



## CHAPTER ONE

---

# CONCEPTUALIZING RESEARCH AND GATHERING STUDIES

---

Evan J. Ringquist

*"For many years, policy makers expressed increasing frustration with social science research. On every issue there were studies arguing for diametrically opposed conclusions. . . . In many areas meta-analysis has now provided dependable answers to the original research questions. Meta-analysis is now increasingly being used by policy makers, by textbook writers, and by theorists to provide the basic facts needed to draw both practical and explanatory conclusions"*

(HUNTER AND SCHMIDT 1996, 325).

In principle, meta-analysis differs little from traditional original research, particularly quantitative research employing survey methods. First, as with original research, conducting a meta-analysis begins with identifying a research question. Good research questions for meta-analysis share characteristics with good research questions for original research; they are theoretically grounded, specific, able to be precisely operationalized, and answered by measuring a particular quantity of interest. Second, as with original research, essential elements of designing a meta-analysis include identifying units of analysis, developing a strategy for sampling those units, and executing the sampling strategy. Of course, this analogy should not be taken too far. After all, the meta-analyst surveys studies, not human subjects. Moreover, the sampling strategy of the meta-analyst aims to identify the population of relevant studies, not a random sample of these studies. Third, similar to those doing original research, the meta-analyst must develop a survey instrument and use this instrument

to extract information from the sample necessary to answer the research question. In this chapter I discuss the essential elements of conceptualizing a meta-analysis and identifying the sample of original studies to be synthesized. Chapter 2 discusses the development of survey (that is, coding) instruments and their application to the original studies.

---

## Conceptualizing a Meta-Analysis

There are two steps in conceptualizing a meta-analysis in public management, public policy, and the social sciences. First, the researcher must identify the *quantity of interest* that will be estimated by synthesizing the results from original studies. Quantities of interest for a meta-analysis may be estimates of a particular program effect; for example, the quantity of interest for the meta-analysis in chapter 8 is the effect from using educational vouchers on student standardized test scores, while the quantity of interest for the meta-analysis in chapter 10 is the effect of public housing deconcentration on the life outcomes of housing assistance recipients. Quantities of interest for a meta-analysis may also be estimates of relationships or associations; for instance, the quantity of interest for the environmental justice meta-analysis used as an example in this book is the association between the residential concentration of poor and minority residents in communities and levels of potential environmental risk in communities, while the quantity of interest for the meta-analysis in chapter 11 is the relationship between the degree of public service motivation among employees in an organization and measures of organizational performance. Whether quantities of interest represent program effects or measures of association, they can only be identified for specific research questions.

Second, *researchers must develop a conceptual model* accounting for variation in estimates of the quantity of interest across studies. A hallmark of research in public management and policy seems to be that original studies provide different and sometimes incompatible answers to the same question. That is, they provide meaningfully different estimates of the quantity of interest. These differences pose a barrier to the accumulation of knowledge and also reduce the value of research for practice. A key difference between meta-analysis in medicine and meta-analysis in public management and policy is that the former places greatest emphasis on estimating the average effect size while the latter ought to place the most emphasis on modeling variation in effect sizes across studies.

## Research Topics, Research Questions, and Quantities of Interest

Quantities of interest (that is, effect sizes), can only be identified for specific research questions. Therefore, a key issue in meta-analysis (as with any empirical research) is the distinction between research *topics* and research *questions*. Research topics are subject focused, are typically vague, and implicitly suggest descriptive analysis. Examples of research topics include

- Environmental justice
- Educational vouchers
- Sustainable development
- Leadership strategies
- Housing discrimination

Research questions, however, are more precise, tend to focus on relationships between variables, and imply explanation or the measurement of effects. Examples of research questions matched to the previous research topics could include

- Are levels of environmental risk inequitably distributed with respect to race?
- Do educational vouchers work?
- Can micro-loans encourage sustainable development?
- Are different leadership strategies more effective under different circumstances?
- What accounts for housing discrimination?

***Research Questions in Meta-Analysis.*** Meta-analysis cannot synthesize evidence regarding research topics. Meta-analysis can only synthesize evidence regarding research questions. Identifying the research question, therefore, is the first step in any meta-analysis. In this respect, designing a meta-analysis is no different from designing a piece of original research. In both instances the investigation begins by articulating a specific and well-bounded research question.

The brief examples just mentioned are simplistic in that they suggest a one-to-one relationship between research topics and research questions. In fact, any research topic can give rise to several specific research questions.

Consider the research topic of educational vouchers. We could address this topic using any of the following questions:

- Does the use of vouchers improve academic performance?
- Why do some jurisdictions adopt voucher programs and others do not?
- Does the use of vouchers affect student attitudes and behavior?
- Why do parents seek vouchers?
- Why do parents or students refuse vouchers when they are offered?
- Are vouchers a less expensive method of providing public education?
- What effects do voucher programs have on private school tuition?
- What effect does competition from vouchers have on public school performance?

Any of these questions, along with others not listed, could serve as the target for a meta-analysis. One of the most important tasks for the meta-analyst is to accurately identify the most important or meaningful research questions on a particular research topic.

***Hypotheses in Meta-Analysis.*** Good research questions can be reconfigured as hypotheses. For example, the research question

- Does the use of educational vouchers improve academic performance?

can be transformed into the hypothesis

- Students using vouchers display higher academic performance.

Other good research questions have several hypotheses embedded within them. For example, the research question

- Why do parents seek vouchers?

becomes the set of hypotheses

- Parents seek vouchers to send their children to academically higher-performing schools.
- Parents seek vouchers to subsidize religious instruction for their children.
- Parents seek vouchers to place their children in a more homogenous school environment.

- Parents seek vouchers to send their children to schools with better athletic programs.

Meta-analysis is most commonly described as a set of techniques for combining the results from individual original studies. But meta-analysis cannot synthesize results from all studies examining a particular research question. To be combined in a meta-analysis, original studies must (1) employ quantitative data analysis and (2) test the same hypothesis or measure the same quantity of interest. Viewed from this perspective, meta-analysis is often a set of techniques for combining the results from a series of hypothesis tests. Properly identifying the most important hypotheses embedded within specific research questions, then, is as important as identifying the proper research questions in the first place.

***Conceptual and Operational Definitions in Hypotheses.*** Research questions can be transformed into two types of hypotheses. *Conceptual hypotheses* articulate the elements of the hypothesis in terms of the theoretical concepts or constructs of interest. For example, researchers in environmental justice investigate the relationship between the residential concentration of members of groups that traditionally have received inequitable treatment and the concentration of sources of environmental risk. When assessing the value of vouchers, we are interested in the relationship between the use of the voucher and student academic achievement. The difficulty with these conceptual hypotheses, however, is that they cannot be tested because the elements of the hypotheses cannot be measured directly. Therefore, researchers take conceptual hypotheses that are formed using theoretically important constructs and transform them into *operational hypotheses*, in which the elements of the hypothesis are represented as specific, measurable indicators. This is no less true for meta-analysis than it is for original research.

Most of the important theoretical concepts in the social sciences are unnatural, or constructed, or “imaginary” in the sense that they are not directly observable in the same ways as are many of the concepts in the physical sciences. Consider, for example, the physical science concept of “temperature” and the social science concept of “achievement.” Both concepts are socially constructed in the sense that physical scientists had to agree that “temperature” meant a quantity of kinetic energy or molecular motion in a sample of matter, and furthermore had to agree on a standard scale for measuring this quantity. “Achievement,” however, is quite different. First, achievement is a multidimensional concept in that it can mean

many things. Most social science concepts share this quality (for example, “economic development,” “representation,” “terrorism”). Second, unlike temperature, achievement does not exist outside of our imagination. Most social science concepts share this quality as well (for example, “organizational mission,” “ideology,” “equity”). Even when the physical and social sciences use the same concepts, they are often one-dimensional and directly measurable in the physical sciences but multidimensional and socially constructed in the social sciences (such as “gender”).

A great challenge for social scientists, then, is developing and implementing *indicators* that can measure these multidimensional and imaginary *concepts* (Cook and Campbell 1979). This *concept-indicator* problem has two elements that are especially relevant for meta-analysis. First, there is always some question as to whether our indicators actually measure the concept of interest. For example, there is a debate in economics as to whether Gross Domestic Product is a valid measure of national economic well-being. Second, since most of our measures are one-dimensional, even if they are a valid indicator of one aspect of the concept they may do a poor job of measuring the overall concept. For example, many educators believe that standardized test scores accurately measure mastery of certain materials and skills by students, but they also believe that these scores are overly narrow and provide an incomplete measure of the concepts of “student achievement” or “educational progress.”

An adequate appreciation of concept-indicator problems is essential when designing a meta-analysis. First, *the meta-analyst often defines the scope of the research by choosing to examine some indicators rather than others*. Original studies are included or excluded from a meta-analysis on the basis of whether the studies employ particular indicators of broader concepts. Table 1.1 offers examples of key concepts and their multiple indicators in environmental justice and school choice research. In the first row we see that educational vouchers are actually one operationalization (or indicator) of the broader concept of “school choice.” Even within the operational definition of educational vouchers there is variation; for example, vouchers may provide direct or indirect subsidies to the parents of school-aged children, and the use of vouchers may be restricted to secular schools or allowed to be used at religious schools. Similarly, the concept of student academic achievement can be operationalized using student standardized test scores, classroom grades, graduation rates, or college admission rates. The use of vouchers may affect any of these operational definitions of “achievement.” The scope of the educational voucher meta-analysis in chapter 8 is defined using these operational definitions. Specifically, we operationalize

**TABLE 1.1. CONCEPTUAL AND OPERATIONAL DEFINITIONS OF ELEMENTS OF HYPOTHESES**

<b>Conceptual Definitions (concepts)</b>	<b>Operational Definitions (indicators)</b>	<b>Sub-Indicators</b>
School Choice	Open enrollment	Inter-district Intra-district
	Charter schools	Private Public
	Educational vouchers	Direct payment Tax credits Religious
Academic Performance	Standardized test scores	Subject test Combined test
	Graduation rates College admission rates	
Target of Inequity	Race and ethnicity	Percentage of black residents Percentage of Hispanic residents Percentage of nonwhite residents
	Class	Median household income Percentage of households below poverty line Percentage of residents with high school degree
Environmental Risk	Risky facilities	Hazardous waste facilities Solid waste landfills Polluting facilities
	Pollution levels	Air pollution Water pollution TRI releases
	Ambient environmental risk	Accidental chemical spills Cancer risk

“school choice” using “educational vouchers” and operationalize “educational vouchers” using direct subsidies. Because of these choices, the meta-analysis can say nothing about the effects of open enrollment, charter schools, or private school tax credit programs (that is, indirect voucher subsidies) on student achievement. Chapter 8 also operationalizes “academic performance” using standardized test scores, so the meta-analysis tells us nothing about the effects from using vouchers on high school graduation or college admission rates.

Second, *meta-analysts often identify important moderator variables by considering carefully the various ways key concepts are operationalized in original research*. The third and fourth rows of table 1.1 illustrate the various ways the concepts of “inequity” and “environmental risk” have been defined in environmental justice research. Unlike in the student voucher example, these operational definitions were not used to exclude original studies from the environmental justice meta-analysis that provides a running example through the book. Instead, I created *moderator variables* (defined in the next section) that identified which effect sizes represented race-based or class-based inequities and which effect sizes measured the inequitable distribution of risky facilities versus pollution levels. These moderator variables were then used in meta-regression models to explain why different studies estimated different levels of environmental inequity.

The great variety of operational hypotheses in original studies presents a double-edged sword to the meta-analyst. On the one hand, meta-analysis can combine the results from original studies that, collectively, employ multiple operationalizations (or indicators) of the same concept. Unlike researchers conducting an original study, meta-analysts rarely have to worry whether their results are an artifact of the one-dimensional indicator used to measure the multidimensional concept. In this way, the meta-analyst can conduct robust hypothesis tests that are more faithful to the conceptual theory that generated these hypotheses. On the other hand, a meta-analysis that combines results from too great a variety of operational hypotheses runs the risk of being labeled with the “apples and oranges” critique. Moreover, the results from meta-analyses that employ a wide variety of indicators are often less useful for policy makers and managers looking to meta-analysis as an input to decision making. Sensible advice is that when researchers conduct a meta-analysis in order to synthesize the evidence regarding a particular theoretical expectation, they ought to include original studies that employ a broad range of operational definitions of the theoretical construct. By contrast, when meta-analysis aims to summarize the evidence regarding a particular policy or management intervention in order to inform decision making, operational definitions ought to be bounded more narrowly.

### Accounting for Variation in Effect Sizes

For most meta-analyses, the quantity of interest can be represented as a simple path diagram linking the focal predictor  $X$  and the dependent variables of interest  $Y$  as in equation 1.1.

$$\boxed{X} \xrightarrow{\Theta} \boxed{Y} \quad (1.1)$$



Meta-analysis measures the relationship between  $X$  and  $Y$ , represented by the arrow, using an effect size, represented as  $\Theta$ . In the environmental justice example,  $X$  represents the percentage of minority residents in communities,  $Y$  represents potential environmental risk in communities, and  $\Theta$  represents the degree of environmental inequity in the sample. In the educational voucher example,  $X$  represents the use of a voucher,  $Y$  represents student performance on standardized tests, and  $\Theta$  represents the differential test score gains attributable to the use of vouchers.

If  $\Theta$  was the same for all original studies, there would be no need for meta-analysis. Estimates of  $\Theta$  do differ between original studies, however, sometimes dramatically, even if these studies ask the same research question and employ the same operational hypotheses. An important benefit from meta-analysis is that it can help us understand the sources of this variation in the effect size  $\Theta$ , as well as help us obtain an estimate of the average effect size (or quantity of interest) across all studies. And an important aspect of designing a meta-analysis is developing a conceptual model that can help account for this variation in effect sizes.

***Moderator Variables as Explanations for Effect Size Variance.*** The effect size  $\Theta$  from path model 1.1 might vary across original studies for several reasons:

1. *Sampling Error.* All studies estimate the same population effect size  $\Theta$ , but effect sizes from individual studies ( $\Theta_i$ ) differ from  $\Theta$  because original studies employ different samples. In meta-analysis this type of variation is addressed using a fixed effects model.
2. *Non-Specific Error Variance.* All studies do not estimate a common population effect size  $\Theta$ , but they estimate the same expected value  $\mu_\Theta$ . Effect sizes from original studies differ both due to sampling error and due to unobservable factors. In meta-analysis, this type of variation is addressed using a random effects model.
3. *Variation in the Measurement of X (different indicators, not different scales).* Studies may employ different operationalizations of the focal predictor  $X$ . For example, in the environmental justice literature, some original studies measure the proportion of minority residents in a community using the percentage of black residents, while other studies measure this quantity using the percentage of nonwhite residents.
4. *Variation in the Measurement of Y (different indicators, not different scales).* Studies may employ different operationalizations of the dependent variable  $Y$ . For example, in the school voucher literature, some original studies measure student academic achievement using standardized math tests, while others measure this quantity using standardized reading tests.

5. Effect sizes may differ between studies due to differences in the research design or quality of the original studies.
6. Effect sizes may differ between studies due to what Rubin (1992) calls “scientifically interesting” factors—program design, implementation context, and so on.
7. Effect sizes may differ between studies due to differences in the model specification used to generate the effect size.

The sources of effect size variance described in reasons 1 and 2 are described in detail in chapter 3. The sources of effect size variance described in reasons 3 through 7 can all be operationalized as *moderator variables*, or measurable factors that can be used to predict or explain differences in the effect sizes from original studies that examine the same research question. Developing a conceptual model to account for variation in effect sizes across original studies, then, means thinking carefully about which factors might cause this variation, and how one might measure those factors using moderator variables.

**Using Moderator Variables in Meta-Analysis.** Chapter 2 describes a set of procedures for creating moderator variables when coding original studies. Most commonly, moderator variables are dichotomous, indicating that a particular effect size does or does not come from an original study possessing a particular attribute. For example, a moderator variable may take on a value of 1 for effect sizes measuring the effect of educational vouchers on student math test scores (reason 4 in the previous section). A different moderator variable might take on a value of 1 for effect sizes from original studies with random assignment of subjects to treatment and control groups (reason 5 in the previous section), and a third moderator variable may take on a value of one for effect sizes from original studies in which the sample is composed only of women (reason 6 in the previous section).

Moderator variables are sometimes used to identify subgroups of effect sizes that might best be analyzed in separate meta-regressions. For example, in my 2005 meta-analysis of the environmental justice literature I conducted separate meta-analyses for effect sizes measuring race-based and class-based inequities (an example of reason 3 in the previous section; see Ringquist 2005). More commonly, however, researchers in the social sciences will use moderator variables as predictors in meta-regression models. To illustrate, if we denote the three hypothetical moderator

variables described in the previous paragraph as  $M_1$ ,  $M_2$ , and  $M_3$ , we might estimate the meta-regression model

$$\Theta_i = b_0 + b_1 M_{1i} + b_2 M_{2i} + b_3 M_{3i} + e_i \quad (1.2)$$

where  $\Theta_i$  is the effect size calculated from the original study. In this model,  $b_1$  estimates how the average effect size from using vouchers is different for math scores,  $b_2$  estimates how the average effect size from using vouchers is different in studies with experimental designs, and  $b_3$  estimates how the average effect size from using vouchers is different for female students.

Chapters 4 and 5 discuss meta-regression in great detail. I introduce the technique here only insofar as it helps illustrate how researchers might consider conceptual models accounting for variation in effect sizes across studies. An important distinction between building empirical models in traditional analysis and building these models in meta-analysis is that while the former focuses on explaining the conditional distribution of the dependent variable  $Y$ , the latter focuses on explaining variation in the relationship between  $X$  and  $Y$ . Meta-analysts do not include moderator variables that help account for the expected value of the dependent variable  $Y$ . Instead, moderator variables are used *only if they affect in a meaningful way the relationship between the focal predictor  $X$  and the dependent variable of interest  $Y$*  (that is,  $\Theta$ ).

Moderator variables in meta-regression are akin to interaction terms in a traditional regression model. When building a conceptual model, then, a meta-analyst interested in the effects of educational vouchers on student performance does not need to identify factors that affect test scores, but rather needs to identify factors that might *moderate* the effect of vouchers on test scores. In environmental justice research, examples of moderator variables might include median household income in a community and property values in a community. Both are plausibly related to the location of polluting facilities, and both are correlated with the focal predictor. Models in original studies that include these variables, then, are likely to estimate a different relationship between the percentage of minority residents in communities and levels of potential environmental risk in communities.

When designing a meta-analysis, it is essential to think carefully about which moderator variables will be used to characterize the sources of variation in effect sizes across studies. This is especially true when using meta-regression, since it is the proper identification, operationalization, and utilization of moderator variables that allows us to combine highly

heterogeneous effect sizes in a single meta-regression model. If we exclude or misspecify important moderator variables—that is, if we conceptualize and operationalize the wrong model—the results from the meta-regression will be less helpful than they might otherwise be, and potentially may be misleading. The consequences of poor model misspecification are no less important in meta-analysis than in the conduct of original studies.

---

## Conducting a Literature Search in Meta-Analysis

A high-quality literature search is necessary for conducting a high-quality meta-analysis. The literature search identifies the original studies that will generate the effect sizes used in the meta-analysis. The design of the literature search also defines the theoretical and empirical universe that the meta-analysis characterizes. A poorly designed or poorly executed literature search, then, generates bad data and limits the theoretical and practical relevance of the research. The meta-analyst has two tasks in the literature search: finding the original studies that address the research question of interest, and judging whether those studies are acceptable for inclusion in the meta-analysis. These tasks must be completed using procedures that are systematic, transparent, and replicable.

While there have been relatively few meta-analyses in public management, public policy, or related fields, a significant proportion of the studies that have been conducted examine only the peer-reviewed or published literature (see, for example, Jarrell and Stanley 1990; Card and Krueger 1995; Stanley 1998; Stanley and Jarrell 1998; Lau, Sigelman, Heldman, and Babbitt 1999; Doucouliagos and Ulubaşoğlu 2008; but see Smith and Huang 1995). This reliance on the peer-reviewed or published literature is unfortunate for two reasons. First, the proportion of relevant original research appearing in the unpublished or grey literature is higher in the social sciences than in the medicine, psychology, and other areas where meta-analysis is more common (Grayson and Gomersall 2003; Rothstein and Hopewell 2009). While I am not aware of any systematic study of the question, the strong presence of research firms and think tanks in the fields of public management and policy makes it likely that the grey literature is probably even more important in these fields than in the social sciences writ large. Second, recent research in the statistics of meta-analysis shows that the tools for diagnosing and correcting publication bias *ex post* perform very poorly, and in general are untrustworthy (see chapters 6 and 7).

Meta-analysts are far better served controlling for publication bias *ex ante* than attempting to remedy this bias *ex post*. All meta-analyses in public management and policy, therefore, should employ an explicit, comprehensive, and systematic search of the grey literature.

## Identifying Original Studies

Before beginning a search of the literature, you need to know what you are searching for. Therefore, a good literature review begins with the design of the meta-analysis discussed in the previous section. Literature reviews in meta-analysis need to be motivated by a specific research question, and the literature search strategy needs to be built around one or more operational research hypotheses. Investigators that have less experience in conducting systematic searches of the empirical literature might consult either *The Oxford Guide to Library Research* (Mann 2005) or the *Information Retrieval Policy Brief* (Rothstein, Turner, and Lavenberg 2004) prior to designing their literature search strategy. It is also often helpful to discuss the research plan with a good reference librarian. Many university libraries have information specialists that focus on the fields of public management and policy, and many more have specialists in the social sciences. In addition, all readers are advised to consult the *Cochrane Handbook* (Higgins and Green 2008) for the most recent advice on best practices in conducting literature searches for meta-analysis. Finally, before beginning a literature search for a meta-analysis, researchers ought to conduct a search to identify any previous meta-analyses of their research question. Because nearly all meta-analysis articles include the phrase *meta-analysis* in the title or abstract, finding existing meta-analyses is usually fairly easy. One simply searches relevant databases using Boolean search terms for the keywords from the research question and the phrase *meta-analysis*. For example, in preparation for the environmental justice meta-analysis, I searched for previous meta-analyses using the Boolean combinations *environmental justice* and *meta-analysis*, *environmental racism* and *meta-analysis*, and *environmental equity* and *meta-analysis*. This strategy is not foolproof, however, so researchers should also use the search phrases *research synthesis*, *meta-analytic*, and *meta-regression* when searching for previous meta-analyses. Also, the Campbell Collaboration ([www.campbellcollaboration.org](http://www.campbellcollaboration.org)) maintains an online library of meta-analyses of topics relevant to public policy.

**Developing a Search Profile.** The first step in conducting a literature search for a new meta-analysis is to craft a *search profile*. A search profile consists of

(1) a set of keywords for identifying relevant original studies; (2) a set of authors that will be used to identify relevant original studies; (3) a strategy for using citation searches for identifying original studies and a set of criteria bounding the literature that will be searched using these keywords, authors, and citations; and (4) a systematic process for executing the search profile.

*Identifying Keywords.* The large majority of the literature search will use online databases, and keywords are necessary for searching these databases. Good initial choices for keywords include words or terms that characterize the specific research question or the operational hypotheses that motivate the meta-analysis. Keywords identified in this manner are often referred to as *natural language* keywords, and natural language keywords are often combined in a literature search using “Boolean Operators.” For example, scholars often use the phrases *environmental justice*, *environmental equity*, and *environmental racism* when considering the question of whether sources of potential environmental risk are distributed inequitably with respect to the race or class of community residents. The relative frequency with which these terms are used in original research varies across researchers and disciplines, and using only one of these terms to guide the literature search would miss a sizable portion of original studies in these fields. The environmental justice meta-analysis, then, used all three sets of keywords in searching electronic databases. In addition, we tried various other keyword combinations, including *race* and *environmental risk*, *class* and *environmental risk*, *race* and *pollution*, *racism* and *pollution*, and so on.

I can offer several lessons that are applicable when identifying natural language keywords for a literature search.

1. The keywords should be identified by reading several original studies that examine the research question of interest, not simply from your own imagination. Use keywords that authors of original studies use to characterize their own research.
2. The process of identifying keywords is iterative. Meta-analysts typically posit a set of keywords, then bring these keywords to the literature search engines and assess the results of the search. If the keyword search returns few relevant studies or many unexpected results, this is a sign that the keywords require revision. Alternatively, if the keyword identifies many of the important studies that examine the particular research question, this keyword is a candidate for the list of final

keywords and phrases. After reading a handful of studies and viewing the results of searches using preliminary keywords, the meta-analyst will posit a new set of keywords and repeat the process. On some occasions, the meta-analyst will identify a useful new keyword or key phrase well into the literature search process. When this happens, the meta-analyst must go back and repeat all previous literature search tasks with the new key phrase. That is, the search profile must be applied when searching all literature sources, with no exceptions.

3. Keywords should almost always be combined with the Boolean operator “and” when conducting natural language searches. For example, above we combined *race* AND *pollution* when conducting the literature search for original studies in the area of environmental justice.
4. Researchers should search for keywords and key phrases in the titles of original studies, in the abstracts of these studies, and in the full text of these studies when the search engine offers these options. One goal of the literature search is to be as comprehensive as possible, and relevant original studies might be missed by searching only study titles and abstracts.
5. In my experience an excellent literature search can be completed using relatively few keywords and key phrases. Often, six keywords and key phrases is enough—so long as the researcher is using the right keywords! One way of assessing the adequacy of the list of keywords and phrases is the number of unique records that are returned when trying out a new keyword. If new keywords do not return records that were not also discovered by previous keywords, the keyword list is probably sufficient.
6. It is vitally important to keep a comprehensive list of all keywords and key phrases that have been tried and the results from searches using these keywords. Good record keeping prevents repetition in the search for good keywords, helps to identify gaps in the types of keywords that have been tried, and encourages researchers to choose the final group of keywords and phrases on the basis of the results that they produce.

Meta-analysts can also conduct literature searches using *constructed vocabulary* as opposed to natural language. Constructed vocabulary keywords refer to the official subject categories used by research databases to organize the studies in their archives. For example, the online reference databases maintained by EBSCO (available at most university libraries) organizes entries using *Sears List of Subject Headings*. Meta-analysts may want to consult the subject headings used by various search engines, select



those that match most closely the research question and operational hypotheses motivating the meta-analysis, and use these subject terms to search the databases.

*Identifying Authors.* A second element of designing a search profile is identifying a relatively large list of prominent authors who have conducted original studies that are relevant to your research question. The meta-analyst then contacts these authors directly in the hope that they might (1) share conference papers, working papers, or other unpublished original studies relevant to the meta-analysis; (2) identify conference papers, working papers, or other unpublished original studies completed by others; (3) identify new, relatively unknown scholars working on the same research question (these new scholars are often the Ph.D. students of the scholars you contact); or (4) identify reports or published studies that you may not have come across in your keyword searches. In conducting the educational voucher meta-analysis in chapter 8 we contacted seven leading authors in the field of educational voucher research.<sup>1</sup>

*Leveraging Citations.* The literature search profile should also identify studies using citation searches. The “ancestry” method can identify additional relevant original studies by examining the sources cited by relevant studies that have already been uncovered using either keyword searches or by contacting authors. Most readers are familiar with this tactic. In addition, the Web of Knowledge (a suite of online literature search tools available at most university and many public libraries) allows researchers to conduct a *forward citation search* that identifies all studies that cite a relevant original study that has already been identified in the literature search. For example, using the Web of Knowledge I identified twenty-four studies that cite Mohai and Saha’s 2007 article examining inequities in the location of hazardous waste facilities. I can then examine these twenty-four studies to determine whether any are relevant candidates for the meta-analysis. While the ancestry method is useful for identifying older studies, the forward citation map tool is helpful for identifying the most recent relevant studies on a research question.

*Bounding the Literature Search.* Finally, the search profile identifies explicitly the criteria that will be used to bound the search for relevant literature. Three very common bounding criteria are time, geographic area, and language. For example, both the environmental justice meta-analysis and the educational voucher meta-analysis considered only studies that examined



environmental inequities or educational vouchers in the United States (a geographic bounding criterion) and studies that were written in English (a language bounding criterion). These bounding criteria were used as a matter of convenience, since many studies have examined both the degree of environmental inequity and the effects of educational vouchers in other countries, and reported the results of these studies in languages other than English. Neither study employed temporal bounding criteria, though some meta-analyses restrict the literature search to studies written after a particular year or within a particular time frame. While the use of bounding criteria makes the literature search more tractable, it does open the meta-analyst to criticisms that her results are unrepresentative of the entire population of empirical literature examining a particular research question.

*Applying the Search Profile.* Once a search profile has been created, it must be applied in an identical fashion across all aspects of the literature search. For example, all keywords and key phrases must be used to search each of the electronic reference databases discussed in the next section. The meta-analyst should not use one set of keywords to search one reference database and a second set to search another. Moreover, it is a good idea to apply the keywords in the same order when searching each research archive or search engine. Similarly, all contacted authors should be asked to provide the same information—author A should not be asked to share information about her own work, while author B is asked only to provide information about the work of others. In addition, the same bounding criteria should be applied to all aspects of the literature search. By crafting a systematic literature search profile and applying this profile in the same manner in all aspects of the literature search, researchers help ensure that their review of the literature will be comprehensive, transparent, and replicable.

*Developing a Search Strategy for the Published Literature.* With a search profile in hand, the meta-analyst brings this profile to the published literature and uses it to identify original studies that are relevant for the meta-analysis. The distinction between “published” and “unpublished” literature is less meaningful in public management and policy than it is in many other fields. While an article that appears in *Public Administration Review*, the *Journal of Policy Analysis and Management*, or the *American Journal of Political Science* is unambiguously part of the “published” literature, what of a report issued by the Government Accountability Office or the U.S. Department of Health and Human Services? These reports are

certainly “published” in that hard copies of the reports can be obtained from the U.S. Government Printing Office. From the perspective of meta-analysis, however, these government reports are part of the “unpublished” literature, since they do not appear in scholarly publication outlets. The same is true for original studies conducted by policy research firms such as Mathematica Policy Research and policy think tanks such as the Brookings Institution. Yet research from sources like these makes up an important part of the relevant empirical literature in public management and policy. In addition, online journals are beginning to make headway in public management, public policy, and the related social science disciplines (for example, *The Economist's Voice*), and it is unclear whether original studies in these outlets ought to be counted as “published” or “unpublished” studies. Rather than labeling studies found in the outlets described here as “unpublished,” I use the broader and increasingly common term *grey literature*. Strategies for identifying original studies in the grey literature are discussed in the next section.

In this book I employ the traditional definition of published research as *studies published in peer-reviewed scholarly journals or books*. It is far easier to search the published literature than it was even ten years ago. A large number of comprehensive and specialty research archives are available and easily searchable online. Many of the most useful of these online research archives are listed in table 1.2. These research archives and search engines range from the general (such as EBSCO and ProQuest) to the field-specific (for example, Psych-INFO and PAIS). Some archives specialize in the most current material (such as Lexis-Nexis Academic Universe) while others restrict access to older studies (for instance, JSTOR). While most of the research archives and search engines listed in table 1.2 place a strong emphasis on domestic (U.S.) publication outlets, a few focus explicitly on providing access to international journals and books (for example, IBSS and BLDSC). Finally, some archives specialize in journal articles (such as JSTOR) while others specialize in books (for instance, WorldCat).

As recent as five years ago, the coverage of many of the research archives in table 1.2 was limited to the most recent decades. Currently, however, nearly all have extended coverage back through the 1960s, and the coverage of many extends back nearly a century. Because meta-analysis summarizes the results from original studies that employ statistical models and other quantitative techniques, the longer time frame covered by these archives is a luxury that we cannot take advantage of, since quantitative empirical research in the fields of public management and policy rarely goes back

**TABLE 1.2. COMMON SOURCES FOR PUBLISHED AND GREY RESEARCH  
IN PUBLIC MANAGEMENT AND POLICY**

Peer-Reviewed Articles	Books	Government Reports	Research Firms and Think Tanks	Working Papers	Conference Papers	Dissertations	Archives of Grey Literature
ProQuest	WorldCat	US GAO	Mathematica	NBER	Proceedings First (EBSCO)	ProQuest	SIGLE
Lexis-Nexis	IBSS	US CBO	MRDC	SSRN	Papers First (EBSCO)	WorldCat	ESRC
SSCI		US CRS	Rand	EconLit	Sociological Abstracts	Index to Theses (UK)	Evidence Network
IBSS		NTIS	IBM Business of Government	NCEE	EconLit	EconLit	C2-RIPE
JSTOR		NRC	Brookings			Sociological Abstracts	
PsychINFO		EBSCO Government Collection	AEI				

TABLE 1.2. (Continued)

Peer-Reviewed Articles	Books	Government Reports	Research Firms and Think Tanks	Working Papers	Conference Papers	Dissertations	Archives of Grey Literature
ERIC		Individual Agencies	Urban Institute				
Wilson Education Index			Heritage Foundation				
EBSCO			Resources for the Future				
PAIS							
International							
BLDSC							
Business Source Premier							
NCJRS							
Sociological Abstracts							
Education Abstracts							
Social Work Abstracts							
Urban Studies Abstracts							
Criminal Justice Abstracts							

more than fifty years. Readers should be aware, however, that many of the larger research archives and search engines offer differing levels of access or service (for example, EBSCO and Lexis-Nexis). This means that the results from a literature search using the same keywords may differ for researchers at institutions purchasing a lower level of access—often smaller public or liberal arts colleges—and researchers at institutions purchasing a higher level of access. Researchers are well advised to be certain of the level of access their institution provides to the online resources, and to conduct their literature searches in an environment that provides the highest level of access possible to these sources.

Finally, one research archive is notably absent from table 1.2: Google Scholar. While Google Scholar is an extraordinarily comprehensive search engine, in my experience it is of limited value to meta-analysts. Google Scholar is comprehensive because it is so undiscerning. When conducting a meta-analysis, a researcher must strike a balance between searches that are sufficiently comprehensive and searches that bury the researcher in a mountain of “false positive” results; in other words, records identified by the search that are not relevant to the meta-analysis. A meta-analyst using Google Scholar will be swamped by the number of false positive “hits” returned by any search. For example, a Google Scholar search using the key phrase *environmental justice* returns over 500,000 records published between 1990 and 2010. Searching a half million records from a single key phrase is simply not feasible in a meta-analysis. The advanced search function in Google Scholar does not improve the situation in practice. Searching for studies only in the social sciences and humanities that include the phrase *environmental justice* in the text returns over 17,000 records. If we remove the restriction to social sciences and humanities, the search returns over 20,000 records. By contrast, this same search in JSTOR returns just over 3,200 records, and a search using Academic Search Premier (EBSCO) returns 2,083 records. While searching through 3,200 records is a large task, it is far more manageable than searching through 20,000 records. Using the advanced search function, we can limit the Google Scholar search to requiring that the search phrase occurs only in the article title. Exercising this restriction reduces the number of records returned to 3,550. Requiring that the phrase *environmental justice* occur in the article title, JSTOR returns 175 records and Academic Search Premier returns 725 records. Moreover, while both JSTOR and Academic Search Premier allow one to search article abstracts, Google Scholar does not. This example gives the readers some sense of the scope of the task facing the meta-analysts in a literature review.

***Developing a Search Strategy for the Grey Literature.*** As discussed previously, many original studies relevant to meta-analyses in public policy and management come from the grey literature. For example, a majority of the effect sizes used in the educational voucher meta-analysis in chapter 8 come from original studies that did not appear in scholarly journals or books. For the environmental justice meta-analysis, just over 20 percent of effect sizes come from studies in the grey literature. A meta-analysis in public management and policy that relies only on the published literature, then, is likely to generate erroneous or nongeneralizable conclusions. In this section I identify nearly one dozen different types of outlets that ought to be considered when searching the grey literature for original studies in public management and policy. Examples of these sources are listed in table 1.2.

- *Government Reports.* The federal government and some state governments release reports that might be considered relevant original studies for meta-analyses in public management and policy. These reports can come from dedicated governmental research entities such as the Government Accountability Office (GAO) or the Congressional Research Service (CRS), from cabinet-level departments or other large federal agencies (for example, Department of Health and Human Services, Environmental Protection Agency), or from quasi-governmental advisory organizations (for instance, National Academy of Sciences, National Academy of Public Administration, Transportation Research board).
- *Reports from Public Policy Research Firms or Think Tanks.* A remarkable number of high-quality original studies are produced by public policy research firms (such as Mathematica Policy Research, the RAND Corporation). Public policy think tanks also generate a remarkable volume of studies that are potentially relevant for meta-analyses in public management and policy. The quality of reports from think tanks is more variable than the quality of reports from policy research firms. Moreover, unlike most policy research firms, many policy think tanks are explicitly ideological (such as the Heritage Foundation), while others are more non-partisan (such as the Brookings Institution). Even the most highly partisan policy think tank, however, can produce a high-quality relevant original study. Meta-analysts should always search the websites of the policy research firms and think tanks most active in studying their research question when conducting a meta-analysis.
- *Working Papers.* The meta-analyst has access to several large archives of working papers in the fields of public management, public policy, and

related disciplines. Some of these working paper archives are sponsored and managed by the federal government (such as National Bureau of Economic Research [NBER] and National Center for Environmental Economics [NCEE]). Other working paper archives are managed by nonprofit organizations (for example, Social Science Research Network [SSRN], EconLIT).

- *Conference Papers.* Proceedings and full-text conference papers from thousands of professional conferences are available via Papers First and Proceedings First, which cover every conference and symposium archived by the British Library Document Supply Center. Access to Proceedings First and Papers First is available through many university libraries in the United States. In addition, professional organizations in some disciplines provide access to conference proceedings in their fields (for example, Sociological Abstracts). Finally, many individual professional organizations are providing access to conference programs, and even conference papers, through their websites (for example, the National Association of Schools of Public Affairs and Administration [NASPAA], the Midwest Political Science Association [MPSA]).
- *Dissertations and Theses.* It was not too long ago when researchers looking for relevant Ph.D. dissertations and master's theses had to pour through microfiche from University Microfilms International at the University of Michigan. Thankfully, the entire UMI database of dissertations and theses is searchable online through ProQuest. One can also search for dissertations online using WorldCat and other sources, and search for dissertations filed in the United Kingdom using the "Index to Theses" listed in table 1.2.
- *Archives of Grey Literature.* Rothstein and Hopewell (2009) identify a small number of online archives that focus explicitly on providing access to the grey literature. All of these archives focus on research released in Europe. These archives are included in table 1.2.

Researchers should employ the search profile when searching online repositories of grey literature in the same manner as when searching online archives of the published literature.

## Judging the Acceptability of Original Studies

**Classifying Studies.** After employing the research profile to conduct a literature search, the next task is to judge the acceptability of the records

returned from that search. That is, the meta-analyst must review the hundreds or thousands of studies identified using the techniques just described and determine which of these studies are acceptable for the meta-analysis. I find it useful to pass the records returned from the literature search through metaphorical “sieves” that employ different criteria to eliminate studies that are not acceptable for the meta-analysis. These “sieves” place records into four different categories.

1. *Hits*. “Hits” means all records returned by the literature search. These include records from the published and grey literatures and records recommended by contacted authors. The literature search for the environmental justice meta-analysis returned over 5,000 hits, while the literature search for the educational voucher meta-analysis returned 6,815 hits. These numbers illustrate that conducting a literature review for a meta-analysis is not for the faint of heart.
2. *Potentially Relevant Studies*. “Potentially relevant studies” are identified using only the titles and other bibliographic information returned in the literature search. Common criteria for excluding hits as not potentially relevant is if they are published in popular outlets (such as newspapers or magazines), are book reviews or opinion pieces, or were identified by the literature search because they contained all of the keywords in the Boolean search but in a nonsensical order. In general, researchers should use liberal or catholic criteria when placing studies into the “potentially relevant” group, because so little information is used to exclude studies at this stage. False positive decisions regarding potentially relevant studies can be corrected in a later stage of the literature search process, but false negative decisions exclude these studies permanently. The environmental justice meta-analysis identified 297 potentially relevant studies, and the educational voucher meta-analysis identified 736 potentially relevant studies.
3. *Relevant Studies*. “Relevant” studies are identified by examining the study abstract or, if available, a study summary. On some occasions determining the relevance of a study requires obtaining the full text of the study and reading the first few pages or examining the tables and figures. Potentially relevant studies are excluded from the smaller group of relevant studies if they (1) are non-analytic (that is, descriptive), (2) are nonquantitative, (3) examine dependent variables or focal predictors that are measured in a manner inconsistent with the operational hypotheses motivating the meta-analysis as defined earlier in the chapter, or (4) do not meet the bounding criteria from the



literature search profile. For example, a potentially relevant study in the area of environmental justice would be excluded if it was a case study of the siting of a solid waste landfill, or a law review article examining the potential present in current environmental statutes to remedy environmental inequities. A potentially relevant study in the area of educational vouchers would be excluded if it examined the effect of vouchers on student satisfaction, or if it examined the effects of vouchers in Chile. Relevant studies, then, are those that examine the specific research question of interest, use statistical analysis to test the operational hypotheses motivating the meta-analysis, and meet the relevant bounding criteria. Potentially relevant studies should be excluded from the relevant category only if it is clear from the abstract or study summary that the study is not relevant. If there is a question as to the relevance of the study, the meta-analysis should obtain more information (in other words, the full text of the study) or classify the study as relevant. The environmental justice meta-analysis identified 88 relevant studies, and the educational voucher meta-analysis identified 84 relevant studies.

4. *Acceptable Studies.* “Acceptable” studies are identified by examining the full text of the study. Relevant studies can be classified as unacceptable for a number of reasons. First, relevant studies might be unacceptable because they were incorrectly categorized as relevant. That is, while the study might have appeared relevant when examining only the study abstract or study summary, it is clear when reading the full text that the study does not meet one or more of the relevance criteria. Second, the study may not report sufficient statistical detail to allow the calculation of effect sizes. As we see in chapters 2 and 3, we actually need very little information to calculate effect sizes. To calculate the most useful  $r$ -based effect size, for example, all that is needed are the sample size and a measure of statistical significance. Still, a surprising number of original studies fail to report even this level of detail in their statistical results. Third, a study can be excluded as unacceptable if the results in that study perfectly duplicate the results in one or more other studies already deemed to be acceptable. It is not uncommon, for example, for a literature search to identify both the grey literature version of a research paper (for example, a conference paper, NBER paper, or SSRN paper) and the published version of that same paper. In many cases, the results in these two versions of the paper are identical. Researchers may employ other context-specific criteria for determining that a relevant study is unacceptable, but these

idiosyncratic reasons for excluding relevant studies should be used sparingly. The full environmental justice meta-analysis identified 49 acceptable original studies (only 48 of these studies are represented in the data set used in subsequent chapters), while the educational voucher meta-analysis identified 33 acceptable studies.

Acceptable studies are subsequently coded to extract the information necessary to calculate effect sizes and to create the necessary moderator (and mediator) variables. The process of coding acceptable original studies is covered in chapter 2, while the process of calculating effect sizes from these studies is addressed in chapter 3.

***Evaluating the Reliability of Classification Decisions.*** It should be evident that conducting the literature search for a meta-analysis requires considerable time and effort. It should also be evident that assigning records (or original studies) to one of the four categories described in the previous section requires the researcher to exercise considerable judgment. For both of these reasons, it is best if the meta-analysis is conducted by a team of researchers. The burden of the literature review can be split among members of the team. More important, members of the team can assign studies to categories independently, and the reliability of these assignments can be assessed. Assessing the degree of intercoder reliability in identifying relevant and acceptable studies is an integral part of a meta-analysis literature search.

Social scientists have developed a handful of measures for assessing the reliability of independent assessments of events or the reliability of repeated measures of a concept. The two most common measures of intercoder reliability in meta-analysis are the *percentage agreement* and *Cohen's kappa*. Consider table 1.3, which represents the decisions of two members of a meta-analysis research team, Coder *A* and Coder *B*. Each coder must make a decision as to whether each of a hundred potentially relevant studies is relevant, or which of a hundred relevant studies is acceptable (intercoder reliability assessments are rarely conducted for identifying potentially relevant studies). Cells *A* and *D* in table 1.3 reflect studies on which Coders *A* and *B* agree, and cells *B* and *C* reflect studies on which Coders *A* and *B* disagree. The percentage agreement is calculated using  $(A + D)/(A + B + C + D)$ , or  $70/100 = .70$ .

While the percentage agreement is the most commonly reported measure of intercoder reliability in meta-analysis, this measure has been criticized for overstating intercoder reliability because it ignores the probability

**TABLE 1.3. OUTCOMES OF HYPOTHETICAL LITERATURE SEARCH DECISIONS**

	Coder <i>B</i> Accept	Coder <i>B</i> Reject
Coder <i>A</i> Accept	<i>A</i> [40]	<i>B</i> [10]
Coder <i>A</i> Reject	<i>C</i> [20]	<i>D</i> [30]

that the raters would agree due to chance. Cohen's kappa ( $\kappa$ ) controls for this chance agreement. The formula for Cohen's kappa is

$$\kappa = (PA_o - PA_c) / (1 - PA_c) \quad (1.3)$$

where  $PA_o$  is the observed relative frequency of agreement, or the standard measure of percentage agreement, and  $PA_c$  is the relative frequencies of agreement predicted due to chance. To calculate  $PA_c$  we note that Coder *A* accepted 50 percent of the studies while Coder *B* accepted 60 percent of the studies. Assuming these decisions are made independently, the probability that both Coder *A* and Coder *B* would agree to accept a study due to random chance is  $(.5)(.6) = .3$ , and the probability that they would agree to reject a study due to random chance is  $(.5)(.4) = .2$ . The overall probability of random agreement  $PA_c = .3 + .2 = .5$ . We would calculate Cohen's kappa as  $\kappa = (.7 - .5) / (1 - .5) = .40$ . By accounting for chance agreement, Cohen's kappa will always be smaller than the simple percentage agreement measure. Orwin (1994) offers the following rules of thumb for evaluating the extent of intercoder agreement using Cohen's kappa:

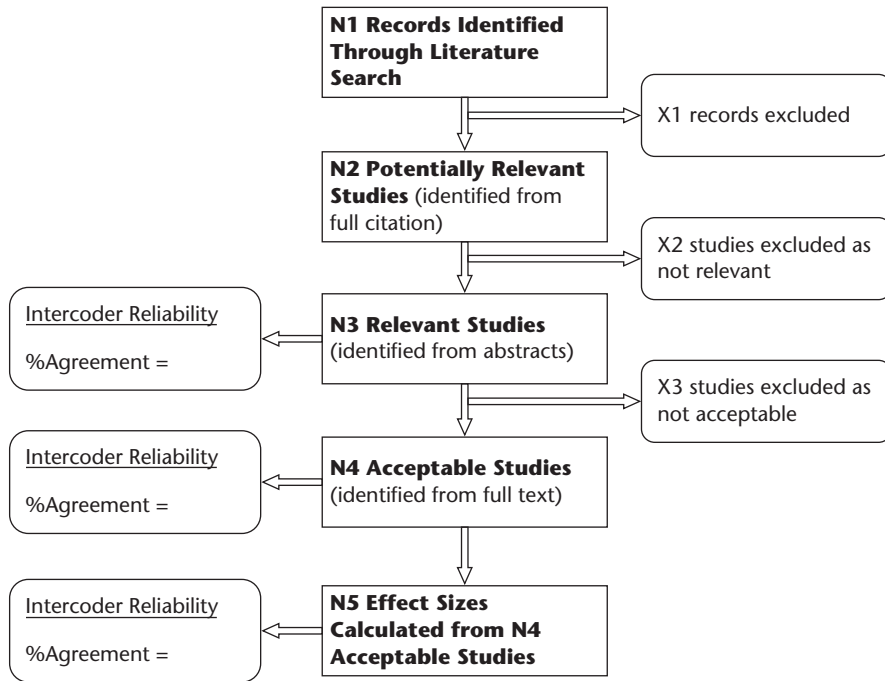
$.40 < \kappa < .59$	fair agreement
$.60 < \kappa < .74$	good agreement
$.74 < \kappa$	excellent agreement

Standard practice in meta-analysis is converging toward the reporting of both measures of intercoder reliability, and this is the approach we follow in chapters 8 through 11.

## Summarizing the Literature Review

The *Cochrane Handbook* (Higgins and Green 2008) recommends that meta-analyses report the results from decisions made during the literature search

**FIGURE 1.1. SAMPLE STUDY FLOW DIAGRAM FOR REPORTING THE RESULTS FROM THE LITERATURE SEARCH**



in a format that is clear and easily followed. Figure 1.1 provides one such format, the study flow diagram. Each step in the literature review is represented by one cell in the flow diagram, and the number of original studies assessed at each stage and the decisions made regarding each study are reported in these cells. In addition, measures of intercoder reliability at each stage are reported on the left-hand side of the study flow diagram. Note that the last cell in the study flow diagram reports the number of effect sizes coded from the acceptable studies. We address coding effect sizes in chapter 2.

## Conclusion

Researchers experienced with conceptualizing and designing original quantitative research will find that these skills transfer directly to meta-analysis. Meta-analysts must begin their inquiry with a theoretically

grounded, specific, and well-bounded research question. Meta-analysts then gather data required to estimate quantities of interest and test hypotheses that allow them to answer this research question. Meta-analysts in public management and policy, more so than their counterparts in medicine and psychology, focus on accounting for variation in the quantity of interest across original studies. As with original research, this variation provides leverage for explanation. Finally, as with original research, meta-analysts should strive to design and execute studies that are replicable. Toward that end, I strongly encourage that researchers maintain an archive of hard copies or electronic copies of all acceptable studies identified in the literature search and used in the meta-analysis (standard practice is to also include a list of these studies as an appendix in the published version of the meta-analysis). Questions that arise later in the meta-analysis, and questions from others after the completion of the meta-analysis, are much more easily addressed when these studies are close at hand.

---

## Note

1. A list of the authors we contacted in this manner includes Joshua Cowen, Jay Greene, Alan Krueger, Jonathan Plucker, Cecelia Rouse, John Witte, and Patrick Wolf.