

No effects in independent prevention trials: can we reject the cynical view?

Manuel Eisner

Published online: 11 March 2009
© Springer Science + Business Media B.V. 2009

Abstract Recent studies suggest that the reported effect sizes of prevention and intervention trials in criminology are considerably larger when program developers are involved in a study than when trials are conducted by independent researchers. This paper examines the possibility that these differences may be due to systematic bias related to conflict of interest. A review of the evidence shows that the possibility of a substantial problem cannot be currently rejected. Based on a theoretical model about how conflict of interest may influence research findings, the paper proposes several strategies to examine empirically the extent of systematic bias related to conflict of interest. It also suggests that, in addition to improved standards for conducting and publishing future experimental studies, more research is needed on the extent of systematic bias in the existing body of literature.

Keywords Conflict of interests · Independent evaluation · Methodological bias · Research synthesis

Introduction

Over the past three decades or so criminological prevention research has been carried by a wave of cautious optimism. In particular, there has been mounting confidence that we can advance crime prevention by learning from the combination of carefully designed experimental or quasi-experimental trials and systematic analyses of progress in the field. Many such systematic reviews have been conducted during the past 10 years. Most of them arrive at the encouraging conclusion that many types of universal, selective and indicated prevention can be effective in

This paper was first presented at the 8th Jerry Lee Crime Prevention Symposium, 5 and 6 May 2008, University of Maryland, USA.

M. Eisner (✉)
Institute of Criminology, Sidgwick Avenue, University of Cambridge, Cambridge CB3 9DT, UK
e-mail: mpe23@hermes.cam.ac.uk

reducing crime and antisocial behavior (Beelmann and Lösel 2006; DuBois et al. 2002a; Farrington and Welsh 2003; Lipsey 1995; Lipsey and Cullen 2007; Sherman et al. 2002; Wilson et al. 2003a, b).

However, in recent years, there has also been concern about whether the two most powerful instruments of evidence-based criminology—experimental primary studies and secondary meta-analyses—produce unbiased estimates of the truth (e.g., Gandhi et al. 2007; Gorman 2005b; Littell 2005; Petrosino and Soydan 2005). The problem is exemplified by a recent meta-analysis by Petrosino and Soydan (2005). They found impressive positive effects in crime-prevention studies conducted by developers-as-evaluators and no effects whatsoever in similar independent trials. The authors discuss two possible explanations for these massive differences: the *high fidelity view* holds that the implementation quality is better in studies in which the program developer is responsible for the implementation. This hypothesis assumes that nothing is substantially wrong with the study results. The only problem is that the findings lack external validity, i.e., they cannot be generalized to routine applications of the same program. In contrast, the *cynical view* assumes that the discrepancy arises from a much more malignant problem. It refers to the possibility that the more positive findings in developer-led studies essentially stem from systematic bias due to problematic decisions during an experimental study. The main assumed cause is a conflict of interest between the role of researchers as disengaged and skeptical truth finders and their roles as enthusiastic advocates of a specific program, a role often linked to significant financial and organizational stakes.

The prospect that some experimental and quasi-experimental studies might be biased could have serious consequences for the idea of evidence-based prevention. Proponents of evidence-based crime prevention want to provide practitioners and politicians with the best possible information about “What works, what doesn’t, and what’s promising” (Sherman et al. 2002). However, if there is hidden research bias in the system—perhaps to a different extent in different prevention areas—the whole program of evidence-based prevention is open to serious criticism, both methodological and ethical.

This paper tries to advance the debate by suggesting ways to examine empirically the extent of bias that arises from conflict of interest. In the first part I summarize the empirical evidence. In the *Is there a smoking gun?* section I suggest a theoretical framework to study bias arising from conflict of interest. In the *Conflict of interest and biased research: a model* section I present ideas about research that might help to illuminate our understanding of research bias.

Is there a smoking gun?

Research in fields other than criminology has conclusively demonstrated that conflict of interest has a significant impact on evaluation outcomes. Perlis et al. (2005), for example, examined the results of 397 clinical trials in psychiatry. They found that among the 162 randomized, double-blind, placebo-controlled studies they examined, those that had disclosed a conflict of interest were 4.9-times more likely to report positive results. Also, Friedman and Richter (2004) analyzed 398

manuscripts submitted to two leading bio-medical journals in 2001. Their findings suggested that papers whose authors experienced a financial conflict of interest were 10–20-times less likely to report negative findings than independent papers. Similarly, Okike et al. (2007) found that, in over 500 presentations at the Annual Meetings of the American Academy of Orthopaedic Surgeons, independent researchers were approximately ten-times more likely to report non-successful outcomes than researchers with a financial interest.

In criminology no direct evidence exists on the effects of conflict of interest on evaluation outcomes because no such research has yet been conducted. However, there is circumstantial evidence suggesting that there might be a substantial problem. It comes from three types of sources. The first are findings from meta-analyses that include some measure of the relationship between the program developers, the implementation personnel, and the researchers. The second are methodological case studies and systematic reviews that search for possible evidence of method bias in published and unpublished studies. A third type comprises findings from independent studies that attempted to replicate results reported in developer-led trials.

The meta-analysis by Petrosino and Soydan

The study by Petrosino and Soydan (2005) should be considered first because it is currently the only meta-analysis in criminology that focused on the effects of developers-as-evaluators on reported prevention effects. It analyzed 300 distinct randomized field trials relevant to individually focused crime reduction. All types of programs were included, so long as the study comprised at least one outcome measure of official crime. For each study the authors coded only one effect size, defined as the first post-treatment effect. Importantly, the authors also collected several variables about the role of the evaluator in the treatment setting.

Overall, this meta-analysis showed a small positive effect size: Cohen's $d=0.11$. Yet, the most important finding was related to the differences in effect sizes by the role of the evaluation team. In studies in which the evaluator was the program developer, the reported effect size was $d=0.47$, by far the largest mean effect for all subcategories. In the studies in which the evaluation team was external to the program delivery, the mean effect size was exactly zero. The authors conclude 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" (Petrosino and Soydan 2005: 444). Indeed, someone looking only at independent evaluations would conclude that offender treatment is, on average, completely ineffective, while someone judging exclusively on the basis of developer-as-evaluator studies would see a highly useful group of prevention strategies.

Of course, these findings are open to different interpretations and do not conclusively demonstrate bias due to conflict of interest. For example, as is discussed in more detail below, involvement of the evaluator in the treatment setting is a variable that is probably indicative of potential conflict of interest, but it is an indirect measure that also comprises other aspects of an experimental study. Also, the study does not control for other moderators of effect size, such as

implementation quality or the number of participants, which may be correlated with the role of the evaluator. It is, therefore, unclear what the impact of developer involvement net of other influences would be.

Other meta-analyses

Petrosino and Soydan (2005) also reviewed 50 meta-analyses of offender treatment programs that included measures of criminal recidivism as outcomes. Of those, 12 operationalized some information that bore on the relationship between the program developer and the investigator. Researcher involvement in the program delivery is not a direct measure of conflict of interest, but it can be assumed to be a reasonable proxy. Petrosino and Soydan (2005) found that a clear pattern emerged from the 12 meta-analyses. In 11 of them, larger mean effect sizes were found when the evaluators were involved or influential in the program setting than when they were not. These differences could be substantial. Beelmann and Lösel (2006), for example, examined social skills programs. They found that programs delivered by teachers or psychosocial professionals yielded an average effect size of $d=0.29$ compared with an effect size of $d=0.49$ if the program had been delivered by the study authors or the research staff.

Methodological case studies and systematic reviews

While meta-analyses demonstrate a statistical pattern, some methodological case studies and systematic reviews have more thoroughly reviewed individual studies and highlighted patterns of methodological irregularities in some developer-led publications.

Littell (2005) and Littell et al. (2005), for example, conducted a systematic review of multisystemic therapy—a standardized treatment program for juveniles with serious behavior problems (Henggeler et al. 1996). They found that all but one study had been conducted by researchers associated with the program developers and that the only independent study had found no positive effects. In their analyses they noted a series of methodological problems that may partly account for the overall positive evaluation of multisystemic therapy (MST). More particularly, they noted a series of methodological weaknesses such as that post-intervention assessments were carried out in a blind fashion, that treatment completion was defined by subjective criteria, and that data suggested post-hoc sample refinement in some cases. Since the publication of Littell's study two new independent evaluations have been published. A study in Norway confirmed the positive results found by the program developers and demonstrated that effects could be maintained 2 years after the intervention (Ogden and Hagen 2006; Ogden and Halliday-Boykins 2004). In contrast, an evaluation in Sweden failed to find significant effects in comparison to treatment as usual (Sundell et al. 2008).

Gandhi et al. (2007) examined the evidence for five school-based drug abuse prevention programs that are recommended as scientifically proven 'model' or 'exemplary' programs on at least one of seven lists compiled by federal agencies in the United States of America. They found that the evidence for program effectiveness was generally mixed and that most positive findings related to increases in knowledge or attitudinal measures. In contrast, despite some claims by

the developers, little was known about the ability of the programs to reduce substance use successfully in the long term. Furthermore, they criticized a series of methodological flaws such as capitalizing on chance by conducting tests on many outcome variables or overemphasizing positive effects in subgroups that were defined post-hoc during the statistical analysis. They recommended that more independent studies should be undertaken, as little is known about “what the results would be if the program were evaluated by individuals who did not develop it” (Gandhi et al. 2007: 61).

Furthermore, Gorman (2003, 2005a, b) and Gorman and Conde (2007) have published several studies on possible methodological bias in developer-as-evaluator trials. They found a series of problematic methodological decisions that are consistent with the assumption of bias resulting from conflict of interest. They included selective reporting on positive results, ignoring problems associated with differential attrition, post-hoc definition of analyzed dataset, inconsistent ad-hoc definitions of the dependent variable, and the unwarranted use of one-tailed significance tests.

Independent evaluations of program effects

The third type of study that bears on the issue of systematic bias is independent evaluation of prevention programs whose results can be compared to the effects found in studies in which conflict of interest is present. Such comparisons differ from meta-analyses because the comparisons are made strictly between different tests of the same product. Of course, independent evaluations may also differ in respects other than being less prone to conflict of interest. For example, evaluations by outsiders tend to be large-scale field trials in which programs are tested under more demanding heterogeneous conditions. It is, nonetheless, instructive to look at some well-known prevention programs in which an independent evaluation has recently been conducted and to compare results with the findings reported by the program developers (see Table 1).

Table 1 Prevention effects in developer-led studies and independent trials

Program	Findings in developer-led studies	Independent replication
Reconnecting Youth (indicated drug prevention program)	Increased GPA; increased self-esteem; increased school bonding; decreased hard drug use; and decreased drug control problems (Eggert et al. 1994)	Negative effects on most outcome measures, no positive effects. Negative effects the stronger the better implementation fidelity (Sanchez et al. 2007)
Triple P (multilevel parenting skills program)	Positive mean effect on child problem behavior of $d=0.35$ in 33 trials (Nowak and Heinrichs 2008)	No positive effects on any aspect of problem behavior evaluated by teachers, parents, or child self-reports (Eisner et al. 2007)
Olweus Bullying Prevention Program	Reductions of up to 50% in bullying in the original study (Olweus 1994)	No overall effects on either attitudinal measures or victimization (Bauer et al. 2007)
ALERT (substance abuse program)	Reduction in cigarette, marijuana and alcohol use by 19–39% (Ellickson et al. 2003)	No effects on mediators or substance abuse itself (St Pierre et al. 2006)

GPA Grade point average

Reconnecting Youth, for example, is a drug abuse prevention program for truant, underachieving, high school students who are at increased risk for drug abuse and related problem behaviors (Eggert et al. 1990, 1994). Based on the positive evaluation results of the program developers it is, amongst others, a Substance Abuse and Mental Health Services Administration (SAMHSA) model program. Positive results mentioned on the SAMHSA factsheet—authored by the program developers—include curbed progression of drug and alcohol use, a 54% decrease in hard drug use, decreased suicidality, a 48% decrease in anger, 18% improvement in grades, and decreased high-school dropout. In contrast, a recent independent evaluation found consistent iatrogenic (i.e., negative) effects on all outcome measures at 6-month follow-up. Also, the negative effects of the program increased as the implementation quality improved (Cho et al. 2005; Sanchez et al. 2007).

Triple P is a multilevel parent training program developed by Sanders (1992, 1999). The program is primarily aimed at improving parenting practices and reducing child problem behaviors. It has been evaluated repeatedly over the past two decades by the program developer and by license holders across the world (de Graaf et al. 2008). A recent comprehensive meta-analysis comprised 55 studies worldwide that measured effects on parenting, child problem behavior, or parental well-being (Nowak and Heinrichs 2008). The mean effect size for the 43 studies that comprised child problem behavior as an outcome variable was Cohen's $d=0.35$. However, almost all studies were conducted by the program developers or researchers involved in the dissemination of the program. In contrast, Eisner et al. (2007) conducted a cluster-randomized field trial with Triple P as a universal prevention program, involving 1,300 children in 56 schools in Zürich, Switzerland. While Triple P staff were involved in the selection of the course providers and the planning of the delivery of the course, the evaluation itself was independent of the program developers. The evaluation confirmed positive effects on parenting according to the participants' self-report. However, it found no positive effect of Triple P on any of the child problem behavior measures (parent assessments, teacher assessments, and child self-assessments) and some evidence of a possible negative effect.

The *Olweus Bullying Prevention Program* is a model program on the Blueprints of Violence Prevention list (Olweus 1994). The US website of the program distributors refers to five trials on the effectiveness of the program. Publications on four of the five trials were authored or co-authored by the program developer. Each of those four publications reported positive effects. Olweus (1994), for example, reported a decrease in school bullying by 50%. A fifth experiment in 12 elementary schools comprising almost 6,000 students in Philadelphia, USA, was an independent evaluation (Bauer et al. 2007). That study found no overall effects of the Olweus program on either relational or physical violent victimization. Also, the study found no effects on student attitudes to intervene or in the perception of school safety.

Finally, St. Pierre et al. (2006) conducted an independent evaluation of *ALERT*, a drug prevention program developed by Ellickson and Bell (1990), Ellickson (1998) and Ellickson et al. (2003) that is on many recommendation lists of evidence-based programs. A recent large field study on a revised version of ALERT that was carried out by the program developers found that the curriculum curbed the initiation of cigarette and marijuana use, current and regular cigarette use, and alcohol misuse

(Ellickson et al. 2003). Yet, contrary to the findings by the program developers, the independent study discovered no evidence for any positive effects of the program, either on mediators or substance use itself.

DuBois et al (2002b), for example, conducted a meta-analysis of 55 studies of youth mentoring programs and found slightly better effects in independent evaluations than in internal evaluations.

Conflict of interest and biased research: a model

It should be emphasized here that the examples given in the previous section should not be interpreted as demonstrating that independent evaluations of a program always fail to show positive effects or that the evidence from systematic reviews invariably reveals a reduced impact in studies where the developers were not involved. Many examples of successful independent replications do exist, and they are rightfully regarded as particularly important supportive evidence for program effectiveness. Also, none of the patterns discussed in the preceding section provides conclusive evidence of systematic research bias in studies in which the program developer is involved. There are several other possible explanations that merit careful consideration. However, there is sufficient suggestive evidence to maintain that criminological prevention science should seriously explore the issue.

In considering the possibility of considerable bias in criminological prevention research, it is useful to develop a theoretical framework that clarifies how it may be caused (for an overview of the literature see, e.g., MacCoun 1998). To this purpose I propose a model that distinguishes between conflict of interest as the underlying cause, intentional misconduct and unintentional cognitive bias as mediating mechanisms, and actual problematic practices as the outcome (Fig. 1).

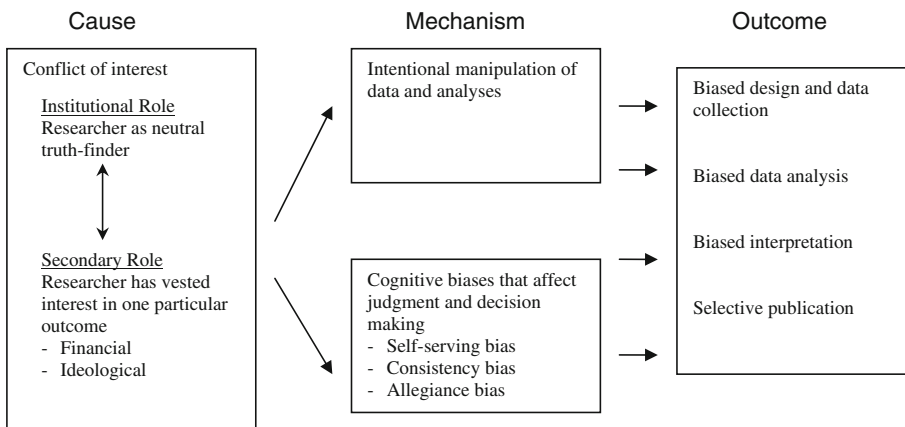


Fig. 1 Theoretical model of conflict of interest leading to research bias

Conflict of interest

In the judicial system many institutional arrangements are in place to guarantee that the judge and the jury are independent and impartial when reaching their verdicts. Similarly, research foundations across the world request that assessors of funding applications have no affiliations, shared financial interests, or other connections with the applicant. Also, ethics codes in medicine increasingly require that physicians should refuse gifts from pharmaceutical companies.

The reasons for these arrangements are easy to see: their purpose is to ensure that actors with an institutionally defined role are not influenced in their decision-making process by any factors other than their duty. More particularly, their judgments should reflect only the impartial assessment of the available evidence and not be influenced by any external factors. Conflict of interest generally refers to situations in which the main, institutionally-defined and publicly-declared interests of a professional compete with a person's secondary private self-interests (Gorman 2006).

The institutional role of researchers is to find the best approximations of truth through the use of scientific methods. Thus, a conflict of interest exists in evaluation research whenever a researcher or a research team benefits from or has an interest in one specific outcome. In criminological evaluation research, such situations seem to be quite widespread. Petrosino and Soydan (2005) report that, in 56% of 281 studies included in their meta-analysis, the evaluation team included staff who also delivered the program. In a review of best-practice drug-prevention programs comprised in the National Registry of Effective and Promising Programs (NREPP), Gorman and Conde (2007) found that 78% of 246 evaluation studies included the program developer as one of the study authors and that only 11% of the evaluations could be regarded as completely independent.

One can distinguish two main sources of conflict of interest in criminological evaluation research, namely financial interests and ideological interests (Gorman 2005b).

Financial interests Probably the most evident interest at stake in program evaluations concerns the financial interests related to a specific product. Such interests may be linked to copyrights, royalties, research funding and income generated from the distribution of programs (Resnik 2000). Conflict of interest arises if at least one member of the research team has a stake in the evaluated product. This entails, but is not limited to, the program developers. Rather, it extends to license holders and other researchers whose employment and future career may depend on the success of a particular program.

Material conflicts of interest are not necessarily associated with the personal enrichment of the program developer. They also encompass interests such as generating income for research staff and maintaining and expanding academic infrastructure. In particular, one should note that the development of new programs requires significant 'upfront' investments of time and resources, benefits the researcher may wish to reap. Furthermore, financial conflict of interest can also arise from commitments to third parties whose economic or political interests are at stake. Government agencies, for example, sometimes support evaluations while

simultaneously preparing the wider dissemination of a program. Hence, unless dissemination plans are suspended until the final research findings have been published, researchers may be under pressure not to compromise the vested interests of their project partners. This latter type of conflict may have become more prevalent to the extent that criminologists increasingly evaluate large-scale field experiments that often require significant initial commitments by government agencies.

Virtually nothing is known about the size of financial interests that are at stake in evaluations of criminological prevention programs. It is likely, though, that the significance of the problem varies between prevention areas. Thus, few criminologists may have direct interests in deterrence-oriented correctional interventions such as boot camps, while many specialist evaluators may be personally involved in offender treatment programs (Lipsey and Cullen 2007).

Also, the importance of financial stakes is likely to have grown substantially over the past two decades as research groups increasingly tend to operate as entrepreneurial “quasi-firms” (Etzkowitz 2003) who have a strong interest in generating income from the dissemination and evaluation of their programs to sustain their research activities. In some cases the income generated by the sale of programs, licenses, and merchandise can sustain sizeable organizations with 20 or more staff members, often associated with academic departments. Increasingly, too, prevention and intervention programs are rolled out internationally, with the consequence of complex networks of interdependence between the initial developers and the holders of national licenses, who are often also associated with academic institutions.

Finally, the spate of ‘best practice’ lists since the late 1990s may have exacerbated the tension between disinterested evaluation and the promotion of particular programs (Gorman 2005b). Many of these lists serve agencies at the state and federal levels to determine whether the introduction of a specific prevention program will receive financial support. As a result, best practice lists create a strong incentive to achieve inclusion on those lists, and there is little doubt that some researchers actively lobby for inclusion of their products in those lists.

Ideological interests Gorman (2006) identifies ideological conflict of interest as a second threat to prevention research. It arises in situations in which researchers hold strong normative views about core issues in their area of research. MacCoun (2005) argues that it is a kind of conflict that is particularly likely in research areas where academic disputes not only center on the means of achieving an agreed goal but where the ends are equally contested and researchers can hardly avoid taking sides in a polarized field of political and public debate. Such allegiances lead them to take on an advocacy role, which conflicts with their role as disinterested scientists and which been shown to result often in poor science associated with systematic bias (in criminology see, e.g., Gilbert 1997).

Criminology, of course, is rife with deep-rooted ideological conflicts related to different views about, for example, the ethical paradigms that should guide penal policy, the extent to which crime should be seen as a psychological disease or a free act, or in what areas of behavior the state is entitled to impose criminal sanctions. Examples of research areas affected by such value conflicts include the death penalty, gun control, early screening of at-risk children, the prohibition of drugs, or

restorative justice. Within crime prevention research, many established specialists are known for their commitment to and association with a particular prevention strategy. To a significant extent, such broader views are based on long careers of evaluation research and can, hence, rightfully claim to be evidence based. However, one should recognize that, even for the most scrupulous researchers, the degree to which wider world views can be derived from empirical evidence is severely limited.

Mechanisms

It is useful to distinguish two main mechanisms through which conflict of interest can lead to biased evaluation results. One is intentional manipulation of research findings, the second is a group of unconscious cognitive biases that influence academic judgment and decision making at various stages of the research process.

Intentional misconduct In some disciplines such as medicine intentional misconduct, including fraud and spin of evaluation findings, is increasingly regarded as a significant threat (Al-Marzouki et al. 2005b; Lock et al. 2001; Ranstam et al. 2000). In criminology, however, it is generally believed to be rare or nonexistent. Thus, Sherman (2006: 400) pointed out that there is no known case of intentional data falsification in anti-crime program evaluations. However, one should bear in mind that (to my knowledge) no-one has, as yet, really looked very closely. Also, one should note that in many evaluation studies the ingredients assumed by rational choice models of crime are present (Cornish and Clarke 1986; Felson 1994): Conflict of interest is a plausible motivating force; there exist opportunities; there are few capable guardians; and there is a very small risk of being discovered.

It therefore seems more prudent to assume that intentional misconduct is a plausible possibility. While wholesale fraud (i.e., the fabrication of significant parts of the data or the findings) is probably very rare indeed, ‘spin’ (e.g., selective reporting, post-hoc massaging of the data or improving outcomes by iterated trial and error analysis) could be more frequent in prevention research than is currently believed.

Cognitive biases Probably the more important mechanisms leading to result bias in the presence of conflict of interest are several types of cognitive biases that influence information processing but are not easily recognized as such by the researcher. Possibly, the most powerful such bias is *self-serving bias* (Babcock et al. 1995; Dana and Loewenstein 2003). In experimental psychology it refers to the tendency where people are not very good at disregarding their own self-interest and evaluating information impartially (Moore et al. 2006). Rather, individuals’ judgments about what is fair are typically biased in favor of their own self-interests.

In experiments people with self-interest tend to allocate more resources to themselves than is considered fair by independent observers (Messick and Sentis 1979). Also, people process information in a selective fashion when they have a stake in reaching a particular conclusion. They tend to pay more attention to evidence that supports the conclusion they would like to reach, they are inclined to disregard information that contradicts their views, and they evaluate supportive

information in a more uncritical fashion. Once they reach a conclusion that conforms to their interests, they tend to forget the contradictory information available to them and develop justifications for their self-serving decisions (Holyoak and Simon 1999).

In the presence of conflict of interest people will even tend towards views that are in favor of their interest, even if they themselves believe that they are fair and impartial (Dana and Loewenstein 2003). Furthermore, if people make successive decisions, self-serving bias can be reinforced by *consistency or confirmation bias*, i.e., the tendency to process information selectively so that they achieve increasing consistency with prior conclusions (Kaptchuk 2003; Russo et al. 2007).

Self-serving bias has been shown to be related to conflict of interest and to affect information processing amongst physicians receiving presents from pharmaceutical companies (Dana and Loewenstein 2003) or amongst auditors who tend to interpret factual information in favor of their clients (Moore et al. 2006). It could also be responsible for the *allegiance or experimenter expectancy effects* found by Luborsky et al. (1999). Essentially, their meta-analysis found that trials that compared the effectiveness of two psychotherapeutic programs usually found better effects for the program that the researcher favored.

Outcomes: where could bias happen?

Unconscious cognitive biases and intentional distortion can potentially affect the decision-making process at every stage of an outcome evaluation. Yet, there is currently limited information about which decisions in a research process are most prone to bias. Of course, an impressive body of literature specifies the methodological requirements for valid causal inference in experimental research (e.g., Farrington 2003; Lösel and Kofler 1989; Shadish et al. 2002), which can be read as instructions on how to avoid bias.

However, it seems valuable to list directly the specific misbehaviors that may lead to biased results. A literature review by Resnik (2000) provides a useful overview about where bias is most likely to occur. Also, Al-Marzouki et al. (2005b) conducted a Delphi study with 40 experts in clinical trials to understand better the techniques that lead to biased evaluation results. Both studies helped to compile the checklist of practices that lead to biased results in the experimental studies listed in Table 2. Behaviors are classified into five domains, namely study design, manipulation of dataset, definition of outcome variables, statistical analysis and reporting and dissemination (see Table 2).

At the study design and data collection phases, researchers make a series of decisions that have an effect on the likelihood of their finding positive effects. Especially experienced researchers are also likely to know how decisions at the design stage can influence study outcomes. In the presence of a conflict of interest they may be more inclined to develop designs that will help to produce positive findings. For example, designs that measure only outcomes as reported by the immediate trial participants tend to capitalize on the chance of expectancy and social desirability effects amongst those who received the treatment. Also, designs that exclusively measure effects immediately after an intervention but fail to include

Table 2 Practices conducive to biased evaluation results

Parameter	Characteristics
Study design and data collection	<ul style="list-style-type: none"> Pre-post design to capitalize on the chance that problem behavior may naturally decrease Immediate post-effects, but no follow up Unrealistic control group (no treatment instead of treatment as usual) Outcome measured from trial participants only, capitalizing on expectancy effects No blinding of outcome assessment Failure to develop written and published protocol for data analysis
Manipulation of data for analysis	<ul style="list-style-type: none"> Post-hoc exclusion of outliers or other cases Post-hoc choice of imputation strategy for missing values Post-hoc reallocation of observations to treatment and control groups Post-hoc definition of implementation quality subgroups
Definition of outcome variables	<ul style="list-style-type: none"> Post-hoc construction of outcome variable (measurement dependence) Post-hoc selection of outcome variables Ignore data on undesirable effects
Statistical analysis	<ul style="list-style-type: none"> Change outcome variables from one study to the other Alter analysis methods until finding a significant result Capitalize on multiple comparisons Selective subgroup analysis Use one-tailed significance tests Inadequate analysis of cluster-randomized data Select covariates until treatment effect is biased in the desired direction Downplaying of problems associated with lack of equivalence of treatment groups
Reporting and dissemination	<ul style="list-style-type: none"> Non-reporting of follow-up results Failure to report unfavorable results Selective reporting of positive results Selective reporting of subgroup analyses Non-publication of studies with negative results (file-drawer problem) Non-publication of outcomes with negative or non-significant results Over-interpretation of positive results in small trials Claim of analysis as “intention-to-treat” when it is not Non-declaration of conflict of interest

follow-up assessments increase the likelihood of finding short-term effects that are practically irrelevant, while reducing the risk of identifying the lack of long-term effectiveness. Furthermore, the chance of finding desirable effects is enhanced if an intervention group is compared with an artificial no-treatment condition instead of a treatment-as-usual condition.

The second domain relates to decisions about the dataset used for the published analyses. Typically, decisions at this level are about the exclusion of outliers, the imputation of missing values, the precise allocation of observations to treatment and control groups, and about assigning subgroups to different levels of implementation quality. Such decisions are necessary in most evaluation trials. However, they become problematic when taken after the outcome data have been collected and post-hoc adjustments become a tempting, albeit illegitimate, possibility. Nothing is known about how widespread this practice is, but some studies have documented problems in individual trials. Thus, Gorman (2005b) found evidence of post-hoc

sample refinement in studies on drug prevention programs. Similarly, Littell (2005) traced the number of cases included in unpublished and published analyses of MST trials. She found that case numbers changed from the unpublished reports to the published versions and that such changes were associated with higher effect sizes. Also, Eisner and Ribeaud (2008) reviewed a trial of Triple P conducted by the German license holders of the program (Heinrichs et al. 2006) and found that non-compliant respondents who had initially been allocated to the treatment group were considered as members of the control group in the statistical analysis, leading to severely inflated effect sizes.

Self-serving bias can also influence decisions regarding the dependent variables. In principle, inferential statistics based on deductive reasoning require that the dependent variable be operationalized before the trial and remain unchanged thereafter. However, dependent variables are usually created from multiple items and/or multiple sources of data. They are, hence, amenable to all kinds of manipulations. The first decision regards which measured variables are presented as outcome variables in a publication. This includes decisions about the number of items included in the outcome measure, cut-off points for creating dummy variables, statistical transformations, etc. Several studies have documented such problematic practices. Analyzing the Life Skills Training (LST) (Botvin et al. 1995) program, Gorman (2005a) found a variety of outcome measures used in different studies that did not appear to be motivated by changes in principal research questions. He therefore argued that the reported findings may have been measurement dependent.

Fourth, the statistical analysis is wide open to various strategies known as ‘capitalizing on multiple comparisons’ or ‘fishing expedition’ strategies, especially if no prior protocol for the data analysis has been specified or if it is not followed. Methodologists have repeatedly warned against failure to control for family-wise error or other kinds of post-hoc adaptations of the statistical analysis.

Separate analyses by implementation quality have become common since research emphasized the importance of high fidelity for achieving desirable results (Elliott and Mihalic 2004). Again, there is nothing to be said against separate analyses by implementation quality if the criteria of how implementation quality is to be measured are defined before the analysis of the data begins. However, if implementation quality is operationalized after the outcome data are known, the risk of post-hoc model fitting increases. This is especially so if fidelity and integrity are measured with multiple variables so that various alternative models can easily be explored.

A fifth set of issues relates to the publication and dissemination process. Possible sources of research bias are selective reporting of positive results, undue emphasis on positive results achieved in subgroups, improper claims about the methodological quality of a study, etc. Hewitt et al. (2008) recently pointed to the problem of interpretive bias. By this they meant the way in which non-significant differences between groups in a randomized controlled trial are interpreted. They argued that authors are more likely to interpret non-significant results that point to the desired direction as positive findings, while non-significant negative results are quickly discounted as being irrelevant.

In conclusion it is worth mentioning that the model outlined here has several implications: first, the assumption that bias may happen at every stage of a research process means that small biases can lead to considerable effects. If, for example, a researcher takes five subsequent problematic decisions, each of which is associated with a small effect size of Cohen's $d=0.10$, the final total bias will be $d=0.61$. If the problematic decisions are associated with an effect size of only $d=0.05$, the total bias would still be $d=0.30$ —an effect that many researchers would consider practically significant if resulting from a genuine treatment effect. Second, the list of problematic practices shows that only some can be detected in published academic work. Many others remain invisible unless a reviewer attempts to trace earlier research reports or to re-analyze the primary data. Third, the assumption of systematic bias can account for patterns found in criminological meta-analyses other than the different mean effect size in developer-led and independent evaluations. For example, meta-analyses regularly show that studies with a small N report much larger effect sizes than large field trials do. This could be explained as a side effect of problematic practices having more serious effects on small samples than on larger samples (e.g., outlier deletion, model fitting, revision of dependent variable measurement). Also, result bias may account for the fact that only a few types of prevention programs have been found to be unequivocally ineffective and that many of those found ineffective have been evaluated predominantly in independent trials.

What is to be done?

Over the past three decades significant effort has gone into the systematic synthesis of knowledge about effective crime prevention. It is reflected in the many meta-analyses that have scanned the evidence in major domains of criminological prevention science (Farrington and Welsh 2003; Lipsey and Cullen 2007; Lösel and Beilmann 2003; Petrosino and Soydan 2005; Wilson et al. 2003a). Most of these studies arrive at a surprisingly consistent conclusion: wherever the researchers look, they find evidence for a background radiation of positive prevention effects with typical mean effect sizes of Cohen's $d=0.20$ to $d=0.40$.

Perhaps these positive effects are unbiased estimates of the truth. However, the evidence presented in the previous sections gives cause for concern. Consider that Petrosino and Soydan (2005) found an effect size of $d=0.47$ for studies conducted by program developers and an effect size of $d=0.00$ for independent trials. The difference between these two values is thus larger than the mean overall effect size reported in most other criminological meta-analyses. If Cohen's $d=0.47$ is a realistic estimate for the mean amount of research bias, many of the positive effects reported in criminological meta-analyses could simply be artifacts.

We currently do not know the extent to which this is the case. Future research should therefore attempt to clarify the relevance of bias in prevention trials. Indeed, unless criminological prevention science understands better where and how conflict of interest leads to biased results, it will find it impossible to disentangle findings that can be trusted from those that cannot. In what follows I therefore suggest a series of strategies that could help us to understand the problem better.

Operationalize conflict of interest in meta-analyses

Many recent meta-analyses in criminology have collected variables on study characteristics that may act as moderators of the outcome. Variables of interest include the sample size, the type of program, the publication date, the target population, and the intensity of the intervention (see, e.g., Lösel and Beelmann 2003; Lundahl et al. 2006; Wilson and Lipsey 2007). However, to date, no criminological meta-analysis has collected information that attempts to measure conflict of interest directly, using multiple indicators of the target construct. Clearly, evaluator involvement in the program delivery—a moderator variable that is collected in many meta-analyses—is an insufficient proxy because it effectively measures a modality of program implementation rather than the self-interests of the researchers.

Although essential, collecting such data is far from easy, because virtually no outlet for criminological prevention trials requires a declaration of conflict of interest from authors of randomized controlled trials. The only approach, therefore, is to reconstruct such information from data that must be collected across a variety of documents. Table 3 provides suggestions for variables that might be used to measure conflict of interest in future studies.

Good quality data on the extent of conflict of interest *and* measures of implementation quality are required to answer the question posed by Petrosino and Soydan (2005), namely whether the larger effects reported in developer-led evaluations are due to bias resulting from conflict of interest or to better implementation quality. Since many existing large-scale meta-analyses already have collected data on the implementation process, the only work needed would be to add information about conflict of interest to existing databases. With such data it would be relatively easy to estimate the effect of bias net of differences in implementation quality.

Matched comparisons with independent trials

A problem highlighted above which causes significant concern relates to the considerable number of independent trials that fail to find significant positive effects and sometimes even find iatrogenic effects in programs believed to be highly effective. Sometimes, this seems to occur in studies with designs that are highly similar to those used in developer-led research. Our understanding of possible bias could be improved by systematically searching for well-designed independent prevention trials and to

Table 3 Some operationalizations of conflict of interest

Parameters
Financial conflict of interest
– Program developer is part of the evaluation team or one of the publication authors
– License holders or collaborators of the program developer are on the evaluation team
– Evaluator is involved in program roll-out after the evaluation
– Evaluator has other demonstrable material benefits from positive outcome
Ideological conflict of interest
– Evaluator is on advisory board of evaluated program
– Evaluator has published relevant work with program developer

match them with developer-as-researcher trials conducted on the same or a comparable program. Such matched-pair comparisons would reduce the problem of many meta-analyses, namely that studies differ on a large number of dimensions. In contrast, comparisons of matched pairs would reduce the number of possible sources for variation in outcomes, with differences in implementation quality and differences in research bias being the most plausible candidates.

Systematic search for problematic practices

Meta-analyses should be combined with more in-depth studies that scrutinize the frequency of problematic practices at various stages of the research. Scholars such as Littell (2005), Gorman (2005a) and Gandhi et al. (2007) have made important contributions to this field, highlighting practices likely to contribute to positive findings. However, work in this area could be greatly enhanced by the development of a comprehensive checklist of problematic, bias-prone practices likely to be found in criminological experiments. The list in Table 2 is a start, but it can certainly be significantly extended. Such a checklist could be used in systematic reviews and meta-analyses to identify the mechanisms that may have caused bias. Unfortunately, several items on this list require comprehensive documentation about a research project and cannot be answered on the basis of published study results alone.

If performed systematically, the administration of such a checklist could help to answer the important question of whether bias-generating practices are more frequent in developer-as-evaluator research than in independent trials. Also, such data could be used to examine which practices contribute most to the size of biased estimates of effect size.

Statistical techniques

Thirdly, several statistical analyses could be useful. For example, one could conduct independent re-analyses of existing datasets that have already been analyzed by researchers with a conflict of interest. If the findings are published, independent researchers would have precise instructions about which hypotheses were tested and which variables were examined. Independent re-analyses could be conducted to examine whether findings on all planned outcomes were published, whether post-hoc modifications of the dataset were methodologically sound, whether the statistical method was adequate and whether the published results can be replicated. Such studies would help to examine whether researchers with a conflict of interest are prone to publish findings that systematically err in favor of the desired outcome. In some areas of medical research there are now initiatives to require full publication of the source data on which clinical trials are based.

Another strategy would be to conduct a meta-analysis that would analyze reported significance levels, especially those on both sides of the conventional cut-off level of $p < 0.05$. If post-hoc manipulation of data occurs, it is likely to take place for positive outcomes that initially marginally failed to reach the threshold of statistical significance or for negative outcomes that were initially just statistically significant. If researchers subsequently 'optimize' their analysis or manipulate the data until significance is reached, then the distribution of significance levels that results from

biased analyses should show proportionally more desired results just below $P < 0.05$ and fewer effects just above $p < 0.05$. The reverse would be true for undesirable effects.

Furthermore, some methodologists have suggested ways to detect fraud by examining peculiarities such as the distribution of single digits either in the primary data or—in the case of a meta-analysis of published outcomes—in the reported statistical tables (Al-Marzouki et al. 2005a; Diekmann 2007). The background is that single digits produced by natural or social processes are distributed in a specific manner described by ‘Benford’s Law’. In contrast, faked data have been shown to deviate systematically from this distribution.

Surveys

Finally, it might be useful to use surveys and experiments as ways to improve knowledge about the incidence of research bias within criminological prevention science. For example, Martinson et al. (2005) surveyed several thousand early- and mid-career scientists funded by the National Institutes of Health (NIH) and asked them to report problematic behaviors. Amongst others they found that more than 15% of the respondents admitted to changing the design, methodology or results of a study in response to pressure from a funding source, while 6% said they had withheld data that contradicted their own previous research. A similar self-report survey could be designed to ask members of the Society of Experimental Criminology and similar bodies about how often respondents have encountered specific types of problematic practices, how often they have personally used such techniques in their research, and how widespread they believe such practices are within the profession. Such a study could shed light on the ‘dark figure’ of systematic bias and its possible link with conflict of interest within our profession.

Conclusions

This paper argues that effect sizes reported in criminological experimental studies should be seen as the sum of a true effect and an effect due to research bias and that the effect size of bias is likely to be higher in studies with a conflict of interest of the evaluators. If this is true, we are unable to estimate the true effect of interventions unless we know the approximate effect size of bias. And, if sizable systematic results bias exists in the field of evidence-based crime prevention research, it has several serious negative effects: first, it leads researchers to draw conclusions and make recommendations that fail to lead to a more efficient allocation of scarce public resources. It therefore obstructs the main goal of evidence-based prevention research. Secondly, if related practices were shown to be widespread, public trust in the research program of evidence-based prevention would be undermined. Thirdly, systematic bias directs research resources into areas that may have less promise than others, where bias is less common. Methodologically sound evaluations will, hence, be disadvantaged and scientific progress obstructed. Finally, such situations create an incentive for new researchers to adopt the same problematic practices that exist in the field.

In recognition of these dangers significant efforts have been made, over the past decade, to improve standards for conducting and publishing experimental and quasi-experimental studies in many domains of research on evidence-based prevention and intervention. The Crime and Justice Group of the Campbell Collaboration, for example, has developed protocols for meta-analyses and systematic reviews. In medical research, the Consolidated Standards of Reporting Trials (CONSORT) statement now provides precise guidance about the information required to report clinical trials (Altman 1996; Campbell et al. 2004). Furthermore, authors of clinical trials in the medical sciences must provide information about conflict of interest, and there are discussions about whether authors should be required to publish the original data so that others can replicate their findings.

Similar improvements are desirable in the field of crime prevention research and can help to reduce research bias, even where conflicts of interest exist. The report of the National Research Council on Improving Evaluation of Anti-crime Programs shows that such developments are underway in criminology (Lipsey et al. 2006).

However, such professional rules will only affect future studies. However, for many years to come, most of our recommendations to policy makers and practitioners will be based on past research. It seems useful, therefore, to clarify the extent of the problem described in this paper.

References

- Al-Marzouki, S., Evans, S., Marshall, T., & Roberts, I. (2005a). Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ*, *331*, 267–270.
- Al-Marzouki, S., Roberts, I., Marshall, T., & Evans, S. (2005b). The effect of scientific misconduct on the results of clinical trials: a Delphi survey. *Contemporary Clinical Trials*, *26*(3), 331–337.
- Altman, D. G. (1996). Better reporting of randomised controlled trials: the consort statement. *BMJ*, *313* (7057), 570–571.
- Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, *85*(5), 1337–1343.
- Bauer, N. S., Lozano, P., & Rivara, F. P. (2007). The Effectiveness of the Olweus bullying prevention program in public middle schools: a controlled trial. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *40*(3), 266–274.
- Beelmann, A., & Lösel, F. (2006). Child social skills training in developmental crime prevention: effects on antisocial behavior and social competence. *Psychothema*, *18*(3), 603–610.
- Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a white middle-class population. *JAMA*, *273*(14), 1106–1112.
- Campbell, M. K., Elbourne, D. R., & Altman, D. G. (2004). Consort statement: extension to cluster randomised trials. *BMJ*, *328*(7441), 702–708.
- Cho, H., Hallfors, D. D., & Sánchez, V. (2005). Evaluation of a high school peer group intervention for at-risk youth. *Journal of Abnormal Child Psychology*, *33*(3), 363–374.
- Cornish, D. B., & Clarke, R. V. G. (Eds.). (1986). *The reasoning criminal: Rational choice perspectives on offending*. New York: Springer.
- Dana, J., & Loewenstein, G. (2003). A social science perspective on gifts to physicians from industry. *JAMA*, *290*(2), 252–255.
- de Graaf, I., Speetjens, P., Smit, F., de Wolff, M., & Tavecchio, L. (2008). Effectiveness of the Triple P positive parenting program on behavioral problems in children: a meta-analysis. *Behav Modif*, *32*(5), 714–735.
- Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics*, *34*(3), 321–329.
- DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002a). Effectiveness of mentoring programs for youth: a meta-analytic review. *American Journal of Community Psychology*, *30*, 157–197.

- DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002b). Effectiveness of mentoring programs for youth: a meta-analytic review. *American Journal of Community Psychology, 30*(2), 157–197.
- Eggert, L. L., Seyi, C. D., & Nicholas, L. J. (1990). Effects of a school-based prevention program for potential high school dropouts and drug abusers. *Substance Use & Misuse, 25*(7), 773–801.
- Eggert, L. L., Thompson, E. A., Herting, J. R., Nicholas, L. J., & Dickens, B. G. (1994). Preventing adolescent drug abuse and high school dropout through an intensive social network development program. *American Journal of Health Promotion, 8*(2), 202–215.
- Eisner, M., & Ribeaud, D. (2008). Markt, Macht und Wissenschaft; Kritische Überlegungen zur Deutschen Präventionsforschung. In E. Marks & W. Steffen (Eds.), *Starke Jugend - Starke Zukunft (Ausgewählte Beiträge des 12. Deutschen Präventionstages, 18. und 19. Juni 2007)* (pp. 173–191). Mönchengladbach: Forum Verlag Godesberg.
- Eisner, M., Ribeaud, D., Jünger, R., & Meidert, U. (2007). *Frühprävention von Gewalt und Aggression: Ergebnisse des Zürcher Interventions- und Präventionsprojektes an Schulen*. Zürich: Rüegger.
- Ellickson, P. L. (1998). Preventing adolescent substance abuse: lessons from the project alert program. In J. Crane (Ed.), *Social programs that work* (pp. 201–257). New York: Russell Sage.
- Ellickson, P. L., & Bell, R. M. (1990). Drug prevention in junior high: a multi-site longitudinal test. *Science, 247*, 1299–1305.
- Ellickson, P. L., McCaffrey, D. F., Ghosh-Dastidar, B., & Longshore, D. L. (2003). New inroads in preventing adolescent drug use: results from a large-scale trial of project ALERT in middle schools. *American Journal of Public Health, 93*(11), 1830–1836.
- Elliott, D. S., & Mihalic, S. F. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science, 5*(1), 47–53.
- Etzkowitz, H. (2003). Research groups as ‘Quasi-Firms’: the invention of the entrepreneurial university. *Research Policy, 32*(1), 109–121.
- Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Sciences, 587*, 49–68.
- Farrington, D. P., & Welsh, B. (2003). Family-based prevention of offending: a meta-analysis. *Australian and New Zealand Journal of Criminology, 36*(2), 127–151.
- Felson, M. (1994). *Crime and everyday life; insight and implications for society*. Thousand Oaks: Pine Forge Press.
- Friedman, L. S., & Richter, E. D. (2004). Relationship between conflicts of interest and research results. *Journal of General Internal Medicine, 19*(1), 51–56.
- Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The devil is in the details: examining the evidence for “proven” school-based drug abuse prevention programs. *Evaluation Review, 31*(1), 43–74.
- Gilbert, N. (1997). Advocacy research and social policy. *Crime and Justice: A Review of Research, 22*, 101–148.
- Gorman, D. M. (2003). Alcohol & drug abuse: the best of practices, the worst of practices: the making of science-based primary prevention programs. *Psychiatric Services, 54*(8), 1087–1089.
- Gorman, D. M. (2005a). Does measurement dependence explain the effects of the life skills training program on smoking outcomes? *Preventive Medicine, 40*(4), 479–487.
- Gorman, D. M. (2005b). Drug and violence prevention: rediscovering the critical rational dimension of evaluation research. *Journal of Experimental Criminology, 1*(1), 39–62.
- Gorman, D. M. (2006). Conflicts of interest in the evaluation and dissemination of drug use prevention programs. In J. Kleinig, & S. Einstein (Eds.), *Intervening in drug use: Ethical challenges* (pp. 171–187). Huntsville: Office of International Criminal Justice.
- Gorman, D. M., & Conde, E. (2007). Conflict of interest in the evaluation and dissemination of “Model” school-based drug and violence prevention programs. *Evaluation and Program Planning, 30*(4), 422–429.
- Heinrichs, N., Hahlweg, K., Bertram, H., Kuschel, A., Naumann, S., & Harstick, S. (2006). Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus Sicht der Mütter und Väter. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*(2).
- Henggeler, S. W., Cunningham, P. B., Pickrel, S. G., Schoenwald, S. K., & Brondino, M. J. (1996). Multisystemic therapy: an effective violence prevention approach for serious juvenile offenders. *Journal of Adolescence, 19*(1), 47–61.
- Hewitt, C. E., Mitchell, N., & Torgerson, D. J. (2008). Listen to the data when results are not significant. *BMJ, 336*(7634), 23–25.
- Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*(1), 3–31.
- Kaptschuk, T. J. (2003). Effect of interpretive bias on research evidence. *BMJ, 326*(7404), 1453–1455.

- Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McGuire (Ed.), *What works? Reducing reoffending* (pp. 63–78). New York: Wiley.
- Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: a review of systematic reviews. *Annual Review of Law and Social Science*, 3, 297–320.
- Lipsey, M. W., Petrie, C., Weisburd, D., & Gottfredson, D. (2006). Improving evaluation of anti-crime programs: summary of a national research council report. *Journal of Experimental Criminology*, 2(3), 271–307.
- Littell, J. (2005). Lessons from a systematic review of effects of multisystemic therapy. *Children and Youth Services Review*, 27(4), 445–463.
- Littell, J., Popa, M., & Forsythe, B. (2005). Multisystemic Therapy for Social, Emotional, and Behavioral Problems in Youth Aged 10–17 (Report for the Campbell Collaboration) (Electronic Version). Accessed 14 January 2008 from http://www.sfi.dk/graphics/Campbell/Dokumenter/MST_Review/MULTISYSTEMIC%20THERAPY%20-%20REVIEW.pdf.
- Lock, S., Wells, F., & Farthing, M. (Eds.). (2001). *Fraud and misconduct in medical research*. London: BMJ Publishing Group.
- Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: a systematic review of randomized evaluations. *The ANNALS of the American Academy of Political and Social Science*, 587(1), 84–109.
- Lösel, F., & Kofler, P. (1989). Evaluation research on correctional treatment in West Germany: A meta-analysis. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system: Psychological perspectives*. New York: Springer.
- Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: a systematic review of randomized evaluations. *The ANNALS of the American Academy of Political and Social Science*, 587(1), 84–109.
- Luborsky, L., Diguier, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: a "Wild Card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, 6(1), 95–106.
- Lundahl, B., Risser, H. J., & Lovejoy, M. C. (2006). A meta-analysis of parent training: moderators and follow-up effects. *Clinical Psychology Review*, 26(1), 86–104.
- MacCoun, R. J. (1998). Biases in the interpretation and use of research statistics. *Annual Review of Psychology*, 49, 259–287.
- MacCoun, R. J. (2005). Conflicts of interest in public policy research. In D. Moore, D. M. Cain, G. Loewenstein, & M. H. Bazerman (Eds.), *Conflicts of interest: challenges and solutions in business, law, medicine, and public policy* (pp. 233–262). Cambridge: Cambridge University Press.
- Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, 435(9 June 2005), 737–738.
- Messick, D. M., & Sentis, K. P. (1979). Fairness and preference. *Journal of Experimental Social Psychology*, 15(4), 418–434.
- Moore, D., Tetlock, P. E., Tanlu, L., & Bazerman, M. H. (2006). Conflicts of interest and the case of auditor independence: moral seduction and strategic issue cycling. *Academy of Management Review*, 31(1), 10–29.
- Nowak, C., & Heinrichs, N. (2008). A comprehensive meta-analysis of Triple P-positive parenting program using hierarchical linear modeling: effectiveness and moderating variables. *Clinical Child and Family Psychology Review*, 11(3), 114–144.
- Ogden, T., & Halliday-Boykins, C. A. (2004). Multisystemic treatment of antisocial adolescents in Norway: replication of clinical outcomes outside of the us. *Child and Adolescent Mental Health*, 9, 77–83.
- Ogden, T., & Hagen, K. A. (2006). Multisystemic treatment of serious behaviour problems in youth: sustainability of effectiveness two years after intake. *Child and Adolescent Mental Health*, 11(3), 142–149.
- Okike, K., Kocher, M. S., Mehlman, C. T., & Bhandari, M. (2007). Conflict of interest in orthopaedic research. an association between findings and funding in scientific presentations. *J Bone Joint Surg Am*, 89(3), 608–613.
- Olweus, D. (1994). Bullying at school: basic facts and effects of a school based intervention program. *Journal of Child Psychology and Psychiatry*, 35(7), 1171–1190.
- Perlis, R. H., Perlis, C. S., Wu, Y., Hwang, C., Joseph, M., & Nierenberg, A. A. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry*, 162(10), 1957–1960.
- Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, 1(4), 435–450.

- Ranstam, J., Buyse, M., George, S. L., Evans, S., Geller, N. L., Scherrer, B., et al. (2000). Fraud in medical research: an international survey of biostatisticians. *Controlled Clinical Trials*, 21(5), 415–427.
- Resnik, D. B. (2000). Financial interests and research bias. *Perspectives on Science*, 8(3), 255–285.
- Russo, J. E., Carlson, K. A., Meloy, M., & Yong, K. (2007). The goal of consistency as a cause of information distortion. *Johnson School Research Paper Series No. 04-07*.
- Sanchez, V., Steckler, A., Nitirat, P., Hallfors, D., Cho, H., & Brodish, P. (2007). Fidelity of implementation in a treatment effectiveness trial of reconnecting youth. *Health Education Research*, 22(1), 95–107.
- Sanders, M. R. (1992). *Every parent: A positive guide to children's behavior*. Sydney: Addison-Wesley.
- Sanders, M. R. (1999). Triple P-positive parenting program: towards an empirically validated multilevel parenting and family support strategy for the prevention of behaviour and emotional problems in children. *Clinical Child and Family Psychology Review*, 2(2), 71–89.
- Shadish, W. R., Cook, T. D., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Sherman, L. W. (2006). “to develop and test:” the inventive difference between evaluation and experimentation. *Journal of Experimental Criminology*, 2(3), 393–406.
- Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (Eds.). (2002). *Evidence-based crime prevention*. London: Routledge.
- St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an independent evaluation of project alert delivered in schools by cooperative extension. *Prevention Science*, 6(4), 305–317.
- Sundell, K., Hansson, K., Löfholm, C. A., Olsson, T., Gustle, L. H., & Kadasjo, C. (2008). The transportability of multisystemic therapy to Sweden: short-term results from a randomized trial of conduct disordered youth. *Journal of Family Psychology*, 22(4), 550–560.
- Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: update of a meta-analysis. *American Journal of Preventive Medicine*, 33(2, Supplement 1), S130–S143.
- Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003a). The effects of school-based intervention programs on aggressive behavior: a meta-analysis. *Journal of Consulting and Clinical Psychology*, 71(1), 136–149.
- Wilson, S. J., Lipsey, M. W., & Soydan, H. (2003b). Are mainstream programs for juvenile delinquency less effective with minority youth than majority youth? a meta-analysis of outcomes research. *Research on Social Work Practice*, 13(1), 2–26.

Manuel Eisner is a Reader in Sociological Criminology at the Institute of Criminology, University of Cambridge, UK. His research interests include the history of violence, developmental criminology, and prevention research. He is the principal investigator of the Zürich Project on the Social Development of Children, a longitudinal study of 1,200 children that is combined with a randomized experiment on the long-term effectiveness of early universal violence prevention programs.