: Analytical survey. *British Medical Journal, 330*(68). doi:10.1136/bmj.38336.804167.47

Naidoo, B., with Warm, D., Quigley, R., & Taylor, L. (2004). *Smoking and public health: A review of reviews of interventions to increase smoking cessation, reduce smoking initiation and prevent further uptake of smoking.* Retrieved from http://www.nice.org.uk/nicemedia/documents/smoking_evidence_briefing.pdf

Nation, M. Crusto, C., Wandersman, A., Kumpfer, K. L., Seybolt, S., Morrissey-Kane, E., & Davino, K. (2003). What works in prevention: Principles of effective prevention programs. *American Psychologist, 58,* 449–456. doi:10.1037/0003-066X.58.6-7.449

National Health Service. (2006). *Critical Appraisal Skills Programme (CASP): Making sense of evidence: 10 questions to help you make sense of randomised controlled trials.* Retrieved from Solutions for Public Health website: http://www.sph.nhs.uk/sph-files/casp-appraisal-tools/rct%20appraisal%20tool.pdf

National Institute for Health and Clinical Excellence. (2009). *Methods for the development of NICE public health guidance* (2nd ed.). London, England: Author. Retrieved from http://www.nice.org.uk/media/2FB/53/PHMethodsManual110509.pdf

National Registry of Evidence-Based Programs and Practices. (2007). Quality of research. Retrieved from http://www.nrepp.samhsa.gov/ReviewQOR.aspx#ROM

Newbury-Birch, D., Gilvarry, E., McArdle, P., Ramesh, V., Stewart, S., Walker, J., . . . Kaner, E. (2009). *Impact of alcohol consumption on young people: A review of reviews.* London, England: Department for Children, Schools, and Families.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177–205). New York, NY: Russell Sage Foundation.

Peters, L. W. H., Kok, G., Ten Dam, G. T. M., Buijs, G. J., & Paulussen, T. G. W. M. (2009). Effective elements of school health promotion across behavioral domains: A systematic review of reviews. *BMC Public Health, 9*(182). doi:10.1186/1471-2458-9-182

Peterson, P. (2001). On the use of college students in social science research: Insights from a second order meta-analysis. *Journal of Consumer Research, 28,* 450–461. doi:10.1086/323732

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184. doi:10.1037/0021-9010.75.2.175

Sampson, M., McGowan, J., Cogo, E., Grimshaw, J., Moher, D, & Lefebvre, C. (2009). An evidence-based practice guideline for the peer review of electronic search strategies. *Journal of Clinical Epidemiology, 62,* 944–952. doi:10.1016/j.jclinepi.2008.10.012

Schünemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Bossuyt, P., Chang, S., . . . Guyatt, G. H. (2008). GRADE: Assessing the quality of evidence for diagnostic recommendations. *Evidence-Based Medicine for Primary Care and Internal Medicine, 13,* 162–163. doi:10.1136/ebm.13.6.162-a

Scottish Intercollegiate Guidelines Network. (2011). *SIGN 50: A guideline developer's handbook.* Edinburgh, Scotland: Author. Retrieved from http://www.sign.ac.uk/pdf/sign50.pdf

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16,* 179–191. doi:10.1037/a0023345

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., . . . Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(10). doi:10.1186/1471-2288-7-10

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, G., . . . Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62,* 1013–1020. doi:10.1016/j.jclinepi.2008.10.009

Smith, V., Devane, D., Begley, C. M., & Clarke, M. (2011). Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology, 11*(15). doi:10.1186/1471-2288-11-15

Swanson, J. M., McBurnett, K., Wigal, T., Pfiffner, L. J., Lerner, M. A., Williams, L., . . . Fisher, T. D. (1993). Effect of stimulant medication on children with attention deficit disorder: A "review of reviews." *Exceptional Children,* 60, 154–162.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81,* 4–28. doi:10.3102/0034654310393361

Thomson, D., Russell, K., Becker, L., Klassen, T., & Hartling, L. (2010). The evolution of a new publication type: Steps and challenges of producing overviews of reviews. *Research Synthesis Methods, 1,* 198–211. doi:10.1002/jrsm.30

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2008). *What Works Clearinghouse procedures and standards handbook* (version 2). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13,* 130–149. doi:10.1037/1082-989X.13.2.130

Valentine, J. C., Cooper, H., Patall, E. A., Tyson, D., & Robinson, J. (2010). A method for evaluating research syntheses: The quality, conclusions, and consensus of 12 syntheses of the effects of after-school programs. *Research Synthesis Methodology, 1,* 20–38. doi:10.1002/jrsm.3

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., & Lux, L. (2002). *Systems to rate the strength of scientific evidence* (AHRQ Publication No. 02-E016). Washington, DC: Agency for Healthcare Research and Quality.

Wigal, T., Swanson, J. M., Regino, R., Lerner, M. A., Soliman, I., Steinhoff, K., . . . Wigal, S. B. (1999). Stimulant medication for the treatment of ADHD: Efficacy and limitations. *Mental Retardation and Developmental Disabilities Research Reviews, 5,* 215–224. doi:10.1002/(SICI)1098-2779(1999)5:3<215::AID-MRDD8>3.0.CO;2-K

Wilson, D. B. (2009). Systematic coding for research synthesis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159–176). New York, NY: Russell Sage Foundation.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6,* 413–429. doi:10.1037/1082-989X.6.4.413

Wong, S. S.-L., Wilczynski, N. L., & Haynes, R. B. (2006). Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. *Journal of the Medical Library Association, 94,* 41–47.