

September 2003

JUSTICE OUTCOME EVALUATIONS

Design and Implementation of Studies Require More NIJ Attention



G A O

Accountability * Integrity * Reliability

GAO
Accountability • Integrity • Reliability

Highlights

Highlights of [GAO-03-1091](#), a report to The Honorable Lamar Smith, House of Representatives

Why GAO Did This Study

Policy makers need valid, reliable, and timely information on the outcomes of criminal justice programs to help them decide how to set criminal justice funding priorities. In view of previously reported problems with selected outcome evaluations managed by the National Institute of Justice (NIJ), GAO assessed the methodological quality of a sample of completed and ongoing NIJ outcome evaluation grants.

What GAO Recommends

GAO recommends that NIJ

- review its ongoing outcome evaluation grants and develop appropriate strategies and corrective measures to ensure that methodological design and implementation problems are overcome so the evaluations can produce more conclusive results;
- continue efforts to respond to GAO's 2002 recommendation that NIJ assess its evaluation process with the purpose of developing approaches to ensure that future outcome evaluations are funded only when they are effectively designed and implemented.

In commenting on a draft of this report, DOJ agreed with GAO's recommendations, and cited several current and planned activities intended to improve NIJ's evaluation program. DOJ also made two substantive comments related to the presentation of information that GAO responded to in the report.

www.gao.gov/cgi-bin/gettrpt?GAO-03-1091.

To view the full product, including the scope and methodology, click on the link above. For more information, contact Laurie E. Ekstrand (202) 512-8777 or ekstrandl@gao.gov.

JUSTICE OUTCOME EVALUATIONS

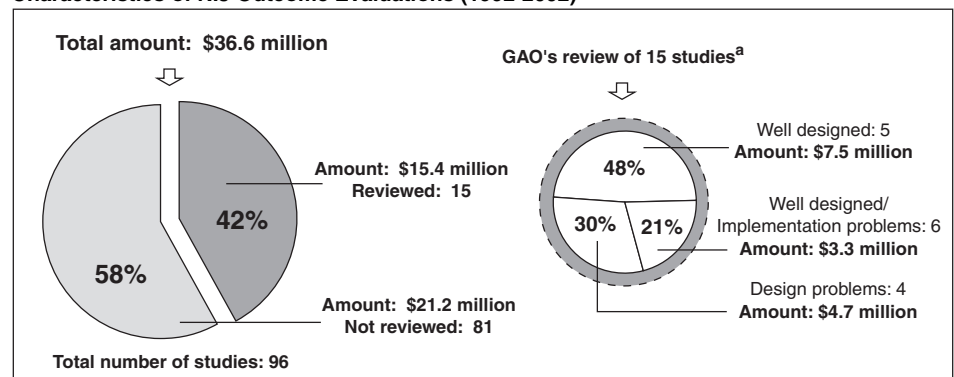
Design and Implementation of Studies Require More NIJ Attention

What GAO Found

From 1992 through 2002, NIJ managed 96 evaluation studies that sought to measure the outcomes of criminal justice programs. Spending on these evaluations totaled about \$37 million. Our methodological review of 15 of the 96 studies, totaling about \$15 million and covering a broad range of criminal justice issues, showed that sufficiently sound information about program effects could not be obtained from 10 of the 15. Five studies, totaling about \$7.5 million (or 48 percent of the funds spent on the studies we reviewed), appeared to be methodologically rigorous in both design and implementation, enabling meaningful conclusions to be drawn about program effects. Six studies, totaling about \$3.3 million (or 21 percent of the funds spent on the studies we reviewed), began with sound designs but encountered implementation problems that would render their results inconclusive. An additional 4 studies, totaling about \$4.7 million (or 30 percent of the funds spent on the studies we reviewed), had serious methodological limitations that from the start limited their ability to produce reliable and valid results. Although results from 5 completed studies were inconclusive, DOJ program administrators said that they found some of the process and implementation findings from them to be useful.

We recognize that optimal conditions for the scientific study of complex social programs almost never exist, making it difficult to design and execute outcome evaluations that produce definitive results. However, the methodological adequacy of NIJ studies can be improved, and NIJ has taken several steps—including the formation of an evaluation division and funding feasibility studies—in this direction. It is too soon to tell whether these changes will lead to evaluations that will better inform policy makers about the effectiveness of criminal justice programs.

Characteristics of NIJ Outcome Evaluations (1992-2002)



Source: GAO analysis of NIJ data.

^aPercentages may not add to 100 percent because of rounding.

Contents

Letter		1
	Results in Brief	3
	Background	5
	Overview of the Evaluations We Reviewed	8
	Most of the Reviewed NIJ Outcome Evaluations Could Not Produce Sufficiently Sound Information on Program Outcomes	9
	Completed Outcome Evaluations Produced Useful Information on Processes but Not on Outcomes for DOJ Program Administrators	23
	NIJ's Current and Planned Activities to Improve Its Evaluation Program	24
	Conclusions	26
	Recommendations for Executive Action	27
	Agency Comments and our Evaluation	28
Appendix I	Objectives, Scope, and Methodology	31
Appendix II	Summaries of the NIJ Outcome Evaluations Reviewed	37
Appendix III	Comments from the Department of Justice	53
Appendix IV	GAO Contacts and Staff Acknowledgments	56
	GAO Contacts	56
	Staff Acknowledgments	56
Tables		
	Table 1: NIJ Outcome Evaluations Reviewed by GAO	9
	Table 2: Characteristics of 5 NIJ Outcome Evaluations with Sufficiently Sound Designs and Implementation Plans	11
	Table 3: Problems Encountered during Implementation of 6 Well- Designed NIJ Outcome Evaluation Studies	15
	Table 4: Design Limitations in 4 NIJ Outcome Evaluation Studies	18
	Table 5: Number and Size of Outcome Evaluation Awards Made by NIJ from 1992 through 2002, and Reviewed by GAO	32

Table 6: Size and Completion Status of the 15 Evaluations Selected for Methodological Review	33
Table 7: Programs Evaluated and Funding Sources for Completed NIJ Outcome Evaluations	36

Abbreviations

BTC	Breaking the Cycle
COPS	Community Oriented Policing Services
DOJ	Department of Justice
GREAT	Gang Resistance Education and Training
NIJ	National Institute of Justice
OJP	Office of Justice Programs
OVW	Office on Violence Against Women

This is a work of the U.S. government and is not subject to copyright protection in the United States. It may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.



United States General Accounting Office
Washington, DC 20548

September 24, 2003

The Honorable Lamar Smith
House of Representatives

Dear Mr. Smith:

The U.S. Department of Justice (DOJ) spent almost \$4 billion in fiscal year 2002 on assistance to states and local communities to combat crime. These funds were used to reduce drug abuse and trafficking, address the problems of gang violence and juvenile delinquency, expand community policing, and meet the needs of crime victims, among other things. In addition, state and local governments spend billions of their dollars annually on law enforcement and criminal justice programs. Given these expenditures, it is important to know which programs are effective in controlling and preventing crime so that limited federal, state, and local funds not be wasted on programs that are ineffective. As the principal research, development, and evaluation agency of DOJ, the National Institute of Justice (NIJ) is responsible for evaluating existing programs and policies that respond to crime. It spends millions of dollars annually to support studies intended to evaluate various DOJ funded programs as well as selected local programs. To the extent that NIJ evaluations produce credible, valid, reliable, and timely information on the efficacy of these programs in combating crime, they can serve an important role in helping policymakers make decisions about how to set criminal justice funding priorities.

Pursuant to our previous reports in which we reported problems with selected NIJ-managed outcome evaluations,¹ in your former position as Chairman of the Subcommittee on Crime, House Judiciary Committee, you asked us to undertake a more extensive review of the outcome evaluation work performed under the direction of NIJ during the last 10 years. Outcome evaluations are defined as those efforts designed to determine whether a program, project, or intervention produced its intended effects.

¹U.S. General Accounting Office, *Justice Impact Evaluations: One Byrne Evaluation Was Rigorous; All Reviewed Violence Against Women Office Evaluations Were Problematic*, [GAO-02-309](#) (Washington, D.C.: Mar. 2002); and *Drug Courts: Better DOJ Data Collection and Evaluation Efforts Needed to Measure Impact of Drug Court Program*, [GAO-02-434](#) (Washington, D.C.: Apr. 2002).

As agreed with your office, we are reporting on the methodological quality of a sample of completed and ongoing NIJ outcome evaluation grants, and the usefulness of the evaluations in producing information on outcomes. Because we learned of changes NIJ has underway to improve its administration of outcome evaluation studies, we are also providing information in this report about these changes.

To meet our objectives, we reviewed outcome evaluation grants managed by NIJ from 1992 through 2002. During this time period NIJ managed 96 outcome evaluation grants. Of these 96 grants, we judgmentally selected and reviewed 15 outcome evaluations chosen so that they varied in grant size, completion status, and program focus. The selected studies accounted for about \$15.4 million, or about 42 percent, of the approximately \$36.6 million spent on outcome evaluation studies during the 10-year period. Although our sample is not representative of all NIJ outcome evaluations conducted during the last 10 years, it includes those that have received a large proportion of total funding for this type of research, and tends to be drawn from the most recent work. Our review assessed the methodological quality of these evaluations using generally accepted social science standards,² including such elements as whether evaluation data were collected before and after program implementation; how program effects were isolated (i.e., the use of nonprogram participant comparison groups or statistical controls); and the appropriateness of sampling, outcome measures, statistical analyses, and any reported results. We grouped the studies into 3 categories based on our judgment of their methodological soundness. Although we recognize that the stronger studies may have had some weaknesses, and that the weaker studies may have had some strengths, our categorization of the studies was a summary judgment based on the totality of the information provided to us by NIJ. We also interviewed NIJ officials regarding the selection and oversight of these evaluation studies. To assess the usefulness of NIJ's outcome evaluations in producing information about program outcomes, we reviewed the findings from all 5 of the completed NIJ outcome evaluations

²Social science research standards are outlined in Donald T. Campbell and Julian Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963); Thomas D Cook and Donald T. Campbell, *Quasi-experimentation: Design and Analysis Issues for Field Settings* (Boston: Houghton Mifflin, 1990); Carol H. Weiss, *Evaluation Research: Methods for Assessing Program Effectiveness* (Englewood Cliffs: Prentice-Hall, Inc., 1972); Edward Suchman, *Evaluation Research: Principles and Practice in Public Service and Social Action Programs* (New York: Russell Sage Foundation, 1967); and U.S. General Accounting Office, *Designing Evaluations*, [GAO/PEMD-10.1.4](#) (Washington, D.C.: May 1991).

in our sample that were funded in part by DOJ program offices, and interviewed program officials at NIJ and program administrators at DOJ's Office on Violence Against Women and Office of Community Oriented Policing Services. Further details on our methodology are provided in appendix I.

Results in Brief

Our methodological review of 15 selected NIJ outcome evaluation studies undertaken since 1992 showed that although most studies began with sufficiently sound designs, most could not produce sufficiently sound information on program outcomes. Specifically, the studies could be characterized in the following ways:

- *Studies that began with sufficiently sound evaluation designs:* Eleven of the 15 studies began with sufficiently sound designs. Some of these well-designed studies were also implemented well, while others were not. Specifically,
 - Five of the 11 studies were sufficiently well designed and implemented—including having appropriate comparison groups or random assignment to treatment and control groups, baseline measures, and follow-up data—so that meaningful conclusions could be drawn about program effects. Funding for these methodologically sound studies totaled about \$7.5 million, or nearly 50 percent of the approximately \$15.4 million spent on the studies we reviewed.
 - Six of the 11 studies began with sufficiently sound designs, but encountered implementation problems that limited the extent to which the study objectives could be achieved. For example, some evaluators were unable to carry out a proposed evaluation plan because the program to be evaluated was not implemented as planned, or they could not obtain complete or reliable data on outcomes. In some cases, implementation problems were beyond the evaluators' control, and resulted from decisions made by agencies providing program services after the study was underway. These studies were limited in their ability to conclude that it was the program or intervention that caused the intended outcome results. Funding for these studies with implementation problems totaled about \$3.3 million, or about 21 percent of the approximately \$15.4 million spent on the studies we reviewed.
- *Studies that did not begin with sufficiently sound designs.* Four of the 15 studies had serious methodological problems from the beginning that limited their ability to produce results that could be attributable to the

programs that were being evaluated. Methodological shortcomings in these studies included the absence of comparison groups or appropriate statistical controls, outcome measures with doubtful reliability and validity, and lack of baseline data. Funding for these studies that began with serious methodological problems totaled about \$4.7 million, or about 30 percent of the approximately \$15.4 million spent on the studies we reviewed.

Outcome evaluations are difficult to design and execute because optimal conditions for the scientific study of complex social programs almost never exist. Attributing results to a particular intervention can be difficult when such programs are evaluated in real world settings that pose numerous methodological challenges. All 5 of the completed NIJ outcome evaluations that focused on issues of interest to DOJ program offices had encountered some design and implementation problems. Nonetheless, DOJ program administrators told us that these evaluations produced information that prompted them to make a number of changes to DOJ-funded programs. The majority of the changes enumerated by DOJ program administrators occurred as a result of findings from the process or implementation components³ of the completed outcome evaluations, and not from findings regarding program results. For example, as a result of NIJ's evaluation of a DOJ program for domestic and child abuse victims in rural areas, DOJ developed a training program to assist grantees in creating collaborative programs based on the finding from the process evaluation that such information was not readily available.

Although outcome evaluations are difficult to design and execute, steps can be taken to improve their methodological adequacy and, in turn, the likelihood that they will produce meaningful information on program effects. NIJ officials told us that they have begun to take several steps to try to increase the likelihood that outcome evaluations will produce more definitive results, including the establishment of an Evaluation Division responsible for ensuring the quality and utility of NIJ evaluations, the funding of selected feasibility studies prior to soliciting outcome evaluations, and greater emphasis on applicants' prior performance in awarding evaluation grants.

³Outcome evaluations can be distinguished from process or implementation evaluations, which are designed to assess the extent to which a program is operating as intended.

We are making recommendations to the Attorney General to improve the quality of NIJ's outcome evaluations. We recommend that NIJ review the methodological adequacy of its ongoing grants and take action to improve, refocus, or limit them, as appropriate; and that NIJ develop approaches to ensure that future outcome evaluations are effectively designed and implemented. In commenting on a draft of this report, the DOJ's Office of Justice Programs' (OJP) Assistant Attorney General agreed with our recommendations. She also provided technical comments, which we evaluated and incorporated, as appropriate. The Assistant Attorney General made two substantive comments on our draft report—one relating to the fact that even rigorous study design and careful monitoring of program implementation do not ensure that evaluation results will be conclusive; the other relating to our purported focus on experimental and quasi-experimental methods to the exclusion of other high quality evaluation methods. We respond to these points in the Agency Comments and Evaluation section of the report.

Background

NIJ is the principal research development, and evaluation agency within OJP. It was created under the 1968 Omnibus Crime Control and Safe Streets Act,⁴ and is authorized to enter into grants, cooperative agreements, or contracts with public or private agencies to carry out evaluations of the effectiveness of criminal justice programs and identify promising new programs. NIJ's Office of Research and Evaluation oversees evaluations by outside researchers of a wide range of criminal justice programs, including ones addressing violence against women, drugs and crime, policing and law enforcement, sentencing, and corrections.

According to NIJ officials, the agency initiates a specific criminal justice program evaluation in one of three ways. First, congressional legislation may mandate evaluation of specific programs. For example, the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 2002,⁵ requires DOJ to conduct independent evaluations of selected programs funded by OJP's Bureau of Justice Assistance and selected projects funded by OJP's Office of Juvenile Justice and Delinquency Prevention. DOJ determined that NIJ would be

⁴42 U.S.C. 3721-3723. NIJ was formerly called the National Institute of Law Enforcement and Criminal Justice.

⁵P.L. 107-77. See H.R. Conf. Rep. No. 107-278, at 88, 108, and 112 (2001).

responsible for overseeing these evaluations. Second, NIJ may enter into an evaluation partnership with another OJP or DOJ office, or another federal agency, to evaluate specific programs or issues of interest to both organizations. In these cases, NIJ, in partnership with the program offices, develops a solicitation for proposals and oversees the resulting evaluation. Third, NIJ periodically solicits proposals for evaluation of criminal justice programs directly from the research community, through an open competition for grants. These solicitations ask evaluators to propose research of many kinds in any area of criminal justice, or in broad conceptual areas such as violence against women, policing research and evaluation, research and evaluation on corrections and sentencing, or building safer public housing communities through research partnerships.

According to NIJ officials, once the decision has been made to evaluate a particular program, or to conduct other research in a specific area of criminal justice, the process of awarding an evaluation grant involves the following steps. First, NIJ issues a solicitation and receives proposals from potential evaluators. Next, proposals are reviewed by an external peer review panel, as well as by NIJ professional staff. The external review panels are comprised of members of the research and practitioner communities,⁶ and reviewers are asked to identify, among other things, the strengths and weaknesses of the competing proposals. External peer review panels are to consider the quality and technical merit of the proposal; the likelihood that grant objectives will be met; the capabilities, demonstrated productivity, and experience of the evaluators; and budget constraints. Reviews are to include constructive comments about the proposal, useful recommendations for change and improvement, and recommendations as to whether the proposal merits further consideration by NIJ. NIJ professional staff are to review all proposals and all written external peer reviews, considering the same factors as the peer review panels. NIJ professional staff are also to consider the performance of potential grantees on any other previous research grants with NIJ. Next, the results of the peer and NIJ staff reviews are discussed in a meeting of NIJ managers, led by NIJ's Director of the Office of Research and Evaluation. Then, NIJ's Office of Research and Evaluation staff meet with the NIJ Director to present their recommendations. Finally, the NIJ Director makes the funding decision based on peer reviews, staff recommendations, other internal NIJ discussions that may have taken

⁶In 2002, the NIJ Director specified that there be an equal number of researchers and practitioners on the review panels.

place, and consideration of what proposals may have the greatest impact and contribute the most knowledge.

NIJ generally funds outcome evaluations through grants, rather than with contracts. NIJ officials told us that there are several reasons for awarding grants as opposed to contracts. Contracts can give NIJ greater control over the work of funded researchers, and hold them more accountable for results. However, NIJ officials said that NIJ most often uses grants for research and evaluation because they believe that grants better ensure the independence of the evaluators and the integrity of the study results. Under a grant, NIJ allows the principal investigator a great deal of freedom to propose the most appropriate methodology and carry out the data collection and analysis, without undue influence from NIJ or the agency funding the program. Grants also require fewer bureaucratic steps than do contracts, resulting in a process whereby a researcher can be selected in a shorter amount of time.

NIJ officials told us that NIJ tends to make use of contracts for smaller and more time-limited tasks—such as literature reviews or assessments of whether specific programs have sufficient data to allow for more extensive process or outcome evaluations—rather than for conducting outcome evaluations. NIJ also occasionally makes use of cooperative agreements, which entail a greater level of interaction between NIJ and the evaluators during the course of the evaluation. According to NIJ officials, cooperative agreements between NIJ and its evaluators tend to be slight variations of grants, with the addition of a few more specific requirements for grantees. NIJ officials told us that they might use a cooperative agreement when NIJ wants to play a significant role in the selection of an advisory panel, in setting specific milestones, or aiding in the design of specific data collection instruments.

NIJ is to monitor outcome evaluation grantees in accordance with policies and procedures outlined in the OJP Grant Management Policies and Procedures Manual. In general, this includes monitoring grantee progress through regular contact with grantees (site visits, cluster conferences, other meetings); required interim reports (semiannual progress and quarterly financial reports); and a review of final substantive evaluation reports. In some cases, NIJ will require specific milestone reports, especially on larger studies. Grant monitoring for all types of studies is carried out by approximately 20 full-time NIJ grant managers, each responsible for approximately 17 ongoing grants at any one time.

Overview of the Evaluations We Reviewed

From 1992 through 2002, NIJ awarded about \$36.6 million for 96 evaluations that NIJ identified as focusing on measuring the outcomes of programs, policies, and interventions, among other things.⁷ The 15 outcome evaluations that we selected for review varied in terms of completion status (8 were completed, 7 were ongoing) and the size of the award (ranging between about \$150,000 and about \$2.8 million), and covered a wide range of criminal justice programs and issues (see table 1). All evaluations were funded by NIJ through grants or cooperative agreements.⁸ Seven of the 15 evaluations focused on programs designed to reduce domestic violence and child maltreatment, 4 focused on programs addressing the behavior of law enforcement officers (including community policing), 2 focused on programs addressing drug abuse, and 2 focused on programs to deal with juvenile justice issues.

⁷A number of these grants included both process and outcome components.

⁸Three of the 15 evaluations were funded as cooperative agreements.

Table 1: NIJ Outcome Evaluations Reviewed by GAO

Grant	Award	Status	
		Completed	Ongoing
Domestic violence and child maltreatment			
National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program	\$719,949	X	
National Evaluation of the Domestic Violence Victims' Civil Legal Assistance Program	\$800,154		X
Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment	\$2,498,638		X
Evaluation of a Multi-Site Demonstration for Enhanced Judicial Oversight of Domestic Violence Cases	\$2,839,954		X
An Evaluation of Victim Advocacy with a Team Approach	\$153,491	X	
Culturally Focused Batterer Counseling for African-American Men	\$356,321		X
Testing the Impact of Court Monitoring and Batterer Intervention Programs at the Bronx Misdemeanor Domestic Violence Court	\$294,129		X
Law enforcement			
An Evaluation of Chicago's Citywide Community Policing Program	\$2,157,859	X	
Corrections and Law Enforcement Family Support: Law Enforcement Field Test	\$649,990		X
Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service	\$399,919	X	
Responding to the Problem Police Officer: An Evaluation of Early Warning Systems	\$174,643	X	
Drug abuse			
Evaluation of Breaking the Cycle	\$2,419,344	X	
Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing	\$187,412	X	
Juvenile justice issues			
National Evaluation of the Gang Resistance Education and Training Program	\$1,568,323	X	
Evaluation of a Juvenile Justice Mental Health Initiative with Randomized Design	\$200,000		X

Source: GAO analysis of NIJ data.

Most of the Reviewed NIJ Outcome Evaluations Could Not Produce Sufficiently Sound Information on Program Outcomes

Overall, we found that 10 of the 15 evaluations that we reviewed could not produce sufficiently sound information about program outcomes. Six evaluations began with sufficiently sound designs, but encountered implementation problems that would render their results inconclusive. An additional 4 studies had serious methodological problems that from the start limited their ability to produce reliable and valid results. Five studies appeared to be methodologically rigorous in both their design and implementation. (Appendix II provides additional information on the funding, objectives, and methodology of the 15 outcome evaluation studies.)

Most of the Reviewed Studies Were Well Designed, but Many Later Encountered Implementation Problems

Our review found that 5 evaluations had both sufficiently sound designs and implementation plans or procedures, thereby maximizing the likelihood that the study could meaningfully measure program effects. Funding for these methodologically sound studies totaled about \$7.5 million, or nearly 50 percent of the approximately \$15.4 million spent on the studies we reviewed. Six evaluations were well designed, but they encountered problems implementing the design as planned during the data collection phase of the study. Funding for these studies with implementation problems totaled about \$3.3 million, or about 21 percent of the approximately \$15.4 million spent on the studies we reviewed.

Five Evaluations Were Sufficiently Well Designed and Implemented

Five of the evaluations we reviewed were well designed and their implementation was sufficiently sound at the time of our review. Two of these evaluations had been completed and 3 were ongoing. All 5 evaluations met generally accepted social science standards for sound design, including measurement of key outcomes after a follow-up period to measure change over time, use of comparison groups or appropriate statistical controls to account for the influence of external factors on the results,⁹ random sampling of participants and/or sites or other purposeful sampling methods to ensure generalizable samples and procedures to ensure sufficient sample sizes, and appropriate data collection and analytic procedures to ensure the reliability and validity of measures (see table 2).

Studies Measured Change in Outcomes Over Time

All 5 evaluations measured, or included plans to measure, specified outcomes after a sufficient follow-up period. Some designs provided for collecting baseline data at or before program entry, and outcome data several months or years following completion of the program. Such designs allowed evaluators to compare outcome data against a baseline measurement to facilitate drawing conclusions about the program's effects, and to gauge whether the effects persisted or were transitory. For example, the National Evaluation of the Gang Resistance Education and Training Program examined the effectiveness of a 9-week, school-based education program that sought to prevent youth crime and violence by reducing student involvement in gangs. Students were surveyed regarding attitudes toward gangs, crime, and police, self-reported gang activity, and

⁹Statistically controlling for external factors that may be related to program outcomes and on which the treatment and comparison groups differ is usually not necessary when there is random assignment of participants to treatment and comparison conditions.

risk-seeking behaviors 2 weeks before the program began, and then again at yearly intervals for 4 years following the program's completion.

Table 2: Characteristics of 5 NIJ Outcome Evaluations with Sufficiently Sound Designs and Implementation Plans

Evaluation study	Sufficient follow-up	Use of comparison groups to control for external factors	Appropriate sampling procedures and reasonable sample sizes	Appropriate data collection and analysis procedures
National Evaluation of the Gang Resistance Education and Training Program	X	X	X	X
Evaluation of Breaking the Cycle	X	X	X	X
Evaluation of a Multi-Site Demonstration for Enhanced Judicial Oversight of Domestic Violence Cases	Planned	X	Planned	Planned
Culturally Focused Batterer Counseling for African-American Men	Planned	X	Planned	Planned
Testing the Impact of Court Monitoring and Batterer Intervention Programs at the Bronx Misdemeanor Domestic Violence Court ^a	Planned	X	Planned	Planned

Source: GAO analysis of NIJ data.

^aAlthough we have categorized this evaluation as having a sufficiently sound design and implementation plan, the grantee's proposal did not discuss how differential attrition from the four treatment groups would be handled if it occurred. Therefore, we do not know if the grantee has made sufficient plans to address this potential circumstance.

Measuring change in specific outcome variables at both baseline and after a follow-up period may not always be feasible. When the outcome of interest is "recidivism," such as whether drug-involved criminal defendants continue to commit criminal offenses after participating in a drug treatment program, the outcome can only be measured after the program is delivered. In this case, it is important that the follow-up period be long enough to enable the program's effects to be discerned. For example, the ongoing evaluation of the Culturally Focused Batterer Counseling for African-American Men seeks to test the relative effectiveness of counseling that recognizes and responds to cultural issues versus conventional batterer counseling in reducing batterer recidivism. All participants in the study had been referred by the court system to counseling after committing domestic violence violations. The evaluators planned to measure re-arrests and re-assaults 1 year after program intake, approximately 8 months after the end of counseling. The study cited prior research literature noting that two-thirds of first-time re-assaults were found to occur within 6 months of program intake, and over 80 percent of

first-time re-assaults over a 2-1/2 year period occur within 12 months of program intake.

Comparison Groups Were Used to Isolate Program Effects

All 5 evaluations used or planned to use comparison groups to isolate and minimize external factors that could influence the results of the study. Use of comparison groups is a practice employed by evaluators to help determine whether differences between baseline and follow-up results are due to the program under consideration or to other programs or external factors. In 3 of the 5 studies, research participants were randomly assigned to a group that received services from the program or to a comparison group that did not receive services. In constructing comparison groups, random assignment is an effective technique for minimizing differences between participants who receive the program and those who do not on variables that might affect the outcomes of the study. For example, in the previously mentioned ongoing evaluation of Culturally Focused Batterer Counseling for African-American Men participants who were referred to counseling by a domestic violence court are randomly assigned to one of three groups: (1) a culturally focused group composed of only African-Americans, (2) a conventional counseling group composed of only African-Americans, or (3) a mixed race conventional counseling group. The randomized design allows the investigators to determine the effect of the culturally focused counseling over and above the effect of participating in a same race group situation.

In the remaining two evaluation studies, a randomized design was not used and the comparison group was chosen to match the program group as closely as possible on a number of characteristics, in an attempt to ensure that the comparison and program groups would be similar in virtually all respects aside from the intervention. For example, the ongoing Evaluation of a Multi-Site Demonstration for Enhanced Judicial Oversight of Domestic Violence Cases seeks to examine the effects of a coordinated community response to domestic violence (including advocacy, provision of victim services, and enhanced judicial oversight) on victim safety and offender accountability. To ensure that the comparison and program groups were similar, comparison sites were selected based on having court caseload and population demographic characteristics similar to the demonstration sites. Only the program group is to receive the intervention; and neither comparison site has a specialized court docket; enhanced judicial oversight; or a county-wide, coordinated system for handling domestic violence cases.

Sufficiently Sound Sampling Procedures and Adequate Response Rates Helped Ensure Representativeness

All 5 evaluations employed or planned to employ sufficiently sound sampling procedures for selecting program and comparison participants. This was intended to ensure that study participants were representative of the population being examined so that conclusions about program effects could be generalized to that population. For example, in the previously mentioned Judicial Oversight Demonstration evaluation, offenders in program and comparison sites are being chosen from court records. In each site, equal numbers of eligible participants are being chosen consecutively over a 12-month period until a monthly quota is reached. Although this technique falls short of random sampling, the optimal method for ensuring comparability across groups, use of the 12-month sampling period takes into consideration and controls for possible seasonal variation in domestic violence cases.

The 5 evaluations also had adequate plans to achieve, or succeeded in achieving, reasonable response rates from participants in their samples. Failure to achieve adequate response rates threatens the validity of conclusions about program effects, as it is possible that selected individuals who do not respond or participate are substantially different on the outcome variable of interest from those who do respond or participate. The previously mentioned National Evaluation of the Gang Resistance Education and Training Program sought to survey students annually for up to 4 years after program participation ended. The grantee made considerable efforts in years 2, 3, and 4 to follow up with students who had moved from middle school to high school and were later enrolled in a large number of different schools; in some cases, in different school districts. The grantee achieved a completion rate on the student surveys of 76 percent after 2 years,¹⁰ 69 percent after 3 years, and 67 percent after 4 years. The grantee also presented analyses that statistically controlled for differential attrition among the treatment and comparison groups, and across sites, and showed that the program effects that were found persisted in these specialized analyses.

Careful Data Collection and Analysis Procedures Were Used or Planned

All 5 well-designed evaluations employed or had adequate plans to employ careful data collection and analysis procedures. These included procedures to ensure that the comparison group does not receive services or treatment received by the program group, response rates are

¹⁰The grantee notes that a 1990 analysis of 85 longitudinal studies reported an average questionnaire completion rate of 72 percent for 19 studies that had a 24-month follow-up period. This is slightly lower than the 76 percent response rate achieved after 2 years in the Gang Resistance Education and Training evaluation.

documented, and statistical analyses are used to adjust for the effects of selection bias or differential attrition on the measured results.¹¹ For example, the Breaking the Cycle evaluation examined the effectiveness of a comprehensive effort to reduce substance abuse and criminal activity among arrestees with a history of drug involvement. The program group consisted of felons who tested positive for drug use, reported drug use in the past, or were charged specifically with drug-related felonies. The comparison group consisted of persons arrested a year before the implementation of the Breaking the Cycle intervention who tested positive for at least one drug. Both groups agreed to participate in the study. Although groups selected at different times and using different criteria may differ in systematic ways, the evaluators made efforts to control for differences in the samples at baseline. Where selection bias was found, a correction factor was used in the analyses, and corrected results were presented in the report.

Six Studies Were Well-Designed but Encountered Problems During Implementation

Six of the 11 studies that were well-designed encountered problems in implementation during the data collection phase, and thus were unable to or are unlikely to produce definitive results about the outcomes of the programs being evaluated. Such problems included the use of program and comparison groups that differed on outcome-related characteristics at the beginning of the program or became different due to differential attrition, failure of the program sponsors to implement the program as originally planned, and low response rates among program participants (see table 3). Five of the studies had been completed and 1 was ongoing.

¹¹Selection bias refers to biases introduced by selecting different types of people into the program and comparison groups; differences in measured outcomes for each group may be a function of preexisting differences between the groups, rather than the intervention. Differential attrition refers to unequal loss of participants from the program and comparison groups during the course of a study, resulting in groups that are no longer comparable. Both may be a threat to the validity of conclusions.

Table 3: Problems Encountered during Implementation of 6 Well-Designed NIJ Outcome Evaluation Studies

Evaluation study	Program and comparison groups differed	Program not implemented as planned	Response rates were low
An Evaluation of Chicago's Citywide Community Policing Program	X	X	
Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing	X	X	
An Evaluation of Victim Advocacy with a Team Approach		X	X
Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service		X	X
Responding to the Problem Police Officer: An Evaluation of Early Warning Systems	X		
Evaluation of the Juvenile Justice Mental Health Initiative with Randomized Design		X	

Source: GAO analysis of NIJ data.

Differences between Program and Comparison Group Characteristics Make it Difficult to Attribute Outcomes to the Program

Three of the 6 studies used a comparison group that differed from the program group in terms of characteristics likely to be related to program outcomes—either due to preexisting differences or to differential attrition—even though the investigators may have made efforts to minimize the occurrence of these problems.¹² As a result, a finding that program and comparison group participants differed in outcomes could not be attributed solely to the program. For example, the Comprehensive Service-Based Intervention Strategy in Public Housing evaluation sought to reduce drug activity and promote family self-sufficiency among tenants of a public housing complex in one city through on-site comprehensive services and high profile police involvement. The intervention site was a housing project in one section of the city; the comparison site was another public housing complex on the opposite side of town, chosen for its similarities to the intervention site in terms of race, family composition, crime statistics, and the number of women who were welfare recipients. However, when baseline data from the two sites were examined, important preexisting differences between the two sites became apparent. These differences included a higher proportion of residents at the comparison site who were employed, which could have differentially

¹²Preexisting differences between the program and comparison groups can be viewed as a design problem. We treat this as an implementation problem in this section because the proposed design for these particular studies appeared to us to be reasonable at the time the funding decision was made. Problems with the comparability of the groups became apparent only after the studies were well underway, and often it was too late to control for the effects of such differences on program outcomes with statistical adjustments.

Program Results Not
Measurable Because Program
Not Implemented as Planned

affected intervention and comparison residents' propensity to utilize and benefit from available services. Additionally, since there was considerable attrition at both the intervention and comparison sites, it is possible that the intervention and comparison group respondents who remained differed on some factors related to the program outcomes. Although it may have been possible to statistically control for these differences when analyzing program outcomes, the evaluator did not do so in the analyses presented in the final report.

In 5 of the 6 studies, evaluators ran into methodological problems because the program under evaluation was not implemented as planned, and the investigators could not test the hypotheses that they had outlined in their grant proposals. For the most part, this particular implementation problem was beyond the evaluators' control. It resulted from decisions made by agencies providing program services that had agreed to cooperate with the evaluators but, for a number of reasons, made changes in the programs or did not cooperate as fully as expected after the studies were underway. This occurred in the evaluation of the Juvenile Justice Mental Health Initiative with Randomized Design, a study that is ongoing and expected to be completed in September 2003. The investigators had proposed to test whether two interventions provided within an interagency collaborative setting were effective in treating youths with serious emotional disturbances referred to the juvenile justice system for delinquency. Juveniles were to be randomly assigned to one of two treatment programs, depending on age and offense history (one for youth under the age of 14 without serious, violent, or chronic offense history, and one for youth ages 14 and older with serious, violent, or chronic delinquencies) or to a comparison group that received preexisting court affiliated service programs. The evaluators themselves had no power to develop or modify programs. The funding agencies¹³ contracted with a local parent support agency and with a nonprofit community-based agency to implement the programs, but the program for youth under the age of 14 was never implemented.¹⁴ In addition, partway through the study, the funding agencies decided to terminate random assignment of juveniles, and shortly thereafter ended the program. As a result, the evaluators had complete data on 45 juveniles who had been in the treatment program, rather than

¹³The treatment programs were to be developed under the funding and oversight of the St. Louis Mental Health Board and the Missouri Department of Mental Health.

¹⁴As a result, juveniles under 14 were randomly assigned to either the program for juveniles 14 and over, or to the comparison group.

Low Response Rates May Reduce the Reliability and Validity of Findings

on the 100 juveniles they had proposed to study. Although the study continued to collect data on juveniles eligible for the study (who were then assigned to the comparison group, since a treatment option was no longer available), the evaluators proposed to analyze the data from the random experiment separately, examining only those treatment and comparison youths assigned when program slots were available. Because of the smaller number of participants than anticipated, detailed analyses of certain variables (such as the type, or amount of service received, or the effects of race and gender) are likely to be unreliable.

Low response rates were a problem in 2 of the 6 studies, potentially reducing the reliability and validity of the findings. In a third study, response rates were not reported, making it impossible for us to determine whether this was a problem or not.¹⁵ In one study where the response rate was a problem, the evaluators attempted to survey victims of domestic abuse, a population that NIJ officials acknowledged was difficult to reach. In *An Evaluation of Victim Advocacy With a Team Approach*, the evaluators attempted to contact by telephone women who were victims of domestic violence, to inquire about victims' experiences with subsequent violence and their perceptions of safety. Response rates were only about 23 percent, and the victims who were interviewed differed from those who were not interviewed in terms of the nature and seriousness of the abuse to which they had been subjected. NIJ's program manager told us that when she became aware of low response rates on the telephone survey, she and the principal investigator discussed a variety of strategies to increase response rates. She said the grantee expended additional time and effort to increase the response rate, but had limited success. In the other study with low response rates—*Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service*—investigators attempted to survey police officers in one city regarding their attitudes about the city's new non-emergency phone system. Only 20 percent of the police officers completed the survey.

¹⁵The Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing reported response rates for both the intervention and comparison sites on a survey at baseline, but did not report response rates for follow-up surveys conducted 12 and 18 months after the intervention began.

Some Evaluation Studies Had Serious Design Limitations from the Beginning

Four of the evaluation studies began with serious design problems that diminished their ability to produce reliable or valid findings about program outcomes. One of the studies was completed, and 3 were ongoing. The studies’ design problems included the lack of comparison groups, failure to measure the intended outcomes of the program, and failure to collect preprogram data as a baseline for the outcomes of interest (see table 4). Funding for these studies that began with serious methodological problems totaled about \$4.7 million, or about 30 percent of the approximately \$15.4 million spent on the studies we reviewed.

Table 4: Design Limitations in 4 NIJ Outcome Evaluation Studies			
Evaluation study	No comparison group	Intended outcomes not measured	Limited pre-program data
National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program	X	X	X
National Evaluation of the Domestic Violence Victims’ Civil Legal Assistance Program	X		X
Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment	X	X	
Corrections and Law Enforcement Family Support: Law Enforcement Field Test	X		

Source: GAO analysis of NIJ data.

Lack of Comparison Groups

None of the 4 outcome evaluation studies had a comparison group built into the design—a factor that hindered the evaluator’s ability to isolate and minimize external factors that could influence the results of the study. The completed National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program did not make use of comparison groups to study the effectiveness of the federal grant program that supports projects designed to prevent and respond to domestic violence, dating violence, and child victimization in rural communities. Instead, evaluators collected case study data from multiday site visits to 9 selected sites.

The other three funded grant proposals submitted to NIJ indicated that they anticipated difficulty in locating and forming appropriate comparison groups. However, they proposed to explore the feasibility of using comparison groups in the design phase following funding of the grant. At

the time of our review, when each of these studies was well into implementation, none was found to be using a comparison group. For example, the Evaluation of a Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment proposed to examine whether steps taken to improve collaboration between dependency courts, child protective services, and domestic violence service providers in addressing the problems faced by families with co-occurring instances of domestic violence and child maltreatment resulted in improvements in how service providers dealt with domestic violence and child maltreatment cases. Although NIJ stated that the evaluators planned to collect individual case record data from similar communities, at the time of our review these sites had not yet been identified, nor had a methodology for identifying the sites been proposed. Our review was conducted during the evaluation's third year of funding.

Intended Outcomes of Program Were Not Measured

Although they were funded as outcome evaluations, 2 of the 4 studies were not designed to provide information on intended outcomes for individuals served by the programs. Both the Rural Domestic Violence and the Multi-Site Demonstration of Collaborations programs had as their objectives the enhanced safety of victims, among other goals. However, neither of the evaluations of these programs collected data on individual women victims and their families in order to examine whether the programs achieved this objective. Most of the data collected in the Rural Domestic Violence evaluation were indicators of intermediary results, such as increases in the knowledge and training of various rural service providers. While such intermediary results may be necessary precursors to achieving the program's objectives of victim safety, they are not themselves indicators of victim safety. The Multi-Site Demonstration of Collaborations evaluation originally proposed to collect data on the safety of women and children as well as perpetrator recidivism, but in the second year of the evaluation project, the evaluators filed a request to change the scope of the study. Specifically, they noted that the original outcome indicators proposed for victim safety were not appropriate given the time frame of the evaluation compared to the progress of the demonstration project itself. The modified scope, which was approved by NIJ, focused on system rather than individual level outcomes. The new 'effectiveness' indicators included such things as changes in policies and procedures of agencies participating in the collaboration, and how agency personnel identify, process, and manage families with co-occurring domestic violence and child maltreatment. Such a design precludes conclusions about whether the programs improved the lives of victims of domestic violence or their children.

Lack of Pre-Program Data Hinders Ability to Show That Program Produced Change

As discussed in our March 2002 report, the Rural Domestic Violence evaluation team did not collect baseline data prior to the start of the program, making it difficult to identify change resulting from the program. In addition, at the time of our review, in the third year of the multi-year National Evaluation of the Domestic Violence Victims' Civil Legal Assistance Program evaluation, the evaluator did not know whether baseline data would be available to examine changes resulting from the program. This evaluation, of the federal Civil Legal Assistance program,¹⁶ proposed to measure whether there had been a decrease in pro se representation (or self-representation) in domestic violence protective order cases. A decrease in pro se representation would indicate successful assistance to clients by Civil Legal Assistance grantees. In May 2003, NIJ reported that the evaluator was still in the process of contacting the court systems at the study sites to see which ones had available data on pro se cases. The evaluator also proposed to ask a sample of domestic violence victims whether they had access to civil legal assistance services prior to the program, the outcomes of their cases, and satisfaction with services. Respondents were to be selected from a list of domestic violence clients served by Civil Legal Assistance grantees within a specified time period, possibly 3 to 9 months prior to the start of the outcome portion of the study. Such retrospective data on experiences that may have occurred more than 9 months ago must be interpreted with caution, given the possibility of recall errors or respondents' lack of knowledge about services that were available in the past.

NIJ Has Funded Outcome Evaluations Despite Major Gaps in Knowledge about the Availability of Data and Comparison Groups

Outcome evaluations are inherently difficult to conduct because in real-world settings program results can be affected by factors other than the intervention being studied. In addition, grantees' ability to conduct such evaluations can depend on the extent to which information is available up front about what data are available to answer the research questions, where such data can be obtained, and how the data can be collected for both the intervention and comparison groups. We found that in 3 of the 15 NIJ evaluations we reviewed, NIJ lacked sufficient information about these issues to assure itself that the proposals it funded were feasible to carry out. These 3 studies totaled about \$3.7 million.

¹⁶Civil Legal Assistance provides grants to nonprofit, nongovernmental organizations that provide legal services to victims of domestic violence or that work with victims of domestic violence who have civil legal needs.

For the Evaluation of Non-Emergency Calls to 911, NIJ and DOJ's Office of Community Oriented Policing Services jointly solicited grant proposals to evaluate strategies taken by 4 cities to decrease non-emergency calls to the emergency 911 system. NIJ officials told us that they had conducted 3-day site visits of the 4 sites, and that discussions with local officials included questions about availability of data in each jurisdiction. The NIJ solicitation for proposals contained descriptions of how non-emergency calls were processed at all 4 sites, but no information on the availability of outcome data to assess changes in the volume, type, and nature of emergency and non-emergency calls before and after the advent of the non-emergency systems. Evaluators were asked to conduct both a process analysis and an assessment analysis. The assessment analysis was to include "compiling and/or developing data" on a number of outcome questions. Once the study was funded, however, the grantee learned that only 1 of the 4 cities had both a system designed specifically to reduce non-emergency calls to 911, as well as reliable data for evaluation purposes.

In the case of the Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment, NIJ funded the proposal without knowing whether the grantee would be able to form comparison groups. NIJ officials stated that one of the reasons for uncertainty about the study design was that at the time the evaluator was selected, the 6 demonstration sites had not yet been selected. The proposal stated that the grantee would explore the "potential for incorporating comparison communities or comparison groups at the site level, and assess the feasibility, costs, and contributions and limitations of a design that incorporates comparison groups or communities." NIJ continued to fund the grantee for 3 additional years, although the second year proposal for supplemental funding made no mention of comparison groups and the third year proposal stated that the grantee would search for comparison sites, but did not describe how such sites would be located. In response to our questions about whether comparison groups would be used in the study, NIJ officials said that the plan was for the grantee to compare a random sample of case records from before program implementation to those after implementation at each of the demonstration sites. Designs utilizing pre-post treatment comparisons within the same group are not considered to be as rigorous as pre-post-treatment comparison group designs because they do not allow evaluators to determine whether the results are due to the program under consideration or to some other programs or external factors.

NIJ also approved the Multi-Site Demonstration of Collaborations proposal without knowing whether data on individual victims of domestic violence and child maltreatment would be available during the time frame of the evaluation. The first year proposal stated that the grantee would examine outcomes for individuals and families, although it also noted that there are challenges to assessing such outcomes and that system outcomes should be examined first. Our review found that in the third year of the evaluation, data collection was focused solely on “system” outcomes, such as changes in policies and procedures and how agency personnel identify, process, and manage families with co-occurring domestic violence and child maltreatment. Thus, although the original design called for answering questions about the outcomes of the program for individuals and families, NIJ could not expect answers to such questions.¹⁷

In the case of the Civil Legal Assistance study, NIJ officials told us that they have held discussions with the grantee about the feasibility of adding comparison groups to the design. According to these officials, the grantee said that a comparison group design would force it to reduce the process sites to be studied from 20 to somewhere between 6 and 8. NIJ advised the grantee that so large a reduction in sites would be too high a price to pay to obtain comparison groups, and advised the grantee to stay with the design as originally proposed. Consequently, NIJ cannot expect a rigorous assessment of outcomes from this evaluation.

¹⁷ NIJ officials told us in August 2003 that the evaluation had been funded for a fourth year, and that the federal agencies funding this evaluation (DOJ and the Department of Health and Human Services) were also considering a fifth year of funding. Four years of funding allows the evaluation to collect data covering about the first 3 years of implementation in the sites. However, data collected from stakeholders at the sites early in the evaluation showed that the sites expected that it would take 3.5 to 4 years to achieve change in key individual level outcomes. At the time of our review, there was no information on whether individual level outcome data would be collected.

Completed Outcome Evaluations Produced Useful Information on Processes but Not on Outcomes for DOJ Program Administrators

Of the 5 completed NIJ studies that focused on issues of interest to DOJ program offices, findings related to program effectiveness were not sufficiently reliable or conclusive. However, DOJ program administrators told us that they found some of the process and implementation findings from the completed studies to be useful.¹⁸

Program administrators from DOJ's Office on Violence Against Women said that although they did not obtain useful outcome results from the Rural Domestic Violence evaluation, they identified two "lessons learned" from the process and implementation components of the study. First, the evaluation found that very little information was available to grantees regarding how to create collaborative programs. Thus, DOJ engaged a technical assistance organization to develop a training program on how to create collaborative projects based on the experiences of some of the grantees examined by the Rural evaluation. Second, program administrators told us that the evaluation found that because Rural grants were funded on an 18-month schedule, programs did not have adequate time to structure program services and also collect useful program information. As a result, Rural programs are now funded for at least 24 months.¹⁹

While shortcomings in NIJ's outcome evaluations of law enforcement programs leave questions about whether the programs are effective and whether they should continue to be funded, program administrators in DOJ's Office of Community Oriented Policing Services said that the studies helped identify implementation problems that assisted them in developing and disseminating information in ways useful to the law enforcement community. These included curriculum development, leadership conferences, and fact sheets and other research publications. For example, as a result of the NIJ-managed study, Responding to the

¹⁸Because of our interest in the effectiveness of criminal justice programs, we limited our review of the usefulness of NIJ outcome evaluations to evaluations of DOJ programs, or evaluations funded by DOJ—a total of 5 evaluations. We did not examine 3 other completed NIJ outcome evaluations focusing on programs funded by agencies other than DOJ.

¹⁹Officials with DOJ's Office on Violence Against Women were not familiar with the findings from the other completed NIJ study focusing on violence against women, the Victim Advocacy with a Team Approach evaluation. This evaluation was funded by a transfer of funds to NIJ for NIJ research and evaluations in the area of violence against women. NIJ officials stated that Office on Violence Against Women officials were consulted in the development of the solicitation.

Problem Police Officer: An Evaluation of Early Warning Systems,²⁰ DOJ officials developed a draft command level guidebook that focuses on the factors to be considered in developing an early warning system, developed an early warning intervention training curriculum that is being taught by the 31 Regional Community Policing Institutes²¹ located across the country, and convened a “state-of-art” conference for five top law enforcement agencies that were developing early warning systems. DOJ officials also said the studies showed that the various systems evaluated had been well received by citizens and law enforcement officials. For example, they said that citizens like the 311 non-emergency number that was established in several cities to serve as an alternative to calling the 911 emergency number. The system allows law enforcement officers to identify hot spots or trouble areas in the city by looking at various patterns in the citizen call data. Officials may also be able to monitor the overall state of affairs in the city, such as the presence of potholes, for example. Similarly, Chicago’s City-Wide Community Policing program resulted in the development of a crime mapping system, enabling officers to track crime in particular areas of the city. Like the non-emergency telephone systems, DOJ officials believe that crime mapping helps inform citizens, police, and policy makers about potential problem areas.

NIJ’s Current and Planned Activities to Improve Its Evaluation Program

NIJ officials told us that they have begun to take several steps to try to increase the likelihood that outcome evaluations will produce more definitive results. We recommended in our March 2002 report on selected NIJ-managed outcome evaluations²² that NIJ assess its evaluation process to help ensure that future outcome evaluations produce definitive results. In November 2002, Congress amended the relevant statute to include cost-effectiveness evaluation where practical as part of NIJ’s charge to conduct evaluations.²³ Since that time NIJ has established an Evaluation Division

²⁰ An early warning system is a data based police management tool designed to identify officers whose behavior is problematic, as indicated by high rates of citizen complaints, use of force incidents, or other evidence of behavior problems, and to provide some form of intervention, such as counseling or training to correct that performance. The NIJ-managed study consisted of a process and outcome evaluation of early warning systems in 3 large urban police departments, as well as a national survey.

²¹ Through the Regional Community Policing Institute network, DOJ’s Office of Community Oriented Policing Services assists local law enforcement agencies with meeting their community policing training needs.

²² [GAO-02-309](#).

²³ Homeland Security Act of 2002, P.L. 107-296 sec. 237.

within NIJ's Office of Research and Evaluation. NIJ officials told us that they have also placed greater emphasis on funding cost-benefit studies, funded feasibility studies prior to soliciting outcome evaluations, and placed greater emphasis on applicants' prior performance in awarding grants.

In January 2003, NIJ established an Evaluation Division within NIJ's Office of Research and Evaluation, as part of a broader reorganization of NIJ programs. According to NIJ, the Division will "oversee NIJ's evaluations of other agency's [sic] programs and...develop policies and procedures that establish standards for assuring quality and utility of evaluations."²⁴ NIJ officials told us that among other things, the Division will be responsible for recommending to the NIJ Director which evaluations should be undertaken, assigning NIJ staff to evaluation grants and overseeing their work, and maintaining oversight responsibility for ongoing evaluation grants. In addition, NIJ officials told us that one of the NIJ Director's priorities is to put greater emphasis on evaluations that examine the costs and benefits of programs or interventions. To support this priority, NIJ officials told us that the Evaluation Division had recently developed training for NIJ staff on cost-benefit and cost-effectiveness analysis.²⁵

NIJ recently undertook 37 "evaluability assessments" to assess the feasibility of conducting outcome evaluations of congressionally earmarked programs prior to soliciting proposals for evaluation.²⁶ In 2002 and 2003, these assessments were conducted to examine each project's

²⁴NIJ Web site (<http://www.ojp.usdoj.gov/nij/about.htm>).

²⁵These analyses compare a program's outputs or outcomes with the costs (resources expended) to produce them. Cost-effectiveness analysis assesses the costs of meeting a single goal or objective, and can be used to identify the least costly alternative to meet that goal. Cost-benefit analysis aims to identify all the relevant costs and benefits, usually expressed in dollar terms.

²⁶Earmarked refers to dedicating an appropriation for a particular purpose. Legislative language may designate any portion of a lump-sum amount for particular purposes. In fiscal year 2002, congressional guidance for the use of these funds was provided in conference report H.R. 107-278. The report specified that up to 10 percent of the funds for the Bureau of Justice Assistance's Edward Byrne Discretionary Grant Program be made available for an independent evaluation of the program (at 88); and up to 10 percent of the funds for the Office of Juvenile Justice and Delinquency Prevention's Discretionary Grants for National Programs and Special Emphasis Programs (at 108) and Safe Schools Initiative be made available for an independent evaluation of the program (at 112).

scope, activities, and potential for rigorous evaluation.²⁷ The effort included telephone interviews and site visits to gather information regarding such things as what outcomes could be measured, what kinds of data were being collected by program staff, and the probability of using a comparison group or random assignment in the evaluation. Based on the review, NIJ solicited proposals from the research community to evaluate a subset of the earmarked programs that NIJ believed were ready for outcome evaluation.²⁸

NIJ officials also stated that in an effort to improve the performance of its grantees, it has begun to pay greater attention to the quality and timeliness of their performance on previous NIJ grants when reviewing funding proposals. As part of NIJ's internal review of grant applications, NIJ staff check that applicants' reports are complete and accurate and evaluate past work conducted by the applicant using performance related measures. Although this is not a new activity, NIJ officials told us that NIJ was now placing more emphasis on reviewing applicants' prior performance than it had in the past.²⁹ NIJ officials told us that NIJ staff may also contact staff in other OJP offices, where the applicant may have received grant funding, to assess applicant performance on those grants.

Conclusions

Our in-depth review of 15 outcome evaluations managed by NIJ during the past 10 years indicated that the majority was beset with methodological and/or implementation problems that limited the ability to draw meaningful conclusions about the programs' effectiveness. Although our sample is not representative of all NIJ outcome evaluations conducted during the last 10 years, it includes those that have received a large proportion of the total funding for this type of research, and tends to be drawn from the most recent work. The findings from this review, coupled with similar findings we reported in other reviews of NIJ outcome

²⁷ Prior to conducting the evaluability assessments, NIJ conducted an initial review of the earmarked programs, and eliminated from consideration those programs that were appearing in legislation for the first time, in order to focus on those programs that were receiving continuation funding.

²⁸ The solicitation deadlines were April 11, 2003, for the Bureau of Justice Assistance programs and July 15, 2003, for the Office of Juvenile Justice and Delinquency Prevention programs.

²⁹ A new requirement of the solicitation for proposals is that applicants report what prior funding they have received from NIJ.

evaluations, raise concerns about the level of attention NIJ is focusing on ensuring that funded outcome evaluations produce credible results.

We recognize that it is very difficult to design and execute outcome evaluations that produce meaningful and definitive results. Real world evaluations of complex social programs inevitably pose methodological challenges that can be difficult to control and overcome. Nonetheless, we believe it is possible to conduct outcome evaluations in real world settings that produce meaningful results. Indeed, 5 of NIJ's outcome evaluations can be characterized in this way, and these 5 accounted for about 48 percent of the \$15.4 million spent on the studies we reviewed. We also believe that NIJ could do more to help ensure that the millions of dollars it spends annually to evaluate criminal justice programs is money well spent. Indeed, poor evaluations can have substantial costs if they result in continued funding for ineffective programs or the curtailing of funding for effective programs.

NIJ officials told us that they recognize the need to improve their evaluation efforts and have begun to take several steps in an effort to increase the likelihood that outcome evaluations will produce more conclusive results. These steps include determining whether a program is ready for evaluation and monitoring evaluators' work more closely. We support NIJ's efforts to improve the rigor of its evaluations. However, it is too soon to tell whether and to what extent these efforts will lead to NIJ funding more rigorous effectiveness evaluations, and result in NIJ obtaining evaluative information that can better assist policy makers in making decisions about criminal justice funding priorities. In addition to the steps that NIJ is taking, we believe that NIJ can benefit from reviewing problematic studies it has already funded in order to determine the underlying causes for the problems and determine ways to avoid them in the future.

Recommendations for Executive Action

- We recommend that the Attorney General instruct the Director of NIJ to:
- Conduct a review of its ongoing outcome evaluation grants—including those discussed in this report—and develop appropriate strategies and corrective measures to ensure that methodological design and implementation problems are overcome so the evaluations can produce more conclusive results. Such a review should consider the design and implementation issues we identified in our assessment in order to decide whether and what type of intervention may be appropriate. If, based on NIJ's review, it appears that the methodological problems cannot be

overcome, NIJ should consider refocusing the studies' objectives and/or limiting funding.

- Continue efforts to respond to our March 2002 recommendation that NIJ assess its evaluation process with the purpose of developing approaches to ensure that future outcome evaluation studies are funded only when they are effectively designed and implemented. The assessment could consider the feasibility of such steps as:
 - obtain more information about the availability of outcome data prior to developing a solicitation for research;
 - require that outcome evaluation proposals contain more detailed design specifications before funding decisions are made regarding these proposals; and
 - more carefully calibrate NIJ monitoring procedures to the cost of the grant, the risks inherent in the proposed methodology, and the extent of knowledge in the area under investigation.

Agency Comments and our Evaluation

We provided a copy of a draft of this report to the Attorney General for review and comment. In a September 4, 2003, letter, DOJ's Assistant Attorney General for the Office of Justice Programs commented on the draft. Her comments are summarized below and presented in their entirety in appendix III.

The Assistant Attorney General stated that NIJ agreed with our recommendations. She also highlighted NIJ's current and planned activities to improve its evaluation program. For example, as we note in the report, NIJ has established an Evaluation Division and initiated a new strategy of evaluability assessments. Evaluability assessments are intended to be quick, low cost initial assessments of criminal or juvenile justice programs to help NIJ determine if the necessary conditions exist to warrant sponsoring a full-scale outcome evaluation. To improve its grantmaking process, the Assistant Attorney General stated that NIJ is developing a new grant "special conditions" that will require grantees to document all changes in the scope and components of evaluation designs. In response to our concerns, NIJ also plans, in fiscal year 2004, to review its grant monitoring procedures for evaluation grants in order to more intensively monitor the larger or more complex grants. NIJ also plans to conduct periodic reviews of its evaluation research portfolio to assess the progress of ongoing grants. This procedure is to include documenting any changes in evaluation design that may have occurred and reassessing the expected benefits of ongoing projects.

In her letter, the Assistant Attorney General made two substantive comments—both concerning our underlying assumptions in conducting the review—with which we disagree. In her first comment, the Assistant Attorney General noted that our report implies that conclusive evaluation results can always be achieved if studies are rigorously designed and carefully monitored. We disagree with this characterization of the implication of our report. While sound research design and careful monitoring of program implementation are factors that can significantly affect the extent to which outcome evaluation results are conclusive, they are not the only factors. We believe that difficulties associated with conducting outcome evaluations in real world settings can give rise to situations in which programs are not implemented as planned or requisite data turn out not to be available. In such instances, even a well-designed and carefully monitored evaluation will not produce conclusive findings about program effectiveness. Our view is that when such problems occur, NIJ should respond and take appropriate action. NIJ could (1) take steps to improve the methodological adequacy of the studies if it is feasible to do so, (2) reconsider the purpose and scope of evaluation if there is interest in aspects of the program other than its effectiveness, or (3) decide to end the evaluation project if it is not likely to produce useful information on program outcomes.

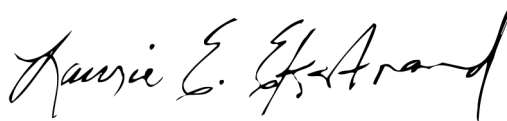
In her second comment, the Assistant Attorney General expressed the view that our work excluded consideration of valid, high quality evaluation methods other than experimental and quasi-experimental design. We believe that our assessment of NIJ's outcome evaluations was both appropriate and comprehensive. We examined a variety of methodological attributes of NIJ's studies in trying to assess whether they would produce sufficiently sound information on program outcomes. Among other things, we systematically examined such factors as the type of evaluation design used; how program effects were isolated (that is, whether comparison groups or statistical controls were utilized); the size of study samples and appropriateness of sampling procedures; the reliability, validity, and appropriateness of outcome measures; the length of follow-up periods on program participants; the extent to which program attrition or program participant nonresponse may have been an issue; the appropriateness of analytic techniques that were employed; and the reported results. Therefore, we made determinations about the cause and effect linkages between programs and outcomes using a myriad of methodological information. In discussing the methodological strengths of experimental and quasi-experimental designs, we did not intend to be dismissive of other potential approaches to isolating the effects of program interventions. For example, if statistical controls can be employed to

adequately compensate for a methodological weakness such as the existence of a comparison group that is not comparable on characteristics that could affect the study's outcome, then we endorse the use of such a technique. However, in those instances where our review found that NLJ's studies could not produce sufficiently sound information about program outcomes, we saw no evidence that program effects had been isolated using alternative, compensatory, or supplemental methods.

In addition to these comments, the Assistant Attorney General also provided us with a number of technical comments, which we incorporated in the report as appropriate.

As arranged with your office, unless you publicly announce its contents earlier, we plan no further distribution of this report until 14 days from the date of this report. At that time, we will send copies to the Attorney General, appropriate congressional committees and other interested parties. In addition, the report will be available at no charge on GAO's Web site at <http://www.gao.gov>.

Sincerely yours,

A handwritten signature in black ink, reading "Laurie E. Ekstrand". The signature is fluid and cursive, with the first letters of the first and last names being capitalized and prominent.

Laurie E. Ekstrand
Director, Homeland Security
and Justice Issues

Appendix I: Objectives, Scope, and Methodology

In response to your request, we undertook a review of the outcome evaluation work performed under the direction of the National Institute of Justice (NIJ) during the last 10 years. We are reporting on (1) the methodological quality of a sample of completed and ongoing NIJ outcome evaluation grants and (2) the usefulness of the evaluations in producing information on program outcomes.

Our review covered outcome evaluation grants managed by NIJ from 1992 through 2002. Outcome evaluations are defined as those efforts designed to determine whether a program, project, or intervention produced its intended effects. These kinds of studies can be distinguished from process evaluations, which are designed to assess the extent to which a program is operating as intended.

To determine the methodological quality of a sample of NIJ-managed outcome evaluations, we asked NIJ, in June 2002, to identify and give us a list of all outcome evaluations managed by NIJ that were initiated during the last 10 years, or initiated at an earlier date but completed during the last 5 years. NIJ identified 96 evaluation studies that contained outcome evaluation components that had been awarded during this period. A number of these studies included both process and outcome components. We did not independently verify the accuracy or completeness of the data NIJ provided.

These 96 evaluations were funded for a total of about \$36.6 million. Individual grant awards ranged in size from \$22,374 to about \$2.8 million. Twenty grants were awarded for \$500,000 or more, for a total of about \$22.8 million (accounting for about 62 percent of all funding for NIJ outcome evaluations during the 10-year review period); 51 grants for less than \$500,000, but more than \$100,000, for a total of about \$11.7 million (accounting for about 32 percent of all NIJ outcome evaluation funding); and 25 grants for \$100,000 or less, for a total of about \$2.1 million (accounting for about 6 percent of all NIJ outcome evaluation funding). Fifty-one of the 96 evaluations had been completed at the time of our review; 45 were ongoing.

From the list of 96 outcome evaluation grants, we selected a judgmental sample of 16 grants for an in-depth methodological review. Our sample selection criteria were constructed so as to sample both large and medium-sized grants (in terms of award size), and both completed and ongoing studies. We selected 8 large evaluations—funded at \$500,000 or above—and 8 medium-sized evaluations—funded at between \$101,000 and \$499,000. Within each group of 8 we selected the 4 most recently

completed evaluations, and the 4 most recently initiated evaluations that were still ongoing, in an effort to ensure that the majority of the grants reviewed were subject to the most recent NIJ grant management policies and procedures. One of the medium-sized ongoing evaluations was dropped from our review when we determined that the evaluation was in the formative stage of development; that is, the application had been awarded but the methodological design had not yet been fully developed. As a result, our in-depth methodological review covered 15 NIJ-managed outcome evaluations accounting for about 42 percent of the total spent on outcome evaluation grants between 1992 and 2002 (see tables 5 and 6). These studies are not necessarily representative of all outcome evaluations managed by NIJ during this period.

Table 5: Number and Size of Outcome Evaluation Awards Made by NIJ from 1992 through 2002, and Reviewed by GAO

Size of grant	All NIJ outcome evaluations		NIJ outcome evaluations reviewed by GAO	
	Number of grants	Total funding	Number of grants (percent reviewed in category)	Total funding (percent reviewed in category)
Large (\$500,000 or more)	20	\$22,801,186	8 (40%)	\$13,654,211 (60%)
Medium (\$101,000-\$499,000)	51	11,687,679	7 (14%)	1,765,915 (15%)
Small (\$100,000 or less)	25	2,110,737	N/A	N/A
Total	96	\$36,599,602	15 (16%)	\$15,420,126 (42%)

Source: GAO analysis of NIJ data.

Table 6: Size and Completion Status of the 15 Evaluations Selected for Methodological Review

Grant title	Award	Size of award		Status	
		Large	Medium	Completed	Ongoing
National Evaluation of Gang Resistance Education and Training Program	\$1,568,323	X		X	
Evaluation of Chicago's Citywide Community Policing Program	\$2,157,859	X		X	
National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program	\$719,949	X		X	
Evaluation of Breaking the Cycle	\$2,419,344	X		X	
National Evaluation of the Domestic Violence Victims' Civil Legal Assistance Program	\$800,154	X			X
Evaluation of a Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment	\$2,498,638	X			X
Corrections and Law Enforcement Family Support: Law Enforcement Field Test	\$649,990	X			X
Evaluation of a Multi-Site Demonstration for Enhanced Judicial Oversight of Domestic Violence Cases	\$2,839,954	X			X
Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing	\$187,412		X	X	
An Evaluation of Victim Advocacy with a Team Approach	\$153,491		X	X	
Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service	\$399,919		X	X	
Responding to the Problem Police Officer: An Evaluation of Early Warning Systems	\$174,643		X	X	
Evaluation of a Juvenile Justice Mental Health Initiative with Randomized Design	\$200,000		X		X
Culturally Focused Batterer Counseling for African-American Men	\$356,321		X		X
Testing the Impact of Court Monitoring and Batterer Intervention Programs at the Bronx Misdemeanor Domestic Violence Court	\$294,129		X		X

Source: GAO analysis of NIJ data.

The evaluations we selected comprised a broad representation of issues in the criminal justice field and of program delivery methods. In terms of criminal justice issues, 7 of the 15 evaluations focused on programs designed to reduce domestic violence, 4 focused on programs addressing the behavior of law enforcement officers, 2 focused on programs addressing drug abuse, and 2 focused on programs to deal with juvenile justice issues. In terms of program delivery methods, 3 evaluations examined national discretionary grant programs or nationwide cooperative agreements, 4 examined multisite demonstration programs, and 8 examined local programs or innovations.

For the 15 outcome evaluations we reviewed, we asked NIJ to provide any documentation relevant to the design and implementation of the outcome evaluation methodologies, such as the application solicitation, the grantee's initial and supplemental applications, progress notes, interim reports, requested methodological changes, and any final reports that may have become available. We used a data collection instrument to obtain information systematically about each program being evaluated and about the features of the evaluation methodology. We based our data collection and assessments on generally accepted social science standards.¹ We examined such factors as whether evaluation data were collected before and after program implementation; how program effects were isolated (i.e., the use of nonprogram participant comparison groups or statistical controls); and the appropriateness of sampling, outcome measures, statistical analyses, and any reported results.² A senior social scientist with training and experience in evaluation research and methodology read and coded the documentation for each evaluation. A second senior social scientist reviewed each completed data collection instrument and the relevant documentation for the outcome evaluation to verify the accuracy of every coded item. We relied on documents NIJ provided to us between October 2002 and May 2003 in assessing the evaluation methodologies and reporting on each evaluation's status. We grouped the studies into 3 categories based on our judgment of their methodological soundness. Although we recognize that the stronger studies may have had some weaknesses, and that the weaker studies may have had some strengths, our categorization of the studies was a summary judgment based on the totality of the information provided to us by NIJ. Following our review, we interviewed NIJ officials regarding NIJ's role in soliciting, selecting, and monitoring these grants, and spoke to NIJ grant managers regarding issues raised about each of the grants during the course of our methodological review.

¹These standards are well defined in scientific literature. See, for example, Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally & Company, 1963); Carol H. Weiss, *Evaluation Research: Methods for Assessing Program Effectiveness* (Englewood Cliffs: Prentice-Hall, Inc., 1972); Edward A. Suchman, *Evaluative Research: Principles and Practice in Public Service & Social Action Programs* (New York: Russell Sage Foundation, 1967); and [GAO/PEMD-10.1.4](#).

²The evaluations varied in the methodologies that were used to examine program effects. Of the 15 evaluations, 14 did not explicitly discuss cost/benefit considerations. The evaluation of Breaking the Cycle estimated cost/benefit ratios at each of the 3 demonstration sites examined.

In the course of our discussions with NIJ officials, we learned of changes NIJ has underway to improve its administration of outcome evaluation studies. To document these changes, we interviewed responsible NIJ officials, and requested and reviewed relevant documents. We are providing information in this report about these changes.

To identify the usefulness of the evaluations in producing information on program outcomes, we reviewed reported findings from completed NIJ-managed outcome evaluations that either evaluated programs administered or funded by the Department of Justice (DOJ), or had been conducted with funding contributed by DOJ program offices (see table 7). Of the 8 completed evaluations that we reviewed for methodological adequacy, 5 had been conducted with funding contributed in part by DOJ program offices, including 2 evaluations funded in part by DOJ's Office on Violence Against Women (OVW) and 3 evaluations funded in part by DOJ's Office of Community Oriented Policing Services (COPS). Of the 2 evaluations funded by OVW, 1 was a review of a national program administered by DOJ, and the other was a review of a locally administered program funded partially by an OVW grant. Of the 3 evaluations funded by COPS, 2 were evaluations of programs funded at least in part with COPS funding, and the other was an evaluation of a program operating at several local law enforcement agencies, supported with local funding. Because of our interest in the effectiveness of criminal justice programs, we limited our review of the usefulness of NIJ outcome evaluations to evaluations of DOJ programs, or evaluations funded by DOJ program offices, and did not examine the 3 other completed NIJ outcome evaluations that focused on programs funded by agencies other than DOJ.

Table 7: Programs Evaluated and Funding Sources for Completed NIJ Outcome Evaluations

Completed NIJ evaluations	DOJ-funded program	Evaluation funded by DOJ program offices
OVW Evaluations		
National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program	Yes	Yes
An Evaluation of Victim Advocacy with a Team Approach	Yes	Yes
COPS Evaluations		
Evaluation of Chicago's Citywide Community Policing Program	Yes	Yes
Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service	Yes	Yes
Responding to the Problem Police Officer: An Evaluation of Early Warning Systems	No	Yes
Other evaluations		
National Evaluation of Gang Resistance Education and Training Program	No	No
Evaluation of Breaking the Cycle	No	No
Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing	No	No

Source: GAO analysis of NIJ data.

We interviewed NIJ officials and relevant DOJ program administrators regarding whether these findings were used to implement improvements in the evaluated programs. At OVW and COPS, we asked officials the extent to which they (1) were involved in soliciting and developing the evaluation grant, and monitoring the evaluation; (2) were aware of the evaluation results; and (3) had made any changes to the programs they administered based on evaluation findings about the effectiveness of the evaluated programs.

We conducted our work at NIJ headquarters in Washington, D.C., between May 2002 and August 2003 in accordance with generally accepted government auditing standards.

Appendix II: Summaries of the NIJ Outcome Evaluations Reviewed

Evaluations with Sound Designs and Sound Implementation Plans

Evaluation	The National Evaluation of the Gang Resistance Education and Training (GREAT) Program
Principal investigator	University of Nebraska at Omaha
Program evaluated	The GREAT program began in 1991 with the goal of using federal, state, and local law enforcement agents to educate elementary school students in areas prone to gang activity about the destructive consequences of gang membership. The program seeks to prevent youth crime and violence by reducing involvement in gangs. According to the evaluator's proposal, as of April 1994, 507 officers in 37 states (150 sites) had completed GREAT training. GREAT targets middle school students (with an optional curriculum for third and fourth graders) and consists of 8 lessons taught over a 9-week period.
Evaluation components	Process and outcome evaluations began in 1994 and were completed in 2001. Total evaluation funding was \$1,568,323. The outcome evaluation involved a cross-sectional and longitudinal design. For the cross-sectional component, 5,935 eighth grade students in 11 different cities were surveyed to assess the effectiveness of GREAT. Schools that had offered GREAT within the last 2 years were selected, and questionnaires were administered to all eighth graders in attendance on a single day. This sample constituted a 1-year follow-up of 2 ex-post facto groups: students who had been through GREAT and those who had not. A 5-year longitudinal, quasi-experimental component was conducted in 6 different cities. Schools in the 6 cities were selected purposively, to allow for random assignment where possible. Classrooms in 15 of 22 schools were randomly assigned to receive GREAT or not, whereas assignment in the remaining schools was purposive. A total of more than 3,500 students initially participated, and active consent was obtained for 2,045 participants. Students were surveyed 2 weeks before the program, 2 weeks after completion, and at 1-, 2-, 3-, and 4-year intervals after completion. Significant follow-up efforts were employed to maintain reasonable response rates. Concepts measured included attitudinal measures regarding crime, gangs and police; delinquency; drug sales and use; and involvement in gangs, gang activities, and risk-seeking behaviors. In addition, surveys were conducted with parents of the students participating in the longitudinal component, administrative and teaching staff at the schools in the longitudinal design, and officers who had completed GREAT training prior to July 1999.
Assessment of evaluation	Although conclusions from the cross-sectional component may be limited because of possible pre-existing differences between students who had been exposed to GREAT and students who had not and lack of detail about statistical controls employed, the design and analyses for the longitudinal component are generally sound, including random assignment of classrooms to the intervention in 15 of the 22 schools, collection of baseline and extensive follow-up data; and statistical controls for differential attrition rates of participant and comparison groups.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Evaluation of Breaking the Cycle
Principal investigator	Urban Institute
Program evaluated	<p>A consortium of federal agencies, led by the Office of National Drug Control Policy and NIJ, developed the Breaking the Cycle (BTC) demonstration program in 3 sites to test the effectiveness of a comprehensive, coordinated endeavor to reduce substance abuse and criminal activity, and improve the health and social functioning of drug-involved offenders. The first site, Birmingham, Ala., received funding in 1997, and the next 2 sites, Tacoma, Wash., and Jacksonville, Fla. received funding in 1998. Participants were adult arrestees (for any type of crime) who tested positive for drug use and had a history of drug involvement. The program was based on the recognition that there was a link between drug use and crime, and it had the support of many criminal justice system officials who were willing to use the authority of the criminal justice system to reduce drug use among offenders. BTC intended to expand the scope of earlier programs such as drug courts and Treatment Alternatives to Street Crime by incorporating drug reduction activities as part of handling felony cases. BTC included early intervention; a continuum of treatment options tailored to participants' needs, including treatment readiness programs in jails; regular judicial monitoring and graduated sanctions; and collaboration among justice and treatment agencies.</p>
Evaluation components	<p>Begun in 1997, and the final report completed in 2003, the evaluation was funded for \$2,419,344, and included both outcome and process components. Comparison groups were selected in each of the 3 sites, and were composed of defendants similar to the BTC participants who were arrested in the year before BTC was implemented. The evaluation examined program success in (1) reducing drug use and criminal activity, as measured by self-reported drug use in the 6 months prior to follow-up interviews and officially recorded arrests in the 12 months after baseline; (2) improving the physical and mental health and family/social well-being of participants, as measured by self-reported interview data on problems experienced in these 3 areas during the 30 days before follow-up; and (3) improving labor market outcomes for participating offenders, as measured by self-reported interview data on employment and social difficulties in the 30 days before follow-up. Survey data were collected at baseline and again at two intervals between 9 and 15 months after baseline. At baseline the sample sizes for the treatment and comparison groups were, respectively, 374 and 192 in Birmingham, 335 and 444 in Jacksonville, and 382 and 351 in Tacoma. Response rates for the follow-up interviews varied across the 3 sites from 65 to 75 percent for the treatment groups, and from 71 to 73 percent for the comparison groups. Method of assessment varied across sites and across samples, with some participants in both the comparison and treatment groups interviewed in person while others were interviewed by telephone. Multiple statistical analyses, including logistic regression, with controls for differences in demographics, offense history, substance abuse history, and work history between treatment and comparison groups were used. BTC's effect on the larger judicial environment was also assessed, using official records on the number of hearings, case closure rates, and other factors.</p> <p>Cost-benefit analyses of the BTC interventions were conducted at the three locations. The costs attributable to the BTC program were derived from budgetary information provided by program staff. The BTC program benefits were conceptualized as "costs avoided" arising from the social and economic costs associated with crime. The estimates of cost avoided in the study were based on (1) the costs (to society) associated with the commission of particular crimes and (2) the costs (to the criminal justice system) associated with arrests. Estimates of these components from the economic and criminal justice literature were applied to self-reported arrest data from the program and comparison group subjects. The derived estimates of benefits were compared to program costs to form cost-benefit ratios for the interventions. An earlier effort to incorporate estimates of savings in service utilization from BTC (as a program benefit) was not included in the final report analysis due to inconclusive results.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Assessment of evaluation	<p>The evaluation was well designed and implemented. The study used comparison groups to isolate and minimize external factors that could have influenced the results. While the comparison groups were selected and baseline data collected 1 year before the treatment groups were selected, the study corrected for selection bias and attrition, using multivariate models that incorporated control variables to measure observed sample differences. The study appears to have handled successfully other potential threats to the reliability and validity of results, by using appropriate statistical analyses to make adjustments. For example, the study relied on both self-reported measures of drug use and arrest histories as well as official records of arrests, to assess the effects of the program. Self-report measures are subject to errors in memory or self-presentational biases, while official records can be inaccurate and/or incomplete. The evaluators made use of both the self-report and official measures to attempt to control for these biases.</p> <p>The methodological approach used in the cost benefit analysis was generally sound. The report specified the assumptions underlying the cost and benefit estimates, and appropriately discussed the limitations of the analysis for policymaking.</p>
--------------------------	--

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Evaluation of a Multi-Site Demonstration for Enhanced Judicial Oversight of Domestic Violence Cases
Principal investigator	The Urban Institute
Program evaluated	<p>The Judicial Oversight Demonstration (JOD) initiative is a multiyear program being implemented at 3 sites (City of Boston/Dorchester District Court, Mass.; Washtenaw County, Ann Arbor, Mich.; and Milwaukee County, Wis.) to address the problem of domestic violence. JOD tests the idea that a coordinated community, focused judicial, and systemic criminal justice response can improve victim safety and service provision, as well as offender accountability. JOD emphasizes uniform and consistent responses to domestic violence offenses, including coordinated victim advocacy and services; strong offender accountability and oversight; rigorous research and evaluation components; and centralized technical assistance. Demonstration sites have developed partnerships with a variety of public and private entities, including victim advocacy organizations, local law enforcement agencies, courts, and other social service providers. The program began in fiscal year 2000, and demonstration sites are expected to receive funding for 5 years.</p>
Evaluation components	<p>A process evaluation began in January 2000. The outcome component of the evaluation began in October 2002 and is to be completed by October 2005. At the time of our review, the evaluation grant amount was \$2,839,954. Plans call for a full outcome assessment to be conducted in 2 sites and, because no appropriate comparison site could be identified, a partial assessment in the third site. The 2 sites with a full assessment were matched with comparison sites having similar court caseloads and population demographics; neither comparison site had a specialized court docket, enhanced judicial oversight, or a countywide coordinated system for handling domestic violence cases. Over 12 months, all domestic violence cases in each site, up to monthly size quotas, will be selected into the following groups: cases where the offender was found guilty and sentenced to jail for 6 months or less and probation or probation only, cases that were dismissed or diverted from prosecution, and cases where the offender received more than 6 months incarceration. Victims and offenders in the first group will be interviewed, and in the second group, victims only will be interviewed. Offender recidivism in both groups will be tracked for 1 year following the intervention using police and court records. For the third group, only offender recidivism will be tracked. In the partial assessment site, subject to data availability, the plan is to compare a sample of domestic violence cases in which the offender was placed on probation in the period before JOD implementation with a sample of cases in which the offender was placed on probation and scheduled for judicial review in the period after JOD implementation. Data about incidents, victims, and offenders are to be obtained from official records, and offender recidivism will be tracked using police and court records. Overall, short-term outcomes for the study are planned to include various measures of offender compliance and victim and offender perceptions of JOD, and long-term outcomes are planned to include various measures of offender recidivism, victim well-being, and case processing changes. In addition, to discern any system level changes due to JOD, aggregate, annual data on all domestic violence cases for the 2 years prior to and 3 years after JOD implementation in all sites will be collected and analyzed.</p>
Assessment of evaluation	<p>The evaluation plan appears to be ambitious and well designed. A quasi-experimental design is planned, and data will be collected from multiple sources, including victims, offenders, and agencies. While lack of sustained cooperation, uneven response rates, and missing data could become problems, detailed plans seem to have been made to minimize these occurrences. The planned approach of selecting cases (choosing equal numbers of cases consecutively until a monthly quota is reached, over a 12-month period) may be nearly as good as random sampling and takes into consideration seasonal variation. However, it could introduce biases, should there be variation as to the time each month when case selection begins.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Culturally Focused Batterer Counseling for African-American Men
Principal investigator	Indiana University of Pennsylvania
Program evaluated	<p>The purpose of this study is to test the relative effectiveness of culturally focused versus conventional batterer counseling for African-American men. It is based on research indicating that conventional counseling dropout and partner re-assault rates are higher for African-American men than they are for white men, and clinical literature in related fields that recommends culturally focused counseling to improve the effectiveness of counseling with African-American men. Culturally focused counseling refers to the counselor recognizing and responding to cultural issues that emerge in group sessions (including such topics as African-American men's perceptions of the police, relationships with women, sense of African-American manhood, past and recent experiences of violence, and reactions to discrimination and prejudice), and a curriculum that includes the major cultural issues facing a particular group of participants. The setting for the evaluation is a counseling center in Pittsburgh, Pennsylvania.</p>
Evaluation components	<p>The evaluation began in September 2001, and the expected completion date is February 2005. At the time of our review, the grant amount was \$356,321. A clinical trial will be conducted to test the effect of culturally focused counseling on the extent to which African-American men drop out of counseling, are accused of re-assaults, and are re-arrested for domestic violence. Plans are for 600 African-American men referred by the Pittsburgh Domestic Violence Court over a 12-month period to batterer counseling at the counseling center to be randomly assigned to either (1) a culturally focused counseling group of only African-Americans, (2) conventional batterer counseling in an African-American only group, and (3) conventional counseling in a racially mixed group. Before assignment, however, the counseling center must recommend the men for participation in the study. Men included in the study will be administered a background questionnaire and two tests of culturally specific attitudes (i.e., racial acculturation and identity) at program intake. The men's female partners will be interviewed by phone 3 months, 6 months, and 12 months after program intake. These structured interviews will collect information on the woman's relationship with the man, the man's behavior, and the woman's help-seeking. Clinical records of program attendance and police records of re-arrests will be obtained for each man. Planned analyses are to include (1) verification of equivalent culturally focused and conventional counseling sub-samples at intake and during the follow-up; (2) comparison of the program dropouts, re-assaults, and re-arrests for the three counseling options at each follow-up interval and cumulatively; and (3) a predictive model of the re-assault outcome based on characteristics, cultural attitudes, and situational factors. Additionally, interviews with a sub-sample of 100 men about their counseling experience are to be conducted.</p>
Assessment of evaluation	<p>This is a well-designed experiment to test the effect of a new approach to provide counseling to perpetrators of domestic violence. The researchers have plans to (1) adjust for any selection bias in group assignment and participant attrition through statistical analysis; (2) prevent "contamination" from counselors introducing intervention characteristics to control groups, or the reverse; and (3) monitor the response rates on the interviews with female partners. The evaluation is on-going. The most recent progress report we reviewed indicated that the evaluation is proceeding as planned, with the recruitment of batterers behind schedule by 1 month, the series of female partner interviews on schedule and very close to expected response rates, and the interviews with the sub-sample of batterers about three-quarters complete. One potential concern we have is that because all men referred by the domestic violence court to the counseling center may not be recommended to participate in the study, any bias in recommending study participants will determine the population to which the study's results can be generalized.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Testing the Impact of Court Monitoring and Batterer Intervention Programs at the Bronx Misdemeanor Domestic Violence Court
Principal investigator	Fund for the City of New York
Program evaluated	Operating since 1998, the Bronx Misdemeanor Domestic Violence Court handles spousal abuse misdemeanor cases. The court has the power to prescribe various conditions of discharge for batterers, including participation in group counseling and/or court monitoring. Given concerns about the effectiveness of these options, it was decided to test the efficacy of batterer counseling programs and court monitoring, alone and in combination with each other. Furthermore, court monitoring was tested based on the frequency of its administration—either monthly or on a graduated basis (less monitoring for fewer incidences of abuse). This was to ascertain whether graduated monitoring might give batterers more incentive to change.
Evaluation components	The evaluation began in September 2001 and is expected to be completed in August 2003. At the time of our review, this evaluation was funded for \$294,129. The proposed study is an outcome evaluation of 4 different treatment alternatives for conditional discharge defendants in domestic violence cases. The treatment options are (1) counseling program and monthly court monitoring, (2) counseling program and graduated court monitoring, (3) monthly court monitoring program only, and (4) graduated court monitoring only. Participants in the evaluation (800 total) are to be assigned randomly to 1 of the 4 treatments at the time of sentencing, and incidents of new crimes are to be measured 6 and 12 months after sentencing. Official crime records at both intervals, and interviews with victims at the 12-month interval are the sources of data. The planned analysis involves looking at the groups as a whole, and subgroups related to age, criminal history, and current charge. Outcome measures are (1) completion of the conditional discharge or imposition of the jail alternative, (2) new arrests for domestic violence, and (3) new reports from victims of domestic violence incidents.
Assessment of evaluation	This is a well-designed approach to measure the comparative efficacy of combinations of program counseling and variations in monitoring. However, at the time of our review, we had some concerns about how well implementation will proceed. One concern is that if one or more of the treatments is less effective, it could result in participants spending time in jail, reducing the possibility of further incidents. This difficulty can be addressed in the analysis, but neither the proposal nor subsequent progress reports discuss this or other differential attrition issues. Also, although the evaluators have a plan to try to ensure good response rates for the victims' survey, it is uncertain how effective they will be. Other surveys of similar populations have been problematic.

Well-designed Evaluations That Encountered Implementation Problems

Evaluation	An Evaluation of Chicago's Citywide Community Policing Program
Principal investigator	Northwestern University
Program evaluated	Chicago's community policing program, known as Chicago's Alternative Policing Strategy (CAPS), began in April 1993. The program reorganizes policing around small geographical areas where officers assigned to beat teams meet with community residents to identify and address a broad range of neighborhood problems.
Evaluation components	<p>There were 2 evaluation efforts in this study, 1 examining the prototype project and the second examining citywide program implementation. The combined evaluations were completed in August 2001, at a total cost of \$2,157,859.</p> <p>The prototype evaluation, conducted between April 1993 and September 1994, compared five areas that implemented CAPS with four areas that did not. Data from the 1990 Census were used to select four sections of the city that closely matched the demographics of the five prototype areas. Residents of all areas were first surveyed in the spring of 1993 regarding the quality of police service and its impact on neighborhood problems. Follow-up interviews occurred in either June or September of 1994 (14 to 17 month time lags). Interviews were conducted by telephone in English and Spanish. The re-interview rate was about 60 percent. A total of 1,506 people were interviewed both times, an average of 180 in each prototype area and 150 in each comparison area.</p> <p>The CAPS citywide evaluation began after the conclusion of the prototype evaluation in July 1994. The purpose of this evaluation was to assess how changing from a traditional policing approach to a community-centered approach would affect citizens' perceptions of the police, neighborhood problems and crime rates. The researchers administered annual citywide public opinion surveys between 1993 and 2001 (excluding 2000). The surveys covered topics such as police demeanor, responsiveness, and task performance. Surveys were also administered to officers at CAPS orientation sessions to obtain, among other things, aggregate indicators of changes in officers' attitudes toward CAPS. Changes in levels of recorded crimes were analyzed. Direct observations of police meetings, surveys of residents, and interviews with community activists were used to measure community involvement in problem solving and the capacity of neighborhoods to help themselves.</p>
Assessment of evaluation	<p>The 1992 crime rates were reported to be similar between prototype districts and their matched comparison areas and the baseline demographic measures used to match the two groups were basically similar. The initial and follow-up response rates of about 60 percent seem reasonable considering the likelihood of community mobility in these areas; however, attrition rates differed for various demographic characteristics, such as home ownership, race, age, and education, raising some concerns about whether the results are generalizable to the intended population. The follow-up time (14-17 months) was the maximum period allowed by the planned citywide implementation of CAPS. A single follow-up survey and the citywide implementation precluded drawing firm conclusions about longer-term impacts of the prototype program.</p> <p>Because CAPS was implemented throughout the city of Chicago in 1995, the CAPS citywide evaluation was not able to include appropriate comparison groups and could not obtain a measure of what would have happened without the benefits of the program. The authors used a variety of methods to examine the implementation and outcomes of the CAPS program, and stated that there was no elaborate research design involved because their focus was on organizational change. However, because the trends over time from resident surveys and crime data were presented without controls or comparison groups and some declines in crime began before the program was implemented, changes cannot be attributed solely to the program.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Evaluation of a Comprehensive Service-Based Intervention Strategy in Public Housing
Principal investigator	Yale University School of Medicine
Program evaluated	The program was an intervention strategy designed to reduce drug activity and foster family self-sufficiency in families living in a public housing complex in the city of New Haven, Conn. The key elements of the intervention were (1) an on-site comprehensive services model that included both clinical (substance abuse treatment and family support services) and nonclinical components (e.g., extensive outreach and community organizing as well as job training and placement and GED high school equivalency certification) and (2) high profile police involvement. The goals of the program were (1) increases in the proportion of residents entering and completing intervention services and (2) a reduction in substance-related activities and crime.
Evaluation components	The evaluation began in 1998 and was completed in 2000. The total evaluation funding was \$187,412. The intervention site was a public housing complex composed primarily of female heads of household tenants and additional family members; the control site was another public housing complex on the opposite side of town, chosen for its similarities to the intervention site. The evaluation design was both process and outcome oriented and involved the collection of both qualitative and quantitative data. At baseline, a needs assessment survey was completed (n=175 at the intervention site and n=80 at the control site), and follow-up surveys with residents took place at 12 and 18 months post-intervention (no response rates reported). All heads of household at the sites were the target population for the surveys. The follow-up surveys, while administered in the same two sites, did not track the same respondents that were surveyed at baseline. Survey measures included access to social services; knowledge and reported use of social services; and residents' perceptions of the extent of drug and alcohol abuse, drug selling, violence, safety, and unsupervised youth in the community. The study also examined crime statistics obtained from the New Haven police department, at baseline and during the intervention.
Assessment of evaluation	The study had several limitations, the first of which is potential selection bias due to pre-existing differences between the sites, as well as considerable (and possibly differential) attrition in both groups, with no statistical control for such differences. Second, respondents may not have been representative of the populations at the housing sites. No statistical comparisons of respondents to nonrespondents on selected variables were presented. In addition, on the baseline survey, the response rates of the intervention and control sites differed substantially (70 vs. 44 percent, respectively). Overall response rates were not reported for the follow-up surveys. Furthermore, implementation did not work smoothly (e.g., the control site received additional unanticipated attention from the police). Finally, the grantee proposed to track data on individuals over time (e.g., completion of services), but this goal was not achieved, in part because of the limited capability of project staff in the areas of case monitoring, tracking, and data management. Thus, although the intervention may have produced changes in the intervention site "environment" over time (aggregate level changes), it is not clear that the intervention successfully impacted the lives of individuals and families at the site.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	An Evaluation of Victim Advocacy with a Team Approach
Principal investigator	Wayne State University
Program evaluated	The program provides assistance to domestic violence victims in some police precincts in the city of Detroit. The domestic violence teams studied included specially trained police officers, police department advocates, legal advocates, and in one police precinct, an on-site prosecutor. The advocates assisted victims by offering information about the legal system, referrals, and safety planning.
Evaluation components	The outcome evaluation began in January of 1998 and the final report was completed in January of 2001. The grant amount was \$153,491. The objectives of the study were to address the relationships between advocacy and victim safety and between advocacy and victims' responses to the criminal justice system, using a quasi-experimental design to compare domestic violence cases originating in police precincts with and without special police domestic violence teams that included advocates. The study focused on assistance provided in 3 police precincts. Precincts not served by in-precinct domestic violence teams, but resembling the precincts with such teams in terms of ethnic representation and median income, were selected as comparisons. Data were collected using police records, county prosecutor's office records, advocate contact forms, and telephone interviews with victims. Cases that met Michigan's legal definition of domestic violence, had adult female victims, and were received in the selected precincts over a 4-month period in 1998 were eligible for the study. The cases were first identified by the police department through police reports and then reviewed for qualification by a member of the research team. A weekly quota of cases was selected from each precinct. If the number of qualified cases for a precinct exceeded the quota, then cases were selected randomly using a random numbers table. Outcomes included rates of completed prosecution of batterers, rate of guilty findings against batterers, subsequent violence against victims, victims' perceptions of safety, and victims' views of advocacy and the criminal justice process.
Assessment of evaluation	The study was severely affected by numerous problems, many of which the researchers acknowledged. First, the sample selection was based on incomplete or unreliable data, since police officers in writing reports often did not fully describe incidents, and precinct staff inconsistently provided complete case information about incidents to the researchers. Second, evaluators were not able to secure cooperation from domestic violence advocates and their supervisors at all service levels in providing reliable reports on service recipients and the type, number, and length of services. Additionally, most domestic violence team members were moved out of the precincts and into a centralized location during the period victims in the study were receiving services, thereby potentially affecting the service(s) provided to them. Further, the researchers were uncertain as to whether women from the comparison precincts received any advocacy services, thereby potentially contaminating the research results between the precincts with the domestic violence teams and the comparison precincts. Finally, low response rates and response bias for data collected from victims were problems. The overall response rate for the initial round of telephone interviews was only about 23 percent and the response rates for follow-up interviews were lower. Response rates were not provided separately for victims from the precincts with the domestic violence teams and the comparison precincts. As a result of the low response rates, the interviewed victims were identified as being less likely to have experienced severe physical abuse, less likely to be living with the abuser, and more likely to have a child in common with the abuser, compared to the victims in the sample who were not interviewed.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Reducing Non-Emergency Calls to 911: An Assessment of Four Approaches to Handling Citizen Calls for Service
Principal investigator	University of Cincinnati
Program evaluated	DOJ's COPS office has worked with police agencies, the Federal Communications Commission, and the telecommunications industry to find ways to relieve the substantial demand on the current 911 emergency number. Many police chiefs and sheriffs have expressed concern that non-emergency calls represent a large portion of the 911 overload problem. Four cities have implemented strategies to decrease non-emergency 911 calls and have agreed to participate in the research. Those cities, each implementing a different type of approach, were Baltimore, Md.; Dallas, Tex.; Buffalo, N.Y.; and Phoenix, Ariz.
Evaluation components	A process and outcome evaluation was conducted between July of 1998 and June of 2000. The grant amount was \$399,919. For the outcome component, the grantee examined whether (1) the volume of 911 calls declined following the introduction of the non emergency call system; (2) there was a corresponding decline in radio dispatches, thus enhancing officer time; and (3) this additional time was directed to community-oriented policing strategies. The bulk of the design and analysis focused on Baltimore, with a limited amount of analysis of outcomes in Dallas and no examination of outcomes in the other two sites. The study compared rates of 911 calls before implementation of the new 311 system to rates of 911 and 311 calls after the system in both cities. In Baltimore, time series analysis was used to analyze the call data; police officers and sergeants were surveyed; the flow of 311 and 911 calls to Neighborhood Service Centers was examined; researchers accompanied police officers during randomly selected shifts in 3 sectors of Baltimore for 2 weeks; and citizens who made 311 calls during a certain 1-month time frame were surveyed.
Assessment of evaluation	The crux of the outcome analysis relies on the study of pre- and post- 311 system comparisons, and the time series analysis done in Baltimore is sound. The rigor of several other parts of this study is questionable (e.g., poor response rates to surveys and short time frames for data from accompanying police officers on randomly selected shifts). In addition, the choice of sites that NIJ required the grantee to examine, other than Baltimore, did not allow for a test of the study's objectives. Although NIJ conducted pre-solicitation site visits to all 4 sites, at the time of the solicitation it still did not clearly know whether outcome data would be available at all the sites. As it turned out, outcome data were not available in Phoenix and Buffalo. Further, since the 311 system in Dallas was not implemented with the goal of reducing or changing call volume, it does not appear to be a good case with which to test the study's objectives.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Responding to the Problem Police Officer: An Evaluation of Early Warning Systems
Principal investigator	University of Nebraska – Omaha
Program evaluated	An Early Warning (EW) system is a data based police management tool designed to identify officers whose behavior is problematic, as indicated by high rates of citizen complaints, use of force incidents, or other evidence of behavior problems, and to provide some form of intervention, such as counseling or training to correct that performance. According to the current study's national survey of local law enforcement agencies (LEA) serving populations of 50,000 or more, about one-quarter of LEAs surveyed had an EW system, with another 12 percent indicating that one was planned. One-half of existing EW systems have been created since 1994.
Evaluation components	<p>Begun in 1998, the study was completed in 1999 and included process and outcome components, as well as a national survey. The total evaluation funding was \$174,643. The outcome portion of the study was composed of case studies of EW systems in 3 large urban police departments (Miami-Dade, Fla.; Minneapolis, Minn.; and New Orleans, La.). Sites were selected judgmentally; each had functioning EW systems in place for a period of 4 or more years and had agreed to participate in the study.</p> <p>Both Miami-Dade and Minneapolis case studies examined official performance records (including citizen complaints in both sites and use of force reports in Miami-Dade) for officers identified by the department's EW system, for 2 years prior to and after departmental intervention, compared to records for officers not identified. The participant groups included officers hired between 1990 and 1992 and later identified by the EW system (n=28 in Miami-Dade; n=29 in Minneapolis); the comparison groups included officers hired during the same period and not identified (n=267 in Miami-Dade; n=78 in Minneapolis). In New Orleans, official records were not organized in a way that permitted analysis of performance of officers subject to EW and a comparison group. The New Orleans case study, therefore, examined citizen complaint data for a group of officers identified by the EW system 2 years or more prior to the study, and for whom full performance data were available for 2 years prior to and 2 years following intervention (n=27).</p>
Assessment of evaluation	<p>The study had a number of limitations, many of them acknowledged by the grantee. First, it is not possible to disentangle the effect of EW systems per se from the general climate of rising standards of accountability in all 3 sites. Second, use of nonequivalent comparison groups (officers identified for intervention are likely to differ from those not identified), without statistical adjustments for differences between groups creates difficulties in presenting outcome results. Only in Minneapolis did the evaluators explicitly compare changes in performance of the EW group with changes in performance of the comparison group, again without presenting tests of statistical significance. Furthermore, the content of the intervention was not specifically measured, raising questions about the nature of the intervention that was actually delivered, and whether it was consistent over time in the 3 sites, or across officers subject to the intervention. Moreover, it was not possible to determine which aspects of the intervention were most effective overall (e.g., differences in EW selection criteria, intervention services for officers, and post-intervention monitoring), since the intervention was reportedly effective in all 3 departments despite differences in the nature of their EW systems. Also, no data were available to examine whether the EW systems had a deterrent effect on desirable officer behavior (e.g., arrests or other officer-initiated activity). Finally, generalizability of the findings in Miami-Dade and Minneapolis may also be limited, since those case studies examined cohorts of officers recruited in the early 1990s, and it is not clear whether officers with greater or fewer years of police experience in these departments would respond similarly to EW intervention.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Evaluation of the Juvenile Justice Mental Health Initiative with Randomized Design
Principal investigator	University of Missouri - St. Louis.
Program evaluated	<p>The Juvenile Justice Mental Health Initiative (JJMI) is a collaborative multi-agency demonstration project funded under an Office of Juvenile Justice and Delinquency Prevention grant, and administered by the St. Louis Mental Health Board, the St. Louis Family Court, and the Missouri Department of Health. The initiative provides mental health services to families of youths referred to the juvenile justice system for delinquency who have serious emotional disturbances (SED). The initiative involves parents and families in juvenile justice interventions, providing coordinated services and sanctions for youths who otherwise might shuttle between criminal justice and mental health agencies. Two new mental health programs were established under JJMI. The first, the Child Conduct and Support Program, was designed for families in which youths under the age of 14 do not have a history of serious, violent, or chronic offending. The second, Multi-systemic Therapy (MST), was designed for families in which youths aged 14 and above have prior serious, violent, or chronic delinquency referrals.</p>
Evaluation components	<p>The evaluation began in October 2001 and is expected to be completed in September 2003. At the time of our review, the evaluation was funded for \$200,000. The study proposed to evaluate the two mental health programs using a random experimental design. Youths referred to the Juvenile Court are first screened for SED. Those who test positive or have prior diagnoses of SED (anxiety, depressed mood, somatic complaints, suicidal ideation, thought disturbance, or traumatic experience) are eligible for the JJMI programs. Eligible youth are randomly assigned to either one of the two treatment programs (depending on age) or to a control group. The evaluation includes a comparison of police contact data, court data, self-reported delinquency, and standardized measures of psychological and parental functioning. Potentially important demographic and social context variables, including measures of school involvement and performance, will be obtained from court records.</p>
Assessment of evaluation	<p>This is an ongoing, well designed study. However, as implementation has proceeded, several problems that may affect the utility of the results have emerged. First, the researchers proposed to sample a total of 200 youths, with random assignment expected to result in approximately 100 juveniles in the treatment and comparison groups. The treatment group turned out to be much smaller than anticipated, however, because the randomization protocol and, subsequently, the MST program itself, were discontinued by the St. Louis Mental Health Board. At the time of termination, only 45 youths had been randomly assigned to the treatment group. The small number of subjects limits the extent of the analyses that can be conducted on this population.</p> <p>The Child Conduct and Support Program designed to address the mental health needs of youth under the age of 14 without a history of serious offending was never implemented by the providers contracted to develop the program. Eligible youth, of all ages, were instead assigned to the MST program. Thus, the evaluation will not be able to compare the relative effectiveness of programs specifically designed for younger and older juvenile offenders with SED.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluations with Design Limitations

Evaluation	National Evaluation of the Rural Domestic Violence and Child Victimization Enforcement Grant Program
Principal investigator	COSMOS Corporation
Program evaluated	The National Rural Domestic Violence and Child Victimization Enforcement Grant program, begun in fiscal year 1996, has funded 92 grants through September 2001 to promote the early identification, intervention, and prevention of woman battering and child victimization; increase victim's safety and access to services; enhance the investigation and prosecution of crimes of domestic violence and child abuse; and develop innovative, comprehensive strategies for fostering community awareness and prevention of domestic abuse. The program seeks to maximize rural resources and capacity by encouraging greater collaboration between Indian tribal governments, rural local governments, and public and private rural service organizations.
Evaluation components	The evaluation began in October 1998 and was completed in July 2002. This evaluation was funded at \$719,949, and included both process and outcome components. Initially 10 grantees (comprising 11 percent of the total number of program grantees) were selected to participate in the outcome evaluation; 1 was unable to obtain continuation funding and was dropped from the outcome portion of the study. Two criteria were used in the selection of grant participants: the "feasibility" of grantees visited in the process phase of the evaluation (n=16) to conduct an outcome evaluation; and recommendations from OVW, which were based on knowledge of grantee program activities and an interest in representing the range of organizational structures, activities, and targeted groups served by the grantees. Logic models were developed, as part of the case study approach, to show the logical or plausible links between a grantee's activities and desired outcomes. The specified outcome data were collected from multiple sources, using a variety of methodologies, during 2-3 day site visits (e.g., multi-year criminal justice, medical, and shelter statistics were collected from archival records where available; community stakeholders were interviewed; and grantee and victim service agency staff participated in focus groups).
Assessment of evaluation	This evaluation has several limitations. First, the choice of the 10 outcome sites was skewed toward the technically developed evaluation sites and was not representative of all Rural Domestic Violence program grantees, particular project types, or delivery styles. Second, the lack of comparison groups makes it difficult to exclude the effect of external factors, such as victim safety and improved access to services, on perceived change. Furthermore, several so-called short-term outcome variables were in fact process variables (e.g., number of clients served, number of services provided, number of workshops conducted, and service capacity of community agencies). Moreover, it is not clear how interview and focus group participants were selected. Finally, pre- and post- survey data were not collected at multiple points in time to assess change, except at 1 site, where pre- and post-tests were used to assess increased knowledge of domestic violence among site staff as a result of receiving training.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	National Evaluation of the Domestic Violence Victims' Civil Legal Assistance Program
Principal investigator	Institute for Law and Justice
Program evaluated	The Civil Legal Assistance (CLA) program is one of seven OJP grants (through OVW) dedicated to enhancing victim safety and ensuring offender accountability. The CLA program awards grants to nonprofit, nongovernmental organizations that provide legal services to victims of domestic violence or that work with victims of domestic violence who have civil legal needs. The CLA grant program was created by Congress in 1998. In fiscal year 1998, 54 programs were funded, with an additional 94 new grantees in fiscal year 1999. Approximately 85-100 new and continuation grants were anticipated in fiscal year 2000.
Evaluation components	The study began in November 2000 and was expected to be completed in October 2003. The proposed evaluation consisted of process and outcome components and the total evaluation funding at the time of our review was \$800,154. The objective of the outcome evaluation was to determine the effectiveness of the programs in meeting the needs of the women served. The researchers proposed to study 8 sites with CLA programs. At each site at least 75 cases will be tracked to see if there is an increase in pro se (self) representation in domestic violence protective order cases, and a total of 240 victims receiving services will be surveyed (about 30 at each site). Focus groups of service providers will be used to identify potential program impacts on the justice system and wider community. Outcomes to be assessed include change in pro se representation in domestic violence protective order cases, satisfaction with services, and legal outcomes resulting from civil assistance.
Assessment of evaluation	The evaluation has several limitations. First, NIJ and the grantee agreed in 2002 not to utilize a comparison group approach whereby data would be collected from a set of comparison sites, due to concerns that investment in that approach would limit the amount of information that could be derived from the process component of the evaluation and from within-site and cross-site analyses of the selected outcome sites. Thus, the study will be limited in its ability to isolate and minimize the potential effects of external factors that could influence the results of the study, in part because it did not include comparison groups in the study design. At the time of our review, it was not yet clear whether sufficient data will be available from the court systems at each outcome site in order to examine changes in pro se representation. In addition, since victims would be selected for the surveys partially on the basis of willingness to be interviewed, it is not clear how representative the survey respondents at each site will be and how the researchers will handle response bias. It also appears that the victim interviews will rely to a great extent on measures that will primarily consist of subjective, retrospective reports.

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Multi-Site Demonstration of Collaborations to Address Domestic Violence and Child Maltreatment
Principal investigator	Caliber Associates
Program evaluated	<p>The Department of Health and Human Services and DOJ's Office of Justice Programs are jointly funding 6 demonstration sites for up to 3 years to improve how 3 systems (dependency courts, child protective services, and domestic violence service providers) work with their broader communities to address families with co-occurring domestic violence (DV) and child maltreatment (CM). Funded sites must agree to implement key recommendations of the National Council of Juvenile and Family Courts Judges' publication, "Effective Interventions in Domestic Violence and Child Maltreatment: Guidelines for Policy and Practice" (aka, the "Greenbook"). At a minimum, the sites need to implement changes in policies and procedures regarding screening and assessment; confidentiality and information sharing; safety; service provision; advocacy; cross-training; and case collaboration. The goals of the demonstration are to generate more coordinated, comprehensive, and consistent responses to families faced with DV and CM, resulting in increased safety and well-being for women and their children.</p>
Evaluation components	<p>The evaluation began in September 2000, and is expected to be completed around September 2004. At the time of our review, this evaluation was funded at \$2,498,638, for both process and outcome components. The original evaluation proposal focused on various process elements as well as the effects of the intervention on perpetrator recidivism and the safety of women and children. In the second year, the evaluator realized that no site considered itself to be in the implementation phase and many of the original outcome indicators for children and families were not appropriate given the initiative time frame. The revised design in the funded third year proposal is therefore a systems-level evaluation. The analytic focus is now on how the 3 systems identify, process, and manage families with co-occurrence of DV and CM.</p> <p>A random sample of case records from before and after the introduction of the intervention will be used to document trends in identification of co-occurring cases of DV and CM over the course of the intervention. Stakeholder interviews conducted during site visits in fall 2001 and later during implementation, and analysis of agency documents, will be used to measure changes in policies and procedures. "Network analysis" of responses on the stakeholder interviews will be performed to measure changes in how key stakeholders work with others within and across systems. Supervisors and workers will also be asked, early in the implementation period and at the end of the initiative, to respond to vignettes describing hypothetical situations involving co-occurrence of DV and CM to see how they might respond to clients.</p>
Assessment of evaluation	<p>This evaluation has several limitations. First, the study objectives changed substantially from year 1 to year 3. The study is no longer examining outcomes for individuals, precluding conclusions about whether the implementation improved the lives of victims of domestic violence or their children. Second, it is not clear whether the evaluator will locate appropriate comparison data at this late stage, and without a comparison group, the study will not be able to determine (a) whether collaboration between systems improved (or weakened) because of the intervention or some extraneous factors and (b) whether collaboration resulted in increased capacity in the 3 systems to identify the co-occurrence of DV and CM, or whether these kinds of cases increased for reasons other than collaboration (e.g., perhaps identification of these cases is improving all over the country). Questions remain about the extent of data available for examining co-occurrence of DV and CM at the 6 sites.</p>

**Appendix II: Summaries of the NIJ Outcome
Evaluations Reviewed**

Evaluation	Corrections and Law Enforcement Family Support (CLEFS) Law Enforcement Field Test
Program evaluated	Since 1996 NIJ has funded, as part of the CLEFS program, 32 grants totaling over \$2.8 million to law enforcement agencies, correctional agencies, and organizations representing officers (unions and membership associations) to support the development of research, demonstration, and evaluation projects on stress intervention methods. The stress intervention methods developed and studied have included stress debriefing and management techniques, peer support services, referral networks, police chaplaincy services, stress management training methods, spouse academies, and stress education programs. While NIJ purports to have developed state-of-practice stress reduction methods through these efforts, it acknowledges that very little outcome data have been generated.
Evaluation components	The evaluation began in June 2000 and is expected to be completed in June 2004. At the time of our review, the grant amount was \$649,990. The study proposes to develop and field test a model to allow for the systematic evaluation of selected program components. The grantee worked with NIJ to identify the test sites and services to be evaluated, based on grant application reviews, telephone interviews, and site visits. Three police departments in Duluth, Minn.; North Miami Beach, Fla.; and Knoxville, Tenn. were selected. Baseline stress correlate data were collected during visits to the 3 sites between January 2002 and March 2002, and baseline officer and spouse/partner surveys were conducted during the same visits. Outcome data were to be collected at baseline (prior to actual program implementation), midway through the implementation, and toward the end of the evaluation. While the original proposal did not specify exactly what stress correlate or outcome data were to be collected, the grantee was considering looking at rates of absenteeism and tardiness, citizen complaints, rule and regulation violations, disciplinary actions, and premature retirements and disability pensions, as stress correlates. These were to be obtained from official agency records. Surveys included questions about program impacts on physical health, emotional health, job performance, job satisfaction, job-related stress, and family related stress. The evaluation also included baseline health screenings. It appears the evaluation plan has been modified to add supervisor surveys (there were none at baseline), and to incorporate group data collection efforts with officers, spouses, supervisors, and administrators.
Assessment of evaluation	The study has several limitations. First, the 3 study sites were chosen on the basis of merits in their proposal to implement a stress reduction or wellness program for officers, from 4 sites that submitted applications. There was no attempt to make the chosen sites representative of other sites with stress reduction programs and police departments more generally. Second, the study will not make use of comparison groups consisting of similar agencies that did not implement stress reduction programs. It is unclear how effects of the interventions in these 3 sites over time will be disentangled from the effects of other factors that might occur concurrently. Third, the grantee will not collect individually identified data, and thus will only be able to analyze and compare aggregated data across time, limiting the extent of analysis of program effects that can be accomplished. Fourth, response rates to the first wave of officer surveys were quite low in 2 of the 3 sites (16 percent and 27 percent).

Appendix III: Comments from the Department of Justice



U.S. Department of Justice

Office of Justice Programs

Office of the Assistant Attorney General

Washington, D.C. 20531

Laurie E. Ekstrand
Director, Homeland Security and Justice Issues
General Accounting Office
441 G Street, N.W.
Mail Stop 2440A
Washington, DC 20548

SEP 04 2003

Dear Ms. Ekstrand:

This letter responds to the General Accounting Office (GAO) draft report entitled, *"JUSTICE OUTCOME EVALUATIONS: Design and Implementation of Studies Require More NIJ Attention"* (GAO-03-1091).

The Office of Justice Programs (OJP) and its component, the National Institute of Justice (NIJ), share GAO's support for high-quality evaluations that can inform criminal justice policy and practice. The GAO's review of 15 NIJ outcome evaluations reveals much about the complexity of real-world evaluation, the challenges to a successful outcome evaluation, and the keys to ensuring a highly successful evaluation.

In the draft report, GAO recommended that NIJ review its ongoing outcome evaluation grants and develop appropriate strategies and corrective measures to ensure that methodological design and implementation problems are overcome. The GAO also recommended that NIJ continue to assess its evaluation process, including considering the feasibility of obtaining more information about the availability of outcome data prior to developing a solicitation for research; requiring proposals to contain more detailed design specifications; and more carefully calibrating NIJ monitoring procedures based on the characteristics of the grant and the knowledge area under investigation.

The NIJ agrees with GAO's recommendations. As GAO noted, in January 2003 NIJ established an Evaluation Division to, in part, improve the quality and utility of NIJ's evaluations. The NIJ has also started, and will continue with, a new strategy of "evaluability assessments." These are quick, low-cost initial assessments of criminal or juvenile justice programs to see if the necessary conditions exist to warrant sponsoring a full-scale outcome evaluation. Also, as part of the grantmaking process, NIJ is developing new grant "special conditions" that will require grantees to document all changes in the scope and components of evaluation designs. For ongoing grants, NIJ is making several changes in response to GAO's concerns. During Fiscal Year 2004, NIJ plans to review its grant monitoring procedures for evaluation grants in order to intensively monitor the larger or more complex grants. The NIJ will also conduct periodic reviews of its evaluation research portfolio to assess the progress of

ongoing grants, document any changes in evaluation design that may have occurred, and reassess the expected benefits of ongoing projects.

A few factual inaccuracies in the draft report have been highlighted in the attachment to this letter. In addition, there are two points that GAO makes in the draft report that we believe require highlighting here.

We strongly agree with GAO that “optimal conditions for the scientific study of complex social problems almost never exist.” (draft report p. 4) The “real-world” conditions with which evaluators must contend often pose substantial challenges to successfully completing even the best designed and carefully planned evaluations. As the draft report notes regarding the six of eleven evaluations that encountered problems, “some evaluators were unable to carry out a proposed evaluation plan *because the program to be evaluated was not implemented as planned, or they could not obtain complete or reliable data on outcomes. In some cases, implementation problems were beyond the evaluators’ control, and resulted from decisions made by agencies providing program services after the study was underway.*” (draft report p. 3; emphasis added) However, GAO did not take this key point into consideration sufficiently when reaching its conclusions about the feasibility of attaining “conclusive results.” The implication is that through rigorous design and careful monitoring, conclusive evaluation results can always be achieved. While we wish it were so, as a practical matter we cannot share that level of optimism.

Second, GAO’s draft report reflects a strong commitment to implementing rigorously designed evaluations. Randomized control trials can provide strong evidence of program effects, and effectively control for spurious factors which, if unchecked, can confound interpretation of the evaluation results. However, randomized trials are not always feasible, and sometimes even non-random comparison groups are unavailable (as GAO notes on p. 40). In these cases, evaluators must choose from among other designs that have sufficient scientific rigor while also taking into account numerous factors such as data availability, cost opportunities for randomization, risk to subjects, likely effect size of the intervention, and the availability of appropriate comparison groups. We do not believe that the GAO sufficiently took this fact into account in its report or recognizes that these other methods of evaluation are valid means of scientific endeavor. In the last two decades, the evaluation field and its theorists have broadened their thinking about what constitutes “quality” evaluation. Increasingly, prominent leaders in the evaluation field are urging researchers to choose methods that are appropriate to the particular evaluation being conducted, and to take into consideration the context of each evaluation rather than utilizing the same set of methods and designs for all evaluations – a direct attack on the experimental design. The GAO report, with its strong emphasis on experimental and quasi-experimental designs, reflects an important view, but one that does not reflect current evaluation theory and practice.

As GAO notes in the draft report, many partner agencies have found NIJ evaluations a rich source of information to inform and guide criminal justice programs and policies. The NIJ will strive to demonstrate even greater value through future outcome evaluations.

The OJP appreciates the opportunity to comment on the draft report. Our additional specific comments are enclosed for GAO's consideration.

Sincerely,



Deborah J. Daniels
Assistant Attorney General

Enclosure

cc: Sarah V. Hart, Director
National Institute of Justice

Cynthia J. Schwimer
Comptroller, OJP

LeToya A. Johnson
Audit Liaison, OJP

Vickie L. Sloan
Audit Liaison, DOJ

OAAG Executive Secretariat
Control Number 20031746

Appendix IV: GAO Contacts and Staff Acknowledgments

GAO Contacts

Laurie E. Ekstrand (202) 512-8777
Evi L. Rezmovic (202) 512-2580

Staff Acknowledgments

In addition to the above, Tom Jessor, Anthony Hill, Stacy Reinstein, David Alexander, Michele Fejfar, Douglas Sloane, Shana Wallace, Judy Pagano, Kenneth Bombara, Scott Farrow, Ann H. Finley, Katherine Davis, and Leo Barbour made key contributions to this report.

GAO's Mission

The General Accounting Office, the audit, evaluation and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through the Internet. GAO's Web site (www.gao.gov) contains abstracts and full-text files of current reports and testimony and an expanding archive of older products. The Web site features a search engine to help you locate documents using key words and phrases. You can print these documents in their entirety, including charts and other graphics.

Each day, GAO issues a list of newly released reports, testimony, and correspondence. GAO posts this list, known as "Today's Reports," on its Web site daily. The list contains links to the full-text document files. To have GAO e-mail this list to you every afternoon, go to www.gao.gov and select "Subscribe to e-mail alerts" under the "Order GAO Products" heading.

Order by Mail or Phone

The first copy of each printed report is free. Additional copies are \$2 each. A check or money order should be made out to the Superintendent of Documents. GAO also accepts VISA and Mastercard. Orders for 100 or more copies mailed to a single address are discounted 25 percent. Orders should be sent to:

U.S. General Accounting Office
441 G Street NW, Room LM
Washington, D.C. 20548

To order by Phone: Voice: (202) 512-6000
 TDD: (202) 512-2537
 Fax: (202) 512-6061

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Web site: www.gao.gov/fraudnet/fraudnet.htm

E-mail: fraudnet@gao.gov

Automated answering system: (800) 424-5454 or (202) 512-7470

Public Affairs

Jeff Nelligan, Managing Director, NelliganJ@gao.gov (202) 512-4800
U.S. General Accounting Office, 441 G Street NW, Room 7149
Washington, D.C. 20548