Journal of Experimental Criminology (2006) 2:363–373
DOI: 10.1007/s11292-006-9015-8

Assessing problematic research: How can academic researchers help improve the quality of anti-crime program evaluations?

JANET L. LAURITSEN*
**Criminology and Criminal Justice, University of Missouri–St. Louis, One University Blvd, St. Louis, MO 63121, USA*
E-mail: janet_lauritsen@umsl.edu

Abstract. In this essay I examine some of the problems that prompted the National Research Council (NRC) report and consider how academic researchers might help resolve them. Many of the problems were found to be associated with research designed to assess program effects on child victimization and violence against women, areas in which research participation by subjects is particularly burdensome and difficult to obtain. Yet, program evaluations often assume that the process of subject participation is well understood and that outcome measures are reliable and valid across all subjects. A multidisciplinary, comprehensive and systematic review of victimization programs and past research is needed to advance the rigor of future evaluations. However, academics should not insist that all victim service programs warrant program evaluation as a condition of continued public support, because the decision to retain a program inevitably involves more than a scientific estimate of its effect.

Key words: interdisciplinary partnerships, victimization research, violence against women

Introduction

The report produced by the National Research Council's Committee on Improving Evaluation of Anticrime Programs provides policy makers, organizations and criminal justice researchers with an excellent discussion of how to foster and conduct useful evaluation research. Policy makers receive advice about what they should expect from program evaluations and the kinds of organizational resources that are necessary for generating good information on the effectiveness of programs. Agencies charged with sponsoring research receive guidance about planning, soliciting, reviewing, and monitoring program evaluation projects. Researchers are encouraged to think carefully about their evaluation methodologies and to plan for the myriad of potential pitfalls that may interrupt their best-laid plans. The report is thorough and should be considered standard reading for each of its three main audiences.

Of course, discussion about how to produce good evaluation research is not new to social scientists. The National Research Council (NRC) report follows a long history of publications devoted to evaluation research, and there have been several excellent and extensive reviews of criminal justice program research, such as the well-known "Preventing Crime" report (see Sherman et al. 1997). Those well versed in this history may interpret some of the recommendations in the report as reminders, rather than as new information. But researchers and others who are new

to the area will find the report especially helpful, as they will be made well aware of the context in which their projects will be judged.

The publication of the NRC report is important, because it outlines the set of challenges that must be confronted by criminal justice evaluators who wish to make their work relevant to the taxpayers who fund their research. It also signals the need to examine the nature of the problems that prompted the report in the first place. In this essay I examine the impetus for the report and consider ways in which academics might help resolve some of these challenges in criminal justice evaluation research. Academic researchers may help advance the state of criminal justice research by focusing not only on methodological issues but by creating useful interdisciplinary partnerships that can bridge some of the barriers that appear to hinder sound program evaluations. We can also pay greater attention to the political context in which our work is assessed.

Criticisms of justice outcome research by the US General Accounting Office

According to the NRC, the main impetus for the report was a set of assessments by the bipartisan US General Accounting Office (GAO) that were "sharply critical" of many of the evaluation projects sponsored by the US Department of Justice (p. 8). The GAO was especially critical of some of the research conducted under the auspices of the Office of Justice Programs (OJP) and the National Institute of Justice (NIJ), the nation's principal research and evaluation agency for criminal justice programs. Many of the details of the GAO assessments are discussed in the NRC report (see pp. 8–11), and the original reports, including responses from the OJP and the NIJ, are available on the GAO website (http://www.gao.gov).

Although GAO reports typically are not thought of as required reading by criminologists, it is important to take a close look at their assessments and the official responses of the Assistant Attorney General, who is charged with overseeing the operations of the OJP. These exchanges clarify the federal government's oversight practices and standards for research quality, illustrate how various agencies have defended some of their decisions, and provide transparent accounting of how these research funding decisions were made. They also reveal where agreements or disagreements about research quality appear most often, which, in turn, may help us find ways to resolve some of the challenges.

A summary of the GAO's assessment of NIJ research is available on-line in a document entitled "Justice Outcome Evaluations: Design and Implementation of Studies Require More NIJ Attention" (US General Accounting Office 2003). This summary review followed several earlier critiques and was prepared in response to a specific request from the Chairman of the Subcommittee on Crime, House Judiciary Committee. To conduct this review of the NIJ's outcome research portfolio, the GAO reviewed a sample of 15 evaluation projects covering the 1992–2002 period. These 15 projects accounted for approximately half of the NIJ's expenditures for evaluation or outcome research during this time period. Of the 15 studies, the GAO concluded that five were "sufficiently well designed and

implemented," six were sufficiently designed "but encountered implementation problems that limited the extent to which the study objectives could be achieved" and four "had serious methodological problems from the beginning" (2003, p. 3). They state their conclusions rather pointedly: "Our in-depth review of 15 outcome evaluations managed by NIJ during the past 10 years indicated that the majority was beset with methodological or implementation problems that limited the ability to draw meaningful conclusions about the programs' effectiveness" (2003, p. 26).

As might be expected, there were few disagreements between the GAO and OJP regarding the five evaluations that were deemed well-designed and implemented. The GAO described these studies as conforming to "generally accepted social science standards for sound design" (2003, p. 10). These standards include sufficient sample size, appropriate data collection, reliable and valid measures of outcomes, random assignment and comparison groups, measurement of change in the outcomes over time, and adequate statistical controls to isolate program effects from other potential factors. These five evaluations accounted for about $3.3 million, or 21%, of the expenditures reviewed.

The GAO expressed strong concerns about six studies they characterized as well designed but seriously hampered by implementation difficulties. Most of these difficulties were characterized as beyond the control of the investigators. In one instance an intervention site and a control group site were chosen to be comparable on the basis of crime rates and race and family composition. Once data collection was well underway, the two sites were found to differ on levels of employment, which posed a problem because the intervention was designed to promote family self-sufficiency. The GAO could have viewed this as a design problem (i.e., it should have been discovered during the site selection process). However, they decided that, given the state of knowledge at the time, the investigators used reasonable criteria in choosing the program and comparison sites (2003, pp. 15–16). Even so, the final report from this project is criticized for failing to include analyses that statistically control for the pre-existing differences that were discovered (2003, p. 16).

Other difficulties that were encountered across these six studies included failures of the programs to be implemented as planned, and low, differential, or unreported response rates among subjects upon whom the outcome data were based. In at least one of these projects, the NIJ consulted with the investigators to develop strategies that would raise response rates (which at the time were less than 25%), but these efforts were unsuccessful. In light of these problems the GAO viewed these studies as providing, at best, inconclusive results and, at worse, biased, unreliable, and invalid conclusions. These six evaluations accounted for approximately $7.5 million, or 48%, of the expenditures reviewed.

Sharp critiques from the GAO were directed toward four evaluation projects described as having "serious design problems that diminished their ability to produce reliable or valid findings about program outcomes" (2003, p. 18). In other words, these were studies that, in the GAO's assessment, should not have been funded in the first place. The major problems included a lack of necessary comparison groups, insufficient measurement of outcomes to assess program

effectiveness, and the lack of baseline or pre-program data upon which program outcomes could be assessed.

One of the studies, designated as a "national evaluation" of a domestic violence and child victimization enforcement grant program, was designed, according to the GAO, as a collection of case study information gathered through site visits to nine program areas. Because there were no comparison groups and few pre-program data, this methodology made it impossible to assess whether the programs operating in the various sites had any effect on violence or on local agencies' response to such violence. In the three other studies the investigators planned to "explore the feasibility of using comparison groups" after funding was made available (2003, p. 18). But the GAO noted that no comparison groups were in use in these studies at the time of their review, even though at least one of the projects was in its third year of funding.

Some studies that were purported to evaluate a program's effectiveness relied on "intermediary results" such as assessments of service providers' knowledge and training rather than on measures of the intended outcomes of the grant program (2003, p. 19). This methodology was viewed as inadequate by the GAO, because the purpose of the programs was to improve the safety of women, children, and other victims more broadly defined. In selecting these intermediary outcomes as the focus of research, the "design precludes conclusions about whether the programs improved the lives of victims of domestic violence or their children" (2003, p. 19). In addition, the lack of pre-program or baseline data was found to be a critical flaw in two studies. One evaluation of whether a domestic violence program resulted in changes in local agency procedures or improved safety of victims had little or no pre-program information upon which to assess the outcomes of the program. In another study, the investigator was still searching for pre-program data several years into the project. These four projects accounted for approximately $4.7 million, or roughly 30%, of the expenditures studied by the GAO.

Responses to GAO criticisms by the Office of Justice Programs

The GAO's summary comments were derived from an earlier series of reports and responses to those reports by the OJP, as well as OJP comments on the draft of the 2003 report (see, e.g., US General Accounting Office 2002). The exchanges between the two offices suggest that there was strong tension over the goals of these evaluation projects as well as differences over the value of the findings produced by the projects. The Assistant Attorney General's (AAG) office had the opportunity to comment on drafts of the various GAO reports, and these comments are included in the GAO reports as Appendices.[1] Commenting on the GAO's draft of the 2002 report, the AAG repeatedly made the argument that it was very difficult for the NIJ or funded evaluators to control the conditions in which those programs operate. This was especially true for Violence Against Women Office programs, which are authorized by the Violence Against Women Act to grant

broad flexibility to program recipients regarding how funds are to be used. They argued that, because of the extent of variation across programs, it would be nearly impossible for any research project to achieve the evaluation research standards used by the GAO.

In fact, the AAG went further and defended the value of these OJP research projects by arguing that the GAO was using standards for evaluation research that were too high and unrealistic. Their comments suggested that the GAO was using its own, unspecified standards for good research:

> In evaluating the NIJ's evaluations, the GAO applied what it said were 'generally accepted social science standards' (p. 6). However, the GAO did not specify the document that contains these standards or directly describe its elements of rigor (US General Accounting Office 2003, p. 34).

It seems odd that the management would challenge the GAO on this issue when, in fact, the methodological concerns they describe are fundamental to conducting good program evaluation research. Their defensive remarks included not only political disagreements over what should be considered useful research but a misunderstanding of what kind of research is necessary to demonstrate whether programs are having an effect on a desired outcome. The AAG's office asserts:

> Without question, randomized trials have their place, but so do comprehensive process evaluations, qualitative studies, and a host of other evaluation designs. We believe that it is possible to glean useful, *if not conclusive*, evidence of the impact of a program from an evaluation which does not rise to the standard recommended by the GAO because of the unavoidable absence of one or more elements (italics added) (US General Accounting Office 2003, p. 34).

The agency further argued that these projects produced useful information about the "likely impact" of the programs because of other data that resulted from the evaluations. For example, they stated that they were able to develop a training program that would help grantees establish greater collaboration between local service and criminal justice organizations, using information gathered from one of the studies deemed inadequately designed by the GAO. They also argued that the lack of baseline data does not reflect a flaw in the design or in the competence of the evaluators or agency but the reality of some of the phenomena for which these programs are designed (e.g., domestic victimization).

In my view this response sidesteps the issues raised by the GAO. It also provoked the GAO to respond by providing a list of well-known texts that discuss the fundamentals of evaluation research (e.g., Cook and Campbell 1990). The GAO reminded the agency that their task was to assess the methodological strengths and weaknesses of impact evaluations and that they "relied on NIJ officials to identify which of the program evaluations of Byrne and VAWO grant programs were, in fact, impact evaluation studies" (2002, p. 43). Agreeing with the OJP that such evaluations are difficult, they nonetheless reassert that there is an

important difference between assessing impact evaluation research and assessing the programs themselves. These exchanges clarify some of the impetus behind the NRC report. An important part of the dispute resides in the loose way that the term "impact evaluation research" was used and interpreted, at least initially, by the OJP.

Equally important to their critique, the GAO believed that some of the NIJ's operational practices were directly responsible for some of the problems they discovered. For example, in several of the studies deemed as having serious design weaknesses, they found that the peer review panels had expressed concerns about the projects. It was not clear to the GAO how the NIJ's evaluation process ensured that reviewers' concerns were adequately handled before such projects were funded. The GAO had also asserted in an earlier report that the general management and oversight of projects by the OJP was inadequate (e.g., missing or late final reports from investigators) and that the NIJ did not pay sufficient attention to ongoing research projects, including those outcome evaluations that were encountering implementation problems (US General Accounting Office 2001). They also argued that many of the initial design problems could be resolved if the agencies encouraged or required investigators to pay greater attention to these issues.

The seriousness of these operational criticisms undoubtedly contributed to the tone of the AAG's early comments. But within a short period of time, it became clear that there was a change in the nature of the OJP's responses to the GAO review. Agreeing that some of its procedures should be strengthened, an Evaluation Division was established within the NIJ Office of Research and Evaluation to establish standards for assessing the quality and usefulness of evaluations. The NIJ also developed training on cost-effectiveness and cost-benefit analyses and conducted "evaluability assessments" to provide information about the feasibility of an impact evaluation prior to proposal solicitation. Greater efforts have also been made to ensure that accurate and timely records have been provided by grantees, and the agency promised greater attention to the reviewing of applicants' prior performance before awarding grants than it had in the past (2003, pp. 25–26). These and other actions undertaken by the OJP and NIJ are consistent with the NRC's recommendations about how agencies charged with sponsoring research should plan, solicit, review, and monitor program evaluation projects.

Academic contributions for improving future research

Although the GAO assessments were focused primarily on the methodological limitations of the NIJ's impact evaluations and the agency's responsibilities for ensuring quality research, implicit in their assessment are concerns about the expertise available for some criminal justice evaluation research. Obviously, their assessment of the agency would have been restricted in scope if all researchers develop, submit, and conduct highly rigorous impact evaluations. It is the responsibility of academic researchers to insist on the highest possible standards when they develop and review research, and as they train the next generation of

researchers. But how else can academics help strengthen impact evaluation research?

Useful suggestions should consider that three of the four proposals that began with serious design limitations were focused on the evaluation of domestic violence and child maltreatment programs, while the fourth was an evaluation of a program designed to reduce stress among law enforcement officers. Of the proposals that were deemed well designed but encountered implementation problems, half were in the area of law enforcement. Although it cannot be assumed from these proportions that research quality varies according to area of evaluation, it does encourage one to look at the unique aspects of these areas of research in order to find possible solutions. My comments focus on the study of victimization-related programs (family violence and other), although they are not necessarily limited to these kinds of programs.

Criminologists and other academics are well aware of the fact that some of the most difficult areas in which to gather sufficient self-report data involve victims of crime (especially child victims and victims of family violence). Compared with research subjects who are more easily recruited (e.g., incarcerated offenders, youth enrolled in schools), victims perceive few incentives and high costs for participating as research subjects. Agencies that provide services to victims also have few incentives to participate. Relying often on shoe-string budgets, volunteers, and underpaid staff, victim services organizations can be easily burdened by data collection and the extensive record keeping it can require. Some domestic violence agencies are ideologically opposed to or highly suspicious of the top-down and intrusive nature of imposed evaluations (Riger et al. 2002; Bennett et al. 2004). And, of course, regardless of whether an agency has been in existence for decades or years, their future existence may be affected by the outcome of an evaluation, especially when government funding is low and must be competitively obtained.

When research participation is this burdensome for subjects and difficult for evaluators to obtain, it would be especially useful for researchers to have access to details about some of the more successful evaluations, to examine not only what the researchers did but how they were able to do it. In the area of victimization services, this is easier said than done. It is possible to find listings of the titles and abstracts of projects funded by various agencies, but, to my knowledge, none of these listings incorporates a peer-reviewed assessment of the methodology of the project using something analogous to the scoring system of the "Preventing Crime" report (Sherman et al. 1997). If such a database were available, researchers could study the most highly ranked evaluations and learn how those projects succeeded. They could also learn what does not work well by studying the difficulties of the lesser-ranked methodologies.

Of course, this idea is not new—it is exactly what the NIJ sponsored when they funded the "Preventing Crime" study, and it is a goal of the Campbell Collaboration's systematic review of government programs (Sherman 2003). The NRC report also encourages this approach and asks sponsoring agencies to accept the reality of criticism that is a productive part of the scientific process (p. 58).

A comprehensive database is not available for service delivery programs for victims of intimate partner or other family violence, or for other forms of victimization. The construction of such a database would be most helpful to researchers if it were developed to include all types of victimization and all types of victim services and programs. Although violence against women and domestic violence research developed independently from mainstream studies of violence and victimization, both areas can be enriched by borrowing from each other's insights and methodological strengths.

This kind of resource would be especially helpful, because violence against women and victimization research represent significant areas of expertise across more academic disciplines than other areas of criminological research (e.g., incarceration, policing). Victimization and its consequences have been studied by specialists in a host of disciplines, including clinical and community psychology, women's and children's psychology, sociology, criminal justice, medicine, economics, and legal studies. Each of these fields maintains a large number of journals in which such research may appear. A recent article discussing outlets for the publication of family violence research uncovered 22 separate journals (Moore et al. 2004), and none of these periodicals was in the areas of criminal justice, criminology, sociology, economics, or law. When publications are this voluminous, the usefulness of a comprehensive database is even greater.

A recommendation for a systematic review of existing research does not mean that there are not already excellent evaluations of victimization services and programs, or domestic violence programs, but simply that potentially relevant research may not be reaching all the interested audiences. It would be very difficult for any given team of researchers to remain well versed in the pertinent literature, given the nature of academic life today (e.g., increasing levels of specialization, publication pressures, and increasing numbers of journals and published works). Support for this systematic review should involve the multiple federal agencies that have sponsored victimization-related research (such as the National Institute of Justice, National Institutes of Health, Centers for Disease Control, National Institutes of Mental Health, and the National Institute for Aging.) Collaboration across agencies and multidisciplinary teams of researchers are more likely to result in a comprehensive systematic review than a team dominated by any one or two agencies or disciplines.

Another way that academics might help to improve the quality of impact evaluation research would be to renew their efforts to understand the limits of their data. For a variety of reasons, phenomena such as low response or participation rates often are glossed over or simply noted and justified as typical for the phenomenon under investigation. Non-response is increasing for all types of survey research, and there is an important literature that considers strategies for investigating and modeling these challenges (e.g., Groves et al. 2002). Because this is especially problematic in studies of violence against women and victimization, more work is needed to investigate how non-response might affect the conclusions drawn from program impact evaluations as well as other types of research.

Other methodological issues in need of better understanding include the validity and reliability of many of the measures used to study program outcomes. In some studies of post-victimization experiences it is common to find that persons who have reported victimization to the police might not report that experience to an interviewer when subsequently contacted. Kilpatrick et al. (1998), for instance, conducted a study of victim participation and satisfaction with the criminal justice system following a report of crime to the police or to a victim compensation agency. They found that nearly 30% of victims told the interviewers that they had not been victimized, despite their previous report. This is not a trivial percentage, but its importance is unknown. Other research has found that estimates of domestic violence are highly sensitive to even small amounts of sample attrition (Ybarra and Lohr 2002). Since the goal of most violence against women and victimization programs is to improve victims' safety and deliver needed services, it is necessary to determine whether commonly used pre- and post- measures are adequate for the purpose of assessing program effectiveness.

In addition, it is commonly believed that there are important group differences in the reliability and validity of measures (especially for race and ethnic groups), but how such differences should be incorporated into program evaluations has not been adequately studied. Similarly, persons who experience repeated victimization often find it difficult to recall offense details such as the number of incidents that occurred and the dates of those events. If victimization and outcome measures are unreliable for high rate victims, evaluation results can be inconclusive or biased, regardless of the rigor of the planned design.

Greater support for basic science is needed to resolve these kinds of challenges, yet the mandate for program impact evaluation assumes that answers to these kinds of questions are known. The early OJP response alluded to this when they argued that low response rates were typical challenges in victimization research (US General Accounting Office 2002, p. 40). Instead of pressing ahead with evaluations, however, the argument should be made that, without additional support for basic science, the usefulness of forced program evaluations is seriously limited. Without greater support from Congress in the form of appropriations, the likelihood that the OJP will be able to rigorously assess the impacts of their various programs is significantly compromised.

It is difficult to remain optimistic that these issues will be addressed, given that mandated evaluations and unfunded mandates are not new to criminal justice research. According to Sherman et al. (1997), this political problem goes back more than 30 years. The bottom line, they argued, is that satisfactory evaluations of grant programs will not result unless sufficient resources are dedicated to the task. Otherwise, program "evaluations" will consist of general descriptions of program implementation, anecdotal evidence from vested interests, and audits of the use of grant program funds.

Finally, the NRC report recommends that agencies carefully prioritize their programs to be evaluated: "Resources should mainly be directed toward programs for which there is (a) the greatest potential for practical and policy significance from the knowledge expected to result and (b) the circumstances are amenable to

research capable of producing the intended knowledge" (p. 61). Academics should also take part in this discussion, because it clarifies the political and social realities that frame the usefulness of scientific work. For instance, although not directly stated, the NRC might have asked whether there is a need to conduct rigorous impact evaluations of some violence against women or other victimization programs. How would such results be used? If several studies were to show that women's safety did not significantly improve after calls to a government-funded hotline, would it then follow that hotlines were an ineffective use of funds that should be discontinued? Or, if the benefits of counseling services to victims of crime are minimal, should we withhold funds dedicated to those services?

One hopes that such consequences are highly unlikely. Not all victim service programs may require rigorous impact evaluation as a condition of continued public support. Clear and systematic standards for impact evaluation are essential when and where such evaluations are warranted, and academic researchers should assist funding agencies and policymakers in identifying programs that do not yield measurable "benefits" subject to evaluation but perform other symbolic or political functions. But they are not always warranted. Some programs receive public support because they produce political benefits for public officials. Some are supported because they embody cherished values or offer diffused forms of assistance expressing social concern and kindness. Academics should not insist that every program be subject to an impact evaluation, because the decision to retain a program inevitably involves more than a scientific estimate of its effect.

Note

1 The Assistant Attorney General in the US Department of Justice is responsible for the operations of the OJP.

References

Bennett, L., Riger, S., Schewe, P., Howard, A., & Wasco, S. (2004). Effectiveness of hotline, advocacy, counseling, and shelter services for victims of domestic violence. *Journal of Interpersonal Violence* 19, 815 – 829.

Cook, T. D., & Campbell, D. T. (1990). *Quasi-experimentation: Design and analysis issues for field settings*, Boston: Houghton Mifflin

Groves, R., Dillman, D., Eltinge, J., & Little, R. (2002). *Survey nonresponse*. New York: Wiley

Kilpatrick, D., Beatty, D., & Howley, S. (1998). *The rights of crime victims — does legal protection make a difference?* Washington, DC: National Institute of Justice.

Moore, T., Rhatigan, D., Stuart, G., Street, A., & Farrell, L. (2004). Where to publish family violence research? *Violence and Victims* 19, 495 – 503.

Riger, S., Bennett, L., Wasco, S., Schewe, P., Frohman, L., & Camacho, J. (2002). *Evaluation of services for survivors of domestic violence and sexual assault*. Thousand Oaks, CA: Sage.

Sherman, L. (2003). Misleading evidence and evidence-led policy: Making social science more experimental. *The Annals of the American Academy of Political and Social Science* 589, 6–21.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress*. Washington, DC: National Institute of Justice.

US General Accounting Office (2001). *Justice discretionary grants: Byrne programs and violence against women office grant monitoring should be better documented*. Washington, DC: US Government Printing Office.

US General Accounting Office (2002). *Justice impact evaluations: One Byrne evaluation was rigorous: All reviewed violence against women office evaluations were problematic*. Washington, DC: US Government Printing Office.

US General Accounting Office (2003). *Justice outcome evaluations: Design and implementation of studies require more NIJ attention*. Washington, DC: US Government Printing Office.

Ybarra, L., & Lohr, S. (2002). Estimates of repeat victimization using the National Crime Victimization Survey. *Journal of Quantitative Criminology* 18, 1–21.

About the author

Janet L. Lauritsen is Professor of Criminology and Criminal Justice at the University of Missouri–St. Louis, USA. Her most recent publications examine how science interacts with social and political factors to influence the measurement of criminal victimization and how race and ethnicity are related to violence against women. Currently, she is Chairperson of the American Statistical Association's Committee on Law and Justice Statistics and Visiting Research Fellow at the Bureau of Justice Statistics.