Journal of Experimental Criminology (2005) 1: 435–450

The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research

ANTHONY PETROSINO*
Learning Innovations at WestEd, 200 Unicorn Park Drive, 4th floor, Woburn, MA 01801-3324, USA
**corresponding author: E-mail: apetros@wested.org*

HALUK SOYDAN
USC School of Social Work, 669 West 34th Street, Los Angeles, CA 90089-0411, USA

Abstract. Using meta-analysis, we report on an investigation of the evaluator's influence in the treatment setting on criminal recidivism outcomes. Many evaluators and users of evaluation of social interventions worry that mixing of the roles of program developer and program evaluator may bias results reported in intervention studies in a positive direction. We first review the results of prior investigations of this issue across 50 prior meta-analyses, finding 12 that tested the impact of investigator influence in the treatment setting. Eleven of these reported that effect size increased positively, sometimes substantially so, when evaluators were influential or involved in the treatment setting. We followed this with a meta-analysis of 300 randomized field trials in individually focused crime reduction, also finding intervention studies in which evaluators who were greatly influential in the treatment setting report consistently and substantially larger effect sizes than other types of evaluators. We discuss two major views – the 'cynical' and 'high fidelity' theories – on why this is consistently the case, and conclude with a further agenda for research.

Key words: meta-analysis, offender rehabilitation, program evaluation, randomized experiments

Introduction

Reducing potential biases that may lead to a wrong result (and therefore a bad decision about a policy, practice, or program) is a primary concern for evaluators in fields such as criminal justice, social work, and education. Most concern about bias is directed toward conduct of actual research, and not surprisingly, there are numerous publications on how to reduce bias in evaluation design, survey instruments, and statistical methods to increase the likelihood of reliable and valid results.

This concern with bias is even more pronounced in recent years, with the greater attention on evidence-based policy. Many advocates of evidence-based policy also advocate randomized experiments when possible to evaluate the impact of interventions (e.g., Sherman 1998). This is because a well-implemented randomized study can produce statistically unbiased results when comparing an experimental group receiving treatment and a control group that does not (e.g., Boruch 1997). As a number of evaluation studies in a particular area are produced, evidence-based policy advocates generally support the use of systematic reviews and meta-analysis (which we will discuss later in this paper) to reduce the potential for bias that may distort findings in a synthesis of separate but similar studies (e.g., Boruch and Petrosino 2004).

Although improving methods to increase the precision of social program evaluation has received a large share of the attention by advocates of the evidence-based approach, the credibility and believability of findings is no less important to considerations about evidence. Findings from even rigorous experimental studies may be disregarded if it is perceived that the evaluators had a conflict of interest that potentially biased their study and report. For example, a rigorous study reported by the National Rifle Association or "Americans Against Guns" on the effects of conceal and carry laws on crime are likely to be viewed with great skepticism, particularly if the report supports the organization's previously held position.

An evaluation reporting a promising new prevention or intervention program may also be greeted with skepticism, if the authors of the study were intimately involved in the program's development and operation. This is because it is commonly believed that researchers intimately involved in the creation and implementation of a new program will have vested interests in the program's success, leading to inflated and positive results. Cook (2005: 12) writes that "Developers are, and should be, passionate advocates for their program, not brokers of honest appraisal." Those who believe program developers assume this role also value the role of outside evaluators who are unconnected to the program setting and are viewed as being more objective.

One immediate question raised by such considerations is whether there really is an observed difference in results of evaluations conducted by program developers, particularly as compared to those conducted by researchers external or less involved in the program setting. The aim of this article is to study the influence of the role of the evaluator on reported outcomes of criminal recidivism. We do this by combing through the results of prior meta-analyses of the offender treatment literature, and follow this up with a separate meta-analysis of 300 randomized trials collected by the first author (Petrosino 1997).

The role of the evaluator: Conflicting advice

The literature seems to encourage two different roles for the evaluator. On the one hand, evaluators are encouraged to be more involved in the program setting. There are many different approaches to evaluation. Some of these, such as 'participatory evaluation' and 'action research,' require more substantive roles for the researcher in the program setting than an outside evaluator would assume.

Moreover, in order to shape and improve their programs, organizations and their personnel are strongly encouraged to conduct formative evaluations. 'Good learning organizations' use the data from formative evaluations to improve their operations and strategies. Practitioners like teachers are encouraged to reflect on their own practice, and use their own qualitative assessments to shape their future efforts with students (e.g., Huebner 2000). This shares similarity with how agencies and organizations should self-evaluate, or what Love (1991) called 'internal evaluation.' Both Shepherd (2003) and Sherman (2004) have argued that one way to improve research in policing is to develop the 'practicing researcher' in law enforcement, similar to the clinician–researcher in medicine. In such a system, law enforcement agencies would conduct studies of their own strategies, and would

also provide learning opportunities for officers. Similarly, Weisburd (1996) advocated for criminologists to 'make the scene' and get more involved in their studies to ensure integrity and develop better understanding of their research.

On the other hand, the results reported by developers evaluating their own programs may be less credible or trustworthy due to a conflict of interest that may potentially bias the results in a positive direction. For example, Drug Strategies (1999) conducted a review of school-based prevention curricula, finding 18 that did have a rigorous experimental study supporting claims of effectiveness. They went on to caution that *external evaluators conducted only four of the 18*. In Latimer's (2001) review of family-based programs for delinquency, he created a three variable composite score for methodological quality: type of assignment to study conditions; attrition, and *whether an external evaluator conducted the study*.

Meta-analysis

Determining the influence of the role of the evaluator on observed effect size is now possible because of meta-analysis. The US Bureau of Justice Assistance (2005) defines meta-analysis as "the systematic analysis of a set of existing evaluations of similar programs in order to draw general conclusions, develop support for hypotheses, and/or produce an estimate of overall program effects." Good meta-analyses will also be good systematic reviews, i.e., they will explicitly describe the research question, study eligibility criteria, search methods, the information that has been extracted from each report, and analysis strategies (e.g., Boruch and Petrosino 2004).

In most meta-analyses, each individual study is treated as a single case in the dataset. For each study, a common metric "effect size" is created to express the observed impact of the program on the outcome measure of interest, when compared to a control or comparison group. By creating a common metric from the difference between the experimental and control groups, studies with different measures can be combined and compared. This effect size then serves as the dependent variable in the meta-analysis. It is the quantitative analysis of effect size that is the main characteristic of meta-analysis.

We note that meta-analysis cannot identify if there was any intentional or subtle distortion by evaluators to inflate the program's impact. It cannot tell us whether findings in any particular study are "right" or "wrong." We also caution that meta-analysis relies almost exclusively upon written reports. It is often the case that evaluation reports do not contain the necessary information on items such as the role of the evaluator to extract for analysis.

But meta-analysis can provide good evidence about patterns, and can help isolate the role of one variable, such as the role of the evaluator, on observed effect sizes. If a researcher conducting meta-analysis has extracted information (i.e., coded) on the role of the evaluator and entered this into the analysis dataset, it can serve as an independent variable in analyses to determine its impact on the dependent variable (effect size).

We relied on the advantages presented by meta-analysis in this study, and conducted a two-stage analysis to ascertain the role of the evaluator on observed effect sizes. First, we examined 50 meta-analyses in the offender treatment area to

determine if other reviewers have examined this issue, and if so, what they reported. Second, we followed up this investigation by analyzing a dataset of 300 randomized field trials of individually focused intervention on criminal recidivism, collected earlier by the first author (Petrosino 1997). Most of these experiments were relevant to offender treatment and delinquency prevention, but several were tests of different criminal justice policies or practices on crime outcomes.

Fifty prior offender treatment meta-analyses

Although our search to find prior meta-analyses of offender treatment was neither systematic nor exhaustive, we were able to identify 50 for this paper (Figure 1). We considered any meta-analysis that included outcomes of criminal recidivism (e.g., arrest, conviction, return to prison). These meta-analyses included 19 focused on juvenile delinquency (e.g., Andrews et al. 1990), six of sex offender treatment (e.g., Loesel and Schmucker, forthcoming), four targeting family-based treatment (e.g., Farrington and Welsh 2003), and three that looked only at studies reported in Europe (e.g., Redondo et al. 2001).

Of these 50 reviews, 12 examined what impact the role of the evaluator in the program setting had on observed effect sizes. None specifically looked at "program developers," but categorized the variable in several ways (e.g., "influence of experimenter/investigator on treatment setting"; "program evaluator independent of program"). Of the 12 meta-analyses that looked at this relationship, regardless of how it was conceptualized, 11 (92%) reported larger effects for "involved evaluators" in the treatment/program setting.

For example, Gensheimer and her colleagues (1986) reviewed 44 experimental and quasi-experimental studies of juvenile diversion programs. They reported that there was a "significant correlation between the investigator's influence in the design and implementation of treatment and the calculated mean effect size" (p. 52). Gensheimer et al.'s (1986) diversion program findings come from a larger meta-analysis of all juvenile treatment evaluations. The findings were very similar, however, when considering this broader sample. Davidson et al. (1990: 35) write that, "It is also clear that not all types of investigators produce the same research results … investigators who had a role in designing or controlling the intervention produced more positive findings."

In the most ambitious meta-analysis project to date, Lipsey (1992) collected nearly 400 experimental and quasi-experimental evaluations of juvenile delinquency treatment. Lipsey (1992: 138) concludes that "treatment provided by the researcher or situations where the researcher was influential in the treatment setting were associated with larger effect sizes." This result was similar to later work by Dowden and Andrews (1998) in their meta-analysis of female offender treatment program evaluations. They (1998: 448) write that "of the methodological considerations, only 'involved evaluator' was significantly correlated with effect size." In a later meta-analysis of offender relapse prevention studies (2003: 552), they also reported "significantly enhanced program effects … with studies that involved … the evaluator in the design and implementation of the program."

- **Boot Camps**
  - *MacKenzie et al (2001)*
- **Cognitive-Behavioral**
  - *Lipsey et al (2001)*
  - *Pearson & Lipton (2002)*
- **Drug Involved Adult Offenders/Drunk Drivers**
  - *Pearson & Lipton (1999)*
  - *Pearson et al (2002)*
  - *Well-Parker et al (1995)*
- **Experiments**
  - *Petrosino (1997)*
- **European**
  - *Egg et al (2000)*
  - *Loesel & Koferl (1989)*
  - *Redondo et al (2001)*
- **Family-based Strategies**
  - *Dowden & Andrews (2003)*
  - *Farrington & Welsh (2003)*
  - *Latimer (2001)*
  - *Wolfendon et al (2003)*
- **Female Offenders**
  - *Dowden & Andrews (1999a)*
- **Juvenile Offenders**
  - *Andrews et al (1990)*
  - *Cox et al (1995)*
  - *Curtis et al (2004)*
  - *Davidson et al (1990)*
  - *Dowden & Andrews (1999b)*
  - *Garrett (1985)*
  - *Gensheimer et al (1986)*
  - *Gottschalk et al (1987)*
  - *Izzo & Ross (1990)*
  - *Kaufman (1986)*
  - *Latimer et al (2004)*
  - *Lipsey (1992, 1999)*
  - *Lipsey & Wilson (1998)*
  - *Loesel & Beelman (2003)*
  - *Mayer et al (1986)*
  - *Petrosino et al (2003)*
  - *Roberts & Camasso (1991)*
  - *Whitehead & Lab (1989)*
  - *Wilson & Lipsey (2000)*
- **Violent Offending**
  - *Dowden & Andrews (2000)*
- **Offender Treatment Generally**
  - *Antonowicz & Ross (1994)*
  - *Cleland et al (1996)*
  - *Pearson et al (1995)*
  - *Pearson et al (1996)*
- **Prison-based Work and Education**
  - *Wilson et al (2000)*
- **Punishment**
  - *Gendreau & Goggin (1996)*
  - *Gendreau et al (1999)*
- **Relapse Prevention**
  - *Dowden et al (2003)*
- **Restorative Justice**
  - *Latimer et al (2001)*
- **Sex Offenders**
  - *Alexander (1999)*
  - *Gallagher et al (1999)*
  - *Hall (1995)*
  - *Hanson et al (2002)*
  - *Loesel (2001)*
  - *Loesel & Schmucker (2004)*

Figure 1. 50 prior meta-analyses in criminal justice.

Although no meta-analysis specifically looked at program developers, a clear pattern emerged from the 12 meta-analyses we examined. Evaluators involved or influential in the program setting report larger effect sizes than evaluators who are not. Program developers would be at the high end of influence and involvement within the treatment setting.

A meta-analysis of 300 randomized experiments in crime reduction

Since 1988, there have been several efforts to identify, and sometimes analyze, the results from randomized experiments relevant to the offender treatment and punishment issue (Weisburd et al. 1990, 1993; Petrosino 1997). To follow up on prior research, we undertook an analysis drawing on a dataset that now includes information on 300 distinct randomized field trials relevant to "individually focused" crime reduction. By individually focused crime reduction, we mean that the experimental intervention was primarily concerned with the reduction of crime through prevention, treatment or punishment delivered to individuals rather than communities or other larger aggregate units. So, while experimental studies testing the crime reduction impact of treatment programs for offenders, different sanctions or punishments, or prevention strategies delivered to at-risk children were included, experiments of security or police strategies in neighborhoods or streets were not.

Evaluation reports were identified if they met the following criteria: (1) used random or quasi-random (e.g., alternation, odd/even case number assignment) methods to allocate participants; (2) as mentioned earlier, individuals were the unit of analysis rather than aggregate units; (3) the results included at least one outcome measure of official crime that could be converted to an effect size (e.g., arrest, conviction); (4) the report had to be published or otherwise available through 1993; and (5) it was available in English.

A variety of search methods were used to find eligible studies. Obviously, our prior research on experiments provided a number of relevant studies to begin with (Weisburd et al. 1990). This was augmented by a number of other search methods, including electronic searches of bibliographic databases (e.g., *Criminal Justice Abstracts, National Criminal Justice Reference Service*); a handsearch (manual visual inspection) of 29 leading social science journals; we sent out letters to hundreds of reviewers and experimental researchers; we published solicitations in association newsletters requesting leads to eligible reports (e.g., in American Society of Criminology newsletter, *The Criminologist*); and we chased down citations to potentially eligible trials from existing reviews and experimental literature. Despite the narrow eligibility criteria, several hundred trials were identified; retrieval methods ended after the first 300 trials were declared eligible following preliminary screening.

Data were extracted for all 300 trials using a 196-item instrument and the data entered into the SPSS-PC statistical software program for management and analysis. The extracted information included a variety of items relevant to the publication (e.g., whether the document was published in a journal or book, or was unpublished), treatment (e.g., the particular modality), methodology (e.g., whether randomization was corrupted), the results (the impact of the intervention on a variety of outcomes), and the investigators (including several variables about the role of the evaluator in the treatment setting).

Each study is represented by a single effect size. Although most justice experiments include a variety of outcome measures reported at multiple time intervals, to remain consistent across studies, we created the effect size from the "first posttreatment effect." This was most often reported at six or 12 months. In those experiments in which multiple outcome measures were reported at the first posttreatment period, we selected the outcome that represented the "earlier point" in the criminal justice system. So we selected police measures such as arrest or contact over bookings, bookings over convictions, and so on.

We used Cohen's *d*, or the difference between the experimental and control groups divided by the pooled standard deviation. Although most experimental reports in criminal justice do not use means nor report standard deviations, formulae exist to convert available test statistics into approximations of effect size (e.g., Wilson 2001). To calculate Cohen's *d* from our data, we used the online effect size calculator created by Wilson and Lipsey (2003) for the Human Services Research Institute. Effect sizes are positive when they reduce crime (positive impact) and negative when they increase crime (negative impact). It is useful to note that Cohen's *d* can be converted to Pearson's correlation coefficient *r* by multiplying by 0.5.

Table 1. Internal versus external evaluation teams–analysis of "first posttreatment effects."

Type of evaluation team	*N*	*Effect size*
Internal	137	0.16
External	124	0.02
Collaboration of internal and external staff	20	0.20
All studies	281[a]	0.10

[a] Nineteen cases were missing.

Results from the meta-analysis of experiments

We looked specifically at three variables in the data set: (1) whether the evaluators were classified as internal or external researchers; (2) the influence of the investigator on the treatment settings, rated as high, moderate or low (a developer, for example, was rated as having a "high degree of influence" and an external academician was rated as having "low" influence); and (3) the more specific role of the evaluator in the setting (e.g., developer, program staff, outside academician).

Table 1 presents the mean effect size for experiments conducted by evaluators classified as internal or external (19 cases were missing information necessary to make this categorization). Note that a third category was created because of 20 studies in which the authors were a combination of outside academicians and internal program staff. This category had the largest mean effect size (0.20), followed by internal evaluators (0.16), and the lowest average *d* for studies conducted by external evaluators (0.02). The average *d* across all experiments was 0.10.

But Table 1 masks some important distinctions because of its gross categorization. For example, government evaluators asked to evaluate a government-administered program are coded as internal evaluators, e.g., a state-level agency researcher conducts an experiment in a state-run correctional institution. But they may have very little influence over the treatment or program setting. Table 2 therefore presents the mean effect size along the continuum of influence in the intervention setting (as rated by the first author).

Consistent with prior meta-analyses in the offender treatment literature, Table 2 shows a much larger effect size for experiments in which evaluators are rated as having "high" influence in the intervention setting (0.40). Experiments in which evaluators have moderate (0.03) or low (0.02) influence had considerably smaller effect sizes.

Table 2. Rating by reviewer of the evaluation team's influence on the design and implementation of the intervention and "first posttreatment effects."

Rating by reviewer of evaluation team influence	*N*	*Effect size*
High	59	0.40
Moderate	69	0.03
Low	152	0.02
All Studies	281[a]	0.10

[a] Nineteen cases were missing.

Table 3. Type of evaluator and "first effects."

Description of evaluation team	*N*	*Effect size*
Internal		
Program Developer/Creator	24	0.47
Program/Agency Staff	51	0.17
Government Evaluator	62	0.02
External		
Academic Researcher/Professor/Graduate Student	73	0.01
Private Research Firm	48	0.04
Foundation/Other Nonprofit	3	−0.05
Collaboration		
Academic/Practitioner	18	0.22
Academic/Government	2	0.04

Finally, we looked at the specific roles that the evaluators had in the intervention setting. As Table 3 demonstrates, program developers/creators conducted 24 experiments. Those experiments reported an average effect size of 0.47, more than twice the next largest category (0.22 for studies conducted by collaborative teams of academicians and practitioners). It also reinforces our decision to report more specifically beyond simple internal and external evaluation categories. Experiments conducted of government programs by government evaluators reported an average effect size of 0.02. Clearly, external evaluators, all things being equal, report very low effect sizes across their studies: In the most substantial category, the 73 experiments conducted by external academicians or graduate students averaged a Cohen's *d* of 0.01.

Discussion

Prior meta-analyses and our analysis of 300 randomized field trials confirmed that the involvement of the evaluator in the development, design, and implementation of a program is a positive and influential factor on effect size. Consider that the average effect size across all 300 (regardless of whether they reported data on role of evaluator) experiments was 0.11, and the mean Cohen's *d* across the 24 trials reported by developers/creators was 0.47. This is not a trivial finding. Using simple conversion statistics available in meta-analytic textbooks (Lipsey and Wilson 2001), and assuming a baseline rate of success of 50% for each group, an effect size of 0.47 corresponds to a rate of success of 61.75% for the experimental group compared to a rate of success of 38.25% for the control group.

These data do not reveal why developers and other evaluators with a high degree of influence in the program setting report substantially larger effect sizes. Lipsey (1995: 76) sets out two competing theories about why this might be:

"A cynical view might attribute this to some biasing or 'wish fulfilling' influence researchers have on the outcomes of the studies they control. I see another interpretation as plausible … when a researcher is closely involved in treatment design … there is likely to be a high level of treatment integrity."

The cynical view

The cynical view sees results such as those presented here as evidence that there are subtle and overt pressures on evaluators to report positive findings. If an evaluator is the program developer, or part of the program staff, (s)he may be under considerable pressure to show that the program works. Such pressure could stem from making sure the program is seen as worthy of retaining, the prestige that comes with developing an effective social program, the financial reward if one's program is adopted by others (e.g., when a developer of school-based violence prevention curriculum gets paid a fee every time another school adopts copyrighted materials), or the increased likelihood of attracting additional or new sources of grants and funding for the program.[1]

Although intentional distortion is hopefully rare,[2] these pressures may lead to subtle strategies to paint their program in the best possible light. Only a few these would appear to impact meta-analysis. For example, developer/evaluators may hold back negative findings about their program and stash them in their file drawer, never to see the light of day. This is a major concern in the pharmaceutical industry, where drug companies report positive findings and squash negative ones, and the medical community has urged reform (e.g., Rincon 2004). Squashing negative results would result in meta-analyses that do not consider the sum total of relevant studies, but only included positive evaluation studies, resulting in inflated program effects. There are statistics to estimate how many 'file drawer' studies of zero impact it would take to downwardly influence effect size estimates in meta-analysis.

Another subtle strategy is to conduct a multitude of statistical analyses but report extensively on the one or two that show statistically significant effects. Evaluators have long been warned about the problems of capitalizing on chance, i.e., that with 20 statistical runs, one will be statistically significant by chance probability alone. Although attempts are sometimes made to contact authors to determine this, meta-analysts almost always is limited by what is reported in the evaluation document.

Bias may also subtly influence reporting by the attention given by developer/evaluators to subgroup analyses in the report. Searching and reporting moderating or subgroup effects is a legitimate empirical strategy, as it may identify hypotheses for further experimental study. But when a program developer emphasizes results for particular smaller subgroups that are positive (e.g., boys 10–14 who completed treatment) but ignores the main effect (i.e., the full experimental versus control comparison), healthy skepticism about the findings seems warranted. In addition, developers may use statistical strategies that skew the chances of a positive result in their program's favor. For example, setting significance levels at 0.10 obviously makes it much easier to report a statistically significant result than the conventional 0.05. Gorman (2003) has consistently shown how these methods used by developers in the drug and violence prevention field have led to certain programs being declared 'model,' 'exemplary,' or 'promising,' by best practice lists, when the evidence for program effectiveness is decidedly less clear. These strategies are less damaging to meta-analysis, unless the main effect data is not reported or cannot be obtained. But in such cases the study would be excluded.

Part of the cynical view is the possibility of an 'experimenter expectancy effect,' or how the hypothesis or expectation of the investigator becomes a self-fulfilling pro-

phecy of participant responses. Following his review of psychological experiments, Rosenthal (1976) reported that investigators influenced the research setting through overt and subtle cues that communicated to the participants about what was 'expected' from them. This experimenter expectancy effect was more pronounced when the outcomes included rating systems and attitudinal tests, and less pronounced with official and administrative data collected outside of the research setting. Whether the experimenter expectancy effect is influential in criminal justice settings (particularly with recidivism data) is an empirical question. It is true that meta-analyses typically report much smaller effects on behavioral rather than attitudinal and psychological measures, but it is difficult to determine whether this is due to the relative weakness of intervention or to experimenter expectancy effects.

High fidelity view

A competing theory holds that the larger effects reported by program developers evaluating their own programs occurs because they are able to achieve high fidelity conditions. They create 'hot house' conditions necessary for strong implementation and sustenance of the program. This includes the special training of program staff, or the direct delivery of treatment by the evaluators, ensuring that protocols are fully adhered to. It is also possible that the engagement and enthusiasm of the developer (sometimes referred to as 'charisma') inspires, leads, and motivates staff in ways that cannot be replicated when the program is more widely disseminated.

Developers are usually able to effectively oversee and monitor their programs, because they are generally smaller in scope and number of participants (Lipsey 2003). Sample size, in general, appears to be a powerful positive influence on effect size, with smaller sample studies reporting larger effects (e.g., Weisburd et al. 1993). Loesel and Beelman (2003) reported that experiments of child skills training had much larger effect sizes when the authors did the training and the total sample size was under 100. Such conditions are difficult to maintain when programs go to scale and are widely disseminated, what one researcher has called going from the 'hot house' to the 'out house' (DeJong, personal communication).

Some support for the high fidelity view comes from a meta-analysis of 'multi-systemic therapy' for offenders (Curtis et al. 2004). In this study, the reviewers compared studies in which the developers were actively involved in training and monitoring, and Ph.D. level students delivered treatment with those evaluations in which the developers were not actively involved and Masters-level students delivered the intervention. They found the average effect size for active involvement and Ph.D. staff was 0.81 versus 0.28 for less involvement and Masters-level program staff.

Conclusion

Our paper, consistent with prior meta-analyses in offender rehabilitation, finds that studies in which evaluators were greatly influential in the design and implementation of treatment report consistently and substantially larger effect sizes than other types of evaluators. This is especially true of program developers. Eleven

of 12 offender treatment meta-analyses that reported testing for this effect and our own quantitative synthesis of 300 randomized field trials support this result.

There is some preliminary evidence from one meta-analysis that the high fidelity conditions that developer/evaluators are able to achieve in their initial studies that may explain these larger effects (Curtis et al. 2004). One immediate question, but one long considered by the research community, is how to ensure greater fidelity as programs are widely disseminated beyond 'hot house' conditions to a larger number of sites.

We recognize that there may be other explanations for these findings. It cannot be ruled out that program developers are simply designing and testing 'smarter interventions.' It is possible that developers, working in an area for a considerable time period, are more likely to implement programs based on scientifically sound theories of offender treatment. Moreover, given the broad range of intervention types considered by the meta-analyses, it would be important to determine whether developers create and implement a particular strategy over others, such as cognitive–behavioral treatment. Could it be the shared use of a more effective modality rather than the influence of the evaluator that explains the results?

Such a question begs for more research using meta-analysis. Follow-up studies should look at the characteristics of investigators, programs and studies in which developers have reported such comparatively large effects. This would not only help us understand why program developers report larger effects, but to determine the lessons that are necessary to assist larger program development, implementation, and dissemination efforts in criminal justice. It would also be relevant to examine the impact of program developers as evaluators in other fields, such as education and social services.

It is probably naïve to think that program staff, government workers evaluating their own programs – or consultants receiving income from an agency for evaluation – are above subtle and overt pressures to report positive findings. Campbell (1969) recognized the difficulties that government administrators faced if an evaluation of a program they supported under their tenure was found to be unsuccessful by a careful study. Campbell (1969) recommended that incentives be built into government so that administrators *are rewarded for producing the information* by the evaluation, and not punished for its results. This is the kind of culture that is necessary, so that program developers and others intimately involved with the treatment setting will be empowered to report accurately and comprehensively on their studies and results.

Given that evaluator involvement in the criminal justice research setting is being encouraged (e.g., Visher and Weisburd 1997), we wonder whether there are any strategies that could be implemented that would help address the concerns raised by both the cynical and high fidelity views. Are there oversight mechanisms that would help to ensure that the data and resulting report from an evaluation study validly represent the intervention's effect? One possible model for this is the U.S. National Institute of Health's requirement that each center or institute conducting a clinical trial create a Data Safety and Monitoring Board (DSMB). This board monitors a clinical trial from beginning to end, providing a further check for integrity of findings.[3]

Acknowledgements

This paper was prepared for the Social Methods Seminar, Institute for Evidence-based Social Work Practice (IMS), National Board of Health and Welfare, Stockholm, Sweden, on December 9, 2004. We appreciate the helpful comments of David Weisburd, Sir Iain Chalmers, and three anonymous peer reviewers on this draft. This work was supported in part by IMS, and earlier grants by the Smith-Richardson Foundation grant to the Jerry Lee Center of Criminology, University of Pennsylvania; the Mellon Foundation to the Center for Evaluation, Initiative for Children Program, American Academy of Arts and Sciences; and the U.S. National Institute of Justice to Rutgers School of Criminal Justice. The paper, however, is solely the work of the authors and does not represent any other person or institution.

Notes

1 Such pressures are not limited to program developers. Although the results are not borne out by our analyses, governmental researchers evaluating a particular agency's program may be under pressure to present results in the best possible light for the agency. This is particularly acute in agencies such as those in criminal justice that are highly politicized. Every researcher within government seems to know of a 'story' in which results were distorted or deleted in order to make the agency's initiative appear better than it was. For example, in a remarkable randomized trial in California, government researchers Berocochea and Jaman (1981) report on the results of incarcerated inmates being released six months early from their sentence with those who were not. Although the results were downplayed, Cook (personal communication) showed how releasing inmates six months early led to a larger and statistically significant negative result. He surmises that this was downplayed because it was not the 'right answer' desired by the government at the time. Also note that the first author was employed by a state justice agency that had received millions of dollars to participate in a widely touted crime prevention program under President George Herbert Bush during the early 1990s. The U.S. Attorney General was coming in for a site visit and wanted to know what the agency's evaluation showed after a year or so of program operation. But there was a big problem: no evaluation had been done. The agency director called the chief of research into his office and demanded that "an evaluation be delivered within 24 hours and it had better be positive." The research staff collected as much anecdotal and descriptive information as possible during the time frame and wrote a suggestive report, highlighting the good things the program was doing. In the morning, the report was handed to the agency director, who only glanced at its cover and felt its weight – and exclaimed, "This is perfect!" The evaluation was descriptive and would not be included in a meta-analysis such as those described here, but highlight the unique pressures faced by researchers in such contexts.

2 Intentional distortion is typically a hidden phenomenon in scientific practice. In recent years, National Science Foundation, the National Institutes of Health, the office of Scientific Integrity, and scientific organizations such as the National Academy of Sciences have explored intentional distortion by scientists.

3 We thank an anonymous reviewer for suggesting this. A report on the DSMB can be found at the US National Institute of Health (1998) website at http://grants.nih.gov/grants/guide/notice-files/not98-084.html

References

Alexander, M. A. (1999). Sexual offender treatment efficacy revisited. *Sexual Abuse: A Journal of Research and Treatment* 11(2), 101–117.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P. & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology* 28(3), 369–404.

Antonowicz, D. H. & Ross R. R. (1994). Essential components of successful rehabilitation programs for offenders. *International Journal of Offender and Comparative Criminology* 38(2), 97–104.

Berecochea, J. & Jaman, D. (1981). *Time served in prison and parole outcome: An experimental study. Report 2*. Sacramento: California Department of Corrections.

Boruch, R. F. (1997). *Randomized experiments for policy and planning*. Newbury Park, CA: Sage.

Boruch, R. F. & Petrosino, A. (2004). Meta-analyses, systematic reviews, and research syntheses Chapter 7. In J. Wholey, H. Hatry & K. Newcomer (Eds.), *Handbook of practical program evaluation*. 2nd edn. San Francisco: Jossey-Bass.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist* 24, 409–429.

Cleland, C. M., Pearson, F. S. & Lipton, D. S. (1996). A meta-analytic approach to the link between needs-targeted treatment and reductions in criminal offending. Paper presented at the Annual Meeting of the American Society of Criminology, Chicago, Illinois.

Cook, T. D. (2005). Why have educational evaluators chosen not to do randomized experiments? Paper presented at Random Evaluation in a Non-Routinized Environment: New Perceptions on R&D in Education and Other Social Sectors, Ecole Polytechnique Federale de Lausanne, Switzerland, April 9th.

Cox, S., Davidson, W. & Bynum, T. (1995). A meta-analytic assessment of delinquency-related outcomes of alternative education programs. *Crime and Delinquency* 41(2), 219–234.

Curtis, N. M., Ronan, K. R. & Borduin, C. M. (2004). Multisystemic treatment: A meta-analysis of outcome studies. *Journal of Family Psychology* 18(3), 411–419.

Davidson, W. S., Redner, R., Mitchell, C. M. & Amdur, R. (1990). *Alternative treatments for troubled youth.* New York: Plenum.

Dowden, C. & Andrews, D. A. (1999a). What works for female offenders: A meta-analytic review. *Crime and Delinquency* 45, 438–452.

Dowden, C. & Andrews, D. A. (1999b). What works in young offender treatment: A meta-analysis. *Forum on Corrections Research* 11(2), 21–24.

Dowden, C. & Andrews, D. A. (2000). Effective correctional treatment and violent reoffending: A meta-analysis. *Canadian Journal of Criminology* 449–467.

Dowden, C., Antonowicz, D. & Andrews, D. A. (2003). The effectiveness of relapse prevention with offenders: A meta-analysis. *International Journal of Offender Therapy and Comparative Criminology* 47(5), 516–528.

Drug Strategies (1999). *Making the grade: A guide to school drug prevention programs. Updated and expanded*. Washington, DC: Drug Strategies.

Egg, R., Pearson, F. S., Cleland, C. M. & Lipton, D. S. (2000). Evaluations of correctional treatment programs in Germany: A review and meta-analysis. *Substance Use and Misuse* 35, 1967–2009.

Farrington, D. & Welsh, B. (2003). Family-based prevention of offending: A meta-analysis. *Australian and New Zealand Journal of Criminology* 36, 127–151.

Gallagher, C. A., Wilson, D. B., Hirschfield, P., Coggeshall, M. B. & MacKenzie, D. L. (1999). A quantitative review of the effects of sex offender treatment on sexual re-offending. *Corrections Management Quarterly* 3, 19–29.

Garrett, C. J. (1985). Effects of residential treatment on adjudicated delinquents: A meta-analysis. *Journal of Research in Crime and Delinquency* 22(4), 287–308.

Gendreau, P. & Goggin, C. (1996). Principles of effective programming with offenders. *Forum on Corrections Research* 8(3), 38–40.

Gendreau, P., Goggin, C. & Cullen, F. (1999). *The effects of prison sentences on recidivism*. Ottawa, ON: Solicitor General Canada.

Gensheimer, L. K., Mayer, J. P., Gottschalk, R. & Davidson, W. S. (1986). Diverting youth from the juvenile justice system: A meta-analysis of intervention efficacy. In S. J. Apter & A. P. Goldstein (Eds.), *Youth violence* (39–57). New York: Pergamon Press.

Gorman, D. M. (2003). The best of practices, the worst of practices: The making of science-based primary prevention programs. *Psychiatric Services* 54(8), 1087–1089.

Gottschalk, R., Davidson W. S. II, Gensheimer, L. K. & Mayer, J. P. (1987). Community based interventions. In H. C. Quay (Ed.), *Handbook of juvenile delinquency* (266–89). New York: Wiley and Sons.

Hall, G. C. N. (1995). Sexual offender recidivism revisited: A meta-analysis of recent treatment studies. *Journal of Consulting and Clinical Psychology* 63(5), 802–9.

Hanson, R. K., Gordon, A., Harris, A. J. R., Marques, J. K., Murphy, W., Quinsey, V. L. & Seto, M. C. (2002). First report of the collaborative outcome data project on the effectiveness of psychological treatment for sex offenders. *Sexual Abuse: Journal of Research and Treatment* 14, 169–195.

Huebner, T. (2000). Theory-based evaluation: Gaining a shared understanding between school staff and evaluators. *New Directions in Evaluation* 87.

Izzo, R. L. & Ross, R. R. (1990). Meta-analysis of rehabilitation programs for juvenile delinquents. *Criminal Justice and Behavior* 17(1), 134–42.

Kaufman, P. (1986). Meta-analysis of juvenile delinquency prevention programs. Unpublished paper. Claremont, CA: Claremont Graduate School.

Latimer, J. (2001). A meta-analytic examination of youth delinquency, family and recidivism. *Canadian Journal of Criminology and Criminal Justice* 43(2).

Latimer, J., Dowden, C. & Muise, D. (2001). *The effectiveness of restorative justice practices: A meta-analysis*. Ottawa: Department of Justice.

Latimer, J., Dowden, C. & Morton, K. (2004). *Treating youth in conflict with the law: A new meta-analysis*. Ottawa: Research and Statistics Division, Department of Justice Canada.

Lipsey M. W. (1992). In Cook, T. C. Cooper, H. Cordray, D. S. Hartmann, H. Hedges, L. V. Light, R. L. Louis, T. A. & Mosteller F. M. (Eds.), *Meta-analysis for explanation* (83–127). New York: Russell Sage.

Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In McGuire J. (Ed.), *What works? Reducing reoffending*. New York: Wiley.

Lipsey, M. W. (2003). The good, the bad, and the ugly: The potential confounding role of moderators in meta-analysis. *Annals of the American Academy of Political and Social Science*.

Lipsey, M. W. & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 313–345). Thousand Oaks: Sage.

Lipsey, M. & Wilson, D. (2001). *Practical meta-analysis*. Newbury Park, CA: Sage Publications.

Lipton, D. S., Pearson, F. S., Cleland, C. M. & Yee, D. (2003). The effects of therapeutic communities and milieu therapy on recidivism: Meta-analytic findings from the correctional drug abuse treatment effectiveness (CDATE) study. In J. McGuire (Ed.), *Offender rehabilitation and treatment: Effective programmes and policies to reduce re-offending* (pp. 39–77). London: John Wiley.

Loesel, F. & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *Annals of the American Academy of Political and Social Science* 587, 84–109.

Loesel, F. & Koferl, P. (1989). Evaluation research on correctional treatment in West Germany: A meta-analysis. In H. Wegener, F. Losel & J. Haisch (Eds.), *Criminal behavior and the justice system* (334–55). New York: Springer-Verlag.

Loesel, F. & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology* 1, 117–146.

Love A. J. (1991). *Internal evaluation: Building organizations from within*. Newbury Park, CA: Sage.

MacKenzie, D. L., Wilson, D. B. & Kider, S. (2001). Effects of correctional boot camps on offending. *Annals of the American Academy of Political and Social Science* 578, 126–143.

Mayer, J. P., Gensheimer, L. K., Davidson, W. S. II & Gottschalk, R. (1986). Social learning treatment within juvenile justice: A meta-analysis of impact in the natural environment. In S. J. Apter & A. P. Goldstein (Eds.), *Youth violence* (24–38). New York: Pergamon Press.

Pearson, F. S. & Lipton, D. S. (1999). A meta-analytic review of the effectiveness of corrections-based treatments for drug abuse. *The Prison Journal* 79, 384–410.

Pearson, F., Lipton, D., Cleland, C. & O'Kane, J. (1995). Meta-analysis on the effectiveness of correctional treatment: Another approach and extension of the time frame to 1994. A progress report. Presentation at the American Society of Criminology Annual Meeting, Boston, Massachusetts, November 15th.

Pearson, F. S., Lipton, D. S. & Cleland, C. M. (1996). Some preliminary findings from the CDATE Project. Presentation at the American Society of Criminology, Chicago, Illinois, November.

Petrosino, A. J. (1997). *What works? Revisited again: A meta-analysis of randomized experiments in individually-focused crime reduction interventions.* Ph.D. dissertation, Rutgers University. Ann Arbor, MI: University Microfilms.

Petrosino, A., Turpin-Petrosino C. & Buehler J. (2003). Scared straight and other juvenile awareness programs for preventing juvenile delinquency: A systematic review of the randomized experimental evidence. *Annals of the American Academy of Political and Social Science* 589, 41–62.

Redondo, S., Sánchez-Meca, J. & Garrido, V. (2001). Treatment of offenders and recidivism: Assessment of the effectiveness of programmes applied in Europe. *Psychology in Spain* 5, 47–62.

Rincon, P. (2004). Secrecy penalizes cancer patients. Online at: http://news.bbc.co.uk/1/hi/sci/tech/3632882.stm (Last accessed on September 10, 2005).

Roberts, A. R. & Camasso, M. J. (1991). The effect of juvenile offender treatment programs on recidivism: A meta-analysis of 46 studies. *Notre Dame Journal of Law, Ethics and Public Policy* 5(2), 421–41.

Rosenthal, R. (1976). Experimenter effects in behavioral research. New York: John Wiley.

Shepherd, J. P. (2003). Explaining feast or famine in randomised field trials: Medical science and criminology compared. *Evaluation Review* 27, 290–315.

Sherman, L. W. (1998). *Evidence-based policing*. Washington, DC: Police Foundation.

Sherman, L. W. (2004). Research and policing: The infrastructure and political economy of federal funding. *The Annals of the American Academy of Political and Social Science* 593, 156–78.

US Bureau of Justice Assistance, Center for Program Evaluation, 2005, Online at: http://www.ojp.usdoj.gov/BJA/evaluation/glossary/glossary_m.htm (last accessed on September 10, 2005).

US National Institute of Health, (1998). NIH Policy for Data Safety Monitoring, online at http://grants.nih.gov/grants/guide/notice-files/not98-084.html (last accessed on September 10, 2005).

Visher, C. & Weisburd, D. (1997). Identifying what works: Recent trends in crime prevention strategies. *Crime, Law and Social Change* 28, 223–42.

Weisburd, D. L. (1996). Preface, In L. Greene (Ed.), *Policing places with drug problems*. Thousand Oaks, CA: Sage.

Weisburd, D., Sherman, L. & Petrosino, A. J. (1990). *Registry of randomized experiments in criminal sanctions, 1950–1983*. Los Altos, CA: Sociometrics Corporation.

Weisburd, D. L., Petrosino, A. J. & Mason, G. (1993). Design sensitivity in criminal justice experiments: Reassessing the relationship between sample size and statistical power. *Crime & justice: An annual review of research* (vol. 17). Chicago: University of Chicago Press.

Wells-Parker, E., Bangert-Drowns, R., McMillen, R. & Williams, M. (1995). Final results from a meta-analysis of remedial interventions with drink/drive offenders. *Addiction* 90, 907–926.

Wilson, D. B. (2001). Effect size determination program. Online at http://mason.gmu.edu/~dwilsonb/downloads/es_calculator.zip (last accessed Nov. 16, 2005).

Wilson, S. J. & Lipsey, M. W. (2000). Wilderness challenge programs for delinquent youth: A meta-analysis of outcome evaluations. *Evaluation and Program Planning* 23, 1–12.

Wilson, D. B., Gallagher, C. A. & MacKenzie, D. L. (2000). A meta-analysis of corrections-based education, vocation, and work programs for adult offenders. *Journal of Research in Crime and Delinquency* 37, 347–368.

Whitehead, J. T. & Lab, S. P. (1989). A meta-analysis of juvenile correctional treatment. *Journal of Research in Crime and Delinquency* 26, 276–295.

Woolfenden, S. R., Williams, K. & Peat, J. (2003). *Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10–17. Cochrane Review.* Oxford, UK: Update Software.

About the authors

Anthony Petrosino is a research and evaluation consultant based in Chelmsford, MA. He received his Ph.D. in criminal justice from Rutgers University in 1997, and received a Spencer Foundation postdoctoral fellowship in evaluation at the Harvard Children's Initiative. He has worked for the New Jersey Division of Criminal Justice, the Massachusetts Executive Office of Public Safety and the American Academy of Arts and Sciences. Anthony also served as Founding Coordinator for the Campbell Collaboration Crime and Justice Group and was recently named Honorary Fellow by the Academy of Experimental Criminology. Recent publications include a biography of Harvard statistician Frederick Mosteller, published by the James Lind Library at: http//www.jameslindlibrary.org/trial_records/20th_Century/1970s/bunker/bunker_biog.pdf.

Dr. Haluk Soydan is a Research Professor and the Co-Director of the Hamovitch Center for Science in the Human Services, University of Southern California, school of Social Work, in Los Angeles. He is founding member and Co-Chair of the international Camplbell Collaboration. Before joining USC, he served as Reserach Director at the National Board of Health and Welfare in Stockholm, Sweden.