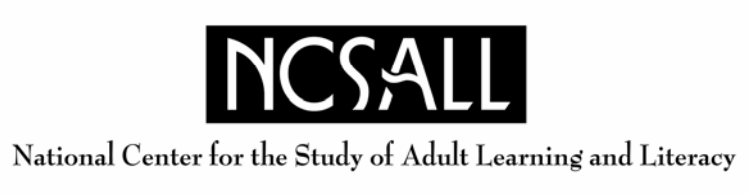
# Evaluating and Interpreting Research Syntheses in Adult Learning and Literacy

Harris Cooper
*Duke University*

NCSALL Occasional Paper
January 2007




Harvard Graduate School of Education
101 Nichols House, Appian Way
Cambridge, MA 02138

National Institute for Literacy

National Institute for Literacy
1775 I Street NW, Suite 730
Washington, DC 20006

*NCSALL Occasional Papers* are funded by the Educational Research and Development Centers program, Award Number R309B960002, as administered by the Institute of Education Sciences (formerly Office of Educational Research and Improvement), U.S. Department of Education. The contents of this paper were developed using funds transferred from the National Institute for Literacy to the Institute of Education Sciences. The views expressed herein do not necessarily represent the positions or policies of the National Institute for Literacy, the Institute of Education Sciences, or the U.S. Department of Education. No official endorsement by the National Institute for Literacy, the Institute of Education Sciences, or the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred.

# CONTENTS

# INTRODUCTION

*A common method of integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies—those remaining frequently being one's own work or that of one's students or friends—and then advance the one or two "acceptable" studies as the truth of the matter.*

Gene Glass (1976, p.4)

As the quote from Gene Glass suggests, research synthesis used to be a subjective process with little protection against bias favoring the perspective of the preparer. However, since Glass's characterization, methods for the retrieval, integration, and interpretation of research literatures have undergone enormous change. Today, the approach suggested by Glass is widely viewed as unacceptable and certainly inappropriate for informing decisions regarding the adoption of social programs and policies. Instead, research synthesis now has its own methodological techniques and decision rules, all meant to help synthesists produce an unbiased estimate of what the cumulative evidence says.

This paper will introduce the methods of research synthesis and meta-analysis to researchers and consumers of research in the field of adult learning and literacy. To begin, the first section of the paper defines key terms and offers a brief history of how the methodologies developed. The second section provides a conceptualization of research synthesis that views it no differently from other research endeavors in the social sciences. Then, the tasks of research synthesis are presented in more detail within the context of a hypothetical example drawn from the literature on adult learning and literacy.

# RESEARCH SYNTHESIS: DEFINITIONS AND HISTORY

There are many terms, often used interchangeably, to label the product that emerges from the activities outlined in this paper (Cooper, 1988). The broadest term is *literature review*. Literature reviews attempt to integrate what other scholars have written and said, to criticize previous work, to build bridges between related topic areas, and/or to identify the central issues that motivate, or should motivate, a field of study. A literature review can summarize and integrate research outcomes, research methods, or theories.

A specific type of literature review has been alternately called a *research synthesis*, *systematic review*, or *research review*. In their general use, these terms are synonymous. Research syntheses primarily focus on empirical studies and seek to summarize past research by drawing overall conclusions from multiple, separate investigations that address related or identical topics. The research synthesist hopes to present the cumulate state of evidence concerning the relation(s) of interest and to highlight important issues that research has left unresolved. The research outcomes that give impetus to a research synthesis can have as their principal focus applied problems or theoretically derived hypotheses.

The term *meta-analysis* often is used as a synonym for research synthesis. In this paper, it will be used in its more precise and original meaning—to describe the quantitative procedures that a research synthesist may use to statistically combine the results of studies. Not all literature reviews or even research syntheses are appropriate for meta-analysis. A discussion of when meta-analysis is and is not appropriate in research synthesis will be presented below when the techniques of quantitative synthesis are described.

## A Brief History of Research Synthesis and Meta-Analysis

In 1904, Karl Pearson conducted what is believed to be the first meta-analysis. Having been asked to review the evidence on a vaccine against typhoid, Pearson gathered data from eleven studies and for each study he calculated a recently developed statistic called the correlation coefficient. Based on the average of the correlations, Pearson concluded that other vaccines were more effective (Pearson, 1904).

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

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

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

These early procedures for statistically combining results of independent studies largely went unused. However, beginning in the 1960s, social science research experienced a period of rapid growth and in the mid-1970s a dramatic change took place. In social psychology, Rosenthal and Rubin (1978) undertook a review of research on the effects of interpersonal expectations on behavior. They found 345 studies that pertained to their research problem. In clinical psychology, Smith and Glass (1977) gathered 833 assessments of the effectiveness of psychotherapy. In education, Glass and Smith (1979) conducted a review of the relationship between class size and academic achievement and found 725 estimated correlations. In organizational psychology, Hunter, Schmidt, and Hunter (1979) uncovered 866 comparisons of the differential validity of employment tests for black and white workers.

These researchers concluded that the traditional research synthesis simply would not suffice. Largely independently, the three research teams rediscovered and reinvented Pearson's and Fisher's solutions to their problem. In discussing his solution, Glass coined the term meta-analysis to describe "the statistical analysis of a large collection of analysis results from individual studies for purposes of integrating the findings" (Glass, 1976, p. 3). Shortly thereafter, other proponents of meta-analysis demonstrated that narrative review procedures that used impressionistic summaries of evidence led to inaccurate or imprecise characterizations of the literature, even when the size of the literature was relatively small (e.g., Cooper & Rosenthal, 1980).

The first half of the 1980s witnessed the appearance of several books primarily devoted to meta-analytic methods. Among them, Light and Pillemer (1984) focused on the use of research synthesis to help decision-making in the social policy domain. Their approach placed special emphasis on the importance of meshing both numbers and narrative for the effective interpretation and communication of synthesis results. With the publication of *Statistical Methods for Meta-Analysis*, Hedges and Olkin (1985) elevated meta-analysis to an independent specialty within the field of statistics. Their book established the legitimacy of many procedures for the quantitative synthesis of research results by presenting rigorous statistical proofs.

Concurrent with the developments in meta-analysis, attempts were being made to frame research synthesis in the terms of a broader scientific process. For example, in 1971, Feldman argued that "systematically reviewing and integrating ... the literature of a field may be considered a type of research in its own right—one using a characteristic set of research techniques and methods" (Feldman, 1971, p. 86). In the same year, Light and Smith (1971) presented a "cluster approach" to research synthesis that was meant to redress some of the deficiencies in the existing strategies. They argued that if treated

properly the variation in outcomes among related studies could be a valuable source of information, rather than a source of consternation as it appeared to be when treated with traditional reviewing methods. Three years later, Taveggia (1974) described six common procedures that were used in all literature reviews: (a) selecting research; (b) retrieving, indexing, and coding studies; (c) analyzing the comparability of findings; (d) accumulating comparable findings; (e) analyzing the resulting distributions; and (f) reporting the results.

Two papers that appeared in the *Review of Educational Research* in the early 1980s brought the meta-analytic and reviews-as-research perspectives together. First, Jackson (1980) proposed six reviewing tasks "analogous to those performed during primary research" (p. 441). His paper employed 36 review articles from prestigious social science periodicals to examine the methods used in syntheses of empirical research. His conclusion was that "relatively little thought has been given to the methods for doing integrative reviews" (p. 459). Cooper (1982) presented a five-stage model of research synthesis that conceptualized it as a data gathering exercise that should be judged by applying criteria similar to those employed to judge primary research. This paper will be discussed in more detail below.

Over the next 10 years, hundreds of meta-analyses were published in the social and behavioral sciences and the development of research synthesis methods progressed unabated. The use of meta-analysis spread from psychology and education (see Hunt, 1997, for a history of these efforts) through many disciplines, especially social policy analysis and the medical sciences (see Chalmers, Hedges, & Cooper, 2002, for a history of meta-analysis in medicine).

In 1994, the *Handbook of Research Synthesis* was published (Cooper & Hedges, 1994). The pursuant decade was a time of enormous growth in the need for and use of rigorous research synthesis in the social and behavioral sciences. Figure 1 presents a chart showing the growth in citations to documents that included the term "research synthesis," "systematic review," "research review," or "meta-analysis" in their title or abstract during the years 1995–2005, according to the Web of Science Citation Index.

# RESEARCH SYNTHESIS AS A RESEARCH PROCESS

Cooper (1982) argued that similar to primary research, a research synthesis involved five stages. The stages demarcated the principal tasks that need to be undertaken when conducting a research synthesis so that the effort produces an unbiased rendering of the cumulative state of evidence on a research problem or hypothesis. To justify this approach, Cooper (1998) wrote:

> …the integration of separate research projects involves scientific inferences as central to the validity of knowledge as the inferences made in primary research.…Most important, the methodological choices at each review stage may engender threats to the validity of the review's conclusions. (pp. 291–292)

For each stage, Cooper codified the research question, asked its primary function in the synthesis and the procedural differences that might cause variation in conclusions.

Cooper then applied the notion of threats-to-inferential-validity (Campbell & Stanley, 1966; also see Shadish, Cook, & Campbell, 2002) to research synthesis. He identified ten threats to validity associated with each stage that might undermine the trustworthiness of a research synthesis' findings. He focused primarily on validity threats that arise from the procedures used to cumulate studies that might influence the outcome of the synthesis, for example, biases in literature searching or the criteria used for including studies in the synthesis. The threats-to-validity approach was subsequently applied to research synthesis by Matt and Cook (1994), who identified 21 threats and Shadish, Cook, and Campbell (2002), who expanded this list to 29 threats. In each case, the authors described threats related not only to potential biases caused by the process of research synthesis itself, but also to deficiencies in the primary research that made up the evidence base of the synthesis, for example, the lack of representation of important participant populations in the primary studies.

Table 1 summarizes a modification of Cooper's (1982) conceptualization. Here, the process of research synthesis is divided into six stages:

Stage 1: Define the problem.

Stage 2: Collect the research evidence.

Stage 3: Evaluate the correspondence between the methods and implementation of individual studies and the desired inferences of the synthesis.

Stage 4: Summarize and integrate the evidence from individual studies.

Stage 5: Interpret the cumulative evidence.

Stage 6: Present the research synthesis methods and results.

These six stages will provide the framework for the remainder of this paper. Different from Cooper's 1982 and 1998 work, the conceptualization used here separates into two stages the processes of (a) summarizing and integrating the evidence from individual studies and (b) interpreting the cumulative findings that arise from these analyses. Also, a list of threats to the validity of conclusions that arise at each stage of synthesis will not be presented. Instead, evaluative questions will be posed that synthesis producers and consumers might ask that relate to the validity of synthesis conclusions. The questions are written from the point of view of a synthesis consumer but this is an arbitrary decision and they can also be used to provide guidance for those carrying out a research synthesis. Each question is phrased so that an affirmative response would mean confidence could be placed in the synthesis conclusions. Each question is preceded by a discussion of some related procedural variations in research synthesis that might enhance or compromise the validity of conclusions and why this is so. While the list is not exhaustive, most of the threats to validity identified in earlier works find expression in the questions.

## An Example from Adult Learning and Literacy

A hypothetical example will be used throughout this paper to assist readers in understanding how to evaluate and interpret research syntheses. This hypothetical research synthesis would seek to answer the question "Do programs meant to assist adults in transitioning from adult basic education (ABE) to postsecondary education (PSE) improve participants' (a) likelihood of success in PSE and (b) subsequent employment?" It is assumed that a researcher of adult learning and literacy is reading such a synthesis and is making a critical evaluation of the trustworthiness of its conclusions.

# THE STAGES OF RESEARCH SYNTHESIS

## Stage 1: Define the Problem

During the problem formulation stage, research synthesists must (a) **define** the variables of interest both **conceptually** and **operationally;** (b) clearly state the type of **relationship of interest;** and (c) place the problem in its theoretical, practical, and/or historical context. This must be done so that relevant and irrelevant studies can be distinguished from one another.

Conceptual definitions. Similar to any empirical investigation, when formulating a problem for research synthesis, the variables of interest must be given a conceptual definition. Conceptual definitions *describe qualities of the variables that are independent of time and space but can be used to distinguish relevant from irrelevant events* (Shoemaker, Tankard, & Lasorsa, 2004). For example, the research problem—Do programs meant to assist adults in transitioning from ABE to PSE improve participants' likelihood of success in PSE and subsequent employment?—requires a definition of "ABE-to-PSE programs" and what is meant by "success" in both PSE and employment. Because these are complex concepts, it might also be useful to define ABE, PSE, and employment.

Conceptual definitions can differ in breadth, or in the number of events to which they refer. Thus, if ABE-to-PSE programs are defined as "any planned attempt to convey information about PSE to participants in ABE programs" the definition encompasses more events than if it is defined as "planned contact between ABE participants and another adult trained in academic and/or employment counseling." The former definition would include informational pamphlets handed out at the end of an ABE class while the latter definition would not. However, the latter definition still permits considerable variation among programs and would include, for example, both programs delivered through one-on-one contact or in groups. So, the first question to ask when evaluating the formulation of a problem in a research synthesis is:

1. Are the variables of interest given clear conceptual definitions?

It might seem that conceptual definitions are more important in theoretical work than in applied settings. However, if the conceptual definition of something as concrete as ABE-to-PSE programs is unclear, it can still cause problems regarding the generalization of findings.

Operational definitions. As in primary research, in order to relate concepts to concrete events, the variables of interest in a research synthesis also must be operationally defined. An operational definition provides *a description of the characteristics of observable events that are used to determine whether the event represents an occurrence of the conceptual variable.* For example, an operational definition of the concept "success in PSE" might include completion of a class with a passing grade and/or completion of a degree program, and "success in subsequent employment" might include obtaining a job, length of tenure in a job, and/or yearly income five years after completing an ABE program.

Research synthesists can begin their work with broad conceptual definitions. However, they may discover that the operations used in previous relevant research have been confined to a narrower conceptualization. For instance, a research synthesis about transition programs might find that past research has only examined the effects of how-to manuals on completing the first year of PSE. If so, it might be inappropriate to label the treatment variable "ABE-to-PSE transition programs" because this implies evidence was found about types of programs that are known to exist but that have never been the focus of research. When such a circumstance arises, the research synthesists need to narrow their conceptual definition to correspond better with the existing operations. Otherwise, conclusions in the synthesis might appear to apply more generally—to more types of programs—than is warranted by the data. So, the synthesists might alter the research problem to "Do ABE-to-PSE transition programs that employ how-to manuals improve the rate of completion of the first year of PSE?"

The opposite problem can also occur—that is, starting with narrow concepts but then finding operations in the literature that could support broader definitions. For example, synthesists might start out with the notion of finding academic performance outcomes but then discover that employment outcomes have been used in research as well. They would then face the choice of either broadening the allowable measures of "success" or excluding many studies that others might deem important to evaluating the impact of PSE programs.

Evaluation of the "fit" between concepts and operations in research synthesis should proceed by considering whether (a) precise conceptual definitions are provided that (b) lead to clear linkages between concepts and operations that are neither too broad nor narrow and that (c) allow meaningful interpretation of results. Thus, the next question to ask when evaluating the definition of the problem in research syntheses is:

2. Do the operations that empirically define each variable of interest correspond to the variables' conceptual definition?

The relationship of interest. In its most basic form, the problem that might motivate a research synthesis includes a clear statement of (a) what variables are to be related to one another and (b) an implication of how the relationship can be tested empirically, that is, the relevant research designs (Kerlinger & Lee, 1999).

In order to be able to determine the appropriateness of different research designs, there are three questions that need to be asked about the problem that motivates a research synthesis:

a. Should the results of the research be expressed in numbers or narrative?
b. Is the problem seeking to uncover a description of an event, an association between events, or a causal explanation of an event?
c. Does the problem or hypothesis seek to understand
    a. how a process unfolds *within* an individual unit over time, or
    b. what is associated with or explains variation *between* units or groups of units?

For a more complete treatment of how these questions relate to research designs, the reader is referred to Cooper (2006).

The research problem "Do programs meant to assist adults in transitioning from ABE to PSE improve participants' likelihood of success in PSE and subsequent employment?" suggests that the research synthesis should focus primarily on summarizing quantitative research looking for a causal connection between participation in transition programs and future success in PSE and employment. Further, while it might be of interest to know what processes the programs set in motion that led individual participants to change their behavior, the focus seems clearly on assessing the average effect of the treatment by comparing participants in transition programs with other ABE participants not receiving the treatment. Thus, the next question to ask when evaluating the formulation of a problem in research synthesis is:

3. Is the problem stated so that the research designs and evidence needed to address it can be specified clearly?

Theoretical, historical, and/or practical context. Research syntheses should be placed in theoretical, historical, and/or practical context. Why are transition programs needed? Where did the idea of such programs come from? Are program components grounded in theory or in practical experience? Are there debates surrounding the utility of transition programs? Do theories predict how, when, and why programs will be effective? Are there conflicting predictions associated with different theories?

As an example, the case for the importance of ABE-to-PSE transition programs could be made by pointing out that the majority of the fastest growing jobs in the U.S. economy requires postsecondary education (U.S. Department of Labor, 2002) and that workers with some PSE on average make $4,700 (for women) and $6,000 (for men) more annually than workers with only a high school education (U.S. Department of Education, 2002). However, while 65% of GED examinees in 1999 said they were obtaining the degree so as to pursue further education (American Council on Education, 2000), only 30–35% obtained any PSE and only 5–10% obtained one year of PSE (Tyler, 2001). Much of this discrepancy between aspirations and accomplishment could be attributed to challenges faced by ABE participants in making the transition to PSE, such as knowing the procedural requirements for enrolling in PSE, developing study and time management skills, and finding the motivation to persist in preparing for and completing PSE coursework (Alamprese, 2005).

Contextualizing a problem in research synthesis does more than explain why a topic is important. Providing a context for the problem also provides the rationale for the search for moderators of the principal findings. It is an important aid in identifying variables that the synthesists might examine for the developer's influence on the outcomes of programs. For example, one group of transition program developers might have used a theory of pedagogy to develop a program delivered through one-on-one instructions, believing it would be more effective than a program employing group instruction of participants. This theoretically derived distinction between programs, along with others, could then be used in the synthesis to test whether research revealed differences in the effectiveness of programs using the different modes of instruction. Also, the synthesists might find that program developers have built their programs so they contain components meant to address the different challenges faced by ABE participants, such as assisting with PSE bureaucratic requirements, developing study and time management skills, and enhancing motivation. Or, they might find in the past literature an empirically based typology of programs that groups the transition programs according to clusters of program components. For example, a National College Transition Network (2006) analysis of transition program data yielded five models of college transition programs. The major focus of each of these was briefly summarized as follows:

1. *Advising:* raise awareness and provide information about postsecondary education options and admission processes.
2. *GED-Plus:* enroll students in GED classes concurrently with transition classes or workshops.
3. *ESOL:* build nonnative English speakers' academic reading and writing skills.
4. *Career Pathways:* provide bridging instruction from ABE that facilitates enrollment in PSE related to specific careers or jobs.

5. *College Preparatory:* provide direct academic instruction designed to address gaps between the knowledge and skills required to complete the GED and those needed for success in PSE.

The research synthesists then can ask which, if any or all, of these distinctions among programs—modes of delivery, inclusion of distinct program components, or packages of components—is associated with program effectiveness.

It is important to point out that both quantitative and qualitative research can be used to place the research problem in a meaningful context. Especially early in the development of a systematic approach to a problem, narrative or qualitative descriptions of the event can be very helpful (Camic, Rhodes, & Yardley, 2003). The narrative descriptions can best be used to discover the salient features of the problem at hand and to assist in deciding what to measure more precisely with numbers. The more open-ended, qualitative approaches to research might focus on questions such as: "What happens when ABE participants encounter the PSE bureaucracy?" and "How do ABE students react to different modes of instruction?" These can be the source of important queries for research synthesists to ask of the quantitative evidence. Quantitative surveys also can be enlightening in the early stages of problem formulation. They can answer specific questions across a broader array of problem instantiations. In addition to establishing the importance of the problem, surveys can answer questions such as: "How available are transition programs with different components?" and "What are the characteristics of ABE participants in transition programs?"

How a problem is contextualized affects the outcomes of syntheses by leading to variation in the way study operations are treated *after* the relevant literature has been identified. Synthesists can vary in the attention they pay to theoretical and practical distinctions in the literature. Thus, two research syntheses that employ identical conceptual definitions of transition programs and that contain the same set of studies can still reach decidedly different conclusions if one synthesis examined information about theoretical and practical distinctions in programs to uncover a moderating relationship that the other synthesis did not test. For example, one synthesis might discover that the effectiveness of transition programs depended on whether they employed one-on-one or group instruction while the other synthesis never examined this feature of the program design. Thus, to evaluate whether (a) the importance of the problem has been established and (b) a list of important potential moderators of findings has been identified, the next question to ask when evaluating research syntheses is:

4. Is the problem placed in a meaningful theoretical, historical, and/or practical context?

## Stage 2: Collect the Research Evidence

The stage of research synthesis that involves gathering the research evidence requires (a) identifying the **sources**—or information channels—and **search terms** that will be used to locate relevant studies while being aware of **publication bias** and (b) delineating the **procedures that will be used to extract information** from the research reports.

Sources of research literature. The decisions about where to look for research literature will influence the types and outcomes of studies that are the basis for the synthesis' conclusions. The studies available through different information channels are different from one another. Thus, the first concern regarding the literature search is that the synthesis may not include—indeed probably will not include—all studies pertinent to the topic of interest. Synthesists who have tapped the broadest and most complementary sources of information are most likely to retrieve a set of results that resembles the entire population of relevant research.

One important feature that distinguishes scientific communication channels concerns *how research gets into the channel*. Channels can have relatively open or restricted rules for entry. Open entry permits primary researchers to enter the channel directly and place their work into its archives. For some channels, for example reference databases, entry occurs without the intervention of the researcher. Restricted entry requires primary researchers to meet the requirements of a third party—some person or entity between themselves and the user of their research—before their work can enter the information channel. For example, the most important of these requirements would be the use of peer review to ensure that research meets certain standards of quality and contribution. It is these restrictions that most directly affect how the research in the channel differs from all relevant research.

A second important feature of information channels concerns *how searchers get into the channel*. Information channels have more or less open or restricted requirements regarding how to access their content. A channel is more restrictive if it requires the literature searchers to identify more specifically what or whose documents they want. A channel is more open if the literature searchers can be more broad or general in their request for information.

Information channels can be grouped under the headings *informal channels*, *formal channels*, and *secondary channels*. The principal forms of informal communication are personal contacts and discussion lists (that have largely replaced the "invisible colleges" of years past). Formal channels of communication typically have explicit rules that primary researchers must follow to enter information into the channels. These rules place restrictions on the kind or quality of information that is admitted to the system. The major formal channels are (a) professional meetings and conferences that have paper presentations and (b) professional journals for published articles. Secondary

channels typically accumulate information about a wide variety of primary research documents, and many now contain the documents themselves. They are constructed by third parties for the explicit purpose of providing literature searchers with relatively comprehensive lists of studies relating to a discipline or topic. The major secondary channels are research registers and reference databases, including citation indexes.

Table 2 briefly describes the entry and access characteristics of each of these channels and also suggests how the research contained in each channel might be different if it was compared to "all research." Regrettably, there is little empirical data on differences in research contained in different communication channels. The problem is complicated further by the fact that the effect of a channel on the information contained in it probably varies from topic to topic. Thus, the assertions in Table 2 should be taken as gross generalizations needing both empirical support and refinement in particular instances.

The question of which and how many sources of information to use in a literature search has no general answer. The appropriate sources partly will be a function of the topic under consideration and partly of the resources of the synthesists. However, as a rule, research synthesists should always access multiple channels with different entry and access restrictions so that they minimize any systematic differences between (a) studies that they evaluate for their relevance to the topic of interest and (b) studies that might have been relevant but never came to their attention.

Generally speaking, reference databases should form the backbone of any comprehensive literature search. These sources probably contain the information most closely approximating all research. However, while they cast the widest net, they may not contain the most recent research because they focus primarily on published documents. That is why it is important to supplement searches of reference databases with searches of informal sources. These will uncover research that was recently completed, under review for publication, and not conducted with publication in mind (for example, contracted evaluations by research firms). Research registers can overcome several of these problems but it is critical for the searcher to know the entry and access criteria for a register, how widely known the registry is in a field, and how frequently it is updated.

Publication bias. An important question faced by research synthesists in regard to their literature search involves whether to include unpublished research. A reason frequently given for excluding unpublished research is that it is often of lesser quality than published research. However, this is too simple a dichotomy. For example, researchers often do not publish their results because publication is not their objective; it does not help them get their work to the audience they seek. Some research associated with degree requirements is conducted by individuals who will not pursue academic careers. Other research is conducted as evaluations for agencies making internal decisions about

program effectiveness. Thus, the decision to publish is not isomorphic with judgments about quality.

Conversely, most researchers would agree that some low quality research does get published. Moreover, research is often turned down for publication for reasons other than quality. In particular, research that fails to achieve standard levels of statistical significance is frequently left in researchers' file drawers, a problem known as "bias against the null hypothesis" (Greenwald, 1975). The concern here is that studies revealing smaller effects will be systematically censured from the published literature. Published estimates of effect may make relationships appear stronger than if all estimates were retrieved by the synthesists. Lipsey and Wilson (1993) compared the magnitudes of effects reported in published versus unpublished studies contained in 92 different research syntheses. They reported that the impacts of interventions in unpublished research were, on average, one-third smaller than published effects.

For these reasons, it is now accepted practice that rigorous research syntheses will include both published and unpublished research. If the synthesis includes only published research, it must be accompanied by a convincing justification. For example, the bias in favor of publishing statistically significant results probably does not extend much beyond the primary research hypothesis. Therefore, a research problem that appears in many articles as a secondary interest of the researchers will be affected by the publication bias to a lesser degree than the researcher's primary focus.

The dictum that synthesists should search for both published and unpublished research should not be taken to mean that the quality of research methods can be ignored in research synthesis. Research quality should be an important consideration in conducting every synthesis, either by (a) excluding studies that do not meet *a priori* quality criteria or (b) dealing with quality in an empirical manner that does not just make assumptions about the relative quality of published and unpublished research but actually tests for differences. These approaches will be addressed more fully below. The point here is that the publication status of a research report is an imprecise measure of quality with known contaminants and should not substitute for more direct appraisal of quality.

Returning to the example of transition programs, the most obvious choices for references databases to search would be ERIC, PsycINFO, and Dissertation Abstracts. These will contain documents relating to adult learning and literacy, and their covered sources do not overlap entirely. In order to locate documents more recent than those contained in the three reference databases, the synthesists might contact the National Center for the Study of Adult Learning and Literacy. This resource might have access to relevant documents, perhaps maintaining a research registry, and might also know of important individuals for the synthesists to contact. The synthesists might also peruse the convention programs of the American Education Research Association (AERA) and other related associations. If they are not already members, they might contact the

officers of AERA's division on postsecondary education and its special interest group on adult literacy and adult education. Each of these groups might maintain a discussion list that could be used to solicit relevant research.

In sum, then, a broad and exhaustive search of the literature is the most important protection against drawing incorrect conclusions in a research synthesis because the included studies are unrepresentative of all studies on a research problem. While the law of diminishing returns applies here, a complete literature search has to include at least a search of reference databases, a perusal of relevant journals, an examination of references in past primary research and research syntheses, and informal contacts with active and interested researchers. The more comprehensive a search, the more confident synthesists and consumers can be that other synthesists using similar, but perhaps not identical, information channels will reach the same conclusions. Thus, one question to ask when evaluating the literature search in research syntheses is:

5. Were complementary searching strategies used to find relevant studies?

Search terms. It is generally agreed that synthesists should begin their search of reference databases or research registers with the broadest conceptual definition in mind. In the early stages of a research synthesis, synthesists should err by being overly inclusive, just as primary researchers collect some data that might not later be used in analyses.

This strategy initially creates more work for synthesists but it has several long-term benefits. First, it will keep within easy reach operations that on first consideration may be seen as marginal but later jump the boundary from "irrelevant" to "relevant." For example, does a course provided by a community college that permits participation by any newly enrolling student, some of whom have traditional high school diplomas and others GEDs, meet the definition of an ABE-to-PSE transition program? Reconstituting the search because relevant operations have been missed consumes far more resources than first putting such studies in the "in" bin but excluding them later.

Second, a good conceptual definition of a variable speaks not only to which operations are considered relevant but also to which are irrelevant. By beginning a search with broad terms, the synthesists are forced to struggle with operations that are at the margins of their conceptual definitions. Ultimately, this results in conceptual definitions with more precise boundaries. For example, if a search for studies begins with the broad keyword "achievement" in the search, it can later be decided to exclude research using course grades as outcomes of transition programs, thus narrowing the achievement construct. But, by explicitly stating these outcomes were excluded, the consumers of the synthesis gets a better idea of where the boundary of the definition lies, and can argue otherwise if they choose.

Third, searches based on broad conceptual definitions allow the synthesis to be carried out with greater operational detail. For example, searching for and including "degree completion" as the outcome of interest only permits an examination of restricted variations in operations, perhaps the types or requirements of degrees. Searching for, and ultimately including, a larger range of operations permits the synthesists to examine broader conceptual issues when they cluster findings according to operations and look for variation in results. Do transition programs produce effects on course grades but not degree completion, suggesting perhaps they have a positive effect on study skills but not motivation? If so, are there plausible explanations for this? Often, these analyses produce the most interesting results in a research synthesis.

Synthesists looking for research on PSE transition programs likely would begin by looking in the ERIC reference databases and conducting a search that looked for the terms "adult basic education" *or* "adult literacy" *or* "adult education." Such a search would likely reveal several thousand documents. If examining the record for each of these was prohibitive, the synthesists might add *and* "transition" to the search. Now, only a few hundred documents might be retrieved.

In sum then, the question to ask with regard to the use of keywords when evaluating research syntheses is:

6. Were proper and exhaustive terms used in searches and queries of reference databases and research registries?

Delineating the procedure to extract information from the research reports. Once the research problem has been carefully and clearly defined and the search has led to a pool of studies to examine for their relevance, it seems as though applying the operational decision rules to separate relevant from irrelevant studies would be relatively straightforward. However, this is a process that is open to more subjectivity and procedural variance than might at first be expected. For example, Cooper and Ribble (1989) found that decisions about the relevance of studies to a topic were influenced by the information provided about the study (relevance judgments were more accurate when bibliographic information included the abstract), the expertise of the searcher (experts identified fewer false positives) and even, perhaps, the personality of the searcher (less dogmatic searchers were somewhat more likely to code studies as relevant).

Because of these sources of variation in relevance decisions, it is important for research synthesists to have more than one person make the initial decision regarding the potential relevance of studies for a topic. The study selection process can be critical to the ultimate conclusions of a synthesis. Unless readers are conversant enough with a literature to recognize that relevant research has been missed, missed reports can be a

serious threat to the validity of a synthesis' conclusions. For this reason, documents that are deemed potentially relevant by any searcher should be carefully examined. When discrepancies about relevance occur, these often can be used to help clarify the problem definition. A fuller discussion of the more general impact of missing data on synthesis conclusions appears below.

Once the decision has been made that a study is indeed relevant to a research problem, the extraction of information about each study takes place. While coders of primary research are fairly reliable in their retrieval of information, it is good practice for synthesists to take steps to ensure that data is reliably extracted from documents. The synthesists should treat the coding of studies as if it were an exercise in data gathering. Coding sheets should be standardized and accompanied by code books explaining the definition and distinctions in each study characteristic that is being extracted. Prior to actual coding, discussions and practice examples should be worked out with coders.

Also, it is often important to obtain numerical estimates of coder reliability. There are many ways to quantify coder reliability and none appears to be without problems (see Orwin, 1994, for a general review of evaluating coding decisions). Two methods appear most often in research syntheses. Most simply, research synthesists will report the agreement rate between pairs of coders. The agreement rate is the number of agreed upon codes divided by the total number of codes. Also useful is Cohen's kappa, a measure of reliability that adjusts for the chance rate of agreement. Kappa is defined as the improvement over chance reached by the coders. Some synthesists will have each study examined by two coders, will compare codes, and then will have discrepancies resolved in conference or by consulting a third coder. This procedure leads to very high reliability and often is not accompanied by a quantitative estimate of reliability. Other synthesists have individual coders mark the codes they are least confident about and discuss these codes in group meetings. This procedure also leads to highly trustworthy codes. Regardless of what techniques are used, the question to ask when evaluating the methods of data collection used to carry out research syntheses is:

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?

Some deficiencies in both retrieval and coding procedures will frustrate synthesists regardless of how thorough and careful they try to be. Some potentially relevant studies do not become public and defy the grasp of even the most conscientious search procedures. With regard to coding studies, it is impossible to remove all subjectivity from the process and some judgments are inherently ambiguous. Perhaps the most frustrating occurrence in collecting the evidence is when synthesists obtain primary research reports, but the reports do not contain the needed information. Reports can be

missing information on statistical outcomes, preventing meta-analysts from estimating the magnitude of the difference between two groups or the relationship between two variables. Or, reports can be missing information on study characteristics, preventing the determination of whether study outcomes were related to how the study was conducted. Some approaches to missing data will be discussed shortly.

## Stage 3: Evaluating the Correspondence Between the Methods and Implementation of Individual Studies and the Desired Inferences of the Synthesis

Above, it was pointed out that research synthesists must provide an explicit statement about the type of relationship under study—is the problem motivating the synthesis seeking to uncover a description of an event, an association between events, or a causal explanation of an event? This aspect of rigorous research synthesis takes center stage when synthesists and consumers evaluate the correspondence between the design and implementation of individual studies and the desired inferences of the synthesis. The research question "Do programs meant to assist adults in transitioning from ABE to PSE improve participants' likelihood of success in PSE and subsequent employment?" suggests that the research synthesis should focus primarily on summarizing quantitative empirical research about a causal relationship between participation in transition programs and future success in PSE and employment. So, how this evaluation would proceed when the purpose of the research synthesis is to make causal statements will be detailed below. Although this is only one type of research question, the logic involved in others types of inferences should be obvious from this example.

Categorizing studies by research design and implementation. Given that causal relationships are under investigation, four important distinctions in quantitative research designs will be encountered in the research literature. First, some primary research is **associational research** and will focus on establishing a simple relationship between participation in transition programs and PSE outcomes. The old dictum that "correlation does not imply causation" applies here. So, studies that simply correlate participation in transition programs with PSE outcomes cannot be taken as evidence of a causal link between the two. While it is unlikely that PSE success caused participation in transition programs, the association may be spurious—that is, both participation and success may have been produced by a third variable, perhaps the motivational level of the participant—with no causal connection between them.

**Modeling research** takes simple correlational studies a step further by examining co-occurrence in a multivariate framework. For example, if researchers wish to know whether transition programs *cause* participants to take more PSE courses, they might construct a multiple regression equation or structural equation model that attempts to provide an exhaustive description of a network of relational linkages (Kline, 1998). This

model would attempt to account for, or rule out, all other co-occurring phenomena that might explain away the relationship of interest. Likely, the model will be incomplete or imperfectly specified, so any casual inferences from modeling research will be very tentative, at best.

Third, **quasi-experimental research** (Shadish et al., 2002) controls the introduction of an intervention but does not control precisely who may be exposed to it. Instead, the researchers use some statistical control in an attempt to equate the groups receiving and not receiving the intervention. It is difficult to tell how successful the attempt at equating groups has been. For example, researchers might be able to offer a class on study skills for ABE students at a community college but might not be able to assign students to the class. Instead, students taking the class of their own volition might be matched on prior grades with students not opting for the class.

Finally, in **experimental research** both the introduction of the event (for example, a study skills class) and who is exposed to it are controlled by the researcher. The researcher uses a random procedure to assign students to conditions, essentially leaving the assignment to chance (Boruch, 1997). Because this approach minimizes average preexisting differences between transition-class participants and nonparticipants, this design makes possible the most confident conclusion that any differences between the participants and nonparticipants was caused by the transition program.

In sum, then, research synthesists are likely to come across a variety of research designs that relate the concepts of interest to one another. A search for studies on ABE-to-PSE transition programs may identify studies that resulted in simple correlations ("Do ABE students who report attending transition courses also report completing more PSE classes?), multiple regressions, perhaps a few structural equation models, some quasi-experiments, and a few experiments. Therefore, it is critical in research synthesis that the type of relationship between the variables of interest be clearly specified. This specification dictates whether the retrieved research uses the appropriate research designs to study the question. Designs appropriate to gather data on one type of relationship may or may not provide information relevant for investigating another type of relationship.

Of course, there are numerous other aspects of the design and implementation of a study—beyond whether the intervention was manipulated or measured and how participants came to be in different conditions—that must be attended to for strong inferences about causality to be made legitimately. For example, researchers might begin by intending to carry out an experiment in which ABE students are randomly assigned to classes that do and do not provide PSE transition assistance. However, as the semester proceeds, some students move out of the school district, others drop out of the ABE program, and still others switch classes. By the end of the semester, it is clear that the students remaining in the transition-assistance classes were not equivalent to nonparticipants, on average, when the study began. Thus, the ability of the experiment to

draw strong causal inferences about the effects of transition assistance has been compromised by severe attrition, perhaps differential attrition in different conditions, and treatment crossover. The synthesists would need to make certain that their codes of the study's research design and implementation characteristics included the data needed to determine whether attrition and treatment crossover should be concerns. Further, internal validity issues are not the only design and implementation issues that needed to be coded. Information about studies related to construct validity (e.g., the fidelity with which the treatment was implemented), measurement validity (e.g., the reliability of the outcome measures) and external validity (e.g., the characteristics of participants) would also need to be extracted from the study reports and used to extend or delimit the studies' power to make causal inferences.

Because these design and implementation variations have implications for the kinds of inferences the synthesists can draw legitimately, the next important question to ask when evaluating a research synthesis is:

8. Were studies categorized so that important distinctions could be made among them regarding their research design and implementation?

Exclusion of research versus examination of design and implementation differences. Once categorized, how should the synthesists treat the variety of research designs? This issue is complex and especially so when the research question deals with uncovering a causal relationship. The different designs produce evidence with different capabilities for drawing strong inferences about the problem. Here, the correlational evidence addresses a necessary but not sufficient condition for causal inference. As such, if this were the only research design found in the literature, it would be legitimate to assert that the causal connection between transition programs and PSE success remained untested. When an association is found, multiple regressions statistically control for some alternative explanations for the relationship but probably not all. Structural equation models relate to the plausibility of causal networks but do not address causality in the generative sense. Well-conducted quasi-experiments may permit weak causal inferences, made stronger through multiple and varied replications. Experiments, with few implementation flaws, permit the strongest inferences about causality.

At one extreme, the synthesists in search of a causal relationship might discard all studies but those using true experimental designs. This approach applies the logic that these are the only studies that directly test the question of interest. All other designs either address association only or do not permit strong inferences. At the other extreme, the synthesists would include all the research evidence but carefully qualify inferences as the evidence for causality moved farther from the ideal. A less extreme approach would be to include some but perhaps not all designs while again carefully qualifying inferences.

There are arguments for each of these approaches. In research areas where strong experimental designs are both relatively easy to conduct and plentiful, excluding designs that permit only weak causal inferences may be an appropriate approach to the evidence. In other areas, experiments may be difficult to conduct and rare—for example, the impact of ABE-to-PSE transition programs. Here, the synthesists may decide that "any evidence is better than no evidence at all" and proceed to summarize the less-than-optimal studies, with the appropriate precautions, of course.

Generally speaking, when experimental evidence on causal questions is lacking or sparse, a more inclusive approach is probably best, assuming that the synthesists pay careful and continuous attention to the impact of research design on the conclusions that they draw. In fact, the inclusive approach can provide some interesting benefits to inferences. Returning to the example, the synthesists might find a small set of studies in which the availability of a transition program has been manipulated and students were assigned randomly to conditions. However, in order to accomplish the manipulation, these studies might have been conducted in courses that involved only a few transition skills—say, those helping participants deal with bureaucracies—and only proximal outcome measures—say, grades during the first year in PSE. Thus, in order to carry out the manipulation of conditions and the random assignment of participants, the researchers found it necessary to study only the short-term impacts of a simple transition intervention. This use of simple manipulations and proximal measures in true experiments is not an unusual circumstance in educational and social policy research. These studies might have demonstrated that participants in ABE transition classes received higher grades, but it could be that bureaucracy training becomes less important as impacts become more distal.

These issues, related to construct and external validity, might go unaddressed if only the experimental evidence were permitted into the synthesis. Instead, the synthesists might use the nonexperimental evidence to help gain tentative, first approximations about how these transition programs play out over time and with broader constructions of achievement. The quasi-experiments found in the literature might use the number of PSE courses completed over three years as the outcome measure. The structural equation models might use large, nationally representative samples of students and relate participation in transition courses to broader measures of success, such as completion of a degree or posteducation income.

By employing these results to form a web of evidence, the synthesists can come to more or less confident interpretations of the experimental findings. If the nonexperimental evidence reveals relationships consistent with the experiments, more confidence can be taken in suggesting that the experimental results generalize beyond the specific operations used in the experiments because the alternative designs with their complementary strengths and weaknesses provide converging evidence. If the different types of designs reveal inconsistent results, it should be viewed as a caution to generalization.

Finally, design considerations—that is, how interventions were introduced and how participants were assigned to receive them—are not the only methodological considerations that might play a role in deciding whether or not to include a study in a research synthesis. As noted above, even experiments using random assignment can encounter problems that diminish their ability to draw strong causal inferences. This raises the more general questions regarding when and how it is appropriate to exclude studies because of flaws in design or implementation.

The point is important because research synthesists often debate whether or not *a priori* judgments of research quality should be used to exclude studies from their work. Proponents of excluding studies based on design and implementation flaws often employ the "Garbage in-garbage out" dictum (Eysenck, 1978). However, studies looking at "quality" judgments suggest that there is great variability in what researchers feel constitutes quality (Valentine & Cooper, 2005). The *a priori* quality judgments required to make the discrete decision about whether to include or exclude studies are likely to vary from judge to judge and be influenced by personal perspectives. So, many research synthesists counter-argue that the impact of design and implementation variation on study results can be viewed as an empirical *a posteriori* question, rather than an *a priori* matter of opinion (Glass, McGaw, & Smith, 1981, p. 222).

Instead of excluding studies, this position holds that synthesists should thoroughly code the design aspects, good and bad, of each study and then demonstrate if, in fact, the outcomes of studies are related to how the studies were conducted. For example, Table 3 presents seven characteristics of studies that might be retrieved from reports and used to determine the amount of participant attrition over the course of the study. For each study, the research synthesists could use the six sample sizes to calculate four estimates of attrition—from the beginning of the transition program or following completion of the program separately for the ABE transition and control groups. These estimates, as well as the response to the dichotomous question about differential attrition, could be used by the synthesist to answer the question "Did studies with different attrition rates, both overall and differential, reveal different outcomes for transition programs?"

Of course, the answers to the questions in Table 3 might also be used to exclude a study from further consideration. For example, the synthesists might decide that studies with overall attrition rates of XX% or differences in attrition rates between transition and comparison conditions of XX% would lead to the exclusion of the study. Here, the synthesists would have to propose and justify the attrition rate cutoff points.

And, attrition rates are only one criterion that might be used to exclude studies. The decision to include or exclude studies on an *a priori* basis requires the synthesist to make an overall judgment of quality that is often subjective, involving assumptions about design priorities that might not be shared by others. Still, several attempts have been recently undertaken to transparently identify *a priori rules* for excluding studies from

research syntheses. For example, the What Works Clearinghouse (2006) provides standards of evidence that rely heavily on issues related to the internal validity of studies. This scheme categorizes studies that are admissible to its research syntheses into two levels of confidence. Studies that meet Clearinghouse evidence standards are randomized controlled trials that do not have problems with randomization, attrition, or treatment disruption, and regression discontinuity designs that do not have problems with attrition or disruption. Studies that meet Clearinghouse evidence standards with reservations are strong quasi-experimental studies that have comparison groups that meet the other evidence standards listed above, randomized trials with randomization, attrition, or disruption problems, and regression discontinuity designs with attrition or disruption problems. The Best Evidence Encyclopedia (2006) standards are similar to the What Works Clearinghouse standards but also make mention of a criteria requiring a minimal duration for treatments. Finally, Confrey (2006) has compared and contrasted the evaluation standards employed by the National Research Council report on the effectiveness of middle school mathematics curricula, which included the use of multiple methods such as content analysis and case studies, with the What Works Clearinghouse standards that focus exclusively on experiments and quasi-experiments.

Instead of excluding studies on an *a priori* basis, a careful enumeration of study characteristics can be devised, such as the example in Table 3 for attrition, by the synthesists, and study characteristics can be compared to study results to determine if they covary with one another. If it is empirically demonstrated that studies with "good" design and implementation features (e.g., low attrition) produce results different from "bad" studies (e.g., studies with severe differential attrition), the results of the good studies can be given more weight when conclusions are drawn. When no difference in results is found, it seems sensible to retain the "bad" studies, because they contain other variations in methods (such as different sample characteristics) that, by their inclusion, will help answer many other questions surrounding the problem area.

Certainly, if *a priori* exclusion of studies occurs in a research synthesis, the criteria for excluding studies must be defined before the literature is examined, so that the rules do not shift based on the outcomes of studies. Likewise, if more lenient criteria for inclusion are employed, the deficiencies of the included research designs need careful attention in the interpretation of results. Further, if a particular area lacks rigorous research, the existence of a research synthesis should not be taken to imply there is no need to conduct well-designed studies in the future and develop a body of stronger research. Rather, the synthesis should be used to highlight such need.

In sum, then, the first obligation of synthesists is to clearly state the approach they have employed for including or excluding studies based on design and implementation considerations and the rationale for it. Therefore, an important question to ask about exclusion criteria when evaluating a research synthesis is:

9. 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?

## Stage 4: Summarize and Integrate the Evidence from Individual Studies

Many approaches are available for the analysis of combined study results. The most obvious distinction between them is whether or not they employ statistical combining procedures, or meta-analysis. There are some convincing arguments suggesting that quantitative synthesis techniques should be used whenever the goal of a synthesis is to formulate a summary statement about the empirical evidence on a relation between variables, such as exposure to a transition program and PSE outcomes. Cooper and Rosenthal (1980) had participants read the same seven articles chosen to suggest that women have greater task persistence than men. Some participants were randomly assigned to a meta-analysis condition and other participants were assigned to a narrative synthesis condition. After the participants completed their assignments, they were asked whether the articles they read suggested the existence of a relationship. Among the narrative reviewers 73% found "definitely" or "probably" no support for the hypothesis compared with only 32% of participants using the meta-analytic technique. In fact, the combined probability that the null hypothesis was true was $p < .005$, indicating (by this standard) that over twice as many inferential errors were made by the narrative than the quantitative synthesists. Other arguments for the use of quantitative techniques to summarize the empirical evidence about relationships are that traditional procedures (a) are unable to generate estimates of the size of the relationships and (b) do not employ strategies for weighting individual studies proportional to their size or quality (see Cooper, 1998).

### *When Not to Use Meta-Analysis*

Using meta-analysis should be the default option when the goal of a synthesis is to summarize a research literature for purposes of making a general statement about the support for, or size of, a relationship between variables. Therefore, it is important to point out instances in which the use of meta-analysis might be less appropriate, or perhaps completely unnecessary.

First, and most obvious, meta-analysis is improper if the goal of the synthesis is to critically appraise a research literature study-by-study or to identify particular studies central to a field. Second, meta-analysis may be inappropriate in cases where conceptual and methodological approaches to research on a topic have changed over time. In such instances, a proper integration would treat the results of studies as an emerging series of

historical events, that is, using a historical approach to organizing the synthesis rather than a statistical aggregation of their cumulative findings. Third, under certain conditions, meta-analysis might not lead to the kinds of generalizations the synthesists wish to make. For example, cognitive psychologists or cognitive neuroscientists might argue that their methodologies typically afford good controls and reasonably secure findings. Thus, the debate about effects usually occurs with reference to the choice of variables and their theoretical, or interpretive, significance. For example, the dispute about massive modularity spawned by evolutionary psychology is a theoretical issue that will not be solved by meta-analysis. Under these circumstances, synthesists might convincingly establish the generalization of a finding using conceptual and theoretical bridges rather than statistical ones. Finally, even if synthesists wish to summate statistical results across studies on the same topic, the studies might have been conducted using decidedly different methodologies, participants, and outcome measures. In such cases, statistical combinations might mask important differences in research findings. In these instances, it may make the most sense not to use meta-analysis, or to conduct several discrete meta-analyses within the same synthesis. In fact, the distinctions proposed above regarding differences in research designs—associational, modeling, quasi-experimental, experimental—represent one category-of-distinction among methodologies many would argue ought not to be crossed by meta-analysis. If the search for research on transition programs found studies in each of these categories, the field would likely be best served by a synthesis that separately examined and interpreted the evidence using each design, although meta-analyses within each design category might be appropriate.

It is also important to point out that the use of meta-analysis is no guarantee that the synthesists will be immune from making all inferential errors. The possibility always exists that the meta-analysts have used an invalid rule for inferring a characteristic of the target population. As in the use of statistics in primary research, this can occur because the target population of studies does not conform to the assumptions underlying the analysis techniques or because of the probabilistic nature of statistical findings.

In sum, then, all research synthesists should provide justification for the methods they use to summarize and integrate the results of the individual studies. They should ensure that the synthesis techniques employed are transparent to the reader. They should provide enough information so that readers can critically assess the strengths and weaknesses of the synthesis methods. Thus, an important question to ask about how results are integrated when evaluating research syntheses is:

| |
|---|
| 10. Was an appropriate method used to combine and compare results across studies? |

### *The Elements of Meta-Analysis*

There are many forms of meta-analysis and many issues that must be addressed when a meta-analysis in conducted. Entire books have been written on this subject so an attempt here to cover the issues exhaustively would be impossible. Instead, the four components common to nearly all meta-analysis—effect size estimation, averaging effect sizes, testing for the homogeneity of effects, and looking for moderators of effects—will be briefly discussed.

Effect size estimation. Cohen (1988) defined an effect size as "the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false" (pp. 9–10). There are many different metrics used to describe an effect size. Generally, each metric is associated with particular research designs.

For studies that employ discrete conditions, regardless of whether participants are assigned to conditions at random or not, the metric typically used to express the magnitude of the treatment effect is the *d*-index. The *d*-index, or standardized mean difference, is a scale-free measure of the separation between two group means. Calculating the *d*-index for any comparison involves dividing the difference between the two group means by either their pooled standard deviation or by the standard deviation of the control group. This calculation results in a measure of the difference between the two group means expressed in terms of their common standard deviation or that of the untreated population.

For example, Table 4 presents the results of eight *hypothetical* studies comparing the grades of ABE students who received assistance in transitioning to PSE with ABE students receiving no transition assistance, using first year grades as the outcome measure. These fictional participants were randomly assigned to either receiving transition assistance or to a no-treatment comparison group. The first two columns of Table 4 provide the hypothetical sample sizes for the treatment and comparison group, respectively. The third column provides the *d*-index for each of the eight fictional studies. In each case, the synthesists subtracted the no-transition-assistance group mean from the transition-assistance group mean, so positive *d*-indexes indicate participants who received the treatment did better on first-year PSE grades than nonparticipants. The *d*-indexes were calculated using the formula:

$$d = \frac{X_1 - X_2}{s_p}$$

where $X_1$ and $X_2$ represent the two group means and $s_p$ is the pooled standard deviation defined as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 + 1)s_2^2}{(n_1 - 1) + (n_2 + 1)}}$$

where $n_1$ and $n_2$ represent the number of participants in each group and $s_1$ and $s_2$ represent the standard deviation of each of the groups.

The *d*-index of .90 for the first study indicates that nine-tenths of a standard deviation separates the two means. This positive *d*-index would mean that the participants receiving assistance had better first-year PSE grades. In the fourth study, the *d*-index of .-18, indicates that the nonparticipants did better than participants by a little less than one-fifth of a standard deviation.

For studies that involve continuous, typically nonmanipulated measures but also include an attempt to statistically equate students on other characteristics, such as multiple regressions or structure equation models, measures of relationship strength can include standardized regression weights ($\beta$), unstandardized regression weights (*b*), or path coefficients (*p*). The standardized beta-weights indicate what change in the criterion measure expressed as a portion of a standard deviation was associated with a one standard deviation change in the predictor variable. For example, if the standard deviation of a measure of time spent in transition programs equaled 4 hours and the standard deviation of the number of PSE courses taken equaled 2 courses, then a beta-weight of .50 would mean that, on average, students in the sample who were separated by 4 hours of time in transition programs also completed 1 additional PSE course.

For studies that involve continuous variables and no attempt to statistically equate students on third variables, the simple bivariate correlation is typically used as the measure of relationship.

Highlighting three different measures of association implies that the synthesists cannot compare across the different types of design. This is not strictly true. Standardized mean differences and correlation coefficients can be transformed one to the other (see Cohen, 1988). A beta-weight equals a correlation coefficient when no other variables appear in the regression equation. However, as noted above, the decision to combine results across these different types of designs is one that should not be taken lightly. So, it makes sense to propose that the choice of an effect size metric ought to reflect the important design characteristics of the studies from which they are derived. Thus, an important question to ask about the effect size metric when evaluating a meta-analysis is:

| |
|---|
| 11. If a meta-analysis was performed, was an appropriate effect size metric used? |

Averaging effect sizes. The most pivotal outcomes of most meta-analyses are the average effect sizes and the measures of dispersion that accompany them. Both unweighted and weighted procedures can be used to calculate average effect sizes across comparisons. In the unweighted procedure, each effect size is given equal weight in calculating the average effect. In the weighted procedure, each independent effect size is first multiplied by the inverse of its variance and the sum of these products is then divided by the sum of the inverses. The weighting procedure is generally preferred because it gives greater weight to effect sizes based on larger samples, and larger samples give more precise population estimates. Also, confidence intervals are calculated for weighted average effect sizes. Hedges and Olkin (1985), Shadish and Haddock (1994), and Lipsey and Wilson (2001) provide procedures for calculating the appropriate weights and confidence intervals.

One important aspect of averaging effect sizes and estimating their dispersion involves the decision about whether a fixed-error or random-error model underlies the generation of study outcomes. In a fixed-error model, each effect size's variance is assumed to reflect sampling error of participants only, that is, error solely due to participant differences. However, sometimes other features of studies can be viewed as random influences. For example, studies that look at the impact of ABE transition programs on PSE success might vary in the types of PSE settings in which the studies were conducted, in the length of the program, and in the level of expertise of the program provider. In addition to being potential moderators of the transition program effect, this variation may suggest it is most appropriate to consider transition programs represented in the synthesis as "randomly" sampled from all programs. That is, in a random-error analysis, study-level variance is assumed to be present as an additional source of random influence.

The question meta-analysts must ask is whether the effect sizes in a data set are affected by a large number of these study-level random influences. If it is the case that the meta-analysts suspect a large number of these additional sources of random error in effect sizes, then a random-error model is most appropriate in order to take these sources of variance into account. If the meta-analysts suspect that the data are most likely little affected by study-level sources of random variance, then a fixed-error model can be applied. Alternatively, Hedges and Vevea (1998; p. 3) state that fixed-error models are most appropriate when the goal of the research is "to make inferences only about the effect size parameters in the set of studies that are observed (or a set of studies identical to the observed studies except for uncertainty associated with the sampling of subjects)." A further statistical consideration is that, in the search for moderators, fixed-error models may seriously underestimate error variance and random effects models may seriously overestimate error variance, when their assumptions are violated (Overton, 1998).

In view of these competing sets of concerns, meta-analysts sometimes apply both error models (e.g., Cooper, Robinson, & Patall, 2006). Specifically, all analyses can be conducted twice, once employing fixed-error assumptions and once using random-error

assumptions. Differences in results based on which set of assumptions is used can be incorporated into the interpretation and discussion of findings.

Formulas to calculate random-error estimates of the mean effect size and confidence intervals are complex and involve a two-stage process. As such, the interested reader should refer to Hedges and Olkin (1985), Raudenbush (1994), or Lipsey and Wilson (2001) for a full discussion of computing random-error models. In addition, several statistical packages have recently been developed specifically for meta-analysis that allow meta-analysts to easily conduct analyses using both fixed and random error assumptions (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2005; Statistics.com, 2006; Stata.com, 2006).

Returning to the fictional example in Table 4, in order to calculate a weighted average *d*-index and its confidence interval (fixed effects), the meta-analyst would first calculate a weighting factor, $w_i$, which is the inverse of the variance associated with each *d*-index estimate:

$$w_i = \frac{2(n_{i1} + n_{i2})n_{i1}n_{i2}}{2(n_{i1}+ n_{i2})^2 + n_{i1}n_{i2}d_i^2}$$

where $n_{i1}$ and $n_{i2}$ represent the sample sizes in columns 1 and 2 and $d_i$ represents the *d*-index in column 3. The next step would require multiplying each *d*-index by its associated weight and dividing the sum of these products by the sum of the weights. The formula is:

$$d_\bullet = \frac{\sum_{i=1}^{N} d_i w_i}{\sum_{i=1}^{N} w_i}$$

where all terms are defined as before. Table 4 shows the average weighted *d*-index for the eight fictional comparisons would be $d = .21$.

The confidence interval around the average effect size estimate would be calculated next. First, the inverse of the sum of the $w_i$s would be found. Then, the square root of this variance would be multiplied by the *z* score associated with the confidence interval of interest. Thus, the formula for a 95% confidence interval is:

$$CI_{d_\bullet 95\%} = d_\bullet \pm 1.96 \sqrt{\frac{1}{\sum_{i=1}^{N} w_i}}$$

where all terms are defined as before. The 95% confidence interval for the eight fictional ABE transition program comparisons with no treatment includes values of the *d*-index .15 standard deviation units above and below the average *d*-index. Thus, we would expect 95% of estimators of this effect to fall between $d = .06$ and $d = .36$. Note that the interval does not contain the value $d = 0$. It is this information that can be taken as a test of the null hypothesis that no effect of transition programs exists in the fictional population. In this example, we would reject the null hypothesis that there is no difference in first-year PSE grades between participants in transition programs and nonparticipants. However, if a random effects model were used, the estimate would be larger, $d = .31$, but the 95% confidence interval would be wider, -.01 to .63.

A problem for meta-analysts arises when a single study contains multiple effect size estimates. This is most bothersome when more than one measure of the same construct is taken and the measures are analyzed separately. Suppose a meta-analysis examining the effects of an ABE transition program finds a study that compared the treatment and comparison group on both a measure of PSE first-year grades and on degree completion. Since the same participants provided both outcomes, these measures are not independent estimates of the transition program effects. Thus, it would be inappropriate to directly combine these two estimates along with a third from a separate study to arrive at an average effect. The first study would be given too much weight. Also, the assumption that effect size estimates are independent underlies the other meta-analysis procedures described above.

There are several approaches meta-analysts use to handle dependent effect sizes. Some meta-analysts do treat each effect size as independent, regardless of the number that come from the same sample of people. If the number of effect sizes from one sample rarely is greater than one, this approach assumes that the effect of violating the independence assumption is not great. Other meta-analysts use the study as the unit of analysis. In this strategy, they calculate the mean effect size or take the median result and use this value to represent the study. So, if a study reports three nonindependent *d*-indexes of, say, .10, .15, and .35, the study might be represented by a single value of .20, if the mean is used, or .15, if the median of the several measures is used.

Another approach is to use a shifting unit of analysis (Cooper, 1998). Here, each effect size associated with a study is first coded as if it were an independent estimate of the relationship. For example, if a single sample of participants permitted comparisons of the transition program's effect on both PSE first-year grades and degree completion, two separate effect sizes would be calculated. However, for estimating the overall effect of

the transition program, these two effect sizes would be averaged prior to entry into the analysis so that the sample only contributed one effect size. To calculate the overall weighted mean and confidence interval, this one effect size would be weighted by the inverse of its variance (based primarily on sample size, which should about be about equal for the two component effect sizes). However, in an analysis that examined the effect of the transition program on grades and degree completion separately, this sample would contribute one effect size to each estimate of a category's mean effect size.

The shifting unit of analysis approach retains as much data as possible from each study while holding to a minimum any violations of the assumption that data points are independent. More sophisticated statistical models have been suggested as a solution to the problem of dependent effects size estimates (Gleser & Olkin, 1994). However, the viability of these procedures lies in whether the meta-analysts can credibly estimate the actual degree of relation among dependent measurements. These procedures have not been used often because this approach is complex and estimating dependencies is tricky.

The issue of dependent estimates is not confined to the problem of multiple measures taken on the same sample. Results in the same study that are reported separately for different samples of people also share other factors that influence their outcomes. Suppose the effects of an ABE transition program is estimated separately for men and women within the same study. Then, the samples are independent but the setting is not, nor the deliverers of instruction, nor the study's design and execution. All these things will likely make these two effect sizes more similar than any two effects drawn at random. Taken a step further, synthesists also might conclude that separate but related studies from the same group of investigators are not independent. In practice, most meta-analysts ignore these study-level interdependencies in effect sizes but not those based on shared samples. So, another important question to ask when evaluating meta-analyses is:

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?

Homogeneity analyses. In addition to the confidence interval as a measure of dispersion, meta-analysts usually carry out what are called homogeneity analyses. Homogeneity analyses allow the meta-analysts to explore why effect sizes vary from one study to the next. A homogeneity analysis provides calculation of how probable it is that the variance exhibited by the effect sizes would be observed if only sampling error was making them different.

If there is greater variation in effect sizes than would be expected by chance, then the meta-analyst can begin the process of examining moderators of outcomes. If the

observed variance is not significantly different from that expected by sampling error alone, many statisticians advise that the meta-analysts stop the analysis and not look for moderators. After all, chance is the most parsimonious explanation for the variation in effect sizes. However, most meta-analysts search for moderators even in the absence of a statistically significant homogeneity analysis if there are good theoretical or practical reasons for doing so.

To test whether the set of *d*-indexes in Table 4 are homogenous, the meta-analysts would calculate a statistic that Hedges and Olkin (1985) called $Q_t$.

$$Q_t = \sum_{i=1}^{N} w_i d_i^{\,2} - \frac{\left(\sum_{i=1}^{N} w_i d_i\right)^2}{\sum_{i=1}^{N} w_i}$$

The *Q*-statistic has a chi-square distribution with N – 1 degrees of freedom, or one less than the number of *d*-indexes. If the obtained value of $Q_t$ is greater than the critical value for the upper tail of a chi-square at the chosen level of significance, the meta-analysts reject the hypothesis that the variance in effect sizes was produced by sampling error alone.

The fictional meta-analysis of the effect of transition programs in first-year PSE grades reveals a highly significant homogeneity statistic $Q(7) = 29.62$, $p < .001$. This suggests that the meta-analysts should reject the hypothesis that the *d*-indexes are all estimating the same underlying population value, or that sampling error alone was responsible for the variation in effects. Therefore, they would continue their analysis of the effect by looking for variables that may moderate the effect of transition programs on first-year PSE grades. Clearly then, another important question to ask about meta-analyses is:

13. If a meta-analysis was performed, was the homogeneity of effect sizes tested?

Testing for moderators of effect sizes. The search for why the outcomes of studies differ is often the most interesting and informative part of conducting a meta-analysis. As previously suggested, homogeneity analysis allows the meta-analyst to test whether sampling error alone accounts for variation in effect sizes or whether features of studies—research designs and implementation, sample and treatment variations, outcome measures—also play a role in making the results of studies different. The meta-analysts calculate average effect sizes for subsets of studies and compare these to determine if they provide insight into what influences the strength and/or direction of the relationship.

In fact, a major strength of research synthesis, especially when meta-analysis is used, is that the synthesists can ask questions about variables that moderate outcomes even if no individual study has included the moderator variable. For example, the fictional synthesists of ABE transition programs might ask whether the effectiveness of programs differed for participants who received one-on-one transition assistance versus group-level assistance, even if no single study has included both types of instruction. The results of such a comparison of average effect sizes can suggest whether mode of instruction would be important to look at in future research.

The last column of Table 4 indicates that four of the fictional studies administered the PSE transition assistance in one-on-one instruction while the other four fictional studies did so in groups. The procedure to test whether mode of instruction explains variance in effect sizes involves several steps. First, a $Q_t$-statistic would be calculated using the formula just presented. Second, a $Q$-statistic would be calculated separately for each subgroup of studies. Third, the values of these $Q$-statistics would be summed to form a value called $Q_w$. Finally, this value is then subtracted from $Q_t$ to obtain $Q_b$:

$$. Q_b = Q_t - Q_w$$

This $Q_b$ statistic is used to test whether the average effects from the groupings of studies are homogenous. It is compared to a chi-square table using degrees of freedom one less than the number of groupings. If $Q_b$ exceeds the critical value, then the grouping variable is a significant contributor to variance in effect sizes and remains a plausible moderator of effect. This test is analogous to conducting an analysis of variance in that a significant $Q_b$ indicates that the group means differ from one another, that is, exhibit more variation than sampling error alone would predict.

Using a fixed-error model, the effect of the fictional transition programs in Table 4 using one-on-one instruction had a significant impact on first year PSE grades, $d$ = .29 (95% CI = .08/.79), but group-administered instruction did not, $d$ = .13 (95% CI = -.09/.35). As noted above, the $Q_t$-statistic for the eight studies was 29.61. The $Q_w$-statistic for one-on-one instruction was 13.89 and for group instruction was 14.72. Thus, the total $Q_w$ for both groupings was 28.61. From here, the $Q_b$-statistic comparing one-on-one to group instruction can be calculated, $Q_b(1) = 1.01$, p = .32. This result is not significant with 1 degree of freedom. Using a random-error model, the impact of mode of instruction does not have a significant effect using either one-on-one instruction, $d$ = .41 (95% CI = -.08/.89), or group instruction, $d$ = .23 (95% CI = -.26/.72). Further, the $Q_b$-statistic comparing modes of instruction using random-error assumptions would indicate there was not a significant difference in the average weighted $d$-index between the two groups of studies, $Q_b(1) = .26$, p = .61.

In this way, the meta-analysts can employ a formal means for testing whether different features of studies explain variation in their outcomes. If reliable differences do exist, the average effect sizes corresponding to these differences will take on added meaning and will help the meta-analysts to guide future research or make policy recommendations. So, the next important question to ask when evaluating meta-analyses is:

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?

## Stage 5: Interpret the Cumulative Evidence

Similar to primary research, proper interpretation of the results of a research synthesis requires (a) careful use of declarative statements regarding claims about the evidence, (b) specification of what results warrant each claim, and (c) any appropriate qualifications to claims that need to be made. Below, five important issues related to the interpretation of results in research synthesis are discussed:

- statistical sensitivity analysis,
- data censoring,
- generalization and specification,
- study-generated and review-generated evidence, and
- the substantive interpretation of effect sizes.

Statistical sensitivity analysis. An important step in many meta-analyses is the performance of statistical sensitivity analyses. Sensitivity analysis is used to determine whether and how the conclusions of an analysis might differ if it was conducted using different statistical procedures or assumptions. There are numerous points at which meta-analysts might decide a sensitivity analysis is appropriate. For example, the calculation of weighted and unweighted effect sizes can be considered a form of sensitivity analysis, as can the use of both fixed-error and random-error models. In each case, the meta-analyst is seeking to determine whether a particular finding is robust across analyses conducted with different sets of statistical assumptions. In the interpretation of evidence, a finding that conclusions do not change under different statistical assumptions means greater confidence can be placed in the conclusion. If results hold under some assumptions but not others, this suggests a caution to interpretation that should be shared with the users of the synthesis. Thus, the question to ask when evaluating the interpretation of results in research synthesis is:

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?

Data censoring. Every study does not have an equal chance of being retrieved by the synthesists, and, regardless of how thorough the literature search, it is likely that some studies were missed. Even after the careful planning, searching, and coding of research reports, missing data can influence the conclusions drawn from research syntheses. When data are systematically missing, not only is the amount of evidence gathered for the synthesists reduced, but the representativeness of the sampled elements may be compromised.

Pigott (1994) described three kinds of missing data that can result from data censoring. First, as noted above, entire studies may be unavailable to include in a data set. In particular, unpublished research findings are frequently missing from research syntheses. This form of data censoring is problematic because it frequently reflects a bias against the null hypothesis found in published research. That is, published articles are more likely to report statistically significant results whereas unpublished research is less likely to include statistically significant results.

Secondly, even if all relevant studies have been uncovered, individual studies may be missing information necessary to calculate an effect size. Missing effect sizes will occur when the primary researcher does not calculate them or does not report adequate information needed for the synthesists to calculate them. The consequence of missing effect sizes can be similar to missing an entire study. That is, a study with a missing effect size cannot be included in the estimate of the average effect. Consequently, the generality of the results may be limited to the sample of studies that had complete data. Further, similar to reasons why entire studies may be missing from a synthesis, effect sizes frequently go unmentioned in study reports when the tested relationship was not significant and the researchers fail to report the precise values of the means, standard deviations, statistical test, and/or *p* values.

Finally, information about study characteristics used to examine moderators of an effect may be missing from individual reports. For example, when examining the effect of ABE transition programs, particular evaluations may fail to report critical features of the program (e.g., how long it lasted, the qualifications of instructors) or characteristics of the participants.

There are a number of strategies that meta-analysts can use to deal with data censoring. Rothstein, Sutton, and Borenstein (2005) provide an in-depth treatment of numerous approaches. One way is to try to estimate the missing values using an imputation technique. Although imputing an estimate for missing values allows the meta-

analysts to include cases with missing data in the synthesis, data imputation methods force them to make assumptions that may not be accurate and can result in other types of bias. In addition, biases against the null hypothesis may affect the available samples of targeted elements as well as the sampled studies of the synthesis. To the extent that more retrievable studies are associated with particular subpopulations of elements (for which the treatment is more effective), the retrieval bias will restrict the accessible elements and this may restrict interpretation in a way that is not addressed through imputation of missing data.

Regardless, research synthesists are obligated to discuss how much data was missing from their reports, how they handled it, and why they chose to treat the missing data the way they did. Finally, it is becoming an increasingly common practice for meta-analysts with large amounts of missing data to conduct their analyses using more than one strategy (a form of sensitivity analysis) and determine whether their findings are robust across different missing data assumptions. Thus, an important question to ask about missing data when evaluating research syntheses is:

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's findings?

Specification and generalization. Research synthesis, like any research, involves specifying the targeted participants, program or intervention types, occasions, settings, and outcomes to which the results are hoped to apply. During interpretation, the synthesists must assess whether and how well each of the target elements is represented in the evidence base. For example, if the research synthesists were interested in making claims about the effectiveness of transition programs for all ABE participants, they would need to note whether important targeted groups were included or missing from the participant samples.

The trustworthiness of any claim about the generality of a research finding will be compromised if the elements in the realized samples are not representative of the target elements, be they people, programs, settings, times, or outcomes. Thus, research synthesists may find they need to respecify their covered elements once their data analysis is complete. For example, if only ESL students were used in studies of PSE transition programs, then any claims about the effectiveness of these programs either must be restricted to this particular type of participant or the rationale for extrapolation beyond the included types of participants must be provided.

The synthesists' influence on permissible generalizations is constrained by the types of elements sampled by primary researchers. Still, generalization in research synthesis injects a note of optimism into the discussion. There is good reason to believe

research syntheses will pertain more directly to the target participants, programs, settings, times, and outcomes—or to more subgroups within these targets—than will the separate primary studies. The cumulative literature can contain studies conducted on participants and programs with different characteristics at different times and in different settings using different outcome measures. For certain problem areas containing numerous replications, participants and circumstances accessible to the synthesists may more closely approximate the targeted elements than does any individual primary study. For example, if some studies of ABE transition programs contain only ESL students and others exclude ESL students, then the synthesist can ask whether transition programs effects were similar or different across the two types of participants, a question unanswerable by any individual study. If some studies examined transition services made available as part of a GED program while others evaluated the same services available during the first semester of community college, the research synthesists can test the robustness of the program effects across the different settings. Thus, the next question to ask when evaluating the interpretation of a research synthesis is:

17. Did the research synthesists discuss the generality and limitations of the synthesis findings?

Study-generated and synthesis-generated evidence. While the potential for testing the generality of findings is improved in research synthesis relative to individual studies, it is important to bear in mind a limitation of this type of evidence. Research syntheses can contain two different sources of evidence about the research problem or hypothesis. The first type is called study-generated evidence. Study-generated evidence is present when a single study contains results that directly test the relation being considered. Research syntheses also contain evidence that does not come from individual studies but rather from the variations in procedures across studies. This type of evidence, called synthesis-generated evidence, is present when the results of studies using different procedures to test the same hypothesis are compared to one another.

Any research problem or hypothesis can be examined through either study-generated or synthesis-generated evidence. However, only study-generated evidence based on experimental research allows synthesists to make statements concerning causality. For example, again suppose the research synthesists are interested in whether ABE transition programs have more of an impact on first-year PSE grades when delivered in one-on-one instruction rather than in groups. Suppose further that the literature search uncovered eight studies that used both types of instructional approaches and randomly assigned participants to one or the other. These studies provide study-generated evidence regarding the effect of mode of instruction. Here, if it is found that outcomes revealed more positive effects for one-on-one instruction, a conclusion would be warranted that the mode of instruction caused the difference.

Suppose instead that no studies manipulated the mode of instruction but, as in Table 4, four studies that experimentally manipulated transition instruction used one-on-one instruction and compared it to a no-treatment control while four others used group instruction. When the magnitude of the effect on PSE success is compared between the two sets of studies, it is discovered that the link is stronger in studies using one-on-one instruction. It could then be inferred that an *association* exists between mode of instruction and first-year PSE grades but it could not be inferred that a causal relation exists between the two.

When groups of effect sizes are compared within a research synthesis, regardless of whether they come from simple correlational analyses or controlled experiments using random assignment, the synthesists can only establish an association between a moderator variable—a characteristic of the studies—and the outcomes of studies. They cannot establish a causal connection. Synthesis-generated evidence is restricted to making claims only about associations and not about causal relationships because it is the ability to employ random assignment of participants that allows primary researchers to assume third variables are represented equally in the experimental conditions. The possibility of unequal representation of third variables across study characteristics cannot be eliminated in synthesis-generated evidence because experiments were not randomly assigned to study characteristics. Thus, it might be the case that the set of studies using one-on-one instruction were also conducted in community college settings, after participants had enrolled. All of the group-level-instruction studies might have been conducted as part of GED courses. Therefore, it is impossible to determine whether it was the variation in instruction or the settings of the instruction that "caused" the difference in the relationship between participation in the transition program and PSE first-year grades. The synthesists cannot discern which characteristic of the studies—or perhaps some unknown other variable related to both—produced the stronger link. Thus, when study characteristics are found associated with study outcomes, the synthesists should report the finding as just an association, regardless of whether the included studies tested the causal effects of a manipulated variable or estimated the size of an association.

Synthesis-generated evidence cannot legitimately rule out as possible true causes other variables confounded with the study characteristic of interest. Thus, when synthesis-generated evidence reveals a relationship that would be of special interest if it were causal, the synthesists should include a recommendation that future research examine this factor using a more systematically controlled design, so that its causal impact can be appraised. Therefore, the next important question to ask when evaluating the interpretation of evidence in a research synthesis is:

18. Did the synthesists make the appropriate distinction between study-generated and synthesis-generated evidence when interpreting the synthesis' results?

The substantive interpretation of effect sizes. Effect size estimates are of little value unless users can understand their substantive or practical, as well as statistical, significance. Cohen (1988) suggested some general definitions for small, medium, and large effect sizes in the social sciences. In defining these adjectives, he compared different average effect sizes he had encountered in the behavioral sciences. Cohen defined a small effect as $d = .2$ or $r = .1$, which he said his experience suggested were typical of those found in personality, social, and clinical psychology research. A large effect of $d = .8$ or $r = .5$ was more likely to be found in sociology, economics, and experimental or physiological psychology.

Because his contrasting elements were so broad and based on his personal reading of the social science literature, Cohen was careful to stress that his conventions were to be used as a last resort. In fact, there is no fixed scale for the interpretation of the size of an effect, and there is no substitute for knowing the research context of the specific substantive question. When interpreting the magnitude of effects, it is most informative to use contrasting elements that are more closely related to the topic at hand. Suppose, as in the fictional data displayed in Table 4, a meta-analysis of ABE transition programs found that the average effect was $d = .21$, indicating participants in ABE transition programs scored about two-tenths of a standard deviation higher on PSE first-year grades than nonparticipants. Using Cohen's guide, we would label this effect "small." However, other contrasting elements might be available to us. These might come from other meta-analyses that looked at entirely different ways to affect PSE success, such as the use of online tutorials. Thus, one way to interpret the effect of transition programs would be to ask whether they were more or less effective than the use of online tutorials. Or, other meta-analyses might share the same treatment, that is, look at transition programs, but vary in outcome measure; for example, some may have examined first-year grades while others looked at degree completion. Then, a good interpretation would consider whether transition programs have a larger effect on first-year grades than on degree completion. Of course, these types of interpretations could occur among results within the same research synthesis as well. When research synthesists cannot find meta-analyses closely aligned with their topic, they might find compendia of meta-analyses on more distant but related topics provided by Lipsey and Wilson (1993), and Meyers, et al. (2001) contain better contrasting elements than the Cohen guidelines.

In addition to multiple related choices of contrasting estimates, synthesists can assess how much any relation might be valued by consumers of research. This assessment involves the difficult task of making practical judgments about significance. So, for example, a *d*-index of .21 on PSE degree completion may be "small" when compared to Cohen's benchmarks and other contrasting elements. Still, the synthesists might argue that this improvement translates into an equivalent measure that suggests a practically important number of ABE students received degrees that would not have done so otherwise, since the increase in a lifetime of income for each student might be great (see Rosenthal, 1990, for a similar argument). It might then be argued that the cost of the

program was minimal relative to its change in the success potential of participants and the impact on their incomes. Levin and colleagues (Levin, 1987; Levin, Glass, & Meister, 1987) have laid out some ground rules for conducting this type of cost-effectiveness analysis for social programs.

Effect sizes also need to be interpreted in relation to the methodology used in the primary research. Thus, studies with more extensive treatments (for example, more frequent transition instruction), more sensitive research designs (within-subject versus between-subject designs), and measures with less random error can be expected to reveal larger effect sizes, all else being equal.

When interpreting the results of research syntheses, it is important that the meta-analysts address the issue of the magnitude of effect not simply by looking at its statistical significance. Thus, the next question to ask when evaluating the interpretation of effect sizes in meta-analysis is:

19. Did the meta-analysts (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?

## Stage 6: Present the Synthesis Methods and Results

Regrettably, a set of definitive guidelines for what needs to be reported in the write-up of a research synthesis, such as that provided for primary researchers by the American Psychological Association's *Publication Manual*, does not yet exist. Some efforts have appeared that help synthesists construct final reports. Rosenthal (1995) presents some sound advice about how to describe the methods and results of research synthesis. Bem (1995), Cooper (1998), Halvorsen (1994), and Light, Singer, and Willett (1994) also present numerous suggestions regarding effective presentation of research syntheses. However, the relative lack of reporting guidelines for synthesists is a problem because different editorial judgments create variation in whether particular aspects and results of syntheses are included in the report.

The division of a primary research report into four sections—introduction, methods, results, and discussion—should serve nicely as a structure for research syntheses, especially ones involving meta-analysis. The division of reports into these four sections serves to highlight the types of information that need to be presented in order for (a) readers to evaluate adequately the validity and utility of the synthesis and (b) those wishing to conduct replications of the synthesis to be able to do so. In fact, the questions presented earlier within each of the stages of research synthesis serve as a general guide for what information should be presented in each section of a synthesis report.

Briefly, the introduction to a research synthesis should present an overview of the theoretical, historical, and/or practical issues surrounding the research problem. It should present the conceptual definitions of the central variables. It should present a general description of the controversies to be resolved, and which of these will be the focus of the synthesis. The methods section should describe operationally how the synthesis was conducted. Most synthesis methods sections will need to present six sets of information:

1. the details of the literature search,
2. the criteria for including studies,
3. a description of the methods used in primary research,
4. how findings were judged to be independent,
5. the details of study coding (including what characteristics of studies were coded and with what reliability), and
6. the statistical procedures and conventions used to conduct the meta-analysis, if applicable.

While the results sections of a meta-analysis will vary considerably depending on the nature of the research topic and evidence, a general strategy for presenting results should include descriptive statistics about the literature, an overall effect size and measures of its dispersion and homogeneity, and the analysis of influences on effect size. Discussions typically contain at least five components:

1. a summary of the major results of the synthesis,
2. a description of the magnitude of the important effect sizes found in the synthesis and interpretation of their substantive meaning,
3. an examination of the results in relation to the predictions and other prior assertions made about relationships,
4. an assessment of the generality and limitations of any findings, and finally,
5. a discussion of topics that should be examined in future research.

So, the final question to ask when evaluating a research synthesis is:

20. Were the procedures and results of the research synthesis clearly and completely documented?

# CONCLUSION

Table 5 presents the twenty questions posed and discussed herein for evaluating the trustworthiness of research syntheses. These questions give expression to many principles of scientific research. The questions are written from the point of view of the reader of a research synthesis. However, synthesists also can use the list to assist them as they consider what procedures to use while carrying out their work. It should be no surprise that the suggestions to readers for evaluating a synthesis are isomorphic with the suggestions to synthesists for how their work should be conducted. Synthesists need simply to change the questions to first person. For example, Question 4 phrased for consumers is, "Is the problem placed in a meaningful theoretical, historical, and/or practical context?" Synthesists considering how to proceed with their work can read the questions as, "Did we place the problem in a meaningful theoretical, historical, and/or practical context?"

## Research Synthesis and Reporting Standards for Primary Research

There is yet a third audience that might find the twenty questions provide important guidance. This audience consists of primary researchers. Primary researchers ought to be concerned with the utility of their evidence for the next users of the data. It was noted above that missing data presents one of the most vexing problems faced by synthesists, and some data goes missing because it is omitted from reports of research. It cannot be expected that primary researchers will be omniscient about what aspects of their studies will be deemed critical by the next users, sometimes many years in the future. However, this does not mean that incomplete reporting of the principal evidence can be excused, for example, reporting F-ratios without accompanying means and standard deviations or reporting regression beta-weights in the absence of a correlation matrix displaying the primary relationships. There are certainly some aspects of research design and implementation that frequently go underreported. These include:

- Specific features of the independent variable or treatment (e.g., the frequency and duration of exposures, training and experience of administrators)
- Characteristics of the study's participants (e.g., age, ethnicity, special statuses)
- Characteristics of the research design (e.g., procedures for allocating participants to conditions, means for recruiting participants)
- Information on the reliability and validity of outcome measures
- Information on statistical outcomes that:
    - Describes all tests that were conducted
    - Includes measures of effect size or permit their calculation

Recently, two efforts have been undertaken to improve the quality of reporting of primary studies and to make reports more useful for the next users of the data. The first is called CONSORT (CONsolidated Standards Of Reporting Trials; http://www.consort-statement.org). CONSORT relates to the reporting of studies that performed random assignment of participants to conditions. It comprises a checklist and flow diagram to help standardize the way researchers report experiments. The flow diagram provides readers with a description of the progress of all participants in the study, from the time they are assigned to conditions until the end of study. The second is called TREND (Transparent Reporting of Evaluations with Nonexperimental Designs; http://www.trend-statement.org/asp/trend.asp). TREND presents a 22-item checklist developed to guide standardized reporting of quasi-experiments.

Both CONSORT and TREND were developed by teams of researchers from the medical and health sciences. However, use of these standards has spread to other disciplines within the social sciences and the application of more stringent rules for reporting primary studies should only become more widespread in the future. While it used to be the case that space limitations in print journals limited reporting of procedures and data, the availability of the internet has removed this barrier to full description. Many journals, including those published by the American Psychological Association, now provide their authors with auxiliary Web sites, where they can place complete descriptions of methods and results, materials used in the study, and even the raw data. Other authors provide this information on Web sites they construct themselves and reference in their written reports.

## The Relative Importance of the Questions about Research Synthesis

The twenty questions presented in Table 5 are not meant to be exhaustive. Other questions could be added. For example, the issue of whether sufficient statistical power to detect effects could be added, as this concern is relevant to meta-analysis as well as individual primary studies. Other questions could depend on the topic under consideration and the unique characteristics of the research being integrated. Also, while it could be argued that these are the twenty most general and important questions, the relative importance of the twenty questions might vary from application to application. For example, the clarity of conceptual definitions *might* be more important in theoretical work than applied work. The four questions specifically related to meta-analysis (Questions 11–13, 15) are important only when a quantitative synthesis is undertaken.

## Conclusion

During the past three decades, great strides have been made to transform research synthesis from a subjective exercise into a systematic, scientific process. When studies on

the same topic accumulate, the next users of this data must be held to the same standards of rigor as were the original data collectors. The twenty questions about research syntheses are based on the implicit assumption that the rigorous summary of findings across studies is no less important to the validity of research conclusions than to the conclusions of individual studies. Ultimately, the value of empirical evidence for guiding future development and implementation of social and educational programs rests on affirmative answers to these questions as well.

# REFERENCES


Alamprese, J. (2005). *Helping adult learners make the transition to postsecondary education*. Cambridge, MA: Abt Associates, Inc.

American Council on Education. (2000). *Who took the GED? 1999 Statistical Report.* Washington, DC: Author.

Bem, D. J. (1995). Writing a review article for *Psychological Bulletin*. *Psychological Bulletin, 118*, 172–177.

Best Evidence Encyclopedia. (2006). *Effective reading programs for English language learners and other language minority students*. Retrieved October 18, 2006, from: http://www.bestevidence.org

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis (Version 2.1) [Computer software]. Englewood, NJ: BioStat.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.

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

Camic, P. M., Rhodes, J. E., & Yardley, L. (2003). Naming the stars: Integrating qualitative methods into psychological research. In P. M. Camic, J. E. Rhodes, & L. Yardley (Eds.), *Qualitative research in psychology: Expanding perspectives in methodology and design* (pp. 3–16). Washington, DC: American Psychological Association.

Chalmers, I., Hedges, L.V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*, 12–37.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Confrey, J. (2006). Comparing and contrasting the National Research Council report on Evaluating Curricular Effectiveness with the What Works Clearinghouse Approach. *Educational Evaluation and Policy Analysis, 28*(3), 195–213.

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

Cooper, H. (1988). The structure of knowledge synthesis: A taxonomy of literature reviews. *Knowledge in Society, 1,* 104–126.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H. (2006). Research questions and research designs. In P. A. Alexander, P. H. Winne, & G. Phye (Eds.), *Handbook of research in educational psychology* (2nd ed., pp.849-877). Mahwah, NJ: Erlbaum & Associates.

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

Cooper, H. & Ribble, R. G. (1989). Influences on the outcome of literature searches for integrative research reviews. *Knowledge: Creation, Diffusion, Utilization, 10,* 179–201.

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76,* 1–62,

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

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

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

Fisher, R. A. (1932). *Statistical methods for research workers*. London: Oliver & Boyd.

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

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills: Sage.

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

Gleser, I. & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–356). New York: Russell Sage Foundation.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Halvorsen, K. T. (1994). The reporting format. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 425-438). New York: Russell Sage Foundation.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

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

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

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

Kerlinger, F. N., & Lee, H. B. (1999). *Foundations of behavioral research* (4th ed.). New York: Holt, Rinehart & Winston.

Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York: Guilford Press.

Levin, H. M. (1987). Cost-benefit and cost-effectiveness analysis. *New Directions for Program Evaluation, 34,* 83–99.

Levin, H. M., Glass, G. V., & Meister, G. R. (1987). Cost-effectiveness and computer-assisted instruction. *Evaluation Review, 11,* 50–72.

Light, R., & Pillemer, D. (1984). *Summing up: The science of research reviewing*. Cambridge, MA: Harvard University Press.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 439–453). New York: Russell Sage.

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

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

National College Transition Network. (2006). *A typology of ABE-to-college transition programs.* Boston: New England Literacy Resource Center/World Education, Inc.

Olkin, I. (1990). History and goals. In K. Wachter & M. Straf (Eds.), *The future of meta-analysis*. New York: Russell Sage Foundation.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). New York: Russell Sage Foundation.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

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

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis*, (pp. 163–176). New York: Russell Sage.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of Research Synthesis*, (pp. 301–322). New York: Russell Sage.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. West Sussex, England: Wiley.

Schmidt, F., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 261–282). New York: Russell Sage.

Shoemaker, P. J., Tankard, J. W., & Lasorsa, D. L. (2004). *How to build social science theories*. Thousand, Oaks, CA: Sage.

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

Stata.com. (2006). *Methods for meta-analysis in medical research*. Retrieved October 18, 2006, from: http://www.stata.com/bookstore/meta.html

Statistics.com. (2006). *Meta-analysis*. Retrieved October 18, 2006, from: http://www.statistics.com/courses/meta

Taveggia, T. C. (1974). Resolving research controversy through empirical cumulation. *Sociological Methods and Research, 2,* 395–407.

Tyler, J. (2001). *What do we know about the economic benefits of the GED? A synthesis of the evidence from recent research.* Providence, RI: Brown University.

U.S. Department of Education, National Center for Educational Statistics. (2002). *Digest of Education Statistics 2001* (NCES 2002–130). Washington, DC: Government Printing Office.

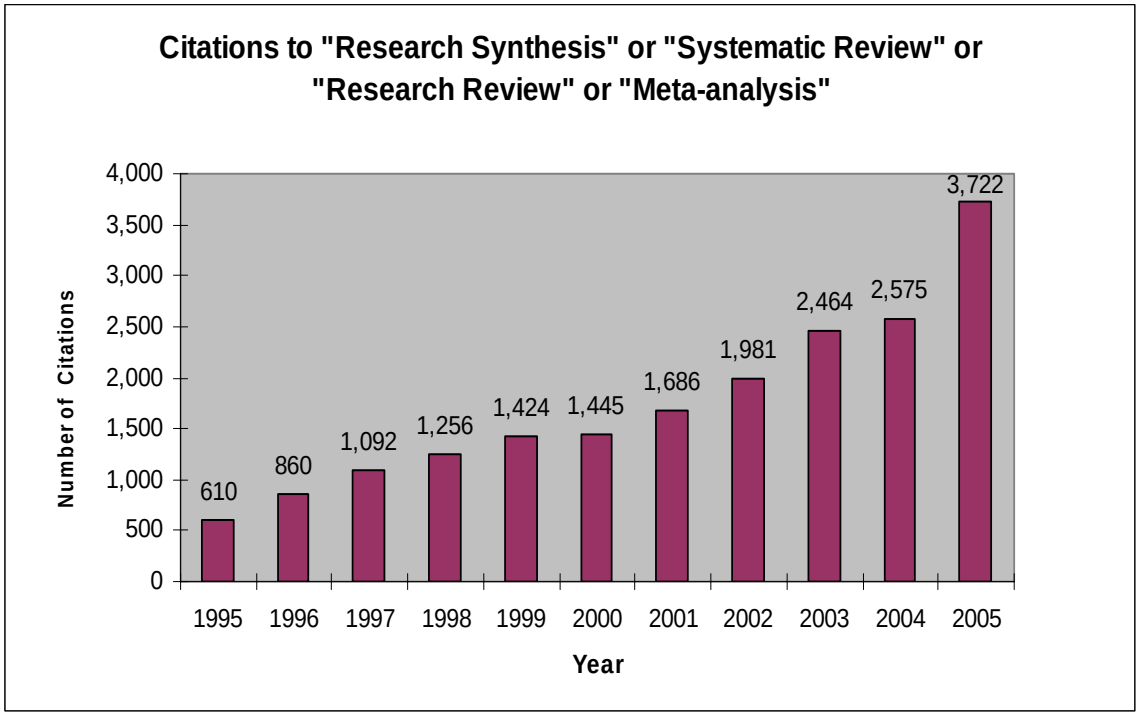
U.S. Department of Labor, Bureau of Labor Statistics. (2002). *Tomorrow's jobs* (Bulletin 2540–1). Washington, DC: Author.

Valentine, J. C., & Cooper, H. (2005). Can we measure the quality of causal research in education? In G. D. Phye, D. H. Robinson, & J. Levin (Eds.), *Experimental methods for educational interventions: Prospects, pitfalls and perspectives* (pp. 85–112). San Diego: Elsevier Press.

What Works Clearinghouse. (2006). *Review process*. Retrieved October 17, 2006, from: http://www.whatworks.ed.gov/reviewprocess/standards.html

# TABLES AND FIGURES

**Figure 1. Citations to "Research Synthesis" or "Systematic Review" or "Research Review" or "Meta-analysis"**




Note: Based on a search for terms included in the Web of Science, June 28, 2006.

## Table 1. Research Synthesis Conceptualized as a Research Process: Some Stage Characteristics

| Stage of Research Synthesis | Research Question to Ask at This Stage of the Synthesis | Primary Function Served in the Synthesis | Procedural Variation that Might Produce Differences in Conclusions |
|---|---|---|---|
| ***Define*** the Problem | What research evidence will be relevant to the problem or hypothesis of interest in the synthesis? | Define the (a) variables and (b) relationships of interest so that relevant and irrelevant studies can be distinguished | Variation in the conceptual breadth and detail of definitions might lead to differences in the research operations (a) deemed relevant and/or (b) tested as moderating influences |
| ***Collect*** the Research Evidence | What procedures should be used to find relevant research? | Identify (a) sources (e.g., reference databases, journals ) and terms used to search for relevant research and (b) extract information from reports | Variation in searched sources and retrieval procedures might lead to systematic differences in (a) the retrieved research and (b) what is known about each study |
| ***Evaluate*** Correspondence between Methods and Implementation of Studies and the Desired Synthesis Inferences | What retrieved research should be included or excluded from the synthesis based on (a) the suitability of the methods for studying the synthesis question and/or (b) problems in research implementation? | Identify and apply criteria to separate "correspondent" from "incommensurate" research results | Variation in criteria for decisions about study inclusion might lead to systematic differences in which studies remain in the synthesis |
| ***Summarize and Integrate*** the Evidence from Individual Studies | What procedures should be used to condense and combine the research results? | Identify and apply procedures for (a) combining results across studies and (b) testing for differences in results between studies | Variation in procedures used to analyze results of individual studies (e.g., narrative, vote count, averaged effect sizes) can lead to differences in cumulative results |

| Stage of Research Synthesis | Research Question to Ask at This Stage of the Synthesis | Primary Function Served in the Synthesis | Procedural Variation that Might Produce Differences in Conclusions |
|---|---|---|---|
| ***Interpret*** the Cumulative Evidence | What conclusions can be drawn about the cumulative state of the research evidence? | Summarize the cumulative research evidence with regard to its strength, generality, and limitations | Variation in (a) criteria for labeling results as "important" and (b) attention to details of studies might lead to differences in interpretation of findings |
| ***Present*** the Synthesis Methods and Results | What information should be included in the report of the synthesis? | Identify and apply editorial guidelines and judgment to determine the aspects of methods and results readers of the synthesis report need to know | Variation in reporting might (a) lead readers to place more or less trust in synthesis outcomes and (b) influence others' ability to replicate results. |

## Table 2. Characteristics of Some Channels for Locating Studies

| CHANNEL | HOW RESEARCH GETS IN | HOW SEARCHER GETS IN | HOW INCLUDED RESEARCH MIGHT DIFFER FROM "ALL RESEARCH"" | | | |
|---|---|---|---|---|---|---|
| | | | Greater Methodological Homogeneity | Greater Outcome Homogeneity | Disproportionate Representation of Published Studies | Missing Most Recent Research |
| INFORMAL CONTACTS | | | | | | |
| Personal Solicitations | Researcher must be known to searcher | Searcher must make inquiry targeted to specific researcher | ✓✓ | ✓(favorable to searchers' point of view) | | |
| Invisible Colleges | Researchers must be known to central and active researchers in a field | Searcher must be a known member of invisible college | ✓✓ | ✓(favorable to searchers' point of view) | | |
| Discussion Lists | Researcher must know about and subscribe to discussion list | Searcher must know about and subscribe to discussion list | ✓ | ✓(favorable to disciplinary point of view) | | |
| FORMAL CHANNELS | | | | | | |
| Professional Meeting and Conference Presentations | Research must pass weak peer review | Searcher must be aware of organization or meeting | ✓ | ✓(favorable to disciplinary point of view) | | |
| Personal Journal Subscriptions | Research may need to pass peer review | Searcher must subscribe to or read journal | ✓✓ | ✓(favorable to searchers' point of view) | ✓ | ✓ |

| CHANNEL | HOW RESEARCH GETS IN | HOW SEARCHER GETS IN | HOW INCLUDED RESEARCH MIGHT DIFFER FROM "ALL RESEARCH"" | | | |
|---|---|---|---|---|---|---|
| Research Report Reference Lists | Research must be known to article's authors | Searcher must subscribe to or read journal | ✓✓ | ✓(favorable to disciplinary point of view) | ✓ | ✓ |
| SECONDARY SOURCES | | | | | | |
| Research Registers | Compiler must be aware of research, by submission or search | Searcher must be aware that register exists | | | | |
| Reference Databases | Research must be in covered source | Must use appropriate search terms | | | ✓ | ✓ |
| Citation indexes | Research must be cited in publication | Must know article that cites research | | | ✓ | ✓✓ |

Note: Checkmarks are used to signify the strength of the potential bias, with two checkmarks denoting greater strength than one checkmark.

## Table 3. An Example of How Participant Attrition Might be Coded from Individual Study Reports

| CODED INFORMATION | INFORMATION CATEGORIES |
|---|---|
| Q1. What was the sample size of the ABE transition group at the start of transition intervention? | ___ ___ ___ (enter sample size) |
| Q2. What was the sample size of the comparison group at the start of the transition intervention? | ___ ___ ___(enter sample size) |
| Q3. What was the sample size of the ABE transition group at the completion of the transition program? | ___ ___ ___(enter sample size) |
| Q4. What was the sample size of the comparison group at the completion of the transition program? | ___ ___ ___(enter sample size) |
| Q5. What was the sample size of the ABE transition group for the analysis of this outcome measure? | ___ ___ ___ (enter sample size) |
| Q6. What was the sample size for the comparison group for the analysis of this outcome measure? | ___ ___ ___ (enter sample size) |
| Q7. Was there evidence that the groups experienced attrition for different reasons?<br><br>Note: Attrition is the loss of participants from groups. This question asks specifically about attrition for different reasons, not about whether attrition occurred at all. Calculation of attrition rates will be based on answers to the six questions on sample size. | 0 = no, the report says that groups did not experience attrition for different reasons<br><br>1 = yes, the report says that groups experienced attrition for different reasons<br><br>99 = NR, the report says nothing about attrition for different reasons |

**Table 4. An Example of *d*-Index Averaging, Calculating a Confidence Interval and Testing of Homogeneity**

| Finding | $n_{i1}$ | $n_{i2}$ | $d_i$ | $w_i$ | $d_i^2 w_i$ | $d_i w_i$ | *Grouping* |
|---|---|---|---|---|---|---|---|
| 1 | 23 | 25 | .90 | 10.88 | 8.81 | 9.79 | One-on-one |
| 2 | 42 | 46 | .51 | 21.26 | 5.53 | 10.84 | Group |
| 3 | 18 | 18 | .13 | 15.96 | 0.28 | 2.12 | Group |
| 4 | 32 | 45 | -.18 | 18.62 | 0.60 | -3.35 | One-on-one |
| 5 | 36 | 24 | .90 | 13.12 | 10.63 | 11.81 | One-on-one |
| 6 | 48 | 48 | .16 | 47.32 | 1.16 | 7.40 | One-on-one |
| 7 | 66 | 64 | -.35 | 32.00 | 3.92 | -11.20 | Group |
| 8 | 27 | 27 | .71 | 12.70 | 6.40 | 9.02 | Group |
| $\Sigma$ | 292 | 297 | 2.78 | 171.89 | 37.34 | 36.44 | |

$$d = \frac{36.44}{171.89} = .21$$

$$CI_{d\,95\%} = .21 \pm 1.96\sqrt{\frac{1}{171.89}} = .21 \pm .15$$

$$Q_t = 37.34 - \frac{36.44^2}{171.89} = 29.62$$

$$Q_w = 13.89 + 14.72 = 28.61$$

$$Q_b = 29.62 - 28.61 = 1.01$$

## Table 5. A Checklist of Questions Concerning the Validity of Research Synthesis Conclusions

DEFINING 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 variables' 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?

COLLECTING THE RESEARCH EVIDENCE

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

EVALUATING THE CORRESPONDENCE BETWEEN THE METHODS AND IMPLEMENTATION OF INDIVIDUAL STUDIES AND THE DESIRED INFERENCES OF THE SYNTHESIS

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

ANALYZING (INTEGRATING) THE EVIDENCE FROM INDIVIDUAL 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?

INTERPRETING THE CUMULATIVE EVIDENCE

15. Were analyses carried out that tested whether results were sensitive to statistical assumptions and, if so, where 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's findings?

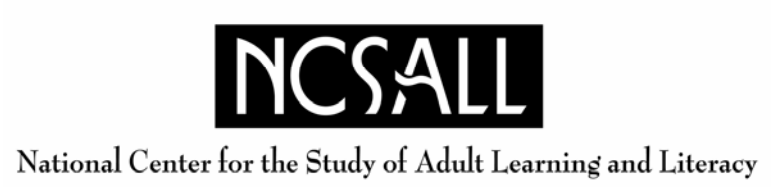
17. Did the research synthesists discuss the generality and limitations of the synthesis findings?

18. Did the synthesists make the appropriate distinction between study-generated and review-generated evidence when interpreting the synthesis's results?

19. Did the meta-analysts (a) contrast the magnitude of effects with other related effect sizes and/or (b) present a practical interpretation of the significance of the effects?

PRESENTING THE RESEARCH SYNTHESIS METHODS AND RESULTS

20. Were the procedures and results of the research synthesis clearly and completely documented?




**NCSALL's Mission**

NCSALL's purpose is to improve practice in educational programs that serve adults with limited literacy and English language skills, and those without a high school diploma. NCSALL is meeting this purpose through basic and applied research, dissemination of research findings, and leadership within the field of adult learning and literacy.

NCSALL is a collaborative effort between the Harvard Graduate School of Education, World Education, The Center for Literacy Studies at The University of Tennessee, Rutgers University, and Portland State University. NCSALL is funded by the U.S. Department of Education through its Institute of Education Sciences (formerly Office of Educational Research and Improvement).

**NCSALL's Research Projects**

The goal of NCSALL's research is to provide information that is used to improve practice in programs that offer adult basic education, English for speakers of other languages, and adult secondary education services. In pursuit of this goal, NCSALL has undertaken research projects in four areas: (1) learner persistence, (2) instructional practice and the teaching/learning interaction, (3) professional development, and (4) assessment.

**NCSALL's Dissemination Initiative**

NCSALL's dissemination initiative focuses on ensuring that practitioners, administrators, policymakers, and scholars of adult education can access, understand, judge and use research findings. NCSALL publishes *Focus on Basics,* a quarterly magazine for practitioners; *Focus on Policy,* a twice-yearly magazine for policymakers; *Review of Adult Learning and Literacy,* a scholarly review of major issues, current research, and best practices; and *NCSALL Reports* and *NCSALL Occasional Papers,* periodic publications of research reports and articles. In addition, NCSALL sponsors the Connecting Practice, Policy, and Research Initiative, designed to help practitioners and policymakers apply findings from research in their instructional settings and programs.

For more about NCSALL or to download free copies of our publications, please visit our Web site at:

**www.ncsall.net**