WESTERN CRIMINOLOGY REVIEW

Western Criminology Review 8(1), 1–15 (2007)

Systematic Reviews and Meta-Analyses: Best Evidence on "What Works" for Criminal Justice Decision Makers*

Anthony Petrosino
Learning Innovations at WestEd

Julia Lavenberg
University of Pennsylvania

Abstract. *With the current emphasis on making "evidence-based policy," criminal justice policymakers today are under more pressure to use research in their decision making. Systematic reviews and meta-analyses can provide policymakers with reliable and comprehensive evidence about what works to reduce crime or improve justice. It is important that decision makers become more familiar with this method. In this article, we present a non-technical summary of systematic reviews. After discussing the need for different evidence to respond to different questions, we examine some of the challenges in locating "evidence." A common method for reviewing literature—the narrative or traditional synthesis—contains a number of methodological flaws that have contributed to the current emphasis on rigorous or systematic reviewing techniques. We consider two policy-relevant examples of systematic reviews addressing popular justice programs (Scared Straight and D.A.R.E.) and conclude with the argument that systematic reviews and meta-analyses offer the most useful information to decision makers who want to base their decisions on "what works" rather than ideology, tradition, politics, or anecdote.*

Keywords: decision making; policy; evaluation; systematic review; meta-analysis

Introduction

Policy initiatives based on rigorous evidence are strongly encouraged within the field of crime prevention today (Coalition for Evidence-Based Policy, 2003). Determining what kinds of evidence should drive decision making about policing, courts, corrections, neighborhood prevention, and other domains for intervention is challenging. Evaluation studies come in all forms, vary on many dimensions, and sometimes conflict. It is tempting to pick out the study that seems most influential or important, and use that to guide decision making. A single experiment certainly can be influential, and may provide good answers to decision makers in the jurisdiction in which it was implemented. If widely publicized, the study may spur other researchers to conduct a new wave of theoretical and methodological studies. But it seems sensible that an evidence-based approach to what works in crime and justice should go beyond the selective consideration of one or a few influential studies.

Systematic reviews can greatly assist policymakers in identifying effective programs and interventions and are considered an important tool among those who advocate evidence-based policy (Davies, 1999; Nutley, Davies, and Tilley, 2000). In systematic reviews, researchers attempt to gather relevant evaluative studies, critically appraise them, and come to judgments about what works using explicit, transparent, state-of-the-art methods. In contrast to traditional syntheses, a systematic review will include detail about each stage of the decision process, including the question that guided the review, the criteria for studies to be included, and the methods used to search for and screen evaluation reports. It will also detail how analyses were done and how conclusions were reached.

Systematic reviews have much to recommend them. Their foremost advantage is that when done well and with full integrity, they provide the most reliable and comprehensive statement about what works. Such a final statement, after sifting through the available research, may be "we know little or nothing—proceed with caution." This can guide funding agencies and researchers toward an agenda for a new generation of evaluation studies. This

* This work was partially supported by a Mellon Foundation grant to the Center for Evaluation, American Academy of Arts & Sciences, a National Institute of Justice grant to the Rutgers, School of Criminal Justice, and a Smith-Richardson Foundation grant to the Jerry Lee Center of Criminology at the University of Pennsylvania. We thank the late Frederick Mosteller, David Weisburd, James Finckenauer, Todd Clear, Ronald V.G. Clarke, Paul Lerman, and anonymous peer reviewers for comments on various renditions of this paper. All comments, however, are solely the responsibility of the authors and not of any institution, funding agency, or other person.

can also include feedback to funding agencies where additional process, implementation and theory-driven studies would be critical to implement.

Systematic reviews have other byproducts. By demonstrating irreconcilable conflicts, they go beyond the obvious "more research needed" to provide a specific research agenda. Because each primary study report is scrutinized, systematic reviews can underscore deficiencies in report writing and lead to better systems for collecting the data that is required by reviewers, including guidelines for editors to use before publishing original research. Reviews also ensure that relevant evaluations—which may have been ignored and long forgotten—are eternally used to respond to inquiries about what works. It is satisfying to investigators to find their study still considered twenty years or more after completion.

In his 1997 book, science writer Morton Hunt explained how the results from meta-analysis contradicted the conclusions drawn by earlier reviewers using traditional methods. For example, he wrote that quantitative estimates from meta-analyses of correctional treatment studies consistently show more positive effects for intervention on recidivism than earlier reviews. One of the reasons that meta-analyses came to different conclusions is that this method took into account the actual size of the effect reported in the study, rather than using statistical significance as the sole criterion for judging whether a program worked or not. In contrast to the pessimistic findings in earlier narrative reviews, such as those reported by Bailey (1966), Logan (1972) and Martinson (1974), meta-analyses across all areas of social, psychological and educational treatment have established that intervention generally has a small, positive—but non-trivial—effect on measured outcomes (Lipsey and Wilson, 1993). Palmer (1994) noted that meta-analyses of correctional interventions have helped to somewhat counter the prevailing pessimism generated by earlier reviews.

This article will present information about why we believe systematic reviews have a distinct advantage over other types of information for making policy decisions related to crime prevention. We begin by acknowledging the need of different evidence for different questions and address some of the challenges policymakers face when attempting to locate the evidence they need. We then propose systematic reviews as a solution to these challenges, discussing the various purposes and types of reviews, along with the limitations of each. Two relevant examples in criminal justice are presented. We conclude with a discussion of the benefits of using systematic reviews and meta-analyses in policymaking and lay out an agenda on how these syntheses could become a centerpiece of evidence-based decision making in criminal justice.

Different Evidence for Different Questions

Policymakers need a wide range of information to inform their decision making. These needs require different types of scientific evidence (Boruch, 1997). To identify the scope and severity of a problem, for example, epidemiological data from sample surveys or trend data from official government statistics and reports (such as crime rates based on the Federal Bureau of Investigation's *Crime in the United States: The Uniform Crime Reports*) are available. These data respond to the question: "What is going on?" To obtain information on risk factors that lead some children to become criminals while others become law-abiding citizens, one would look to etiological studies (e.g., longitudinal studies that follow children to adulthood). These types of studies answer questions such as "how did this problem occur?" Determining "what works" to reduce crime, however, requires a different type of scientific evidence. In this case, data from outcome or summative evaluations, or those studies that have tested the impact of some intervention on an outcome measure of crime, are necessary. The question under consideration, then, drives the type of evidence required for an answer.

Evaluation studies have an advantage over drawing conclusions about whether a program works based on anecdotal evidence. 'Stories' are important, but they are prone to bias. It is rather easy, for example, to find a compelling story or anecdote to demonstrate that an intervention worked, or conversely, that it failed miserably. Personal experience with a program might also result in skewed views about what works. One of Rossi's (1987) lessons from his experience with program evaluation was that staff and clients invariably will love the program they are participating in. The objective data will not support their enthusiasm, and when the report is issued, the evaluator will not be invited to dinner! Evaluation studies, however, aim to reduce bias by systematically testing the effects of an intervention using social science methods. Thus, evaluation reports provide the evidence policymakers should seek when requiring information on the effectiveness of crime prevention and other justice programs.

Challenges to Finding Evidence

Although an evidence-based approach is strongly endorsed within the field of crime prevention today, sev-

eral challenges face the policymaker who desires to use evidence in decision making. Among these challenges are information overload, fragmentation of research across fields, difficulty of locating reports beyond those published in peer review journals, unevenness in methodological quality among studies, and selective use of evidence by advocacy groups.

Research, like other information, is now being disseminated through various outlets. The Internet and World Wide Web make a wide range of research information and reports from around the globe—some of it of questionable quality—available in seconds. Although the indisputable benefit of this progress is that more information is easily accessible to a broader audience, the negative impact is that too much information is produced for anyone to comprehend and stay abreast of. In addition to the challenges inherent in sifting through the enormous volume of information available, research relevant to criminology is often found in divergent fields. Besides criminological journals, periodicals in sociology, public health, psychology and education routinely publish studies relevant to criminal justice.

Yet, despite these technological advances and burgeoning publication sources, some evaluation reports remain difficult to find. A rather large number of evaluation studies are located in what Sechrest and his colleagues (1979) call the "fugitive literature" or in what Hopewell and colleagues (2006) refer to as the "grey literature." The term "fugitive literature" is especially appropriate to use in criminological circles because the documents are so difficult to identify and retrieve, much like criminals on the lam. Such studies, however, are part of the "evidence" to consider. Examples are governmental reports, doctoral dissertations and master's theses, conference papers, technical documents, studies done in foreign countries, and other literature that is not published in readily accessible sources. Lipsey (1992), in his review of delinquency prevention and treatment studies, found that approximately four of ten were reported outside of journals or academic presses.

Some may argue that unpublished studies are of lesser quality because they were not published in peer-reviewed scientific journals. Such an assertion, at the very least, ignores the high quality evaluations done by private research firms. For example, Greenberg and his colleagues (1999) reported that Abt Associates, a private research firm in Cambridge, Massachusetts, conducted over 25 percent of the randomized trials in social market effects (e.g., employment programs). Many of these were never published in journals, most likely due to the fact that evaluators in entities such as Abt Associates do not have the organizational incentives to publish in peer-review journals as professors and university-based researchers do.

Even if all evaluation reports were easily accessible, there is great variation in the type of design and quality of methods used. More often than not, the results across studies of the same intervention will differ, sometimes substantially so, and it is likely that some of that difference is due to methodological characteristics of the studies (Lipsey, 1992). DiIulio (1991) suggests that this methodological variation provides easy fodder for special interest groups and politicians to exploit; he contends that rigorous evaluations such as randomized experiments provide far less leeway and are not as easily exploited. This point dovetails with Hacsi's (2002) finding from case studies in educational evaluation research, in which proponents and opponents selectively used evidence, regardless of its quality, to support their presupposed positions on matters such as whether the federal government should support Head Start (preschool) or reductions in average class size.

Policy-relevant questions such as "what works to reduce crime in communities?" or "are there effective programs in reducing offender recidivism?" are not easily answered. The studies that bear on these questions are often scattered across different disciplines, are sometimes disseminated in obscure or inaccessible outlets, and can be of such questionable quality that interpretation is risky at best. Compounding these challenges, political and special interest groups selectively use evidence to promote a particular position. How then can policy and practice be informed by such a fragmented knowledge base, comprised of evaluative studies that range in quality? What study, or set of studies, if any at all, ought to be used to influence policy? What methods should be used to appraise and analyze a set of separate studies bearing on the same question? We believe that systematic reviews and meta-analyses provide policymakers with the best evidence, however imperfect, to guide their decision making.

Sources and Presentation of Evaluation Studies

Policymakers often obtain their information on what works from media outlets such as daily newspapers, weekly periodicals such as *Newsweek*, and television news shows such as *60 Minutes* (Weiss and Singer, 1988; Forsetlund and Bjorndahl, 2002). Media outlets typically report findings from a single study (the latest and, assumedly, the greatest). In reality, only a few of the presumably thousands of studies relevant to crime and justice conducted each year receive any media atten-

tion. Their notoriety may be due to a number of factors: a well-known investigator or research firm may have conducted the study, study results are controversial or go against conventional wisdom, powerful advocacy groups have seized the findings to advance their agenda, or the researchers have used a public relations approach to disseminate their work (Weiss and Singer, 1988). This study can then become the definitive work among the policy and practice community about "what works." A study reported in the news media and reaching a wide audience is more likely to change perceptions about the nature of the problem and the effectiveness of the intervention than is one reported in obscure scholarly journals reaching a narrow set of academicians (Weiss, Murphy-Graham, and Birkeland, 2005).

Regardless of whether a study is reported in the media or in a peer-reviewed journal, it may be possible that decision makers will have to act upon a single study because that is all the scientific evidence available. For example, many advocate quality preschool programs as a crime prevention strategy (Wilson and Hernnstein, 1986), but these recommendations are based primarily on the results of a single long-term evaluation that examined the effects of the Perry Preschool curriculum on the arrest records of the children who participated (Schweinhart, 1987). Following children from preschool for up to 25 years is an expensive proposition, and this is most likely the reason why only one study on the effects of quality preschool on crime has been reported.

Relying upon one or even a few studies *if others are available* is problematic. For example, if only one study (e.g., the Perry Preschool experiment) has been reported and we rely on it to make judgments about what works, we are relying on 100 percent of the available evaluation research. If five similar studies have been conducted, relying on only one study means that we draw on only 20 percent of the available evidence (Cook et al., 1992). Increase it to 20 relevant studies—and we would rely upon only 5 percent of the available evidence!

It is possible, however, that one study does represent the other studies quite well. Or, it may be that the one study is the very best of all those conducted. Studies in a particular area sometimes do converge, but in other cases, they conflict. A particular study, or even a few studies, may be unrepresentative of all the evidence. Any assertion that a study represents the 'norm' remains unsupported unless all relevant studies are examined.

Roberts (2000) underscores the importance of taking all studies into account in his review of medical evaluations of a blood plasma solution, known as human albumin, in treating the critically ill. He reviewed randomized trials testing the effects of albumin on the subsequent mortality of patients. Some of the studies, particularly those that were publicized in the medical literature, showed that albumin was successful in reducing mortality among patients. But Roberts (2000) made a concerted effort to locate all of the relevant clinical trials, particularly those that never reached the journals. His review indicated that, on average, albumin increased the mortality of seriously-ill patients relative to doing nothing at all. He notes that British newspapers soon carried stories about his review, and estimated that the use of albumin cost 500 lives a year in the United Kingdom (Roberts, 2000). The pharmaceutical companies that manufactured albumin were unhappy, as the U.K. government soon issued guidelines against the treatment, leading to plummeting sales of albumin. Using only one or a few of those published studies might have led Roberts or anyone else to conclude that albumin was effective.

In criminal justice, Sherman and Berk (1984) conducted the seminal Minneapolis Domestic Violence Experiment, reporting that arresting misdemeanor domestic violence offenders was the most effective option for police, compared to the traditional strategies of separating the offender and victim for eight hours or attempting an informal mediation between the parties at the scene. If policymakers were to rely solely upon the Minneapolis study, many jurisdictions would continue to mandate arrest for police officers responding to misdemeanor (non-felony) domestic violence calls. In fact, the number of departments adopting such a policy after the Sherman and Berk (1984) report was staggering (Sherman and Cohn, 1989). There have now been five replications of the Minneapolis study and serious questions have been raised about whether arrest is an effective response to all misdemeanor domestic violence cases (Sherman 1992). To conclude that arrest "works" on the basis of the earlier Minneapolis experiment without taking into account the results of these subsequent replications seems misinformed.

Learning what works requires more than examining the isolated results of one or two evaluations. Lipsey (1997) noted that each evaluation study is part of a cumulative "brick-building" process in constructing knowledge about interventions and implementation. The only way this information can be mined is by identifying the accessible studies, analyzing them for what they tell us, and gleaning new discoveries from them. In short, this process is known as knowledge building or accumulation. But how do we accumulate knowledge from separate but similar studies? The method used to systematically examine separate but similar studies is the research review.

Research Reviews

Reviews typically summarize a number of different reports to draw conclusions (Khan et al., 2001). Of course, almost every individual report contains some type of review to frame the current study or argument. These literature reviews are typically rather brief, as they are not meant to be the focus of the report. Our definition of a research review is a report that goes beyond a cursory synthesis and focuses on the results of prior studies in order to draw conclusions from them.

We note that reviews may serve many different purposes. For example, researchers may conduct *critical reviews* in which they use a series of available empirical studies to highlight certain important issues upon which they would like to focus. Canadian researchers Ross and Price (1976) did exactly this in their review of research on behavioral modification programs. They covered a multitude of issues, including the lack of evidence on effectiveness, the type of clients who should be served, and how institutions were currently managing behavioral modification. In critical reviews, the research evidence is selectively used to highlight crucial issues.

Reviews can be written to provide *state-of-the-art* reports. In contrast to reviews in which critical issues are identified, state-of-the-art reviews often take the form of a discussion of recent studies in order to document advances made in dealing with a particular problem. Farrington (1994) provides an example of such a review in the area of early developmental and childhood prevention. He drew upon the findings of several recent evaluation studies to show that programs that featured components like visiting the homes of expectant mothers from impoverished areas can be beneficial. State-of-the-art reviews can bring us up to speed on policy and practice innovations, and inform us about recent program victories or failures. The focus of this type of review is to illustrate what is possible, and what successes have been reported.

Comprehensive reviews cover a wide range of studies in order to address multiple, related issues. Textbooks for college studies often contain this type of review, skimming the most influential studies in a variety of areas but not delving into any one too deeply (Oxman and Guyatt, 1988). Some of the more influential reviews in criminology and justice are like this. For example, the University of Chicago publishes the annual volume, *Crime and Justice: An Annual Review of Research*. Although published by a different press, this series is very similar to the *Annual Review* publications in psychology, sociology, and public health (see www.annualreviews.org). In short, each volume usually contains a series of comprehensive, multi-interest articles that summarize research to draw conclusions about a number of different issues. Rarely is the focus of those papers solely on the effectiveness of a particular intervention, or set of interventions.

Government task forces, or quasi-government bodies such as the National Academy of Sciences, often issue large, comprehensive syntheses on a wide range of interests. The goal of such reviews is to discuss pertinent policy, practice and research issues relevant to the topic, rather than summarily conclude what works. Effectiveness may be one of the score of issues addressed. Available studies are used to selectively highlight certain points. These reviews can be important. In the case of the National Academy of Sciences, they are approved by a panel of diverse members, including experts on opposing sides of issues (Weiss, personal communication). They sometimes represent a strong consensus statement, and politicians are comfortable using them for agenda setting. But since reports like those issued by National Academy panels cover lots of ground, they are not normally designed to provide a definitive answer about "what works." When they do include such material, just as textbooks, it is given only a very cursory treatment, sometimes relying on the most recent or well-known evaluation studies.

It is important for policymakers to identify the purpose of a review before using it as a source of information and decision making, as well as understand the types of research reviews commonly seen in the literature. Research reviews designed to find out "what works" generally are one of two types: traditional or systematic.

The Traditional or Narrative Review of "What Works"

There is a half-century of history in criminology of trying to pull together scientific evidence from separate but similar studies into a single review (see Kirby, 1954, for an early example). The earliest reviews, though sometimes remarkable in their exhaustiveness, generally used narrative or qualitative methods in coming to conclusions. Reviewers often read studies and used some type of unknown and inexplicit process of reasoning (i.e., what Bushman and Wells [2001] called 'mental calculus') to determine what works or did not. This is not to say that the process was based on nefarious motives, ill will, or unscientific principles. Usually these reviewers made judgments on the basis of whether the study was believable according to methodological factors like internal validity.

Methods for analyzing separate but similar studies

have a century of application. It was not until the 1970s, however, that methods for conducting reviews were scrutinized the same way that methods for surveys and experiments have always been. This was ironic, as some of the most influential and widely-cited papers across fields were literature reviews (Chalmers, Hedges, and Cooper, 2002). But from this increased scrutiny, three major areas of criticism of the traditional or narrative review evolved.

One set of criticisms focused on the narrative review's *lack of explicitness*. Most reviews suffered from a lack of details about how the reviewer conducted the research. Information about why certain studies were included and others excluded was often missing. The report of the review often did not describe what searches were done in order to find evaluation studies. Reviewers sometimes provided more weight to a few studies over others, but did not provide the criteria for making such judgments. Ironically, these same reviewers rarely tolerated the same lack of explicitness in reports they included in their own reviews! In the end, the reader of most narrative reviews was forced to accept and trust the reviewer's expertise rather than put the conclusions to test.

Because of the lack of explicitness, it was difficult for the serious reader to determine how the reviewer reached conclusions about what works. This includes the criteria used to judge an intervention's success. Consider the debate over the conclusions in the Lipton, Martinson, and Wilks (1975) summary of over 200 correctional program evaluations, briskly reported first by Martinson (1974). Despite finding that nearly half of the evaluations reported in Martinson's paper had at least one statistically significant finding in favor of treatment, his overall conclusions were gloomy about the prospects of correctional intervention. The criterion for success was not readily known, but it must have been incredibly strict (Palmer, 1975).

A second set of criticisms focused on the *methods used in the reviews*.[1] Most of the reviewers did not attempt to control for problems that could potentially bias their review toward one conclusion or another. At its worst, a reviewer advocating a position could selectively include only those studies favoring that viewpoint. For example, a reviewer in favor of strict gun control laws could ignore evaluations that report little effect for such laws. Far more likely than intentional distortion was how narrative reviewers failed to deal with *potential* biases. For example, some reviewers examining what works may rely on easy-to-get journal articles as the only source for reports of evaluations. But research in other fields suggests that relying on journal articles can bias the results toward concluding that interventions are more effective than they really are (Berlin, Begg, and Lewis, 1989). This is because researchers in those fields were found to be more likely to submit their manuscripts to journals when they find a positive result—and more likely to bury the manuscript in their file drawer when they do not (Berlin et al., 1989).

Because the rules of scientific rigor and explicitness are not applied with equal force to the narrative review, the reviewer runs the risk of selectively including and excluding studies (Cooper, 1989; Wolf, 1986). Selection bias in literature reviews can lead to different published conclusions, as illustrated by the vast differences across sex offender treatment outcome studies:

> Vernon Quinsey's (1984:101) conclusion in his review of recidivism studies of rapists applies to this broader review as well: 'The differences in recidivism across these studies are truly remarkable; clearly by selectively contemplating the various studies, one can conclude anything one wants' (Furby, Weinrott, and Blackshaw, 1989:22).

Another set of criticisms has to do with *practicality*. Traditional reviews have difficulty coping with the growth of research. Relying on available journals in a library or the papers collected in office files will no longer ensure coverage of the available studies. In the same way that it would be difficult to make sense of a large, growing and scattered collection of police reports or prison folders without orderly methods, it is also difficult to make sense of the burgeoning number of relevant evaluation studies without some systematic process for doing so.

Although narrative reviews of program evaluations can be influential (Martinson, 1974), they are considerably more difficult to conduct as the number of studies under review increases. Reaching conclusions from the results of multiple studies is risky when the populations, settings, study characteristics, and interventions vary widely across research reports (Wolf, 1986). It is difficult to examine interaction effects under such conditions without statistics. As Glass and his colleagues (1981) noted, accurately summarizing a considerable number of outcome studies is just as difficult without quantification as a large number of survey responses or case files. Cooper stressed the need for rigor to cope with the increased numbers of scientific studies (1989:145):[2]

> Because of the growth in empirical research, the increased access to information, and the new tech-

> niques for research synthesis, the conclusions of research reviews will become less and less trustworthy unless something is done to systematize the process and make it more rigorous. Because of the increasing role that research reviews play in our definition of knowledge...adjustments in procedures are inevitable if social scientists hope to retain their claim to objectivity.

The narrative or traditional review, therefore, has a significant number of methodological limitations that compromise its ability to provide sound evidence to decision makers in criminal justice. The alternative method for synthesizing studies, now referred to as the systematic review, has its roots in the creation of quantitative reviewing or meta-analysis in the psychology and education fields in the 1970s.

A Brief History and Overview of Meta-Analysis

About the same time that the traditional review was coming under heavy criticism, the modern statistical foundation for quantitative reviewing was being developed (Glass, McGaw, and Smith, 1981; Hedges and Olkin, 1985). In 1976, Gene Glass coined the term *meta-analysis* to describe quantitative approaches to reviewing studies. He and Mary Lee Smith deserve much credit for popularizing this approach by applying this technique to research on the effects of psychotherapy (Smith and Glass, 1977) and class size (Glass and Smith, 1978). Glass (1976) popularized a standardized effect size measure for expressing the difference between experimental and control groups in standard deviation units. Using this numeric effect size as a dependent variable, Smith and Glass (1977) were able to quantify over 400 psychotherapy experiments. They concluded, in contradiction with some of the notable narrative reviews on the issue (Eysenck, 1961), that subjects exposed to psychotherapy experienced—on average—a strong, beneficial effect when compared to control group subjects.

Using the standardized effect size measure—or common metric—moved the emphasis of the review from statistical significance, which can be misleading, to the actual magnitude of effect the experimental treatment achieved. The common metric expresses the difference between the groups in a manner that is independent of statistical significance.

The Smith and Glass (1977) findings led to extensive use of meta-analysis in the fields of psychology and education. Its popularity soon spread to other fields, particularly medicine and business, with the technique receiving national press coverage (Mann, 1994; Strauss, 1991). Other researchers were simultaneously developing their own statistical approaches to synthesis (Hunter, Schmidt, and Jackson, 1982; Rosenthal, 1991; Hedges and Olkin, 1985).

Most meta-analyses of research on the effects of social or educational interventions follow a similar path. After identifying eligible studies, the researchers create a measure of "effect size" for each experimental versus control contrast of interest in the study. Most commonly, reviewers do this by standardizing the difference between scores of the experimental and control groups, placing outcomes that are conceptually similar but measured differently (such as rearrest or reconviction) on the same common scale or metric. Though these are different indices, they do measure a program's effect on some construct (e.g. "criminality"). These effect sizes are usually averaged across all similar studies to provide a summary of program impact. The effect sizes also represent the "dependent variable" in the meta-analysis, and more advanced syntheses explore the role of potential moderating variables, such as sample size or other characteristics on effect size. Many texts on meta-analysis have been produced over the past two decades and can be consulted for further details on the methods involved (Lipsey and Wilson, 2001; Cooper and Hedges, 1994; Hedges and Olkin, 1985).

The Lexicon of Research Reviews

Although there is some confusion about the term, meta-analysis involves the quantitative analysis of prior research results. Khan and his colleagues (2001) define meta-analysis as "the use of statistical techniques to combine the results of studies addressing the same question into a summary measure." The term "systematic review" became popular in the 1990s in medicine to overcome inadequacies in the term meta-analysis. First, researchers may sometimes have very good reasons for not using meta-analytic or quantitative methods to summarize studies. This does not mean that their reviews were unsystematic. For example, a reviewer may find that there are few studies meeting the eligibility criteria for inclusion into the review. Such was the case in a systematic review of treatment of sexual offenders reported by White and his colleagues (1999). They conducted a vigorous search and retrieval effort to locate randomized experiments testing interventions for that population. They located only three experiments that met their eligibility criteria, and attempted no quantitative synthesis. The review was systematically performed, and was important in pointing

out that the evidence base in this area is scant, requiring vigorous investment in experiments. But it was not a meta-analysis.

Another shortcoming of the term meta-analysis is that it could include quantitative reviews that used inexplicit or biased methods. For example, a quantitative review that does not describe the search methods used would still be called a meta-analysis.

Using the term "systematic review" seems to get us out of some of those quandaries but may lead us into others. One general rule used to define a systematic review is that it will usually include a "methodology and results" section. But a review could use systematic methods to summarize evaluation studies, and then rely on "statistical significance" to make judgments about "what works." This definition would classify such a review as systematic even though there are empirical reasons undermining its conclusions. The definition of systematic review created by Khan and his colleagues at the NHS Centre for Reviews and Dissemination (2001:1) would also treat vote counting, a formerly popular method of summarizing studies within a review, in similar fashion:

> A review of the evidence on a clear formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyze data from the studies included in the review.

Given the definitional problems, we prefer to think of systematic reviews as ranging on a continuum of quality. At one end, a systematic review may include a methods and results section with very brief details provided and rudimentary analysis. At the other end, the review may be written in very explicit fashion with state-of-the-art statistical techniques applied. The Cochrane Collaboration is an international organization specializing in such upper-end systematic reviews, though most of their syntheses are relevant to health care issues (Chalmers and Altman, 1995). They have developed a list of steps in conducting systematic reviews at the upper end of this continuum:

1. The question guiding the work is explicit and can be answered by a systematic review;
2. The eligibility criteria for studies to be included is explicit;
3. The search methods are comprehensive and designed to reduce potential bias;
4. Each potentially eligible study is screened against the criteria with exclusions justified and recorded
5. The sample of eligible studies and the corresponding data set is the most complete possible;
6. If meta-analysis is possible, the methods are technically appropriate;
7. If statistical analyses are used to examine subgroup effects, they are technically appropriate; and
8. A structured and detailed report, explicitly reporting each stage of the review, is produced.

Systematic reviews, therefore, include reviews in which rigorous methods are employed regardless of whether meta-analysis is undertaken to summarize, analyze and combine study findings. When meta-analysis is used, however, estimates of the average impact across studies, as well as how much variation there is and why, can be provided. By using meta-analysis, we can generate clues as to why some programs are more effective in some settings and not others.

Criticism of Meta-Analyses and Systematic Reviews

Meta-analyses and systematic reviews are not without criticism. The most frequent criticism leveled is commonly referred to as the "apples and oranges" critique (Lipsey and Wilson, 2001). This criticism charges systematic reviews and meta-analyses for mixing vastly different studies together (e.g., by including heterogeneous study findings [Eysenck, 1994] or by including studies of differing methodological quality) to produce a single estimate of treatment effect. Gorman (1995) criticized a meta-analysis of eight outcome studies of Drug Abuse Resistance Education [D.A.R.E.] by claiming that the review team mixed together apples, oranges and a few poorly-done studies, or lemons! But some have argued that the apples and oranges criticism is not appropriate if the goal of the review is to broadly analyze "fruit" (Rosenthal and DiMatteo, 2001).

There have been a number of advances in methods to address the apples and oranges criticism, specifically regarding heterogeneity and methodological variability issues. Setting sensible eligibility criteria can reduce some of this variability before the sample of studies is collected and analysis begins. Moreover, reviewers now code the methodological, contextual, and treatment characteristics—often in excruciating detail—and explore how these variations impact estimates of treatment effect in the meta-analysis (Lipsey and Wilson, 2001). Another common method in meta-analysis is to conduct statistical tests of homogeneity to determine if the effect sizes obtained from the sample of studies is significantly different from what would be expected by chance or sampling error. If the "test of homogeneity" is significant, then the meta-analyst should assume that there are meaningful

subgroups or moderating influences in the database of studies (Cooper and Hedges, 1994). It is now uncommon, because of these methods, to uncover meta-analyses that report only a single overall effect size to represent a heterogeneous sample of studies. Note that systematic reviews and meta-analyses attempt to address the apples and oranges criticism with explicit and transparent methods. Narrative and traditional reviews are also subject to the apples and oranges criticism but lack an arsenal of methods to respond to it.

An Example of a Systematic Review of a Single Program: Does 'Scared Straight' Work?

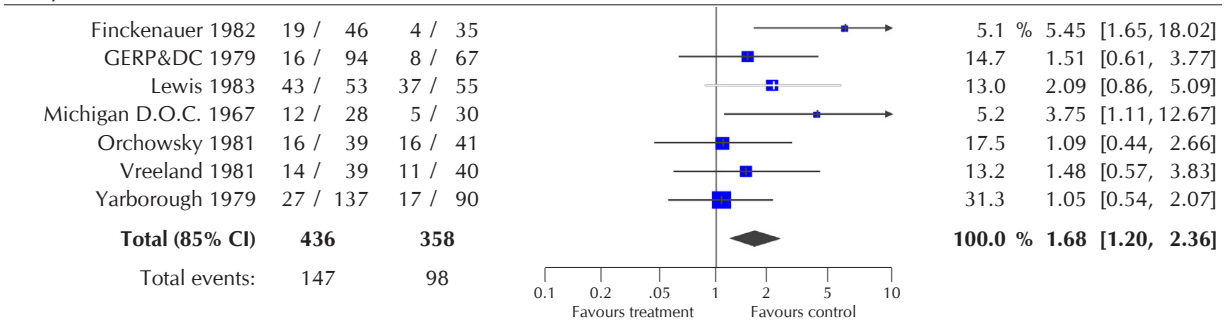
Petrosino and his colleagues (2003) reported on the effects of Scared Straight and other juvenile awareness programs. These "kids visit prisons" programs are meant to deter juvenile delinquents or children at risk by making them aware of the grim realities of prison life. Many of these programs feature a "rap session" in which prisoners brutally describe what institutional life is like, in an attempt to deter youngsters from committing crimes. Although researchers have long believed that this type of program was ineffective and possibly harmful, it has remained in use and has even experienced something of a revival in recent years. Although other reviewers had included Scared Straight as one of several programs included in their reviews, there was no existing systematic review focusing solely on evaluations of this program.

Petrosino and his colleagues (2003) conducted a rigorous search for randomized experiments that examined the effects of the Scared Straight program on subsequent measures of crime. Their methods included electronic searches of abstracting or bibliographic databases, contact with colleagues and research centers, visually examining the contents of bound criminological journals (i.e., "handsearch"), and tracking citations listed in existing reviews. Their techniques located nine randomized experiments reported between 1967 and 1992, including five unpublished studies. All of the experiments included a no-treatment control group, and seven of the nine reported data that could be statistically combined in the meta-analysis.

A common approach to analyzing data in meta-analysis is to use a forest plot of the odds ratio for each study. An odds ratio is simply the number of events (such as the number of juveniles failing or being arrested) divided by the number of "non-events" (number of juveniles succeeding or not being arrested). An odds ratio of "1.0" means that the program did not increase or decrease a juvenile being successful (not arrested). A 1.0 is a precise "no difference" effect, or effect of zero. Odds ratios above 1.0 mean that the program increased the failure rate; similarly, odds ratios below 1.0 mean the program was successful in reducing subsequent arrests.

Figure 1 presents the forest plot for the seven experimental studies of Scared Straight and other juvenile awareness programs. All seven report negative effects for the treatment group. In other words, children participating in the juvenile awareness program did worse than juveniles who did not. Petrosino et al. (2003) concluded that Scared Straight methods were not effective

Figure 1. First Effect of Intervention, Official Crime Measures, Fixed-Effects Model



Test for heterogeneity chi-square = 8.50 df = 6 p = 0.20 I^2 = 29.4%

Test for overall effect z = 3.01 p = 0.003

Note: n = number of participants reoffending; N = number assigned to group; CI = confidence interval; weight = amount of weight given to study in analysis; GERP&DC = Greater Egypt Regional Planning and Development Commission; D.O.C. = Department of Corrections.

Source: *Petrosino, et al. (2003)*

in deterring subsequent crime, and likely had a backfire or toxic effect on juveniles. What is remarkable is that a meta-analysis of nearly 400 experimental or well-controlled quasi-experimental evaluations of preventative or treatment interventions for juvenile delinquency showed that nearly two-thirds (64%) were positive in direction (Lipsey, 1992). The Scared Straight studies clearly go against the trend of most juvenile intervention showing positive effects. This meta-analysis underscored that presumably beneficial interventions can go against conventional wisdom and best intentions and have a negative impact on the very juveniles and citizens policymakers and practitioners desire to help.

An Example of a Systematic Review Comparing One Program With Others: Does 'D.A.R.E.' Work?

One of the most popular school-based drug prevention programs in the world is Drug Abuse Resistance Education, or D.A.R.E. Initiated in 1983 as a joint project between the Los Angeles Police Department and Unified School District, the core program used uniformed police officers to deliver a 17-week curriculum (lasting one hour per week) to 5th and 6th grade students (i.e., 10-12 year olds). Several early evaluations were positive, and the program quickly expanded with federal funding throughout three-fourths of the nation's school districts (Rosenbaum and Hanson, 1998).

Given the federal investment in the program, it was only natural that decision makers would wish to know whether D.A.R.E. worked to reduce drug use and led to better attitudes toward the police. The National Institute of Justice issued a solicitation for an evaluation of the research on D.A.R.E., and after a peer review process, selected the Research Triangle Institute (RTI) in North Carolina to conduct the study (Ennett et al., 1994). RTI followed the tenets of systematically reviewing evidence. They were explicit in their procedures, used methods to reduce bias, and presented a detailed report outlining what they did and why they did it. Although there were many uncontrolled studies on D.A.R.E., their extensive searches turned up only eight evaluations that used either a randomized field trial or rigorous quasi-experimental procedures. They examined the outcomes of self-reported drug use, attitudes toward police, attitudes toward drugs and knowledge about drugs. For each of these measures, they created a standardized effect size expressing the difference between the experimental and control groups.

Their results showed that D.A.R.E. had positive impacts on knowledge, but the findings were less persuasive when it came to attitudes or behavior. Given that the researchers at RTI used effect size rather than odds ratios, it was difficult to understand how D.A.R.E. was faring without a basis for comparison. They did not collect a sample of evaluations of other types of drug prevention programs to compare to D.A.R.E. To remedy this, they worked with Nancy Tobler, who had conducted several earlier meta-analyses of school-based drug prevention programs. Using the Tobler database, the RTI researchers identified programs delivered to 5th and 6th graders (like the core D.A.R.E. curriculum) and classified them as "interactive" or "non-interactive." Interactive programs were those that involved role-playing and modeling and did not rely on straight lectures providing information. Non-interactive programs involved little more than providing information to youngsters about the harm of drugs. Although the authors did not attempt to define how interactive D.A.R.E. was, the program was weighted toward the officer delivering a standardized curriculum in the classroom and likely fell somewhat in-between the interactive and non-interactive groupings.

The comparison data were telling. Although D.A.R.E. did better on some measures than "non-interactive" programs, the evidence showed that drug prevention defined as "interactive" was far more effective with 5th and 6th grade students than D.A.R.E. This was true across measures of attitude, knowledge and self-reported drug use. Even though self-reported drug use (which included tobacco, alcohol and marijuana) were small for all groups, the positive impact for interactive programs was three times the size of D.A.R.E. Without this comparison data, it is unlikely that the review would have generated much controversy (Elliot, 1995). But given the results, some questioned whether the federal investment in D.A.R.E. was really worth it all, and whether these more effective alternatives should be supported.

A Modest Agenda for Improving the Policy-Review Connection

What if a wide range of systematic reviews could be produced on a large scale, and made available in rapid fashion to decision makers in criminal justice? This electronic archive could provide a resource for federal, state, and local decision makers to access so they can determine "best evidence" on what works for a variety of interventions relevant to reducing crime and making the justice system fairer and more effective. Inspired by the success of the Cochrane Collaboration in health care (www.cochrane.org), the international Campbell Collaboration (www.campbellcollaboration.org) was inaugurated in 2000 to prepare, update and disseminate systematic

reviews in social science. The Campbell Collaboration (C2) initiated review groups to supervise work in three substantive areas: education, social welfare, and crime and justice.

The Campbell Crime and Justice Group (CCJG) now oversees a portfolio of over 40 titles. The Scared Straight example, mentioned earlier, was initiated as a pilot review for the C2 and is available online (Petrosino et al., 2003). Completed reviews also exist on the effects of boot camps (Wilson et al., 2005) and the effectiveness of counter-terrorism strategies (Lum et al., 2006). The pace of producing reviews has been somewhat unsteady, likely reflecting the difficulty in both the organization and the individual teams in obtaining funds to leverage time and resources toward the review. Nonetheless, with sufficient funds the CCJG archive (and C2 in general) should become an important source of rigorous evidence on the effects of criminological and justice interventions. Long-term investment in the C2 and CCJG is needed to expand the archive so that it contains a large number of reviews, each addressing particular policy or practice questions.

The CCJG is only one of many entities producing systematic reviews and meta-analyses like the aforementioned Scared Straight and D.A.R.E. examples. Petrosino (2000) located 205 systematic—or possibly systematic—reviews of research on the effects of interventions relevant to crime, drugs or alcohol. More recently, Petrosino (2005) found 50 meta-analyses in correctional intervention alone, and these were located without a comprehensive search. Rigorous syntheses likely number in the hundreds across areas relevant to crime and justice, and represent a form of "criminological intelligence" that has not been mined or exploited in any way. No organized collection of existing reviews currently is available, and interested users have to locate them as they would any other literature, through bibliographic database searches of Criminal Justice Abstracts and the National Criminal Justice Reference System (NCJRS) abstracts.

As a parallel resource to the CCJG reviews, we propose that an electronic archive be created that would provide short, structured abstracts of existing (already available) systematic reviews and meta-analyses. There is at least one important precedent for such a database. The Cochrane Collaboration's main product is an electronic publication known as the *Cochrane Library*. Though the main part of the Cochrane Library provides access to over 1,500 completed systematic reviews of research on the effects of health care interventions, the publication also makes available other relevant databases. For example, the UK-National Health Service Centre for Reviews and Dissemination at the University of York produces the *Database of Abstracts of Reviews of Effectiveness* (www.york.ac.uk/inst/crd/darehp.htm). This Centre produces structured abstracts of reviews relevant to health care, and includes mostly non-Cochrane reviews (e.g., reviews from the *British Medical Journal* or *Journal of American Medical Association*). Subscribers to the Cochrane Library also get access to this database of structured abstracts to other reviews. Such a resource, as an ancillary to CCJG reviews, could cover a range of policy and practice questions and provide fertile ground for future research and directed funding.

Conclusion

Though careful studies on the use of systematic reviews in decision making have not been reported in academic journals, Weiss (1978) suggested over twenty years ago—before review methods were the object of considerable attention—that policymakers would find syntheses more compelling than single studies. This is because a good review would presumably reconcile different studies that are often used by competing sides in policy debates, at least where reconciliation of distinct studies is possible. A good review would also pull together the relevant information so that policymakers or their aides (or agency staff to whom they would delegate such responsibility) would not have to spend time tracking and synthesizing data. Such syntheses would be most important when decisions about appropriations were made, particularly when governments were looking for new programs or strategies to fund.

Nonetheless, we do not wish to overzealously sell evidence, and we recognize the constraints faced by the justice policymaker (Petrosino et al., 2001). Lipton (1992) underscored the multitude of inputs into any decision, including budgetary restrictions, constituent wishes, public opinion, and reappointment or reelection concerns. Research necessarily is but one input into that process, and Weiss (1998) argues that is as it should be in a democratic society. But research evidence can and should be an important consideration in policy and practice choices made by decision makers in criminal justice. Given the explicitness, comprehensiveness, and rigor of a systematic review and meta-analysis, they should be the starting point for considerations about "what the science says" about what to do to reduce crime and increase fairness in the criminal justice system.

Endnotes

1. Research also confirmed that studies in a review are part of a sampling distribution. As such, variation

in studies can be due to sampling error as well as real distinctions between context, intervention delivery and other factors (Cooper and Hedges, 1994). Hedges and Olkin (1985) and others have shown how quantitative techniques can be used to determine how much variation across studies is likely due to sampling error and how much is likely due to subgroup differences. Narrative or traditional reviews do not use such techniques and run the risk of attributing differences that occur because of sampling error to conflict between studies (when they actually may be in convergence).

2. Although the number of evaluations and other research is cumulatively and annually increasing, the number done relative to the funds invested in programming is infinitesimal.

References

Bailey, Walter C. 1966. "Correctional Outcome: An Evaluation of 100 Reports." *Journal of Criminal Law, Criminology and Police Science* 57:153-160.

Berlin, Jesse, Colin B. Begg and Thomas A. Louis. 1989. "An Assessment of Publication Bias Using a Sample of Published Clinical Trials." *Journal of the American Statistical Association* 84:381-392.

Boruch, Robert F. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Thousand Oaks, California: Sage.

Bushman, Brad J. and Gary L. Wells. 2001. "Narrative Impressions of Literature: The Availability Bias and the Corrective Properties of Meta-Analytic Approaches." *Personality and Social Psychology Bulletin* 27:1123-1130.

Chalmers, Iain, Larry V. Hedges and Harris Cooper. 2002. "A Brief History of Research Synthesis." *Evaluation and the Health Professions* 25:12-37.

Chalmers, Iain and Doug Altman, eds. 1995. *Systematic Reviews*. London, UK: BMI.

Coalition for Evidence-Based Policy. 2003. "Bringing Evidence-Driven Progress to Crime and Substance Abuse Policy: A Recommended Federal Strategy." March 11, 2006. (http://www.excelgov.org/admin/FormManager/filesuploading/Final_report_-_Evidence-based_crime__subs_abuse.pdf?PHPSESSID=ce5173719edfe1b9a5e21861109edc53).

Cook, Thomas D., Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis and Frederick Mosteller, eds. 1992. *Meta-Analysis for Explanation*. New York: Russell Sage Foundation.

Cooper, Harris C. 1989. *Integrating Research: A Guide for Literature Reviews.* Second Edition. Beverly Hills, CA: Sage.

Cooper, Harris C. and Larry V. Hedges. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Davies, Philip. 1999. "What is Evidence-Based Education?" *British Journal of Educational Studies* 47:108-121.

DiIulio, John. 1991. *The Future of American Corrections*. New York: Basic.

Elliot, Jeff. 1995. "Drug Prevention Placebo: How D.A.R.E. Wastes Time, Money and Police." *Reason* (March): 14-21.

Ennett, Susan T., Nancy S. Tobler, Christopher Ringwalt and Robert Flewelling. 1994. "How Effective is Drug Abuse Resistance Education? A Meta-Analysis of Project DARE Outcome Evaluations." *American Journal of Public Health* 84:1394-1401.

Eysenck, Hans J. 1961. "The Effects of Psychotherapy." Pp. 697-725 in *Handbook of Abnormal Psychology*, edited by H. J. Eysenck. New York: Basic.

Eysenck, Hans J. 1994. "Systematic Reviews: Meta-Analysis and its Problems." *British Medical Journal* 309:789-792.

Farrington, David P. 1994. "Early Developmental Prevention of Juvenile Delinquency." *Criminal Behaviour and Mental Health* 4:209-227.

Forsetlund, Louise and Arild Bjørndahl. 2002. "Identifying Barriers to the Use of Research Faced by Public Health Physicians in Norway and Developing an Intervention to Reduce Them." *Journal of Health Services and Research Policy* 7:10-18.

Furby, Lita, Mark Weinrott and Lynn Blackshaw. 1989. "Sex Offender Recidivism: A Review." *Psychological Bulletin* 105:3-30.

Glass, Gene V. 1976. "Primary, Secondary and Meta-Analysis of Research." *Educational Researcher* 5:3–8.

Glass, Gene V., Barry McGaw and Mary L. Smith. 1981. *Meta-Analysis in Social Research.* London: Sage.

Glass, Gene V. and Mary L. Smith. 1978. *Meta-Analysis of Research on the Relationship of Class Size and Achievement.* San Francisco: Far West Laboratory for Educational Research and Development.

Gorman, Dennis. 1995. "The Effectiveness of DARE and Other Drug Use Prevention Programs." *American Journal of Public Health* 85:873-874.

Greenberg, David, Mark Shroder and Matthew Onstott. 1999. "The Social Market Experiment." *Journal of Economic Perspectives* 13:157-172.

Hacsi, Timothy A. 2002. *Children as Pawns.* Cambridge, MA: Harvard University Press.

Hedges, Larry V. and Ingram Olkin. 1985. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press, Inc.

Hopewell, Sally, Steve McDonald, Mike Clarke and Matthias Egger. 2006. "Grey Literature in Meta-Analyses of Randomized Trials of Health Care Interventions." *The Cochrane Database of Methodology Reviews* Issue2: Art. No.: MR000010. pub2. DOI: 10.1002/14651858.MR000010.pub2.

Hunt, Morton. 1997. *The Story of Meta-Analysis.* New York: Russell Sage Foundation.

Hunter, John E., Frank L. Schmidt and Greg B. Jackson. 1982. *Meta-Analysis: Cumulating Research Findings Across Studies.* Beverly Hills, CA: Sage.

Khan, Khalid S., Gerben ter Riet, Julie Glanville, Amanda Sowden and Jos Kleijnen, eds. 2001. *Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for Those Carrying Out or Commissioning Reviews. CRD [Centre for Reviews and Dissemination] Report Number 4.* 2nd Edition. York, UK: York Publishing Ltd.

Kirby, Bernard C. 1954. "Measuring Effects of Criminals and Delinquents." *Sociology and Social Research* 38:368-374.

Lipsey, Mark W. 1992. "Juvenile Delinquency Treatment: A Meta-Analytic Inquiry into the Variability of Effects." Pp. 83–127 in *Meta-Analysis for Explanation: A Casebook*, edited by T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis, and F. Mosteller. New York: Russell Sage Foundation.

Lipsey, Mark W. 1997. "What Can You Build with Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation." *New Directions for Evaluation* 76:7-23.

Lipsey, Mark W. and David B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis." *American Psychologist* 48:1181-1209.

Lipsey, Mark W. and David B. Wilson. 2001. *Practical Meta-Analysis.* Thousand Oaks, CA: Sage.

Lipton, Douglas, Robert Martinson and Judith Wilks. 1975. *The Effectiveness of Correctional Treatment: A Survey of Treatment Valuation Studies. New York:* Praeger Press.

Lipton, Douglas S. 1992. "How to Maximize Utilization of Evaluation Research by Policymakers." *Annals of the American Academy of Political and Social Science* 521: 175-188.

Logan, Charles H. 1972. "Evaluation Research in Crime and Delinquency: A Reappraisal." *Journal of Criminal Law, Criminology, and Police Science* 63:378-87.

Lum, Cynthia, Leslie Kennedy and Alison Sherley. 2006. "The Effectiveness of Counter-Terrorism Strategies. A Campbell Collaboration Systematic Review." January. Available at: http://www.campbellcollaboration.org/CCJG/reviews/CampbellSystematicReviewOnTerrorism02062006FINAL_REVISED.pdf

Mann, Charles. 1994. "Can Meta-Analysis Make Policy?" *Science* 266:960-962.

Martinson, Robert. 1974. "What Works? Questions and Answers about Prison Reform." *Public Interest* 10:22-54.

Nutley, Sandra M., Huw T.O. Davies and Nick Tilley. 2000. "Editorial: Getting Research into Practice." *Public Money and Management* 20:3-6.

Oxman, Andrew D. and Gordon H. Guyatt. 1988. "Guidelines for Reading Literature Reviews." *Canadian Medical Association Journal* 138:697-703.

Palmer, Ted. 1994. *The Effectiveness of Correctional Intervention*. Albany, NY: State University of New York Press.

Palmer, Ted. 1975. "Martinson Revisited." *Journal of Research in Crime and Delinquency* 12:133-152.

Petrosino, Anthony. 2000. "Crime, Drugs and Alcohol." In Contributors to the Cochrane Collaboration and the Campbell Collaboration, *Evidence from Systematic Reviews of Research Relevant to Implementing the 'Wider Public Health' Agenda*. York, U.K.: University of York, National Centre for Reviews and Dissemination (available at http://www.york.ac.uk/inst/crd/wph.htm).

Petrosino, Anthony. 2005. "From Martinson to Meta-Analysis: The Role of Research Reviews in the US Offender Treatment Debate." *Evidence and Policy* 1:149-172.

Petrosino Anthony, Carolyn Turpin-Petrosino and John Buehler. 2003. "'Scared Straight' and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency" (Updated C2 Review). In *The Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE),* November. Philadelphia, Pennsylvania: Campbell Collaboration.

Petrosino, Anthony, Robert F. Boruch, Haluk Soydan, Lorna Duggan and Julio Sanchez-Meca. 2001. "Meeting the Challenges of Evidence-Based Policy: The Campbell Collaboration." *Annals of the American Academy of Political and Social Science* 578:14-34.

Quinsey, Vernon L. 1984. "Sexual Aggression: Studies of Offenders Against Women." Pp. 84-121 in *Law and Mental Health: International Perspectives*, edited by D.N. Weisstub. Vol. 1. New York: Pergamon.

Roberts, Ian. 2000. "Randomised Trials or the Test of Time? The Story of Human Albumin Administration." *Evaluation and Research in Education* 14:231-236.

Rosenbaum, Dennis P. and Glenn S. Hanson. 1998. "Assessing the Effects of School-Based Drug Education: A Six-Year Multi-Level Analysis of Project D.A.R.E." *Journal of Research in Crime and Delinquency* 35:381-412.

Rosenthal, Robert. 1991. *Meta-Analytic Procedures for Social Research.* Second Edition. Beverly Hills, CA: Sage.

Rosenthal, Robert and M. Robin DiMatteo, 2001. "Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews." *Annual Review of Psychology* 52:59-82.

Ross, Robert R. and M.J. Price. 1976. "Behavior Modification in Corrections: Autopsy Before Mortification." *International Journal of Criminology and Penology* 4:305-315.

Rossi, Peter. 1987. "The Iron Law of Evaluation and Other Metallic Rules." *Research in Social Problems and Public Policy* 4:3-20.

Schweinhart, Lawrence J. 1987. "Can Preschool Programs Help Prevent Delinquency?" Pp. 137-153 in *From Children to Citizens: Families, Schools and Delinquency Prevention,* edited by J.Q. Wilson and G.C. Loury. New York: Springer-Verlag.

Sechrest, Lee, Susan O. White and Elizabeth D. Brown. 1979. *The Rehabilitation of Criminal Offenders: Problems and Prospects*. Washington, D.C.: National Academy of Sciences.

Sherman, Lawrence W. 1992. *Policing Domestic Violence*. New York: Free Press.

Sherman, Lawrence W. and Richard A. Berk. 1984. "The Specific Deterrent Effects of Arrest for Domestic Assault." *American Sociological Review* 49:261-272

Sherman, Lawrence W. and Ellen Cohn. 1989. "The Impact of Research on Legal Policy: The Minneapolis Domestic Violence Experiment." *Law and Society Review* 23:117-144.

Smith, Mary Lee and Gene V. Glass. 1977. "Meta-Analysis of Psychotherapy Outcome Studies." *American Psychologist* 4:753-760.

Strauss, Stephen. 1991. "Meta-Analysis: Lies, Damn Lies and Statistics." *Globe and Mail*, November 2nd, p. D10.

Weiss, Carol H. 1998. *Evaluation Methods for Studying Programs and Policies*. Upper Saddle River, NJ: Prentice.

Weiss, Carol H. 1978. "Improving the Linkage Between Social Research and Public Policy." Pp. 23-81 in *Knowledge and Policy: The Uncertain Connection*, edited by L. E. Lynn. Washington, D.C.: National Academy of Sciences.

Weiss, Carol H. and Eleanor Singer. 1988. *Reporting of Social Science in the National Media*. New York: Russell Sage Foundation.

Weiss, Carol H., Erin Murphy-Graham and Sarah Birkeland. 2005. "An Alternate Route to Policy Influence: How Evaluations Affect D.A.R.E." *American Journal of Evaluation* 26:12-30.

White, Paul, Caroline Bradley, Mike Ferriter and Luke Hatzipetrou. 1999. "Managements for People with Disorders of Sexual Preference and for Convicted Sexual Offenders." (Cochrane Review). *Cochrane Library*. Issue 4. Oxford: Update Software.

Wilson, David B., Doris Layton MacKenzie and Fawn Ngo Mitchell. 2005. *Effects of Correctional Boot Camps on Offending.* A Campbell Collaboration systematic review, available at: http://www.aic.gov.au/campbellcj/reviews/titles.html.

Wilson, James Q. and Richard Hernnstein. 1985. *Crime and Human Nature*. New York: Simon and Schuster.

Wolf, Frederic M. 1986. *Meta-Analysis: Quantitative Methods for Research Synthesis*. Beverly Hills, CA: Sage.

About the authors:

Anthony Petrosino is Senior Research Associate at Learning Innovations at WestEd and Associate Director of Research at the Northeast and Islands Regional Educational Laboratory (NEIREL). He was the Founding Coordinator for the Campbell Collaboration's Crime and Justice Group and served as a consultant on projects for the Harvard Graduate School of Education, the Canadian Department of Justice, the Netherlands Ministry of Justice, and the UK Home Office. After working as a researcher for state justice agencies in New Jersey and Massachusetts, he received his Ph.D. in criminal justice in 1997 from Rutgers University and was Spencer Post-Doctoral Fellow in Evaluation at the Harvard Children's Initiative. Anthony also was Research Fellow at the Center for Evaluation, Initiatives for Children Program, at the American Academy of Arts and Sciences. He recently published a biography of Harvard statistician Frederick Mosteller for the James Lind Library, available at: http://www.jameslindlibrary.org/trial_records/20th_Century/1970s/bunker/bunker_biog.pdf.

Julia Lavenberg is a doctoral candidate in the Graduate School of Education at the University of Pennsylvania, located in Philadelphia. She serves on the Steering Committee of the Campbell Collaboration's Information Retrieval Methods Group and co-authored the policy brief used to guide researchers in conducting the information retrieval phase of a systematic review. She is planning to conduct a systematic review as her dissertation project.

Contact information:

Corresponding author: Anthony Petrosino, Learning Innovations at WestEd, 200 Unicorn Park Road, Fourth Floor, Woburn, MA 01801. Phone: (781) 481-1117. Email: apetros@wested.org.

Julia Lavenberg, Graduate School of Education, University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104. Email: jlavenbe@dolphin.upenn.edu.