J Exp Criminol (2009) 5:185–200
DOI 10.1007/s11292-009-9073-9

Testing for Analysts' Bias in Crime Prevention Experiments: Can We Accept Eisner's One-tailed Test?

Lawrence W. Sherman · Heather Strang

Published online: 8 April 2009

Abstract Eisner (Journal of Experimental Criminology, this issue, 2009) suggests that developer-led evaluations often make programs look better than independent evaluations do because the former suffer systematic biases in favor of prevention success. Yet, his proposed remedies suffer their own systematic bias, constituting a 'one-tailed' test of bias in only one direction. In this response we suggest that a more objective assessment of 'analysts' effects' requires a 'two-tailed' test of bias, in which reviewers would measure indications of bias *both for and against success* in evaluations reported by both developers and independent evaluators. After exploring the full complexity of the distinction between developers and evaluators, we report on one case in which independent evaluations were more favorable than those of developers. We then suggest possible indicators of analysts' biases against finding success that may characterize the work of developers who "bend over backwards" to find harm in their programs, and of independent evaluators who may seek to "get a scalp" of a developer or a program.

Keywords Bias · Developers · Evaluators · Experiments

Introduction

Eisner's (2009) analysis of the growing evidence of 'developer effects' in crime prevention evaluations is a valuable contribution on an important issue for experimental criminology. The credibility of evidence-based policy depends entirely

L. W. Sherman (✉) · H. Strang
Cambridge University, Cambridge, Cambridgeshire, UK
e-mail: ls434@cam.ac.uk

L. W. Sherman
The University of Pennsylvania, Philadelphia, PA, USA

H. Strang
Australian National University, Canberra, Australian Capital Territory, Australia

on our ability to produce unbiased estimates of policy effects, regardless of the source of bias. A robust discussion of what we would prefer to call 'analysts' biases' is no less important than the classic discussions of the research design features needed to remove specification bias in causal inference (Campbell and Stanley 1963).

Eisner starts where the evidence to date leaves off: a choice between two possible interpretations of positive correlations between developer involvement in an experiment and the effect size of its success in crime prevention. Those interpretations are described as the 'high fidelity view' and the 'cynical view.' The high fidelity view is that developer-involved evaluations find better results because the program is more likely to be implemented correctly than when the program developer is not involved. The cynical view is that the difference is due to an analytic bias that developers introduce in favor of the evaluation showing the success of their program.

This choice, however, fails to describe all logically possible interpretations of the correlation reported in the literature. The evidence Eisner reports, after all, ranks merely as a 'level 1' on the Maryland Scale of Scientific Methods (Sherman et al. 2002). Because it lacks any temporal sequence, the correlation can run in either direction. The recursive nature of the correlation can actually apply to both options that Eisner poses. For the 'high fidelity view,' developers may be attracted to research sites where they can tell in advance that operational partners are likely to deliver their programs with competence and fidelity. For the 'cynical view,' it is just as plausible that independent evaluators introduce a systematic analyst's bias *against* finding crime prevention benefits.

Independent evaluators: a cynical view Eisner suggests a causal model for his 'one-way' theory of the 'cynical view' that relies on a conflict of interest by the developer, whether financial or ideological. Yet, it is just as plausible to have a 'cynical view' that independent evaluators have a conflict of interest in evaluating a program, motivated by a variety of possible emotions. Independent evaluators may feel a natural hostility to those they evaluate, just as prosecutors may resolve cognitive dissonance by treating criminal defendants as 'the enemy,' or as journalists may treat people in power as someone they would like to 'get,' like scalps on a belt.

It is instructive, for example, that in 2008, in the aftermath of London's Mayor forcing out the Metropolitan Police Commissioner after a wave of negative news media 'spin' on his leadership, several newspaper reporters admitted to a reporting bias. They said that they had decided to try to 'get' the Police Commissioner because they had let his predecessor 'get away with' too many things for which he could have been attacked. If, as Eisner suggests, independent evaluators (like journalists) should be "disengaged and skeptical truth finders," then a cynical view of their work is that they may become *too* skeptical and overly engaged, corrupted by a desire to disprove the value of a respected or popular program. This conflict may even occur for no other reason than that an evaluator predicts that a negative result will draw more attention to the evaluation than a positive result.

The Kansas City preventive patrol experiment (Kelling et al. 1974) is a classic example of 'independent' evaluators showing analysts' biases against finding a crime prevention benefit. As Sherman (1986) pointed out, the experiment found substantial effect sizes in the reduction of robbery by extra police patrol, or increased robbery by reduced police patrol. Yet, because the experiment ($n=15$ patrol beats)

lacked adequate statistical power for the test of patrol effects on specific crimes, the substantial effect sizes were not statistically significant. The finding that patrol dosage did not have any effect on crime attracted enormous publicity and misled the field with a biased estimate for almost two decades, until a larger, adequately powered, randomized trial (n=110 'hot spots') focused on high-crime locations reached the opposite conclusion (Sherman and Weisburd 1995).

Our suggestion of a possible two-way interpretation of the so-called developer effect correlation means that it could also be called a prosecutor's effect, where independent evaluators see their role (even subconsciously) as proving that a program or policy does not work. This 'two-tailed' view of potential analysts' biases has an important substantive implication for Eisner's diagnosis of the problem we face and his recommendations for curing it. First, his diagnosis assumes that conflict of interest is only an issue for developers, and not for independent evaluators. We disagree. Second, his recommendations for ways that systematic reviewers can guard against developer's bias constitute a 'one-tailed' test of the hypothesis that developers manifest systematic biases in analytic choices *increasing* the chances of finding success. We cannot accept that test. It is logically just as possible to look for (and find) systematic bias *reducing* the chances of finding crime prevention benefits in a crime prevention evaluation.

A four-way test: positive and negative, developers and independents We further suggest that a 'bias for success' is not limited to developers and that a 'bias against success' is not limited to independent evaluators. 'Independent' evaluators may well face financial or organizational pressures from their funding sources to 'prove' that a program works, especially if they need repeat funding from the source of funds for the current study. It would make just as much sense to require independent evaluators to disclose their prior history of funding from a government agency as it does to disclose their prior history of funding from a for-profit corporation. Government agencies, as any watcher of the British TV series "Yes Minister" may observe, may be just as biased and cynical in their use of research as the most rapacious corporations.

There is also some evidence to support the logical possibility of an 'anti-cynical' or 'self-sacrificing' view. Some developers may bend over backwards to test for possible harms that their program may cause. They may go well beyond independent evaluators in looking for negative evidence, simply out of an over-developed superego. As a counterpoint to the high fidelity view of implementation of a program theory, this view of developer evaluations implies a high fidelity to the science of casual inference. Or, one step further, it could be motivated by high fidelity to the principle of 'First, do no harm.'

Overview Using the foregoing introductory framework, this comment presents three responses to Eisner's diagnosis and recommendations. First, it explores the complex nuances of key concepts that both he and other commentators fail to define adequately. These include the concepts of developer, independent evaluator, program and conflict of interest. Second, it presents a fifth case for Eisner's (2009) Table 1, documenting the 'negative developer effect' across the results of 12 randomized controlled trials in restorative justice, in which eight trials analyzed by independent evaluators found more benefits from a program than did four trials analyzed by developers. Third, this comment proposes a set of indicators of analysts' biases *against* finding crime prevention effects, which could be detected by careful coding

of any randomized controlled trial, regardless of how the role of the trial report authors might be coded. We conclude by suggesting that, if any assessment of analysts' bias is added to systematic reviews, it should be conducted in the two-tailed framework we offer here. That said, it is not clear to us that such assessments can be conducted to a reasonable degree of reliability.

Defining key concepts

Eisner (2009) requires four concepts to sustain his argument: (a) developers; (b) independent evaluators; (c) program; and (d) conflict of interest. As he notes, in order for reviewers to use these concepts in testing for analysts' biases, each of them must be susceptible to reliable coding of written reports of crime prevention program evaluations. We suggest that reliable coding would be problematic at best, and often impossible, given current reporting practices. Our own reading of that literature suggests that these concepts have not even been fully explored, at least in the context of crime prevention evaluations. Absent a consensus on how to define these concepts and require reporting by primary research authors, it will be difficult to undertake an unbiased test for analysts' biases.

One problem in developing reliable definitions of these four concepts is the literature's underlying metaphor of pharmaceutical evaluations. As we note at several points in this section, Eisner's use of concepts derived from that industry does, in fact, fit at least some examples of program evaluation in crime prevention. However, it does not fit all of them, or even most. Our discussion seeks merely to illustrate crime prevention testing that falls far away from the pharmaceutical model of post-test marketing of a product. Space does not allow us to develop a full typology of experimental criminology in terms of these ideas. What we can do is to falsify Eisner's premise of homogeneity in the context and structure of analysts' biases for or against reporting crime prevention success.

Program developer In the pharmaceutical industry, a developer is a corporation that may invest many years and millions in any currency in developing and testing a new product for patent protection. The very survival of the developers in that industry depends upon their finding that at least some of the drugs they develop are efficacious against a disease, thereby yielding billions in sales revenues to recoup the company's investment in research and development. This includes recouping the costs of tested products that fail to show evidence of the benefits hypothesized. In this economic context, a developer is reliably identified as the owner of a license or patent from which revenue may be derived.

As Eisner notes, some crime prevention programs have been established under license, by which revenues may depend on the research literature reporting consistent benefits in return for those revenues paid to the license holders. Here, again, such developers may be reliably identified by their exclusive legal right to market a product. Yet, as we survey the many tests of crime prevention programs reported in discussions of experimental criminology (Farrington 1983; Farrington and Welsh 2005), we conclude that revenue-yielding 'licensed' programs constitute a minority of practices tested by experimental criminology.

Once we examine crime prevention developers who do not hold legal rights to charge others for using their program content, it becomes much more difficult to define what a 'developer' is (Sherman 2006). We can illustrate this point with the history of The Jerry Lee Program on Randomized Controlled Experiments in Restorative Justice. During the course of the 12 randomized controlled trials that we designed or directed, we have considered ourselves to be both developers and 'independent' evaluators at different times, even within the same projects. A chronological narrative of our changing roles can provide the kind of thick detail that is necessary to understand the concept of a crime prevention "program developer."

Canberra, 1992–1995 John Braithwaite, author of the theory of reintegrative shaming (Braithwaite 1989), invites Sherman to design and implement an experimental evaluation of restorative justice conferences on the New Zealand juvenile justice model. During this time, Braithwaite identifies prospective operational partners and funding sources, while Sherman designs a series of options for randomized controlled trials. Sherman then submits these designs as grant proposals to Australian funders, with Braithwaite listed as project director and Sherman listed as scientific director. Braithwaite negotiates with several Australian trainers and consultants familiar with restorative justice and includes their contracts in the project funding. Several consultants are set up to work in a program office of the Australian Federal Police in Canberra, which delivers the restorative justice conferencing (RJC) treatment in the form of four randomized controlled trials (RCTs). Sherman maintains an arm's length distance from those negotiations, while working with Strang to establish a separate research operation that she would direct at the Australian National University, where Braithwaite was her administrative supervisor and Sherman was her scientific supervisor. Sherman makes it clear that he has no preference for restorative justice working or not working, except insofar as he would like to find better ways to prevent crime. He repeatedly says that he would oppose the program if and when it found that it increased crime and that the possibility of such effects is a major reason to do the experiment. He receives no salary from any source except his regular salary from the University of Maryland, in part to retain institutional independence from the program being evaluated.

Indianapolis, USA, 1995–2000 Sherman meets with prosecutors and police officials to persuade them to replicate the Canberra RCTs of RJC. Several of them visit Canberra, and Sherman arranges for others to be trained in the Canberra methods of RJC. Sherman writes a grant proposal for National Institute of Justice funding through the Indianapolis-based Hudson Institute, which receives the grant. Indiana University professor Edmund McGarrell then takes over to re-design, direct and complete the project with no further involvement by Sherman.

Canberra, 1995–1997 Sherman and Strang train the entire Canberra police department in the operational procedures for four randomized trials, which he launches with Strang on 1 July 1995. Sherman takes the first (2 a.m.) telephone call from a police officer to report an eligible case, while telling police that his job is to be the independent evaluator. Sherman repeatedly uses the example of Braithwaite's being like Jonas Salk, who 'developed' a polio vaccine that was then 'independently'

evaluated by someone like Sherman, someone few people remember (Thomas Francis of the University of Michigan, USA). Despite this statement, Sherman, Strang and Braithwaite appear as co-authors of progress reports and consult each other on matters of program delivery—thus suggesting, by Eisner's standards, that Sherman and Strang were, in fact, embedded in the process of development of the program.

Canberra, 1997–2000 As the frequency of case referrals from police for random assignment slows down, Sherman and Strang increasingly engage with police to improve the program's delivery. They develop and propose alternative organizational arrangements for the delivery of the RJC treatments, which had been less than 100%. Against police commander preferences, they also attempt to communicate directly with police officers to remind them to keep referring cases and to use the program. They conclude the first 5 years by publishing an initial report in which only one of the four trials is deemed a success in reducing repeat offending (Sherman et al. 2000).

London, England, 2000–2008 Invited by the UK Government to bid on a contract for testing RJC in the UK, Sherman and Strang are forced to choose between funding as developers or as evaluators. Given the key role of developers in delivering the RCT research design, they choose to be developers. Joanna Shapland at Sheffield University is selected as 'independent evaluator.' The two operations work at arms' length for most of 8 years, until the Government asks the developers to comment on a draft of the final report by the evaluators. At that time Sherman suggests to Shapland that the use of prevalence, rather than frequency, measures of repeat offending create an analytic bias against showing program success. He recommends that the report be revised to include a forest plot of effect sizes of frequency of reconviction in the seven RCTs that Sherman and Strang have developed. Shapland then decides to add that particular analysis, which shows a statistically significant pattern of crime prevention, with 27% fewer reconvictions across all seven RCTs. None of the separate RCTs attained significance, but all effects are in the desired direction and the *P* value of the pattern was 0.01 (Shapland et al. 2008). While Shapland remains at all times an independent evaluator, standard protocol allowing developers to comment on evaluators' research leads to an important aspect of the evaluation's analysis.

The most important point about the details provided above is that almost none of them has ever appeared in any written report on the research. Moreover, it is neither typical nor required for the authors to supply such details, especially in peer-reviewed journals where space is limited. The Consolidated Standards of Reporting Trials (CONSORT) statement does not require them, nor does it even use the concept of 'developer.' It therefore seems unlikely that systematic reviewers would have been able to code the findings of our first four experiments as "developers" or "evaluators," at least not without contacting us (which goes beyond standard protocols for systematic reviews).

'Independent' Evaluator As the last section illustrates, the definition of an independent evaluator may be, empirically, a matter of degree rather than a clear set of boundaries. The evaluation literature is replete with advice about communication between evaluators and operators, much of which might compromise some notion of independence. 'Formative' evaluation, for example, is inherently a process by which

evaluators interpret what they are finding for the benefit of program operators, often making suggestions for better management operations to improve the program. Is that independent, or developmentally engaged, evaluation? Both answers would seem to be plausible, so much so that reliable coding would be difficult.

There are even situations in which evaluators suggest ideas, operating agencies agree to test them, and then the evaluators 'judge' whether the agency can achieve the results hypothesized. In those circumstances, there may still be a large social gulf, or even hostility, between the evaluators and the operators. In the Houston fear reduction experiments (Pate et al. 1986), the police in the program team expressed disdain for the evaluators who had helped plan the strategy they were carrying out by saying "we don't care what your report card (school examination marks) shows about our program." So we can ask, in this case, whose program was it? Were the evaluators independent of the program, even though the idea came from them, given that the evaluators had virtually no control over the program's administration?

Program And just what is a program? The definition of an evaluator's independence becomes even more complicated when the definition of a program or policy is taken into account. Even in what appears to be the most rudimentary set of program outputs, there can be enormous variability from one delivery of those outputs to the next. A policy of arrest for domestic violence, for example, varied widely across six experiments in terms of the eligible offenses for which arrests could be made, the length of time spent in custody, and the rate at which the arrests resulted in prosecution (Sherman 1992). A policy of restorative justice in the same country varied widely within a single evaluation in terms of basic features, such as whether victims and offenders met face to face, how long any meetings lasted, how many people were present, whether their friends or family were involved in communications or negotiations, whether there was a negotiated agreement among the parties, and whether the process was led by a police officer, a mediation professional, or a lay person (Shapland et al. 2008).

The reason these variations reduce the clarity of 'independence' is that evaluators may send messages to operators (or developers) about what is being measured. Hence, if any of the variables creating heterogeneity in the program being evaluated are tracked by evaluators—and even fed back to operators—that, in itself, could become part of that particular version of the program being evaluated. When an evaluator sends no such feedback to operators or developers, then that could yield a different kind of program. It could also yield a test result with greater external validity for operating in a world without independent evaluators at one's side, or one biased against the program's potential success because of a highly developed concern about offering such advice.

The extent to which programs vary in terms of subtle nuances, such as tone of voice or cultural significance of different actions, may also explain differences in results. Each of Eisner's examples in his Table 1, for example, could have had a variety of program-variation explanations. When an RCT led by a developer yields a different result from that of an RCT in a different country or region by an independent evaluator, there are at least two plausible rival hypotheses for explaining that difference in result: region and bias. There are probably even more. Even then, however, the bias may not be driven by conflict of interest, especially if the nature of that conflict is purportedly ideological.

Conflict of interest Here again, the pharmaceutical model (rarely found in crime prevention) differs from the more common situation in which there is no monetary revenue at stake. We agree with Eisner that any financial conflicts of interest should be reported in each publication by a developer–evaluator who has money to gain or lose. Yet, we do not see how the same can be done for allegedly 'ideological' conflicts of interest. In terms of crime as the dependent variable, surely most criminologists share an ideological commitment to less crime and suffering. Even if we are all committed to discovering how that might be done, however, that does not mean we are prepared to use biased analytic techniques in order to claim that we have found how to make a better world.

With respect to independent variables, experimenters may have predispositions toward more or less coercion, more or less use of prison, guns, nurses or even police. Yet, it is hard to see how such predispositions constitute a conflict of interest that can be reliably measured and monitored. Medical researchers may vary in their preference for or against surgery, pharmaceuticals, or lifestyle changes. Yet, there is no precedent, to our knowledge, for reporting such preferences in medical journals. Nor has anyone, to our knowledge, suggested that predispositions for different treatments would bias an analyst's results.

At a metaphysical level, Eisner's implicit premise is that self-interest (or 'tastes' for different kinds of crime prevention) must conflict with public interest, or at least an interest in finding the truth. This premise seems to suffer the same limitations as neo-classical economics. It requires a limited range of 'utilities,' or preferred outcomes, of the kind that economists assume everyone shares: money, influence, fame. It allows no room for the growing body of evidence in behavioral economics, in which people behave irrationally according to economic principles but quite rationally in relation to their perceptions of gain and loss.

For people whose preference is for the utility of truth, there may be a conflict of interest between doing no harm, on the one hand, and doing good on the other. That is why there is so much attention paid to false negative results and false positive results. The concern for doing no harm may lead to a bias against finding success, and vice versa. This is not an ideology, but perhaps more a matter of personal attitude. It is also one that would be extremely difficult to measure reliably, even if principal investigators were asked to fill out a personality assessment inventory. Yet, there may be indicators of evaluators, both developers and independents, who go out of their way to try to challenge the good results of the programs they evaluate—simply as a means of being very careful. They may bend over backwards so that if there is anything bad to say about a program that may have some good results, the first evaluator is the one to discover it. If there is any conflict of interest here, it is between wanting to see a successful program do good and wanting to avoid a future evaluator being the first to detect a harmful effect.

A case of a negative developer effect

Eisner's Table 1 reports four cases of conflicting evidence on "Prevention effects in developer-led studies and independent trials." In each of the four cases, the developers' evaluations were positive, while the independent trial results were

negative. Eisner does *not* say that this set of results constitutes a systematic review of all developer versus independent results on the same programs, although some readers may gather that impression. What he does say is that

> "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."

Even with this qualification, however, he only addresses two possibilities: either developers report better results than independent evaluations do, or they report the same results as independents. He does not address a third possibility: that developer–evaluators could report worse results for their programs than independents could.

Because Eisner's Table 1 uses four cases to illustrate the one-tailed problem he describes, we wish to add a fifth case to report a "black swan" (Popper 1959) observation that falsifies an apparently universal pattern: a case in which a developer did report worse results than independents did. That case is our own reports (Sherman et al. 2008) on four of the developer-evaluated RJC trials in Australia described above, as compared with independent reports on eight of the other trials we designed ($n=1$) in the USA or directed ($n=7$) in the UK. These results are compared between Figs. 1 and 2 and between Figs. 3 and 4. In all four figures the outcome measure is the after-only difference in 2-year criminal conviction rates between the intent-to-treat RJC and control groups.

In Fig. 1 we show a bar graph of the results of our four Canberra tests, depicting the percentage difference in the number of convictions per year per offender.[1] Both the youth property crime experiment and that for adults driving with excess prescribed content of alcohol (PCA) in the bloodstream showed substantially higher conviction rates among offenders randomly assigned to RJC. The experiment on youth violence and that for youth department store shoplifting both showed substantially fewer reconvictions.[2]

In Fig. 2 we show a bar graph of the results of the independent evaluation of the seven RCTs we developed and directed in the UK. These include post-conviction, pre-sentencing RJC for burglars, robbers, adult assaulters and adult property offenders, post-sentencing RJ for serious assault offenders in prison or just beginning probation, and a comparison of diversion from prosecution with and without RJC for juveniles arrested for violence or property crimes. In sharp contrast to the developer–evaluator results in Fig. 1, the independent results in Fig. 2 are uniformly positive. The pooled average across all cases was 27% fewer convictions in the RJC-assigned group than in

[1] Offenders are each coded by intention to treat as randomly assigned, although random assignment was performed at the case level. Eight of 12 experiments had no co-offenders allowed, but four youth experiments in Canberra, England and Indiana featured co-offenders.

[2] These results differed from those of the report by Sherman et al. (2000) on the before–after difference in the differences in arrest rates in order to make them comparable to the results of both Shapland et al. (2008) and McGarrell and Hipple (2007). The latter two studies used convictions rather than arrests, and used after-only differences rather than before-after difference-of-differences.

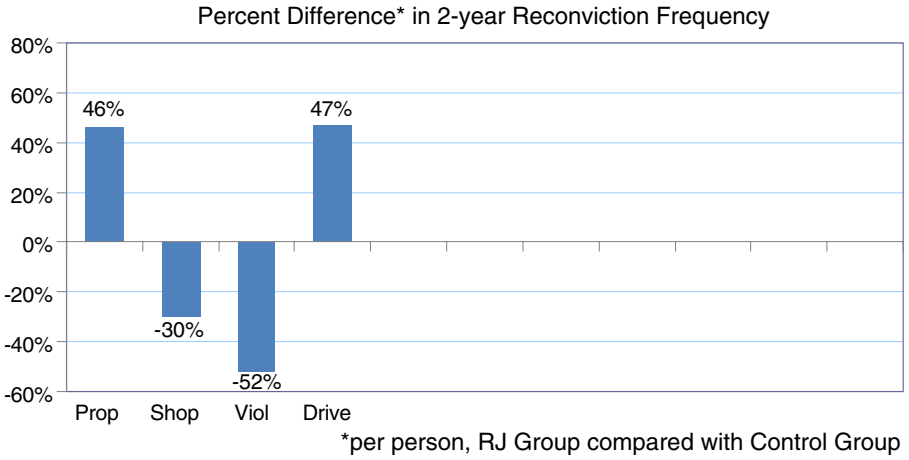

Fig. 1 Percent difference* in 2-year reconviction frequency in developer-evaluated RCTs of restorative justice conferencing

the control groups. None of the bars in the graph reflects a statistically significant difference, nor can the findings be aggregated when displayed in this way. Hence, for further comparison, we next employ forest plots, which are specifically designed to detect statistically significant patterns of results across tests.

In Fig. 3 we display a forest plot of the standardized mean difference (Cohen's D) in 2-year post-intervention convictions between cases randomly assigned to RJC and to controls in the four Canberra experiments. The plot shows that the average effect across the four trials is exactly zero, with $P=0.95$. In contrast, Fig. 4 shows the forest plot of the same outcome differences in the eight independent tests of RJC in the UK ($n=7$) and Indiana ($n=1$). The mean difference across these eight tests is −0.19, almost exactly at a level Cohen describes as a 'small' but clearly cost-effective

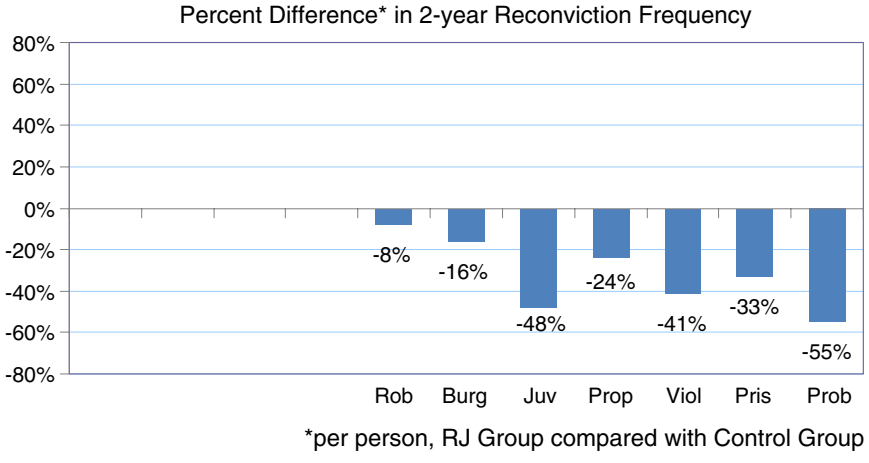

Fig. 2 Percent difference* in 2-year reconviction frequency in independent-evaluator RCTs of restorative justice conferencing

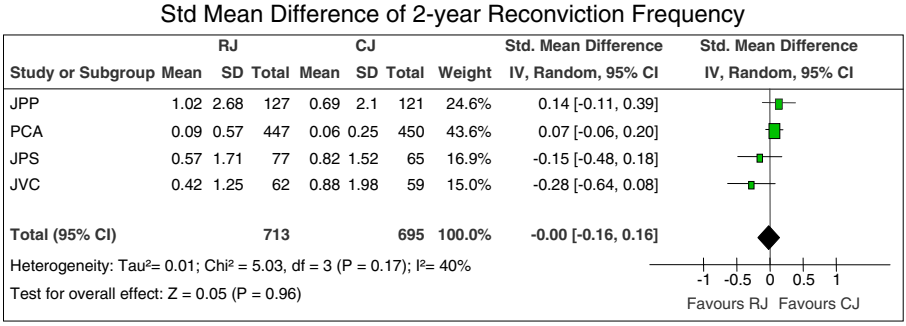

Fig. 3 Standard mean difference in 2-year reconviction frequency in developer-evaluated RCTs of restorative justice conferencing

difference, as Shapland et al. (2008) reported—with an average return of £8 for each £1 spent on the running costs of delivering RJC in the seven UK trials. The pattern of effects in Fig. 4 is statistically significant (P=0.0002) with the confidence interval ranging from −0.09 to −0.29. In sum, a review of the developer-led evaluations shows no effect overall. A review of independent evaluations shows a clear effect.

As Eisner notes, such comparisons necessarily entail many other confounding variables. Figures 3 and 4, for example, were originally designed to display the difference between the use of RJC as diversion (versus prosecution), on the one hand, and the use of RJC as an add-on to criminal justice, on the other. The two sets of studies also compare a highly disrupted chain of different police leaders and systems for delivering RJC in Canberra, compared with much greater stability and competence of delivery systems in the other (independently evaluated) sites. Differences of nationality and offender background also matter, with the very particular effects of Aboriginal offenders in the Australian studies.

The Aboriginal issue also prompted the developers to compare results for white and Aboriginal offenders in the Canberra experiments, as there were good theoretical reasons to anticipate a differential response by race. As we have reported (Sherman and Strang 2006, 2007; Barnes 2006; Woods 2006), the Aboriginal offenders reacted

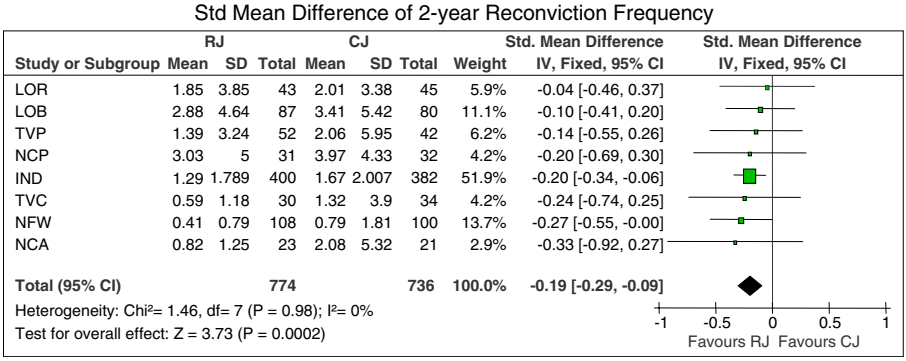

Fig. 4 Standard mean difference in 2-year reconviction frequency in independently evaluated RCTs of restorative justice conferencing

extremely badly to RJC, with before–after increases in their arrest frequency exceeding 300% in the RJC-assigned group but not in the controls who were prosecuted in court. None of the independent evaluators reported even attempting such sensitivity tests by race, let alone negative results from them.

It may be argued that our negative results should be associated with our original self-image as independent evaluators and not as the developer–evaluators that we might be coded as under the criteria Eisner (2009) suggests. Yet we have, for example, been co-authors with an openly normative advocate of restorative justice, Project Director John Braithwaite. We were, for example, deeply engaged with the local police in trying to make sure the program was delivered. We even reported one officer for misconduct in an RJ meeting that we attended, as part of our attempts to observe 100% of court and RJC events. For all these reasons perhaps we should code ourselves as developers in Canberra as well as evaluators, at least as Eisner seems to use the term. However, the fact that the coding is unclear reinforces our concern about the reliability of any test for analysts' biases using this key concept.

We do not know whether this set of findings is the only case in which evaluators who were (arguably) developers reported worse results than independent evaluators did. We do know that we have been very concerned not to avoid either overly positive or negative 'spin' on RJC in general, with a premise that it may have very different effects under different circumstances. Taking our roles to be neither advocates nor judges, but rather part of a process of invention (Sherman and Strang 2004; Sherman 2006), we have simply tried to pay detailed attention to the truth about RJC effects as our major tool for re-inventing justice (Sherman 2003), if not with RJC, than with something else.

Practices conducive to analyst's bias against program success

In the case of the 12 RCTs in restorative justice that we have designed since 1993, we can report repeated efforts to try to find harmful results from the program we were testing. We can confess to probing the data extensively with sensitivity analyses and to using the conservative procedure of intention-to-treat analysis rather than distinguishing treatment as delivered in any way. We have also delayed full publication of positive results of any one experiment or group of experiments, in the belief that the soundest conclusions would come from a prospective meta-analysis of all the RCTs. On the other hand, we also delayed full publication of some negative results for the same reason.

Our history of reporting results of these randomized trials can, therefore, be criticized for delay, in similar proportion to the criticism of Sherman and Berk (1984) for undue haste in publishing a single experimental result on the effects of arrest for domestic violence (Lempert 1989). Whether this constitutes a bias against finding positive results is arguable.

Other practices might be more clearly applied to the detection of an analyst's bias *against* interpreting the results as successful. Perhaps the most powerful indicator is the one described in the preceding section: developer reports that are less favorable

than those of independent evaluators. The suggestion here is that when this indicator is present, it can be used to outweigh many of the lesser indications of bias in favor of beneficial results in Eisner's (2009) 'one-tailed' list of indicators.

The case of negative developer effects, however, would only apply to a test for developer bias against program success. In order to detect potential analysts' biases among independent evaluators as well, we suggest a broader checklist. To some extent, this checklist should be the mirror image of Eisner's checklist for bias in favor of success. It should also reflect ways in which analysts may, *without any conflict of interest*, unwittingly or mistakenly analyze results in a way that tends to minimize evidence of success. Some of the practices we list have even been official policy of government agencies at various times—a misguided attempt to use a one-size-fits-all policy for all crime prevention program evaluations. What we suggest here is a flexible and open-ended inquiry into whether research evidence has been misinterpreted in ways that obscure a true success in preventing crime.

Our Table 1 summarizes the other tail of the practices Eisner describes as indicating 'biased' results. What he appears to mean is 'results biased in favor of success.' Hence, we present our indicators as practices conducive to a bias towards finding negative program results.

Table 1 Practices conducive to evaluation results biased against program success

Parameter	Characteristics
Developer versus independent evaluations	Developer-led results more negative than independent results
Manipulation of data	Failure to discuss outlier effects
Definition of outcome variables	Limitation to prevalence of conviction
	Failure to test for cost effectiveness
Statistical analysis	Failure to report power tests when results are not significant
	Failure to use forest plots for multiple positive but non-significant results
	Failure to conduct sensitivity analyses by outcome measures
	Failure to conduct subgroup sensitivity analyses
	Use of bright-line significance tests when *P* values are marginal but effect sizes are large
Reporting and dissemination	Failure to report favorable results
	Selective reporting of negative results
	Selective reporting of negative subgroup analyses
	Delay in publication of studies with positive results (reverse file-drawer problem)
	Over-interpretation of negative results in small trials
	Use of intention-to-treat without reporting non-treatment or crossover rates

Discussion

When Jonas Salk was asked whether he would patent his polio vaccine, he replied with incredulity that it would be impossible. It would be, he said, like "patenting the sun" (Smith 1991). Perhaps we should say the same about programs that work to prevent crime.

Criminology was founded in the Age of Enlightenment, which Benjamin Franklin called "the age of experiments." The issues Eisner (2009) raises about conflict of interest in evaluating crime prevention programs were familiar to 18th century inventors and experimenters, who should be well known to experimental criminologists. The Enlightenment was a time, as now, in which experiments were used in both unselfish quests for solutions to major problems and for profit-driven quests to make fortunes. Franklin himself was often asked why he did not try to earn more from his inventions, such as the lightning rod, the Franklin stove and bifocal glasses. He brushed off the questions by saying he hoped merely that they would be of benefit to humanity.

Others were not so selfless. Anti-slavery campaigner Josiah Wedgwood (Charles Darwin's grandfather), for example, made a fortune by conducting some 5,000 confidential experiments to perfect his method of glazing china, at least one of which he later patented. James Watt, who was Wedgwood's fellow member of the Lunar Club, a group of inventors working in the Midlands in the UK, made a fortune by developing and testing a better steam engine through repeated experiments (Uglow 2003). Other, less famous (or scrupulous) inventors made money by selling fraudulent medicines to the credulous, prompting the pharmaceutical industry to re-invent itself as selling 'ethical' drugs.

More famous medical inventors shared Franklin's disinterest in profits. Dr. Edward Jenner, for example, discovered a vaccine for smallpox by conducting experiments with cowpox inoculations, then promoted it at the cost of his medical practice. The British government gave him £30,000 in grants to compensate for his lost income over a decade (www.jennermuseum.com). Naval ship's doctor James Lind conducted experiments in the treatment of scurvy by various liquids, including fruit juices. His reward for curing scurvy by the ingestion of lime juice was to wait more than 40 years until the British Navy adopted his recommendations (Trohler, U. at www.jameslindlibrary.org).

The first criminologist, the playwright and magistrate Henry Fielding, actually promoted new ideas for crime prevention in both *pro bono* and profit-making arenas. In his 1751 treatise on the causes of robbery, he invented and developed a wide range of new policies, some of which he was later able to test. As the founder of the modern professional police, he received a Home Office grant to pay their salaries; some of them were put on horseback for rapid response to crime emergencies like street robberies. Yet, he also established the first crime statistics in the guise of a newspaper sold by the copy, reporting each crime that occurred in London—at least those of which he was notified.

Some 250 years later, we now grapple with a far more nuanced set of issues than the founders of our field. Now, as then, there may be no easy answers. Now, as then, there were social forces and moral philosophies attempting to restrict the scope of

experimentation, just as the US Congress restricted research on gun crime in the 1990s. The justifications for these restrictions and condemnations may vary, yet all are seen by their proponents as driven by the best of motives.

It is in this context that we should be wary of any attempts to restrict or regulate data analyses. At the very least, we should not assume that analysts have an ideological conflict of interest, unless there is clear evidence for that assumption on a case by case basis. The eighteenth century was also the age of liberty, with major advances in human freedom. Then, as now, we rely heavily on the marketplace of ideas and suffer from restraint of free expression.

We take Eisner's contribution as a contribution to that marketplace: a one-tailed test for sniffing out a potential success bias in crime prevention evaluations. We also attempt to contribute to that marketplace, with the other half of what we think should be a two-tailed test for bias in assessing crime prevention success. Whether anyone should, or can, use these tests with reliable estimates of bias remains to be seen. Perhaps it is useful just to place this discussion in the public square, so that data analysts of the future may use it as a benchmark against which to make their own decisions about reporting results and conclusions.

Acknowledgments From the Jerry Lee Program on Randomized Controlled Experiments in Restorative Justice, a consortium of the Regulatory Institutions Network, Research School of Asian and Pacific Studies, Australian National University; the Jerry Lee Center of Criminology, University of Pennsylvania; and the Jerry Lee Centre for Experimental Criminology, Institute of Criminology, University of Cambridge. We wish to thank Jerry Lee, the Australian National University, our research partners in the Australian Federal Police, the Metropolitan Police of London, The Thames Valley Police, Northumbria Police, and the many funders of the experiments discussed in this paper, including the Home Office (UK), the National Institute of Justice (USA), the Criminology Research Council (Australia), the Smith Richardson Foundation (USA), the Esmée Fairbairn Foundation (UK), and the Jerry Lee Foundation (USA).

References

Barnes, G. (2006). *Race and characteristics of cases assigned to restorative justice vs prosecution*. Paper presented to the American Society of Criminology, Los Angeles.

Braithwaite, J. (1989). *Crime, Shame, and Reintegration*. Cambridge: Cambridge University Press.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Eisner, M. (2009). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology*. doi:10.1007/s11292-009-9071-y

Farrington, D. P. (1983). Randomized experiments on crime and justice. In: Morris, N. and Tonry, M., eds., *Crime and Justice: An Annual Review of Research, vol. 4.*, pp. 257–308.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: what have we learned in the last two decades? *Journal of Experimental Criminology, 1*, 9–38.

Kelling, G. L., Pate, T., Dieckman, D., & Brown, C. (1974). *The Kansas City Preventive Patrol Experiment: Summary Report*. Washington, D.C.: Police Foundation.

Lempert, R. (1989). Humility is a virtue: on the publicization of policy-relevant research. *Law & Society Review, 23*, 145–161.

McGarrell, E. F., & Hipple, N. K. (2007). Family group conferencing and re-offending among first-time juvenile offenders: the Indianapolis experiment. *Justice Quarterly, 24*, 221–246.

Pate, A. M., Wykoff, M. A., Skogan, W., & Sherman, L. W. (1986). *Reducing fear of crime in Houston and Newark: a summary report*. Washington, D.C.: Police Foundation.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Shapland, J. M., Atkinson, A., Atkinson, H., Dignan, J., Edwards, L., Hibbert, J., et al. (2008). *Does restorative justice affect reconviction?: the fourth report from the evaluation of three schemes*. London: Ministry of Justice.

Sherman, L. W. (1986). Policing communities: what works? In: Reiss, A. J., Jr., and Tonry, M., eds., *Communities and crime. Crime and justice: a review of research*. Chicago: University of Chicago Press.

Sherman, L. W. (1992). *Policing domestic violence: experiments and dilemmas*. N.Y.: Free Press.

Sherman, L. W. (2003). Reason for emotion: reinventing justice with theories, innovations and research. *Criminology, 41*, 1–38.

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., & Berk, R. A. (1984). The specific deterrent effects of arrest for domestic assault. *American Sociological Review, 49*, 261–272.

Sherman, L. W., & Weisburd, D. (1995). General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial. *Justice Quarterly, 12*, 625–648.

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (eds). (2002). *Evidence-based crime prevention*. London: Routledge.

Sherman, L. W., & Strang, H. (2004). Verdicts or inventions? Interpreting results from randomized controlled experiments in criminology. *American Behavioral Scientist, 47*, 575–607.

Sherman, L. W., & Strang, H. (2006). *Race and restorative justice: differential effects for Aboriginals and Whites in the Canberra RISE project*. Los Angeles. Paper presented to the American Society of Criminology.

Sherman, L. W., & Strang, H. (2007). *Restorative justice: the evidence*. London: Smith Institute.

Sherman, L. W., Strang, H., & Woods D. J. (2000). *Recidivism patterns in the Canberra reintegrative shaming experiments*. Canberra: Australian National University, Centre for Restorative Justice. http://www.aic.gov.au/rjustice/rise/recidivism/report.pdf

Sherman, L. W., Strang, H. & Woods D. J. (2008). Effects of restorative justice conferences on criminal convictions. Paper presented to the 3d Stockholm Symposium on Criminology, June.

Smith, J. S. (1991). *Patenting the sun: polio and the salk vaccine*. New York: Anchor Books.

Uglow, J. (2003). *The Lunar men*. London: Faber and Faber.

Woods, D. J. (2006). *Race and repeat offending in the Canberra RISE project*. Los Angeles. Paper presented to the American Society of Criminology.

Lawrence W. Sherman is the Wolfson Professor of Criminology at Cambridge University, UK, and Director of its Jerry Lee Centre of Experimental Criminology at the Institute of Criminology. He is also Professor of Criminology at the University of Pennsylvania, USA. The founding President of the Academy of Experimental Criminology, he is the author of the forthcoming book *Experimental Criminology* and has designed or directed over 30 randomized field experiments.

Heather Strang is Director of the Centre for Restorative Justice in the Regulatory Institutions Network (Regnet), Research School of Pacific and Asian Studies, at the Australian National University. She is also a Lecturer in Criminology at the University of Pennsylvania and Senior Research Fellow in the Institute of Criminology at Cambridge University. Elected a Fellow of the Academy of Experimental of Criminology in recognition of her book *Repair of Revenge: Victims and Restorative Justice*, she has led twelve randomized trials of restorative justice conferencing. She is currently directing an Australian Research Council study of both offenders and victims in the ten-year aftermath of four RCTs of restorative justice in Canberra.