

“To develop and test:” The inventive difference between evaluation and experimentation

LAWRENCE W. SHERMAN

Jerry Lee Center of Criminology, University of Pennsylvania, 3809 Walnut Street, Philadelphia, PA 19104, USA
Email: lws@sas.upenn.edu

Abstract. The National Research Council (NRC) Report on Improving Evaluation of Anticrime Programs raises a fundamental question about the mission of evaluation research. The implicit premise of the report is that the mission of evaluation is to answer questions about programs developed by others; in short, to *test* anti-crime programs. In contrast, the mission of experimental criminology has, historically, been to *develop* anti-crime programs as well as to test them. There are times when an arm’s-length relationship between program and evaluation may be appropriate. Yet, such a separation necessarily produces a courtroom-like adjudication role for evaluators, rather than the laboratory-like, participant–inventor role that has characterized the best of experimental criminology. The recent case of the Chicago police’s “evaluating” the use of sequential suspect identification methods developed by academic psychologists shows the many flaws of the “testing-only” model. This suggests that providing “effective guidance of criminal justice policy and practice,” as the NRC report defines its focus [Lipsey, M. ed (2005). http://newton.nap.edu/pdf/0309097061/pdf_image/R1.pdf] will not only require evaluation research (defined as arm’s-length testing) but the full toolbox of experimental criminology to develop *and* test anti-crime programs.

Key words: crime prevention, evaluations, experiments, innovation, inventions, lineups, program development, sequential identification, verdicts

Introduction

If experimental criminology had a succinct mission statement, what would it be? Could it be something as “sticky” (Gladwell 2000) as “To Protect and Serve,” the mission long emblazoned on the side of Chicago police cars? Or would it reflect the recent—and terribly flawed—effort of the Chicago Police Department to evaluate the effects of a new method for identifying crime suspects? The premise of that research project might be this: “To Test—At Arm’s Length from the Developer.”

While the culture of science may scorn laconic statements of purpose, it also values the elegance of comprehensive brevity. My own suggestion for the mission of experimental criminology draws on the premise that science is a never-ending process of building knowledge, and rarely a final adjudication. For experimental criminology, this process is best summarized: “To Develop and Test.”

The four words in this mission statement could describe the mission of a wide range of scientific or evidence-based fields; medicine, engineering, agriculture and education are examples. The four words could also summarize the mission of the *Journal of Experimental Criminology*, which describes its aims and scope as publishing “.high quality experimental and quasi-experimental research in the

development of ...crime and justice policy” [emphasis added]. What the words could not be adopted for is the mission statement implied by the National Research Council report on *Improving Evaluation of Anticrime Programs* (Lipsey, ed., 2005).

From the first page of the executive summary—all that many practitioners might ever read—the report defines the mission of evaluation research as providing evidence for “effective guidance of criminal justice policy and practice...about their effects on populations and conditions they are intended to influence.” That opaque language alone cries out for a “sticky” summary, but language is not the point. The point is the report’s silence on the role of criminology (or any science) in the *development* of anti-crime programs. A search for the word “development” in the PDF file of the executive summary, for example, reveals no reference to the role of evaluators or social scientists in the design of programs or innovations subject to evaluation; the word “development” is used exclusively with reference to the development of research designs and tools.

Experimental criminology is not just the testing of “other people’s programs.” Its mission also includes the design of our own “crucial” experiments, major discoveries that have been achieved by theoretical synthesis and pragmatic demonstrations. Most of the Fellows of the Academy of Experimental Criminology have contributed in this way to what we know about preventing crime—including two members of the National Research Council (NRC) report’s committee. Yet, the funding for what experimental criminology can learn in the future may be limited to money earmarked as “evaluation” research. The Home Office has long followed this firewall distinction, forcing implementers of randomized controlled trials to be defined as “developers”, while funding an evaluation by independent, arm’s-length “evaluators.” When experimental criminologists are thereby excluded *either* from program development, *or* from program testing, the potential value of our contributions is greatly reduced. Moreover, the expense to government becomes far greater.

The dangers of arm’s-length evaluation may be further compounded when evaluations are conducted by organizations with a stake in the outcome. The recent reports on an Illinois evaluation of sequential versus simultaneous witness identification (Zernike 2006; Mecklenburg and Sheri 2006) shows what can happen when the evaluator is able to exclude the developer from a test, with the evaluator undermining the integrity of the program in the course of the evaluation (Diamond 2006; Wells 2006).

My principal comment on the NRC report, then, is this: it is half a loaf, better than none, but still in need of the “development” side of the verb “to test.” My corollary comment is that we would discover how to prevent crime much more quickly and effectively by *investing more in criminology that “develops and tests”* than by investing in independent, arm’s-length testing. The latter may be essential in the case of programs developed from non-scientific models of crime prevention that consume large amounts of money, such as “Scared Straight” or Drug Abuse Resistance Education (DARE). But such examples are few and far between. More

important is the continuing challenge of developing and testing new, high-integrity programs that work better than current practices, such as David Olds' home visits by nurses to high-risk mothers. His successful integration of development and testing is an exemplar of experimental criminology, a model that was unfortunately neglected from the NRC report.

Granting that both arm's-length and development-linked testing are needed, how can we decide to allocate funding between them? Perhaps the best answer is to fund both of them equally, with a year-to-year review of which method is producing more cost-effective evidence on how to reduce crime. Such assessments are not simple, but they could be done if each method were to be given a fair chance to compete. They cannot be done if the develop-and-test model is neglected altogether. The following analysis reviews some evidence on the value of develop-and-test criminology.

Criminologists as inventors

In the history of experimental criminology, criminologists have been actively involved in the design, development and implementation of anti-crime programs (Sherman 2005). This work has been heavily informed by criminological theory and data, in contrast to anti-crime programs designed solely by non-criminologists. That history began with what is arguably the first published work in criminology, Henry Fielding's *An Enquiry Into the Causes of the late Increase of Robbers*.

In this book Fielding (1751) used pre-theories of environmental criminology to describe the epidemiology of armed robbery–murder and to sketch out a plan for reducing robberies in London. The book led the Prime Minister of England to ask Fielding to establish the first paid police service in English history, the Bow Street Runners. Fielding's "level 2" evaluation design of his own invention was a before–after, no-control group interrupted time-series, which may only have found a regression to the mean: the wave of murders prompting the "experiment" stopped shortly after the police force went to work, and months passed with no murders in London. He then managed to arrange for the Runners' survival long after he died, until they grew into the Metropolitan Police some 80 years later.

In what Benjamin Franklin called the "age of experiments," Fielding was a classic example of an Enlightenment inventor. His contemporaries invented—and tested—smallpox vaccine, organ transplant surgery, paper money, the steam engine, constitutional democracy, lightning rods, and many other useful inventions. Rather than standing back to evaluate what others invented, the Enlightenment "evaluators" tested their own inventions, usually revising them continuously on the basis of the results. Josiah Wedgwood's notebooks reflect, for example, over 5,000 tests of different ways to achieve a smooth glaze on china plates (Uglow 2002) p. 53). Fielding developed his inventions on a smaller scale, adding crime statistics, rapid response "911" capacity on horseback, and other ways of improving his design. His own motto may have been the advice of transplant surgeon John Hunter, given to vaccine inventor Edward Jenner: "Don't [just] think, try!"

Thinking and trying

The relationship between thinking and trying is central to the question of the role of evaluation. If “trying” is divorced from “thinking”—in the sense of incorporating criminological theory and data in the design of an anti-crime program—then there is far less reason to believe that a program is likely to succeed, no matter how rigorous the evaluation of its effects. If, by contrast, criminologists are, at least, part of the team of inventors of an anti-crime program, or even the lead inventors, then they may be able to design the most knowledge-based policies that can be tested.

Useful evaluations must begin with useful programs. If the National Research Council has concluded that criminological knowledge cannot improve the design of anti-crime programs, it would be a strong indictment of criminology. Yet their intent seemed more modest: to limit the potential conflict of interest between invention and testing, in the same Weberian march of social differentiation and rule-making that may be strangling the European economies. This may, in turn, build on the criticisms of evaluation research by the General Accounting Office (GAO) and others that prompted the commissioning of the NRC report. We must not forget that the GAO itself has no record of discovering new ways to solve problems. The GAO is an entirely adversarial institution and sees evaluation research as the necessary bloodletting of a democracy. This view has many virtues to be praised. That is not, however, the only role of experimental criminology, or the best use of its strength.

Imagine what the state of early infancy would be like today if the experimental criminologist David Olds had been told three decades ago that he could not design his own program for improving the treatment of infants by high-risk mothers (Olds et al. 1986). We would be deprived of an extremely effective anti-crime program. Under the GAO’s adversarial model that is implicit in the NRC report, Olds would have been limited to the role of evaluating existing parent-training programs. If so, he might have spent the rest of his life carefully documenting what *does not* work. Instead, available funding then allowed him to design his own program that built on early childhood research, even while taking a bold shot in the dark that nurse (R.N.) home visits would be the best way to deliver the training. After three tests on different samples in different parts of the country, his evaluations of the nurse home visitation program show consistent anti-crime results. All of those evaluations have been designed and led by Olds himself. As of 2006, further replications led by others are now under way in seven European cities, but Olds remains closely involved in the design of the evaluation and the integrity of the program delivery.

Now imagine if Olds had been allowed to design the program but had been forbidden to conduct the evaluation. Suppose, in addition, that he had been allowed to deliver the program on a random assignment basis, but he had been barred from collecting or analyzing outcome data. Suppose that all of the publications resulting from the evaluation were then authored by the “independent evaluator” who had been retained by competitive bidding on a request for proposal (RFP), and that

Olds would have had no access to the data as the “subject” of the evaluation. How much incentive would Olds have had to replicate the program and to carry it forward? How much conflict over data analysis would have arisen between the program “developer” and the program “evaluator?” Who would have been the authoritative scientist in the room whenever matters of research design, measurement or analysis were on the table?

As it happened, Olds was able to both think and try, to design and evaluate the program, to invent and re-invent it as he went along. It is this creative control that provides the inspired “flow” needed to make great science. It is no different, in one sense, from the preference of some movie-makers to write, direct, and produce their own movies. It is what allows strong leadership to take a holistic vision of the work to be done, rather than dividing it up into disconnected parts.

The fact that Olds could package the program delivery as a randomized trial gave him, as well as his funders, a rigorous way to evaluate the program. The idea that he had a conflict of interest in evaluating his own program seems, in retrospect, preposterous, if not insulting. In theory, he could have fudged the data; scientists in all fields have done that on occasion. But they have done it even when they did not design their own program. There is no evidence, to my knowledge, of higher rates of data corruption in developer-evaluated programs than in independently evaluated programs.

By holding himself accountable with his own randomized trial, Olds gave himself a great incentive to insure very high levels of integrity and fidelity in the administration of the program—an essential element of an “efficacy trial,” as distinct from the “effectiveness” trial (under typical field conditions) that should come only after efficacy has been established. Thinking, and trying, and thinking some more seem to be essential in getting a program to efficacy. It is not, however, what usually happens when programs are developed without benefit of theory or prior data—as they have been, all too often.

The central problem would seem to be getting more science into the design and development of programs, not just the evaluation. Any model of evaluation that divorces the thinking from the trying runs the risk of weakening program development. Yet that is just what the NRC report may be recommending.

Verdicts or inventions?

The NRC report places anti-crime evaluation research squarely in the camp of rendering verdicts on other people’s programs, like the judge or jury (as “fact-finder”) in a courtroom. This view is critically different from the problem-solving vision of evaluation research, in which the program evaluator tries to improve on a program rather than rendering a verdict on whether it “works” or not (Sherman and Strang 2004). While many people have read much recent evaluation literature (e.g., Sherman et al. 1998) as taking a “death penalty” verdict approach to programs that “don’t work,” former Attorney General Janet Reno and others have called on criminologists to be more constructive. Why can we not do further testing,

Attorney General Reno asks, to see whether a program might be more effective if modified in some way or if used with different populations. Certainly, the contemporary view of evaluation as an independent “court of evidence” conflicts with General Reno’s request, which itself reflects the much longer tradition of evaluation as an integrated (and repeated) step in the process of invention.

The NRC report’s vision of how evaluators relate to programs is seen in its repeated references to RFPs in a way that implies that the program comes first, then the evaluation. This is underlined by the recommendation that the National Institute of Justice (NIJ) maintain a separate evaluation unit to issue such RFPs and to insure judicious independence of an evaluation from those with a stake in the program’s success. These procedures reflect a valid concern for potential conflicts of interest. The report also reflects the way in which much federal (and UK) funding of anti-crime program evaluation has been spent.

This vision of evaluation is not, however, evidence based. The report fails to reflect a systematic testing of the hypothesis that conflicts of interest pose a greater risk than spending money on ineffective programs. The report fails to describe fully and accurately the way in which many (if not most) successful demonstrations of effective anti-crime programs have actually been achieved: by an “embedded,” collaborative, partnership between researchers and practitioners. For such an environment, in which social scientists can both frame and test hypotheses, retaining an independent evaluator by competitive bidding is intrusive and disruptive to the process of invention.

While the prescriptions in the report are useful in many ways, they miss an important opportunity to encourage the vital partnerships between criminologists and public (or non-profit) agencies that have yielded so many successes. As a purely empirical matter, the report fails to examine the evidence on how effective anti-crime programs have actually been developed—as well as rigorously evaluated. That failure is all the more striking in light of the extraordinary contributions by three of the committee members, including its chair, Mark Lipsey, who first documented the strong positive correlation between program success and evaluator involvement in program development—in a paper that was not even cited in the NRC report (Lipsey 1995). The members also include Denise Gottfredson and David Weisburd, whose respective leadership was central to the development of programs which they have also evaluated with rigorous scientific methods. If one is looking for the arm’s-length, you-design-and-we-evaluate model of program evaluation that is implicit in this NRC report, it will not be found in Gottfredson’s successful collaborations with schools (Gottfredson 1986, 1987) or in Weisburd’s successful partnerships with police agencies (Weisburd and Green 1995).

One can only conclude that this was, indeed, a report written by a committee, in which all that could be reported was the common denominator of what was agreeable to all members and their subsequent reviewers. How much insight and vision was lost in the process we will never know. This is an unfortunate, if perhaps necessary, characteristic of committee reports. In this case it also omitted a review of the evidence on what the report did not recommend: the engagement of evaluators in program development.

What works in anti-crime evaluations: Some evidence

The most important hypothesis for this comment is that *anti-crime programs are more likely to be found effective if the same experimental social scientists contributed to the program development and the program evaluation.*

The hypothesis itself has been tested before. Each time it has been supported (i.e., failed to be falsified). Yet, its survival can be subject to differing interpretations (Lipsey 1995). One is that author affiliation improves the integrity of program delivery. Another is that author affiliation reduces the integrity of data analysis—biasing the “spin” on the results in favor of program effectiveness. While it is impossible to choose between these two hypotheses based solely on the evidence, it is worth reviewing what evidence we have.

In 1992, a review of 400 delinquency treatment evaluations (Lipsey 1992) reported that delinquency reduction effects were greater in evaluations in which the evaluators had been able to influence the program. Lipsey (1995) later was the first to suggest the “cynical” versus “program integrity” explanations for this correlation. These competing explanations remain unresolved to this day, even as new evidence of the correlation is reported.

In a systematic review of 62 evaluations of sex offender treatment programs Losel and Schmucker (2005) found that the affiliation of the evaluation’s author with the design and delivery of the project almost doubled the tested effectiveness of the program in preventing recidivism. Author affiliation was also one of the most powerful predictors of large effect size for anti-crime effects of the program. Losel and Schmucker (2005, p. 137) interpret this as evidence of greater program integrity associated with affiliated authors than with “arm’s-length” authors. They support this interpretation with independent measures of the extent to which the programs were delivered as designed, which also show association with author affiliation. While a cynic might reply that the authors can spin program integrity data just as easily as outcome data, the integrity correlation may, at least, lower the plausibility of the author “spin” hypothesis somewhat.

Most recently, and also in this journal, a review of 12 previous meta-analyses and a new meta-analysis of 300 randomized field trials tested the developer-evaluator hypothesis (Petrosino and Soydan 2005). The results were even more definitive, if no more interpretable. When evaluations are authored by people who clearly participated in the development of the program, they are more likely to show anti-crime effects than when authors are not involved in the development and implementation of the programs. This was the finding for 11 of the 12 previous meta-analyses. In the new meta-analysis of a data set on 300 evaluations collected by Petrosino (1997) that began under an NIJ grant co-directed by L. Sherman and D. Weisburd, the largest effect sizes were associated with evaluators who had the greatest influence on the program implementation—and in the role of program developer.

Petrosino and Soydan also conclude that there is no way to choose between the hypotheses with the currently available data. Yet, they make the case that, in principle, there is a need to be concerned about developers-as-evaluators. That

principle is based on empirical evidence not from criminology but, rather, from the business of medicine. As we so often do in experimental criminology, those in medicine draw on the metaphor of pharmaceutical development to demonstrate the clear danger of a financial incentive to demonstrate the safety and efficacy of a program. This also seems to be the basis on which the Home Office and the GAO seem to prefer independent evaluations of programs, regardless of the nature of the field.

Many academics are passionately concerned about conflict of interest among criminologists. One Campbell Collaboration reviewer recently—and anonymously—suggested that the Australian National University’s Centre for Restorative Justice had an obvious conflict of interest in evaluating restorative justice programs. (Perhaps he or she would prefer that the Centre, which has reported more negative evaluations of restorative justice than positive ones, should change its name to the inelegant “Centre for the Independent Evaluation of Hypotheses About Restorative Justice”).

The evidence does not support these concerns. While integrity problems have been well documented in pharmaceutical research, there is no documented case (to my knowledge) of intentional falsification of data in anti-crime program evaluations. The virtue of experimental criminology in this regard is that few of us have any financial incentive at all to make a result come out one way or another. If one program does not work, there are always others to be tried. The incentive for a criminologist is in the discovery and publication of internally valid results, not the marketing of a product for a profit.

In my own work, for example, which the NRC report cites at the outset of Chap. 2, I have reported more program failures than successes. As of this writing, I have reported, in print or in public meetings, on 15 anti-crime field experiments I have directed in which my staff were in constant control of the process of random assignment—a characteristic not examined in any meta-analysis to date, but one which is central to the program integrity of the test. In each of these tests we were also influential in the development and administration of the program, usually in partnership with a police department. Of the 15 tests, eight show no effect on crime and seven show a crime reduction effect of the program. My own view is that my preferences had no effect on the results, but that the successes were only made possible by the high degree of integrity in program delivery achieved by our teams in all 15 experiments.

Adversary evaluation: The Illinois lineup “experiment”

The potential for “independent” evaluations to invoke a clear bias against the developer is clearly shown by the (Mecklenburg and Sheri 2006) Report of the Illinois State Police to the Illinois Legislature. This evaluation was undertaken in the aftermath of a Gubernatorial suspension of executions of criminals sentenced to death. The reason for the suspension was the discovery of post-sentence DNA evidence of the innocence of the death row inmates in question. This decision was

an implicit criticism of both prosecutors and police, since they had decided to arrest and prosecute these innocent people with the death penalty in view. In that context of implied criticism it is not clear that Illinois police agencies are any more “independent” of the hypothesis that their procedures produce eyewitness identification errors than the most prominent developer of an alternative method, Iowa State University psychology professor Gary Wells.

Nonetheless, based on a recommendation of the Report of the Illinois Governor’s Commission on Capital Punishment, the Illinois Legislature in 2003 commissioned the Illinois State Police—and not Professor Wells—to test the effectiveness of Wells’ sequential, double-blind, identification procedure in the field (Mecklenburg and Sheri 2006, p. i). Wells had developed this method as an alternative to the standard simultaneous police lineups of a suspect surrounded by people “similar” in physique and demographic characteristics, or a simultaneous array of photographs including at least one person police suspected of the crime. An estimated 77,000 people are charged with crimes in the USA each year when these standard methods are used, with at least 150 convictions based on such evidence overturned on the basis of DNA tests (Zernike 2006). In repeated laboratory experiments Wells had demonstrated that these standard methods yielded much higher rates of error than did an alternative method he developed. That method had two elements:

1. The sequential, one-at-a-time presentation of potential suspects to a witness.
2. Administration of the lineup by a police officer who did not know who investigators suspected (“double-blind”).

Wells had demonstrated these results in using randomized trials in laboratories but not with real criminals. He was ideally suited to take the method into the field and see whether the results could be replicated with real crime victims, suspects and police investigators. Such an experiment would have been difficult, but possible—especially when led by a highly experienced scientist who understood the nature of experiments. Instead, the test was led by an attorney with no reported experience in the conduct of field experiments. The attorney was a lawyer for the Chicago Police Department, retained by the Illinois State Police to coordinate a simultaneous field test in Chicago, Joliet and Evanston. The report claimed that the cases in the sample were chosen in a way that was “random and predetermined, i.e., it could not be within an officer’s discretion or within an officer’s control” (Mecklenburg and Sheri 2006, p. 25). It also claimed that the same officers would be conducting both procedures, so that there would be no officer effects on the results.

The report concluded that sequential lineups actually backfired, identifying more innocent people than did simultaneous lineups. The report was taken as a major blow by those attempting to change state laws to require sequential lineups. As a “verdict” on a new innovation, it was, perhaps, not final, but it is a major precedent that must now be dealt with by future research, in effect, “on appeal.” Unfortunately, the test suffered many scientific flaws that might have been avoided by working in partnership with the major developer of the method rather than working with those whose advanced positions were openly critical of the idea.

The main flaws of the research design were as follows:

1. Complete absence of random assignment, despite the report's claim.
2. Failure to insure that the same investigators used the two different methods equally.
3. Failure to disentangle the use of double-blind and sequential procedures

Moreover, the rate of false identifications in the standard-methods group was so much lower in Illinois than in all other tests that several commentators called on the researchers to explain this anomaly—a finding on which the entire conclusion was based.

The lack of random assignment was evident in the report but was pointed out neither by its author nor the *New York Times*. On pp. 25–26, the report describes three methods by which the treatments were assigned to cases, none of them even close to randomized assignment. In the Evanston design, the assignment was based on even versus odd case numbers. Now I have witnessed police investigators wait for a case to come in so they could get the outcome they prefer. I have also led an experiment that randomly assigned arrest based on officers' advanced knowledge, where we found clear evidence that officers violated the random assignment sequence to get the result they preferred (Gartin 1995). Using odd versus even case numbers does not provide any credible evidence that treatment assignment is removed from an officers' control, let alone "random."

In Chicago and Joliet the assignment was not done by case but by geographic area. All cases in certain geographic areas—i.e., investigative offices—were handled with one method, while all cases in the others were handled with the other method. The reports on these cases were turned in to the author, apparently without audit as to whether investigators "shopped" for the station using the method they preferred. Neither the basis of assignment of a case to an area (location of the crime, of the suspect's residence, of the victim's residence?) nor any means of detecting subversion of the experimental design were discussed. Regardless of issues of integrity, assignment by area is manifestly not random assignment. It is a rival hypothesis explaining the results, in the absence of any evidence to the contrary. The report provides no data on the characteristics of the crime or populations in these different areas but asks the reader to accept assurance that the cases were all equivalent. They were not. The design was at best a level 3 on the Maryland Scale (Sherman et al. 1998), which is known to be subject to many threats to internal validity.

A corollary to the lack of random assignment is a lack of measured consistency in the skills and experience of the investigators using the different methods. All of the difference in results could, in principle, be explained by those differences. They could also be explained by the fact that the non-blind lineups allow the investigator to use subtle cues to steer the witness away from a "filler" person not suspected of the crime and towards the person the police believe to be guilty. When the lineup is double-blind as to the identity of the person the investigators suspect (both lineup administrator and witness unaware of investigator hypothesis), there is no chance for the lineup administrator to steer the witness away from fillers.

This problem could have been dealt with in a straightforward experimental design to disentangle the two elements of Wells' program: a four-cell factorial design (blind-sequential, blind-simultaneous, non-blind simultaneous, non-blind sequential). The results of such a design would have provided a fairer estimate of the inaccurate identification rate between sequential and simultaneous methods, controlling for the presence or absence of double-blind administration. Blind-sequential cases could be compared to blind-simultaneous cases. Non-blind-sequential cases could be compared to non-blind simultaneous cases. Each comparison would be as similar as possible, except for the one difference being compared. As similar as is possible, of course, without random assignment—which is not very similar at all.

The report on the project in a front page *New York Times* story (Zernike 2006) stressed the fact that this was the first field test of the Wells program. It made no mention of the fact that the Illinois test failed to use random assignment, whereas Wells' (and other) laboratory experiments were consistently based on randomized controls. Thus, the difference between field and laboratory experiments could have been due to the far greater threats inherent in the level 3 designs used by the Illinois study. The point should not have been that field tests trump laboratory experiments. They do not, unless the field experiments are just as rigorous as the laboratory experiments. That may not always be possible. But in this case they could have been.

The counterfactual to this terribly flawed but highly influential Illinois State Police "evaluation" is that Gary Wells had been invited to set up a field experiment in partnership with Illinois police agencies. Whether such partnership would have been possible in the context of antagonism over innocent people on death row is unclear. What is clear is that the develop-and-test model in the Gary Wells case differs substantially from the model in the David Olds case. In the Olds case, develop-and-test were unified. Can we imagine what would have happened to Olds' model if it had been given over to an independent evaluator to assess? By the standards of the Illinois project, Olds would have had no random assignment left by the time the evaluation was completed, probably because nurses might have complained about it or said it was inconvenient. Absent random assignment, there is no way to tell whether the Olds program would have been just as negatively evaluated in the field as the Wells program.

The NRC written report shared with the workshop I attended a sense of unreality of the problems, politics and possibilities of evaluations of such questions. Evaluation policy is not just a matter for OMB, GAO or NIJ. It is far broader, affecting the ways in which state legislatures, police leaders and others think about evidence-based policy across the nation and the world. That thinking is largely uninformed about basic principles of scientific method. The Illinois report's misuse of the concept of random assignment shows that clearly. The histories of medicine, agriculture and other fields suggests that the introduction of science may work best on a face-to-face basis: partnerships between scientists and administrators, rather than arms' length evaluations.

Conclusion: To develop and test

The GAO and the NRC may be losing sight of the baby in the bath water. The baby we are trying to raise is more effective public programs to reduce crime and improve justice. The bath water is the way in which public programs are evaluated. GAO and NRC want better evaluations. Who doesn't? But it is not nearly as important that we do every evaluation right as that we solve problems. Better evaluations do not guarantee better solutions. Solving problems requires better ideas and theories, the most promising of which should then be tested rigorously. That is the promise of experimental criminology: coming up with better ideas and testing them rigorously.

Experimental criminology is a major player in the arm's-length evaluation of anti-crime programs, when such an adversarial approach is needed. The highest and best use of experimental criminology, however, is to develop and test *better* programs. It is a strategic mistake to divert funding from investigator-initiated grants to massive evaluations of atheoretical but popular spending programs. These programs are generally designed in the absence of science and have little expectation of being found effective—no matter how good the evaluation research. As the NRC report observes, not every program deserves to be evaluated. The money wasted on ineffective programs could be recovered, not by better program evaluations but by better use of existing evaluation results by the Congress and the Executive Branch.

The NRC report does a valuable service by raising the question of what our mission is. In the long run it is to develop and test. In the short run it is to persuade our colleagues and our funders that developer-evaluators are far more useful than verdict-evaluators. If the cynical view of the strong benefits reported for developer-evaluator programs is correct, a massive effort of secondary analysis should reveal that to be so. No one has yet been able to support the cynical view directly by re-analysis of publicly archived data. In principle this can be done at any time, and developer-evaluators know that.

There is no way, however, to impose program integrity in an experiment after it is all over. It is hard enough to measure it, even as the experiment is in progress. The difference experimental criminologists make is that they provide leadership for program integrity. That leadership is far more important than any conventional concern about conflict of interest. Just ask David Olds.

References

- Diamond, Shari Seidman (2006). Police lineups and eyewitnesses. Letter to the New York Times. April 24, 2006. Correction appended Sect A; Column 6; Editorial Desk; p. 18.
- Fielding, H. (1751). *An enquiry into the causes of the late increase of robbers, etc., with Some proposals for remedying this growing evil*. London: A. Millar.
- Gartin, Patrick R. (1995). Dealing with design failures in randomized field experiments: analytic issues regarding the evaluation of treatment effects. *Journal of Research in Crime and Delinquency* 32, 425–445.

- Gladwell, M. (2000). *The tipping point: How little things can make a big difference*. Boston: Little, Brown.
- Gottfredson, D. C. (1986). An empirical test of school-based environmental and individual interventions to reduce the risk of delinquent behavior. *Criminology* 24, 705–731.
- Gottfredson, D. C. (1987). An evaluation of an organizational development approach to reducing school disorder. *Evaluation Review* 11, 739–763.
- Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In H. C. T.D. Cook, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 83–127). New York: Russell Sage Foundation.
- Lipsey, M. W. (1995). What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? In J. McQuire (Ed.), *What works? Reducing reoffending*. New York: Wiley.
- Lipsey, M. ed., (2005). http://newton.nap.edu/pdf/0309097061/pdf_image/R1.pdf.
- Losel, F. & Schmucker, Martin (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology* 1, 117–146.
- Mecklenburg, Sheri H., on Behalf of the Illinois State Police (2006). Report to the Legislature of the State of Illinois: The Illinois Pilot Program on Sequential Double-Blind Identification Procedures. Chicago: Sheri Mecklenburg, March 17, 2006. Downloaded on March 30 from http://www.psychology.iastate.edu/FACULTY/gwells/Illinois_Report.pdf.
- Olds, D. L., Henderson, C. R., Chamberlin, R. & Tatelbaum, R. (1986). Preventing child abuse and neglect: A randomized trial of nurse home visitation. *Pediatrics* 78, 1436–1445.
- Petrosino, A. (1997). What Works? Revisited Again: A Meta-Analysis of Randomized Experiments in Individually-Focused Crime Reduction Interventions. *School of Criminal Justice*. Newark, Rutgers University. PhD.
- 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, 435–450.
- Sherman, L. W. (2005). The use and usefulness of criminology: Enlightened justice and its failures. *Annals of the American Academy of Political and Social Science* 600, 115–134.
- Sherman, L. W. & Strang, H. (2004). Verdicts or inventions? Interpreting results from randomized experiments in criminology. *American Behavioral Scientist* 47(5), 575–607.
- Sherman, L. W., Gottfredson, Denise C., MacKenzie, Doris L., Eck, John, Reuter, Peter & Bushway, Shawn D. (1998). *Preventing crime: What works, what doesn't, what's promising*. Washington, DC: U.S. Department of Justice.
- Uglow, J. (2002). *The lunar men: Five friends whose curiosity changed the world*. New York: Farrar, Straus & Giroux.
- Weisburd, D. & Green, Lorraine (1995). Policing drug hot spots: The Jersey City drug market analysis experiment. *Journal of Criminal Justice* 12(4), 711–735.
- Wells, Gary (2006). Comments on the Illinois Report. Downloaded May 28, 2006 from http://www.psychology.iastate.edu/FACULTY/gwells/Illinois_Project_Wells_comments.pdf.
- Zernike, Kate (2006). Questions Raised Over New Trends in Police Lineups. New York Times, April 19, p. 1.

About the author

Lawrence W. Sherman is the Director of the Jerry Lee Center of Criminology and Albert M. Greenfield Professor of Human Relations at the University of Pennsylvania, where he is also Chair of the Department of Criminology. Since 1979 he has designed and co-directed over 25 randomized controlled field experiments in the United States of America, Australia, and the United Kingdom, including the first randomized trials of arrest (with Richard Berk), restorative justice (with Heather Strang), and police patrols of crime hot spots (with David Weisburd). The founding President of the Academy of Experimental Criminology, he has also been elected president of the American Society of Criminology, the International Society of Criminology, and the American Academy of Political and Social Science.