The effects of a group batterer treatment program: A randomized experiment in Brooklyn

Bruce G. Taylor , Robert C. Davis & Christopher D. Maxwell

THE EFFECTS OF A GROUP BATTERER TREATMENT PROGRAM: A RANDOMIZED EXPERIMENT IN BROOKLYN*

BRUCE G. TAYLOR**
National Institute of Justice

ROBERT C. DAVIS***
Vera Institute of Justice

CHRISTOPHER D. MAXWELL****
Michigan State University

** Bruce Taylor is the deputy director of the Arrestee Drug Abuse Monitoring (ADAM) Program in the Office and Research and Evaluation at the National Institute of Justice (NIJ). His current intramural research projects at NIJ include evaluating a jail-based batterer treatment program through an experimental design; pilot testing a new domestic violence addendum for the ADAM Program; a multicity study examining HIV testing, risk, and prevention behaviors of arrestees as they relate to HIV and other sexually transmitted diseases; studies comparing the prevalence of illicit drug use among arrestees in seven countries; and studies examining different drug-testing kits and the validity of self-reports of drugs and alcohol. His most recent published research appears in *Women and Criminal Justice, Criminology, Violence and Victims, Journal of Personality and Social Psychology, and Journal of Drug Issues.* He received his doctorate from Rutgers University in 1996.

*** Robert Davis is a senior research associate at the Vera Institute of Justice, and a consultant for the American Bar Association. He is co-author of a recent book on crime prevention and principal editor of a recent book on crime victims. Currently he is working on a project funded by the Ford Foundation to promote democratic policing in emerging democracies. He is also conducting NIJ projects on innovative approaches to domestic violence prosecution and a national evaluation of victim services.

**** Christopher D. Maxwell is an assistant professor in the School of Criminal Justice at Michigan State University. His research interests include the social control and criminal justice processing of intimate violence, the efficacy of aggression and delinquency prevention programs, and the impact of social and ecological contexts on patterns of delinquency, crime, and criminal justice decision making. Current projects include several experiments testing criminal justice-based treatment programs for spouse abusers; a reanalysis of six experiments, each of which tested for the deterrent effect of arrest on spouse abusers; an assessment of how police handle domestic conflicts in the context of community policing; an evaluation of a community collaborative approach for response to domestic violence; and a study of the effects of various ecological contexts on individual-level judicial decision making. Dr. Maxwell is a graduate of Rutgers University (PhD, 1998; MA, 1994) and Indiana University-Bloomington (BA, 1990).

Despite numerous evaluations of batterer treatment programs, most lack sufficient methodological rigor to yield valid answers about the programs' effectiveness. This paper presents results from an experimental evaluation in which 376 adult males convicted of domestic violence were randomly assigned to either a 40-hour batterer treatment program or 40 hours of community service that did not include any therapeutic treatment. We examined both official records and victims' reports of recidivism. Those assigned to the treatment program showed significantly lower recidivism, on the basis of all outcome measures from official records. Although victims' reports also recorded fewer failures among the batterers assigned to the treatment group, the differences in failure rates were not large enough to be statistically significant. Overall results suggest that therapeutic treatment for batterers may reduce domestic violence among convicted batterers who agree to this sentence.

Therapeutic treatment programs for batterers became a popular court sanction beginning in the early 1980s. As state and national policies increasingly promoted arrest and prosecution to control domestic violence (Buzawa and Buzawa 1996), this sanctioning method has become more important. Because of these changes, criminal courts have sanctioned an expanding pool of batterers, and judges have relied increasingly on group treatment programs as their sanction of choice despite victims' desires or degree of willingness to cooperate (Hanna 1996; Rebovich 1996). By the late 1990s nearly every state used batterer treatment programs; administrators estimated that nearly 80 percent of their clients were referred by the courts (Healey, Smith, and O'Sullivan 1998).

Because of the substantial growth in the number of batterer treatment programs, it is important to understand how effective these programs are in changing batterers' behavior. Although the number of female victims of intimate violence declined in the 1990s, an estimated 840,000 women annually were assaulted by their current or former spouses or boyfriends (Greenfeld et al. 1998). Furthermore, by 1996 nearly 25 million women — one-fourth of the adult female population — reported that they had experienced some type of physical victimization during their lifetime by a family member or intimate partner (Tjaden and Thoennes 1998). Therefore an intervention that also reduces the likelihood of future domestic violence will benefit many women over their lifetime. In addition, because many victims stay with their partners even after the batterer's arrest and conviction, it is essential to use effective programs that can change abusive behavior rather than simply delaying it during a period of incarceration. Also, some observers have argued that batterer treatment groups have the potential to create a "ripple effect" throughout the criminal justice system (Dutton 1986). For example, if such groups are effective, police may be more willing to make arrests and take domestic violence seriously,

prosecutors may be more willing to proceed with domestic violence cases, and judges may be less willing to use incarceration as the only sanction.

More than three dozen evaluations of batterer treatment programs have been conducted (for a recent review, see Davis and Taylor 1999). Although these evaluations show an evolution toward more rigorous science since the first studies in the early 1980s, many still lack sufficient methodological rigor to yield valid answers about the effectiveness of the programs. Our study is one of the most recent attempts to test batterer treatment using an experimental design that randomly assigns court-mandated batterers to treatment or to a control condition. In this study we address some methodological problems found in prior research, such as disentangling the effects of treatment from sample selection effects.

PRIOR RESEARCH ON BATTERER TREATMENT

There is no lack of empirical studies on batterer treatment programs, as shown by at least six published reviews of more than three dozen published single-site evaluations (e.g., Eisikovits and Edleson 1989; Gondolf 1995; Rosenfield 1992; Saunders 1996; Tolman and Bennett 1990) and eight research reviews (Crowell and Burgess 1996; Davis and Taylor 1999; Dobash et al. 1995; Dutton 1988; Hamberger and Hastings 1993; Rosenbaum and O'Leary 1986; Saunders and Azar 1989; Tolman and Edelson 1995). This situation is deceptive, however, because only a handful of investigations can make any valid claim about differences between treated and untreated batterers. In the following section we review three generations of batterer treatment studies: Those which failed to use control groups and examined only batterers assigned to treatment programs, quasi-experiments, and those studies which assigned batterers randomly to treatment.

Studies Without Control Groups

The oldest and largest portion of empirical literature consists of studies that examine outcomes only among batterers assigned to treatment programs. This set includes studies that assess violence or other outcomes among participants in treatment programs only after treatment (single-group posttest-only designs), studies that measure violence among treatment participants both before and after treatment (single-group pretest/posttest designs), and studies that compare the violence of batterers who completed treatment with that of batterers who did not complete treatment, but were assigned to receive it. This literature contains more than two dozen examples. Studies of this type were important in developing this

area of research, and provided a foundation for the newer, stronger designs.

It is difficult, however, to interpret results from these nonexperimental studies. Single-group, posttest-only designs provide no reference point by which to judge whether treatment programs reduce violence. Single-group pretest/posttest designs that show a reduction in violence after batterers participate in a treatment program are problematic because research repeatedly has shown that domestic violence declines after the police are called, even if nothing else is done. In fact, research suggests that only about one-third of batterers commit another act of domestic violence within six months after the police intervene (e.g., Davis and Taylor 1997; Fagan et al. 1984; C.D. Maxwell 1998). In addition, among studies that compared outcomes between batterers who completed treatment and those who did not, the treated and the untreated (dropout) groups were not comparable before treatment. Palmer, Brown, and Barrera (1992) suggest that better attendance indicates a greater motivation to change, even before treatment.

Quasi-Experiments

The next generation of evaluation studies includes quasi-experimental designs using nonequivalent matched groups. In at least four studies, batterers whom the court mandated to treatment were compared with batterers who received other interventions (Chen et al. 1989; Dobash et al. 1996; Dutton 1986; Harrell 1991). These studies usually are more rigorous than those without control groups because they examine larger samples, do not rely only on batterers' self-reports to determine new violence, and include follow-up periods of at least one-year.

Among the four quasi-experimental studies, three reported lower violence among treated than among untreated batterers (Chen et al. 1989; Dobash et al. 1996; Dutton 1986). The effect sizes, however, were not always statistically significant. Also, they depended on which outcome measures were examined and whether comparisons involved all men assigned to treatment or only those who completed a requisite number of sessions. In contrast to these three studies, Harrell (1991) found that men who completed treatment were more abusive after treatment than men in a control sample.

Although quasi-experiments are a step forward from studies without control groups, Palmer et al. (1992) point out that they are not reliable, unbiased estimates of treatment effects because we cannot know whether batterers assigned to treatment and those in control groups were equivalent before the treatment. In some

quasi-experiments (such as Dutton 1986 or Harrell 1991), we know that the control group comprised batterers more prone to recidivate than those in the treated group; thus the study favored the finding of treatment effects. Researchers may try to control statistically for any pretreatment differences between groups; these controls, however, may not include key factors related to outcomes.

Randomized Experiments

The safest way to ensure that sample mean estimates are unbiased is by assigning batterers randomly to treatment and control groups. In 1992 Palmer and colleagues conducted the first batterers' treatment experiment, randomly assigning batterers to a true no-treatment control group. The number of subjects, however, was far smaller than was needed to detect treatment effects: 59 probationers were assigned by a "block random" procedure to either a 10-session psychoeducational group (combining group discussion with information) or a no-treatment control group. To measure outcomes, Palmer and her colleagues examined police reports six months after treatment; they found that recidivism rates (domestic physical abuse or serious threats) for the treatment group were just one-third those of the control group. Even with the small *N*, this difference was statistically significant. Although Palmer et al. (1992) attempted to measure violence on the basis of surveys of victims and batterers, low response rates and a small sample size precluded any analysis of recidivism based on interview data.

Dunford (1999) completed another randomized study. He assigned 861 domestic violence perpetrators who were in the Navy to one of four programs: group treatment for batterers, couples counseling, rigorous monitoring (periodic calls to victims and record checks), or safety planning for the victim (intended to approximate a no-treatment control). The results have not yet been published, however. Florida Atlantic University is conducting an experiment, but results are not yet available. In this study, the Broward County (Florida) domestic court judges randomly assigned more than 400 convicted male misdemeanor domestic violence offenders to either one-year probation and Duluth-like batterer treatment or a control group that received only one year of probation (Feder 1999).

PURPOSES OF THE PRESENT STUDY

We agree with Fagan (1996) that randomized experiments raise fewer questions about internal validity than other research designs. Our study adds to the literature on randomized experiments in several ways. First, our planned sample size was based on

an examination of effect sizes described in the literature. We designed the study to test the treatment's effects on several indicators of violence and attitudes, including victims' reports, which were not used by Palmer et al. (1992) because of the small number of surveys completed by victims.

Second, whereas the Palmer et al. (1992) experiment included all batterers sentenced to probation, regardless of the batterers' willingness to enter treatment, our study involved only cases in which prosecutor, defendant, and judge agreed that treatment was appropriate.[1] Although our results may not be as generalizable as Palmer's, our study did not include batterers who were completely unmotivated.[2] This is a key point because it has been argued (e.g., Rosenfeld 1992) that treatment cannot be expected to work for individuals treated against their will. Essentially the Palmer et al. (1992) study tested cases more typically referred to court-mandated batterer groups. Our study tests a subset of court-ordered batterer treatment cases: that is, men who agreed to a sentence including batterer treatment.

Finally, our experiment included both six- and 12-month follow-up interviews with the victims. The six-month interviews were important because any treatment effect may last only while a batterer is in treatment. The 12-month follow-up interviews were held to determine whether any initial effect of treatment would lessen after batterers were no longer in treatment but still were under court control.

METHOD

The study used an experimental design: 376 male criminal court defendants charged with assaulting their intimate female partners were assigned randomly to a 40-hour batterer treatment program or to 40 hours of community service without batterer education or treatment groups. In the community service program, participants were required to help clean local parks and public buildings.

---

1 The victims in these cases were not consulted as to whether they felt treatment was appropriate. In our study, however, the victim agreed in all cases to cooperate with the prosecutor and provide testimony. If they were not willing to testify against the defendants, the prosecutor generally dropped their cases and they did not reach our sample.

2 The men in our sample were not pure volunteers; they were offered a plea agreement that included batterer treatment. These men could have accepted or rejected the offer, but they would have gone to trial if they had rejected it. If found guilty, they would have faced a more severe sentence; if found not guilty, they might have received no sanction at all. Our recruitment process probably filtered out many batterers who would have been unmotivated treatment participants. Accordingly, we mean by the term *motivated* that the men in our sample were sufficiently motivated to choose treatment over an alternative, possibly more severe sanction.

Assignments were made during sentencing, after all parties (judge, prosecutor, and defendant) agreed to batterer treatment if it was available through the random assignment process. To provide several assessments, we interviewed the batterers and their victims about new violence on three occasions: at the time of sentencing, six months after sentencing, and 12 months after sentencing. In addition, 12 months after sentencing we searched police records for new complaints to the police by the victims and for arrests of the batterers.

Sample Characteristics

The sample consisted of males convicted of a misdemeanor domestic violence offense (although some originally were charged with felonious assault) in Kings County (Brooklyn, NY), in which the judge, the prosecutor, and the defendant agreed in principle to accept batterer's treatment if the Alternatives to Violence (ATV) program accepted the defendant. This sample represented a small percentage of about 11,000 domestic violence defendants adjudicated (i.e., with dispositions other than dismissal) during intake between February 19, 1995 and March 1, 1996. We selected a total of 376 cases for the experiment, about 1.5 cases per day.

Nearly two-thirds (64 percent) of the selected defendants were charged with third-degree assault, a class A misdemeanor. Another 19 percent initially were charged with felonious assault, although these individuals later pled to a misdemeanor charge. The remaining 17 percent were charged with violating restraining orders, menacing, harassment, and other offenses. Conditional discharges were the most common disposition (76 percent of the individuals in the sample).[3] Twenty-two percent of the cases were adjourned in contemplation of dismissal, a form of pretrial diversion in which defendants' cases are dismissed and their records are expunged if they avoid arrest and adhere to judicial conditions for six months. Two percent of the sample received probation. Conditional discharges and probation place defendants under court control for one year, compared with a period of six months for most adjournments in contemplation of dismissal. We also investigated whether the treatment and the control groups received the same quality and quantity of court supervision and found no difference in the supervision received by the two groups.[4]

---

[3] Overall, in misdemeanor domestic violence cases in Kings County Criminal Court, about 25 percent are dismissed, 45 percent receive an adjournment in contemplation of dismissal, 25 percent receive a conditional discharge, and about 5 percent receive probation or a jail sentence.

[4] We found no statistically significant difference between the groups in the percentage of cases receiving different types of supervision. Among those assigned

All of the batterers (see Table 1) were males with a mean age of 33 years. Thirty-six percent were African-American, 28 percent Hispanic, and 21 percent West Indian. The remaining 16 percent were either white, Asian, or other. About one-third reported not having a high school diploma, another one-third reported earning a high-school diploma or GED, and the other one-third reported some formal education beyond high school. Approximately two-thirds (64 percent) of the men reported employment (either part-time or full-time) at the time of their arrest; just 40 percent said they had been employed continuously during the past year. The mean household income was approximately $16,000 per year, 41 percent personally earned less than $7,000, and 10 percent earned more than $27,000.

Table 1. Batterers' Characteristics, by Assigned Treatment

	Total	ATV	Control
Total N of Assigned Batterers	376	190	186
Age (mean)	33.0	32.7	33.3
Race/Ethnicity			
African-American	36%	32%	40%
Hispanic	28%	30%	26%
Western Carribean	21%	19%	22%
White, Asian, or other	16%	19%	12%
Education			
No diploma/GED	38%	36%	40%
High school diploma/GED	31%	31%	31%
More than high school	31%	33%	30%
Employed at Time of Arrest	64%	64%	63%
Employed Throughout Year	40%	45%	34%
Household Income (Mean)	$16,300	$16,500	$16,100
Personal Income			
No income	19%	20%	19%
$1.00 to $7,500	25%	28%	22%
$7,501 to $17,000	28%	28%	28%
$17,001 to $27,000	17%	16%	19%
Over $27,000	10%	9%	11%
Relationship Status			
Married	43%	44%	41%
Separated/divorced	8%	8%	7%
Living together	19%	18%	20%
Intimate	9%	9%	9%
Past intimate	25%	26%	25%
Other	2%	2%	3%
Prior Arrest	39%	42%	37%

Note: All reported data (except for information on prior arrest) are taken from the baseline interview with the batterers.

to community service, about 77 percent received a conditional discharge, 22 percent received an adjournment in contemplation of dismissal, and 1 percent received traditional probation. For those assigned to the Alternatives to Violence (ATV) treatment groups, about 77 percent received a conditional discharge, 21 percent received an adjournments in contemplation of dismissal, and 2 percent received traditional probation.

Finally, slightly more than one-third (43 percent) of the men said they were married to the victim, another 25 percent reported that the victim was a past intimate partner, and 19 percent reported that they were currently living with, but not married to, the victim. The remaining men reported that they were married, but separated or divorced from the victim (8 percent), or that the victim was a current intimate partner with whom they did not live (9 percent). The remaining 2 percent did not specify a relationship status with the victim. Overall, a typical subject in this sample was a male around age 30, with no prior criminal history, little more than a high school diploma, some but not consistent employment, and a household and personal income of about $16,000 per year. Typically, he was married to or living with his victim at the time of his arrest.

All the victims were females, with a mean age of 29 years. Six in 10 were African-American (59 percent), 30 percent were Hispanic, 9 percent were white, and 2 percent identified themselves as belonging to some other racial group. The proportion of victims who graduated from high school (66 percent) was comparable to the proportion of high school graduates among batterers. Fewer victims, however, were employed (38 percent), and a large proportion (43 percent) received public assistance. Just 9 percent of the victims reported in the initial interviews that the batterer was their primary source of financial assistance. Victims were poorer than batterers: close to half (46 percent) reported annual household incomes under $10,000. About two-thirds of the victims and the defendants lived together at the time of arrest (62 percent according to victim interviews, but 70 percent according to batterer interviews). A majority of the victims reported having children with the batterer (79 percent according to victims; 63 percent according to batterers).

The median length of the relationship with the batterer was 5 1/2 years. Victims reported in the initial interview that the violence began, on average, in the second year of the relationship. Sixty-two percent of victims said they had previously called the police because of their perpetrator's abuse, 48 percent had filed a police complaint against their perpetrator in the past, 34 percent reported having had an order of protection against their perpetrator, and 23 percent stated that the perpetrators had been arrested in the past for abusing them. Thus the majority had been abused previously by the batterer. Until the current incident, however, only about one-third had witnessed their batterer formally sanctioned with a restraining order or arrest.

Assigned Treatments

The effects of batterer treatment programs can be assessed in several ways. One way is to compare treated groups with those not treated. For example, some judges leave it to probation officers' discretion to assign treatment or not at the time of probation intake. This method is used in Feder's (1996) experiment. That option was not available to us because probation for misdemeanor spouse abuse charges is rare in New York City (only 2 percent of our sample received a regular probation sentence), and judges mandate batterers to treatment; program completion usually is the only condition of plea arrangements. It was not possible to suggest to criminal justice officials that they allow recruited defendants to receive no formal sanctions beyond arrest and conviction. We needed an alternative sanction for the control group, preferably one which does not appear to be related to preventing domestic violence and one without a therapeutic regimen. Community service, as defined below, was such a sanction, and criminal justice officials agreed to use it as an alternative to batterer treatment for men whom we designated as controls. All participants in our experiment were assigned to receive either 40 hours of group batterer treatment or 40 hours of community service.

We evaluated Victim Services' Alternatives to Violence (ATV) program, which operates in New York City and is based on the Duluth model. This model, rooted in a feminist perspective, assumes that domestic violence is a by-product of male and female sex roles, which result in an imbalance of power. The tested curriculum included defining domestic violence, understanding the historical and cultural aspects of domestic abuse, and reviewing criminal/legal issues. Through a combination of instruction and discussion, participants were encouraged to take responsibility for their anger, actions, and reactions. One male and one female leader conducted sessions in either English or Spanish. The program mandated 40 hours of attendance at weekly group meetings over at least eight weeks.

Judges required defendants rejected by lottery from batterer treatment to participate in 40 hours of community service over a two week period.[5] Participants were assigned to renovate housing

---

5 We found variation in the amount of time participants needed to complete either community service or batterer treatment. Community service usually was completed in about a month; batterer treatment took at least two months. Participants, however, still were under some court supervision before their six-month or one-year sentence expired, and could face a harsher penalty if they violated the terms of their sentence (e.g., a protection order). One could argue that even though the numbers of mandated hours were equal, attending the batterer program over a longer period might make a difference. If this theory were correct, we would expect the difference in recidivism rates between the two groups to dissipate over time.

units, clear vacant lots for community gardens, paint senior citizen centers, and clean up playgrounds; none of these tasks would be expected to affect abusive behavior.

Participants in both batterer treatment and community service programs were expelled from the programs if they developed a pattern of nonattendance: three misses constituted grounds for dismissal from the ATV program. Such cases were referred to the prosecutor's office for action. At the district attorney's discretion, delinquent cases were returned to the court calendar, and new sentences could be imposed.

Assignment Process and Case Intake

Cases were drawn from three of eight postarraignment parts in Kings County Criminal Court. Two of the parts were devoted specifically to domestic violence cases. The third was the jury trial part, to which domestic violence and other cases were transferred if the parties could not reach a negotiated disposition. When the judge, the prosecutor, and the defense agreed the batterer treatment was appropriate, the prosecutor called the ATV office in the court building. The ATV intake person or the research assistant picked up the defendant in court and brought him to the ATV office for an intake interview.

After the defendant completed the interview, his name and case identifier were written on a new line of a logbook. Each line carried a preassigned designation (batterer treatment or community service), set by using a random number table. The use of the log with predetermined treatment assignments and the presence of a research assistant on the three busiest days of the week helped ensure the integrity of the random assignment process. Defendants assigned to batterer treatment were given a start date (usually within a week of intake) and received directions to the class. The defendant was accompanied back to the courtroom, and the prosecutor was told of the lottery assignment. The prosecutor then informed the judge, who in turn disposed the case in keeping with the assignment.

To test this possibility we examined a slightly different version of our time-to-failure/survival model (see Table 4). In a model the same as that in Table 4 but with the addition of a time-dependent covariate, this additional variable of time was not significant. Also, in our other prevalence models for officially recorded failures (see Table 4), significant findings are present at six and 12 months postsentencing. If the length of time to complete the program exerted an effect, differences that existed at the six-month follow-up might disappear at 12 months.

Experimental Misassignments

Treatment misassignments are a major problem in conducting randomized experiments in the field (Berk, Smyth, and Sherman 1988): that is, cases which the subjects do not receive the treatment or control condition to which they are randomly assigned. The judges in our study could "override" cases when they felt that following the random assignment process could cause egregious harm. Judges overrode 14 percent of our random community service assignments, and instead mandated the ATV program. In no case did the judge override a treatment assignment and reassign the defendant to a control group. Another problem is the situation of those who start a treatment program but do not finish and do not receive full treatment. This occurred in our sample. When either or both these problems arise, one is tempted to make comparisons based on delivered treatment instead of assigned treatment. This solution may appear to be fairer because the treatment of interest will be tested with only those who receive the full dosage.

However, as in procedures used in analyzing the Spouse Assault Replication Program data, we chose to "analyze as randomized" rather than according to the treatment received by perpetrators (see Garner, Fagan, and Maxwell 1995). Analysis of cases according to actual treatment received defeats the purpose of randomizing cases: that is, creating equivalent groups of cases before treatment. In our case, the crossovers were created because judges intervened in the random assignment process. Gartin (1995) makes another argument for analyzing as randomized. In policy studies such as ours, he argues, the issue is not the effect of the treatment per se, but the effect of a policy to apply treatment.

Nevertheless, we considered this a serious issue and performed several tests to assess the extent of possible bias in our results. As one potential test of bias, Sherman (1992) proposed following the "analyze as randomized" dictum as long as the proportion of treatment crossovers does not exceed the proportion of cases with negative outcomes. In our study the crossover rate was 14 ($n = 53$), which did not exceed our one-year combined crime report rate of 17 percent nor the one-year, victim-reported recidivism rate of 19 percent. Berk et al. (1991) recommended conducting tests for any pattern to misassignments, so as to understand what types of subjects the judges may have shifted from the assigned treatments. Similarly, Berk et al. (1988) suggested that researchers generate a statistical model of the misassignment process.

In Table 2, under the heading "Treatment Overridden," we report our effort to predict misassignment through a logistic regression model with five predictor variables (age, ethnicity, marital

status, employment, and prior arrest history). None of the predic-
tor variables approached statistical significance; the percentage of
override cases, however, was fairly small (14 percent). Thus, al-
though the judge's abrogation of the random assignment in a mi-
nority of cases was not likely to be a random process, we believe
that the number of missassignments did not measurably skew the
equivalence we had sought through randomization.

Interview Procedure

We attempted to interview victims on three occasions: at case
intake or date of court disposition, at six months after intake, and
at 12 months after intake. We made the first attempt at all victim
interviews by telephone. When telephone attempts failed, inter-
view teams went to victims' homes. If the home interview attempts
failed, we mailed letters offering first $25 and then $50 for an inter-
view. At the third stage, we also turned over 70 difficult cases to a
licensed private investigator; the investigator found five more
victims.

We took a number of steps to ensure participants' safety. First,
to guarantee confidentiality, the research staff described the project
to nonparticipants as a health study. Second, batterers were not
present during interviews. The research staff conducted interviews
at times that were convenient and safe for the victim, and at local
community centers or at Victim Services offices if it was deemed
unsafe to conduct the interviews in the victim's home or over the
phone. Finally, professional counselors were available to assist the
participants if needed.

Our completion rate for the victim surveys was 51 percent for
the first interview (not shown in Table 2), 48 percent for the second,
and 50 percent for the third. In addition, completion rates by as-
signed treatment groups did not differ significantly (51 percent vs.
50 percent at time 1; 41 percent vs. 50 percent at time 2; 52 percent
vs. 48 percent at time 3). We were unable to contact 131 victims (35
percent) at any time during the follow-up period. In many cases,
the victims had moved. Research in other cities (see Davis, Smith
and Nickles 1997) with court-involved domestic violence victims
has shown that this population is highly transient: many of those
who were staying with the batterer or with family members at the
time of arrest moved shortly thereafter. Among the victims we lo-
cated, 7 percent refused to take part or terminated the interview
early. Inaccurate or outdated information obtained from prosecu-
tors' files was the primary reason for not conducting interviews
with victims.

Table 2. Implementation of Experimental Design

	Treatment Overridden		Completed Victim Interviews			
			6 Months		12 Months	
Base Rate	14%	(53)	48%	(171)	50%	(186)
Logistic Regression	b	Exp(B)	b	Exp(B)	b	Exp(B)
Age	.01	1.01	−.00	1.00	−.01	.99
Ethnicity (African-American)						
Hispanic	−.28	.76	.54	1.72*	−.46	.63
West Indian/Caribbean	−.06	.95	−.27	.76	−.46	.63
Other race	−.35	.70	−1.26	.29***	−.78	.46*
Married	.52	1.68	.41	1.51	.22	1.25
Employed	−.45	. 64	.22	1.24	−.07	.93
Number of Prior Arrests	.07	1.07	.02	1.02	−.01	.99
ATV Treatment Assigned			−.32	.73	.23	1.26
Police Recorded Failure Before Interview			−.22	.80	.07	1.07
Intercept	−2.18***		−.52		.11	
Initial Log-Likelihood	305.84		518.17		521.24	
Final Log-Likelihood	298.96		496.23		511.59	
P Value	.44		.01		.38	

*$p < .05$; **$p < .01$; ***$p < .001$

We also examined variation in victims' interview completion rates (see Table 2) by the batterer's age, ethnicity, marital status, employment status, and prior arrests. We used batterer characteristics because they were available for virtually the entire sample and because they have been the primary control variables used in other research on interventions to prevent domestic violence. In addition, we examined variation in victim interview rates by assignment to treatment versus control group and by whether the police had recorded any failure before the interview. We found no significant differences in interview completion rates as a function of any these variables, except for ethnicity (see Table 2). We were more successful in interviewing Hispanic victims ($b = .54$; $p < .05$) than African-American victims for the six-month victims interviews, but the completion rate was higher for African-American victims than victims from one of the "other" racial groups (mostly whites and Asians) at both six months ($b = -1.26$; $p < .001$) and 12 months ($b = -.78$; $p < .05$). We also examined the effects, on interview completion rates, of the batterer's income and education, and the relationship of the victim to the batterer. These variables, not shown in Table 2, were not significant.

Reviewers of the batterer treatment literature (e.g., Edleson 1996; Gondolf 1997) have stressed the importance of obtaining high response rates with respondents. Our follow-up attrition rate of

about half the participants clearly falls short of this goal. To investigate the impact of the attrition, we conducted a number of analyses and found no measurable differences between participants and nonparticipants other than the victim's ethnicity. Neither this analysis nor any of our statistical corrections, however, can rule out the possibility of unmeasured differences between these groups, which might interact with treatment assignment to produce a biased comparison test.

Measures of Recidivism

We collected data from several sources to develop multiple indicators of new violence by the batterer against the victim, including arrest reports, crime complaints (which may or may not have resulted in an arrest), and victims' reports of violence by the batterer. These sources have been used in other domestic violence studies that track batterers, such as The Spouse Assault Replication experiments conducted by the National Institute of Justice (see Garner et al. 1995). Other research has shown that it is important to capture outcomes from a variety of sources and along several dimensions because different outcome measures do not always behave similarly (Davis and Taylor 1997). From our data we constructed four recidivism measures: prevalence, rate or frequency of failures, severity, and time to the first failure. Our study included both six- and 12-month postsentencing measures. This combination of data sources, measurement dimensions, and time intervals potentially could produce 24 basic outcomes or recidivism models. Because of limitations in data collection, however, we obtained only enough information to construct 13 outcome models (see Tables 4 to 6).

Victim self-report surveys. To assess frequency and severity of violence through victim interviews, we employed Harrell's (1991) revision of the Conflict Tactics Scale (Straus 1979).[6] Harrell's scale measures the frequency of 11 violent acts.[7] In the outcome models, we examined the combined prevalence and frequency of the 11 violent acts and the prevalence and frequency of the six most violent of

---

[6] This study was limited by budgetary constraints on survey length; thus we could not measure the psychological and emotional abuse that some batterers may have adopted to avoid future arrest.

[7] (1) Forced you to have sex; (2) choked or strangled you; (3) threatened to kill you; (4) beat you up; (5) threatened you with a knife, gun, or other weapon; (6) used a knife, gun, or other weapon against you; (7) threw something at you; (8) pushed, grabbed, or shoved you; (9) slapped or spanked you with an open hand; (10) kicked, bit, or hit you with a fist; and (11) hit or tried to hit you with something.

these acts (our severe victimization measure).[8] The previous two months were the reference period for the scale (as opposed to the previous six months for the criminal justice measures). We reasoned that, if treatment did make a difference, the effect would take some time to appear. Thus, by asking victims to report at the six-month interval about the entire period, we could include reports of violent incidents committed shortly after batterers were assigned to treatment. By deciding on the two-month reference period, we also ensured that any reported violence would have occurred after batterers had completed most of their 40 hours of treatment. Unfortunately, as in other studies, we interviewed victims involved in the triggering court case, but not any new partners the batterers may have started to see.[9]

Criminal justice records. We searched computerized records of the Criminal Justice Agency (CJA) and the New York City Police Department (NYPD) to determine whether a new crime report was filed or the batterer was arrested during the study period. We accessed CJA's database of New York City arrests via the court docket numbers of cases. Docket numbers led us to defendants' state criminal identification numbers, which we used to determine whether the defendants were arrested during the 12 months after sentencing. When new incidents were found, we recorded the arrest date and the charge. In addition, we searched the district attorney's computer database, using the docket number to determine whether the victim in the new incident was the same as the victim in the original incident. Because the searches used ID numbers, we are confident that our information on new arrests is highly accurate.

We also searched the NYPD's computerized records to determine whether new crime complaints had been filed against the defendant since sentencing. These searches, conducted by NYPD personnel, used the batterer's name and incident addresses. Therefore the searchers were subject to some error whenever the batterer's name or the street names were spelled incorrectly in the database. Also, each police precinct maintains its own database. When batterers commit a crime outside their home precinct, their home precinct is supposed to receive a record, but we do not know

---

8 (1) Forced you to have sex; (2) choked or strangled you; (3) threatened to kill you; (4) beat you up; (5) threatened you with a knife, gun, or other weapon; (6) used a knife, gun, or other weapon against you.

9 The failure to interview all partners may result in an underreporting of domestic violence incidents. Such underreporting, however, should affect only the treatment comparisons if batterers in one of the treatment groups were more likely to find new partners than batterers in the other group. The major obstacle to interviewing these new partners is the great expense of finding them and persuading them to agree to the interview.

how reliably information is transferred. When we found new incidents, officers recorded the date, the nature of the complaint, and whether the complainant was the victim in the original case. As a result of these limitations, we suspect that the NYPD data undercounted victims' complaints. We have no reason to believe, however, that the undercounting would vary according to treatment assignment.

To further reduce measurement error, we merged the CJA arrest reports and the police complaint data into one measure that captured the number of documented criminal justice incidents (e.g., arrests or crime complaints) involving both the defendant and the victim after the treatment was assigned. In this step we followed the procedure used by Maxwell (1998) in the most recent reanalysis of the Spouse Assault Replication Program experiments.

Correcting for Missing Information

Much (if not all) research in behavioral, economic, and social science is plagued by missing information (Winship and Mare 1992). There are two types: item nonresponse and unit nonresponse (Little and Schenker 1995). In the former, missing information is unobserved or unmeasured information on one or more variables for a subset of cases in a database; often it occurs because researchers do not record certain responses or because participants fail to provide certain responses (Weisberg 1985). In the second type of missing information, cases included in a study represent nonrandom samples of a population. Unit nonresponse is often known as sample selection bias (Dubin and Rivers 1989). Unlike the first type, unit nonresponse typically is created when subjects act so as to make it impossible for the researchers to obtain their responses. The nonrandom selection of cases from the entire population into a study is itself a social process and an aspect of social science that is often overlooked (Winship and Mare 1992). Our experiment presented two occasions for sample selection bias: one at the six-month victim interview and another at the 12-month interview.

We addressed both types of missing information. The first problem was missing information for some cases on the nonexperimental covariates. These covariates in the recidivism models were taken from our baseline batterer data set. Although we had conducted baseline interviews with 95 percent of the batterers, not all these men answered all the questions: 22 cases were missing age information; 24 lacked employment status; and 25 were missing ethnicity and marital status. To address this problem, we used imputed quasi-valid values for the missing data, employing a multiple

regression procedure. We constructed a regression model that computed a predicted value for all cases on the basis of cases with valid data, and then used these predicted values to replace the remaining missing data.

To handle sample selection bias or missing victim interviews, we tested a two-step process proposed by Heckman (1979). The first step was to specify a model using a multiple regression of the selection process that would be captured in a single latent measure. We used two different models, one for the six-month interview and another for the 12-month interview (see Table 2). As stated earlier, the batterer's ethnicity was the only statistically significant measure in either the six- or the 12-month victim interviews. We estimated the experiential outcome models with a traditional covariate for sample selection bias included along with the assigned treatment covariate. The addition of this covariate did not change the relationship between the treatment and the outcome measures in any of the experimental models.

To address differences across the treatment groups and potential sample selection bias, we used another method: replacing the sample selection covariate with a set of substantively interesting control measures. Introducing covariates in analyzing data from a randomized experiment is not strictly necessary: randomization should ensure that other (known and unknown) measures that are related to the failure measure, such as the suspect's age or prior criminal record, are distributed similarly across the study groups. In our study, this similarity apparently was achieved. (See our discussion of pretreatment comparisons between the treatment and the control group.)

Nevertheless, introducing covariates is increasingly common in analyzing data from randomized experiments (Patel 1996) for several reasons. First, statistical controls for other factors tend to improve the precision of the treatment comparisons and to correct for any major imbalance in the distribution of these measures across treatments that may have occurred by chance (Armitage 1996). Second, because the batterers assigned randomly to the same treatment group are not exactly alike, statistical controls can address the natural variations between batterers within each study group (Gelber and Zelen 1986). Third, an experimental analysis typically tests only for the average effects of treatment across all batterers, whatever their characteristics. Additional nonexperimental hypotheses, however, can specify other expected direct effects on the outcomes, such as age, and how treatment effects may vary across dimensions of other uncontrolled extraneous factors such as marital status, employment level, or prior criminal record.

The tests for the additional direct effects follow each of the regression models that test for only the direct effects of treatment. We used the following nonexperimental measures in our models: defendant's age, ethnicity, relationship status with the victim, employment level, and prior arrests. All of these measures have been associated with general offending patterns (Blumstein et al. 1986) as well as with violence between intimates (Fagan and Browne 1994; Fagan, Garner, and Maxwell 1997; Hotaling and Sugarman 1986).

RESULTS

We begin by examining the official criminal justice measures of prevalence (new police complaints or arrests involving the same perpetrator and the same victim) at six months and at 12 months postsentencing. Later we examine our victim-based measure of prevalence — the prevalence of new reports of violence made by victims during research interviews — at the six month and the 12-month follow-up postsentencing. We conducted the prevalence comparisons using logistic regression.

In the next set of analyses we examine incidence/frequency of failure, using a count model (negative binomial regression) that tests whether the distribution of failures (cases in which new violence occurred) differed according to treatment. We used negative binomial regression to model the distribution that we observed: a large majority of the sample did not fail at all during the time observed, some failed once, fewer failed twice, and only a handful failed more often. This kind of highly skewed distribution violates the normality assumption of ordinary least squares, even with log or other data transformations.

In the final set of analyses, using proportional hazard models, we examined differences between treatment conditions in time elapsed to first failure. Using a time-to-failure model, we could address the question of timing not only at six and at 12 months, but at any point during the follow-up period. We have only one time-to-failure model based on the official recidivism measures, because we did not collect all of the information during the victim interviews that we needed to construct this measure using the data reported by victims.

In addition, for the final three tables (4, 5 and 6) we estimate two models. The top rows in each table present the experimental model, or our type 1 models. These initial models specify only the treatment assignment measure. The second set of rows presents the full model, or our type 2 models. These final models specify a

more complete set of structure measures, along with the treatment assignment measure.

Table 3 presents the distribution of officially recorded offenses and the victim-reported incidents by assigned treatment. As reported there, the failure rate in the ATV treatment group is significantly smaller (about 50 percent), based on officially recorded incidents, at six months ($M = .38$ for the controls, and $M = .16$ for the treatment group; $f = 8.09$; $p < .01$) and again at 12 months ($M = .55$ for controls, and $M = .28$ for treatment group; $f = 6.82$; $p < .01$). On the basis of victims' reports, four out of four comparisons again show lower failure rates among the ATV treatment group; not one of the four rate differences was significant, however. Thus, on the basis of these six comparisons, the treatment group showed positive change (e.g., reduction in the mean number of failures), but only two of the six positive effects reach statistical significance. In the remaining analysis we focus more closely on these experimental comparisons by using appropriate regression techniques, modeling different outcome dimensions, and controlling statistically for sample differences.

Table 3.   Frequency of Outcome Incidents, by Assigned Treatment

	Officially Recorded Incidents				Victim-Reported Incidents							
	6 Months		12 Months		6 Months				12 Months			
					Any		Severe		Any		Severe	
	% Control	% ATV	% Control	% ATV	% Control	% ATV	% Control	% ATV	% Control	% ATV	% Control	% ATV
0	79	90	74	84	79	78	80	86	78	85	86	93
1	12	6	13	10	4	8	3	6	7	7	6	2
2	4	3	7	5	5	6	9	3	4	4	1	2
3	4	1	3	1								
4	1	1	2	1								
5	1		1		7	5	4	3	8	3	4	2
6			1									
7			1	1								
8					5	3	4	3	1		1	
9												
10+									2	1	2	1
Mean	.38	.16	.55	.28	.90	.67	.76	.45	.99	.46	.74	.32
N	186	190	186	190	93	78	93	78	90	99	90	99
Chi-Square	11.10*		8.97		1.86		4.50		3.86		4.53	
F-Test	8.09**		6.82**		2.37		4.20		12.96		8.36	

$*p < .05$; $**p < .01$; $***p < .001$

Criminal Justice Recorded Failures

Prevalence of official failures. The first results presented in Table 4 detail the prevalence of officially recorded failures based on a logistical regression using the classic experimental analysis (our type 1 models). That is, we compare the proportion of the treatment group

failing with the proportion of the control group failing, without controlling for other independent variables. Our six-month type 1 model shows, as expected, that treatment assignment was associated with a lower prevalence of officially recorded failures (b = $-.97$); this difference is statistically significant with a two-tailed test ($p < .001$). The difference or effect size is also substantively large: the model estimated that 59 percent fewer than the expected number of subjects assigned to the ATV treatment group committed a recorded offense against the victim ($Exp(B) = .38$).

The full model (type 2) for the six-month results builds on this first model by adding additional control measures to account for the natural heterogeneity between and within the two comparison groups. Again, in terms of the proportion of officially recorded failures, we found a statistically significant difference between the treatment group and the control group ($b = -.86; p < .01$). The sizable reduction in the proportion of batterers with any incident within six months remains the same in the experimental group. No other batterer characteristics were related significantly to a change in the prevalence of officially recorded complaints.

In addition, whether a victim interview was completed before we searched the criminal history files was not statistically significant in this six-month model (nor in any of the models reported in Tables 4, 5, and 6). Some observers voiced concerns that interviewed victims may constitute a different kind of sample, or that interviews could lead to more violence; thus we were pleased to find no significant effect from the interview measure. Therefore this set of nonsignificant findings (in Tables 4 to 6) tends to support the idea that the victims who completed an interview were not significantly different, in terms of victimization risk, from those who did not .

In the 12-month follow-up using the experimental model, we found results nearly identical to those of the six-month model. This later model showed again that treatment was associated with a significantly lower prevalence of officially recorded failures ($b = -.61, p < .01$). At 12 months, however, the difference in incident rates between the two experimental groups had diminished (from $Exp(B) = .41$ to $Exp(B) = .55$). This suggests that the early effect of treatment may lessen with time. Similar results from the type 2 model for the 12-month failure measure showed a statistically significant difference between the treatment group and the control group ($b = -.59; p < .05$). In another parallel with the six-month analyses, none of the additional control measures introduced in the type 2 model were significantly correlated with the likelihood of any failure within 12 months.

Table 4.   Officially Recorded Incidents Since Assigned Treatment

Type 1	Prevalence[a]				Rate[b]				Time to Failure[c]	
	6 Months		12 Months		6 Months		12 Months		12 Months	
	b	Exp(B)	b	Exp(B)	b	Exp(B)	b	Exp(B)	b	Exp(B)
ATV (assigned)	−.97	.38***	−.61	.55*	−.84	.43**	−.70	.50*	−.55	.58*
Alpha					−.98***		−.43			
Intercept	−1.29***		−1.03***		4.24***		3.46***			
Initial Log-Likelihood	−163.38		−194.62		−262.99		−350.09		924.42	
Final Log-Likelihood	−158.59		−191.78		−229.07		−295.26		918.54	
P value	.00		.02		.00		.00		.02	
Type 2										
ATV (assigned)	−.86	.42**	−.59	.55*	−.84	.43*	−.70	.50*	−.53	.59
Prior victim interview	−.02	.98	−.19	.83	.13	1.14	.02	1.02	−.20	.81
Age	.02	1.02	.01	1.01	.00	1.00	.00	1.00	.01	1.01
Ethnicity (African-American)										
Hispanic	−.61	.54	−.33	.72	.56	.1.75	−.41	.66	−.35	.70
West Indian/Caribbean	−.68	.51	−.64	.53	−.66	.52	−.64	.53	−.55	.58
Other race	−.75	.47	.78	2.18	−.27	.76	−.33	.72	−.70	.50
Married	.23	1.26	.12	1.13	.34	1.40	.28	1.32	.12	1.13
Employed	−.54	.58	−.28	.76	−.25	.78	−.27	.76	−.28	.76
Prior arrest	.25	1.28	.36	1.43	.22	1.25	.18	1.20	.38	1.46
Alpha					−.93		−.43			
Intercept	−1.46		−.94		3.74**		3.46***			
Initial Log-Likelihood	−163.38		−194.56		−254.68		−350.09		924.42	
Final Log-Likelihood	−152.03		−186.56		−225.51		−295.26		906.10	
P value	.01		.06		.00		.00		.03	

[a]Logistic Regression
[b]Negative Binomial Regression
[c]Cox Regression
*$p < .05$; **$p < .01$; ***$p < .001$

Frequency of official failure. As in the earlier logistical regression, the type 1 or experimental model of the frequency of failure at the six-month follow-up showed results that were both in the expected direction (treatment lowers frequency of failures) and statistically significant (b = -.84; p = < .01). Once again, the difference was also substantively large. Over the first six months, the batterers assigned to the treatment group experienced 57 percent fewer incidents than the control group (Exp(B) = .43). The six-month type 2 model (see Table 4) similarly showed a statistically significant difference between the treatment group and the control group in the frequency of officially recorded failures (b = -.84; p < .05). Also, as in the early logistic regression models, no other significant control variables correlated with a change in the frequency of officially recorded complaints.

Corresponding results also emerged for the 12-month count models (see Table 4). The type 1 model for the 12-month results showed that treatment is associated with a significantly lower frequency of officially recorded failures (b = -.70; p = .05). As in the prevalence models, however, the size of the effect was diminished somewhat by 12 months: batterers assigned to the treatment group experienced a 50 percent reduction in incidence compared with the control group (Exp(B) = .50), as opposed to a 57 percent reduction at

six months. Finally, the type 2 model for the 12-month results continued to show a statistically significant difference between the treatment group and the control group in the frequency of officially recorded failures ($b = -.70$; $p < .05$). Once again, none of the covariates were statistically significant.

Time to first official failure. Our final analyses using the official data, reported in Table 4, examine the timing of the first officially recorded incident after the treatment assignment. These results are based on a Cox regression that modeled the hazard of time to first new officially recorded offense if it occurred within the first 12 months of follow-up. As in the earlier results, the type 1 model showed a significant decrease in the hazard rate for the treatment group compared with the rate for the control group ($b = -.55$; $p < .05$). In other words, the batterers assigned to the treatment group registered a 42 percent reduction in the likelihood of a new incident on any given day after the treatment assignment ($Exp(B) = .58$).[10] The type 2 model also showed a statistically significant difference between the treatment group and the control group in time to first failure ($b = -.53$; $p < .05$). Once again, the control measures were not related significantly to the hazard rate.

Overall, the analysis based only on the official recorded data reveals consistently less recidivism among batterers assigned to the ATV treatment group. Batterer treatment seemed not only to reduce the prevalence and frequency of officially recorded recidivism over 12 months, but also to create a consistent period of greater safety for victims during the first year of follow-up. In other words, the positive treatment effect found at six months did not diminish significantly at 12 months. Also, the batterer treatment was not trivial in reducing the number of incidents between the batterer and the victim known to the police. Depending on the measure and the length of follow-up, the effect of the ATV treatment varies between 41 percent and 59 percent less recidivism than would be expected. Across the 10 models based on criminal justice records, the weighted average was 50 percent less than expected.

Incidents Reported by Victims to Research Interviewers

Early in Table 3 we saw that the treatment and the control groups had virtually the same prevalence and incident rates for

---

[10] We conducted a Cox regression with a treatment by time-dependent interaction term to test for the proportionality of the hazard rates over the first 12 months of follow-up. The coefficient for the interaction term was about .0001 and was nonsignificant in both the experimental and the full models. Therefore we removed the interaction term from the analysis to simplify the interpretation and presentation of the results.

"any" victimization as reported by the victims at both six months (treatment = 78 percent no victimization, .67 mean; controls = 79 percent no victimization, .90 mean) and 12-months (treatment = 85 percent no victimization, .46 mean; controls = 78 percent no victimization, 0.99 mean). We also found similar nonsignificant differences between the control and the experimental groups for the "severe" victimization measure (see Table 3). In this final set of analyses we examine the victim interview data, using a similar set of regression models to those used for the official recorded failures.

Six-month treatment effects. Table 5 presents results of logistic and negative binomial regression for the prevalence and frequency of victim-reported failures (any victimization or severe victimization) by the six-month follow-up interview. The results from the type 1 model, which estimated the prevalence of any victimization, showed that the effect of the treatment measure was slightly positive but not significant. The type 2 model, with the additional control measures, also showed that the treatment variable was nonsignificant, although the estimated reduction in the prevalence of victimization was about 16 percent. Yet in contrast to the earlier models based on the official data, with the type 2 model we found that prior arrest was statistically associated with a rather sizable increase in the risk of any repeat victimization ($b = 1.03$; $p < .01$). The batterers with a prior arrest history were 2.8 times more likely to fail on this measure than batterers without any such history.

With regard to severe victimization, the type 1 and type 2 models in Table 5 produce results similar but not identical to those found in the "any victimization" models. The treatment assignment measure was still nonsignificant in both models, but in both instances the effect was negative and the reduction was not small. In the full model, the reduction in prevalence was estimated to be slightly more than 50 percent of the base rate. In addition, the results from the full model for severe victimization showed not only that prior arrest was related significantly to an increased risk of victimization ($b = 1.15$; $p < .01$), but also that Hispanics were 3.4 times more likely than African-Americans ($b = 1.21$; $p < .05$) to severely assault their intimate partner.

Table 5 also presents the victim-reported frequency (negative binomial) of "any victimization" and "severe victimization" at six months. The two type 1 models show that treatment assignment was associated with a lower frequency of victim reports of both "any victimization" ($b = -.30$) and "severe victimization" ($b = -.53$); neither of the two coefficients was statistically significant, however. We found nearly identical results for the two full models. Findings

Table 5. Victims' Reports of Incidents at Six Months (n = 171)

| | Prevalence[a] | | | | Frequency[b] | | | |
| | Any | | Severe | | Any | | Severe | |
Type 1	b	Exp(B)	b	Exp(B)	b	Exp(B)	b	Exp(B)
ATV (assigned)	.02	1.02	−.45	.64	−.30	.74	−.53	.59
Alpha					7.94	.00***	9.85	
Intercept	−1.29***		−1.36	.00***	−.10	.00	−.27	
Initial Log-Likelihood	−89.31		−79.81		−294.67		−349.92	
Final Log-Likelihood	−89.31		−78.81		−170.97		−156.05	
P value	.96		.28		.00		.00	
Type 2								
ATV (assigned)	−.17	.84	−.75	.47	−.69	.50	−.52	.59
Prior victim interview	−.54	.59	−.34	.71	−.51	.60	−.19	.83
Age	.04	1.04	.04	1.04	.06	1.06	.07	1.08
Ethnicity (African-American)								
Hispanic	.75	2.11	1.21	3.36*	1.01	2.75	1.42	4.13
West Indian/Caribbean	−.03	.97	.24	1.27	−.58	.58	−.42	.65
Other race	1.09	2.97	1.36	3.88	.63	1.87	.47	1.60
Married	−.66	.52	−.70	.49	−1.38	.25	−1.61	.20
Employed	−.25	.78	−.32	.73	.19	1.21	−.00	1.00
Prior arrest	1.03	2.80**	1.15	3.17**	.90	2.45	.61	1.85
Alpha					6.30	.00***	7.97	.00***
Intercept	−2.46	.00***	−3.08	.00**	−2.05	.00	−2.64	.00
Initial Log-Likelihood	−89.31		−79.41		−269.99		−234.44	
Final Log-Likelihood	−89.31		−69.43		−164.75		−138.72	
P value	.96		.02		.00		.00	

[a]Logistic regression
[b]Negative binomial regression
*p < .05; **p < .01; ***p < .001

for the treatment assignment measures for both "any" and "severe" victimization were nonsignificant. Nevertheless, in both models, the reduction in the expected frequency of victimization estimated by the full model was not small. For "any victimization," the expected frequency of failure among those assigned to the treatment was reduced by about 50 percent. The reduction was about 40 percent for "severe victimization."

Twelve-month treatment effects. Table 6 displays the 12-month results for the prevalence and incidence of victim-reported incidents ("any" and "severe" victimization). Essentially these 12-month results mirror the six-month findings: we found all negative relationships between treatment assignment and victimization rates, but all still were nonsignificant. Two control measures in the full models were significant, however. The prevalence models showed that for batterers married to their victim, the risk of any victimization one year later was reduced significantly. Similarly, the risk of severe victimization was lower when the batterer was employed at the time of the treatment assignment. These two significant effects, however, did not remain in the frequency models.

As with the official criminal justice measures, consistent results emerged. For 15 of the 16 victim-based recidivism models, the

Table 6. Victims' Reports of Incidents at 12 Months (n = 189)

| | Prevalence[a] | | | | Frequency[b] | | | |
| | Any | | Severe | | Any | | Severe | |
Type 1	b	Exp(B)	b	Exp(B)	b	Exp(B)	b	Exp(B)
ATV (assigned)	−.47	0.63	−.80	.45	−.76	.47	−.83	.43
Alpha					9.87***		21.75**	
Intercept	−1.25***		−1.78***		−.11		.30	
Initial Log-Likelihood	−90.56		−63.82		−327.30		−288.34	
Final log-likelihood	−89.78		−62.46		−164.79		−111.98	
P value	.21		.10		.00		.00	
Type 2								
ATV (assigned)	−.43	.65	−.71	.49	−.83	.44	−.61	.54
Prior victim interview	.95	2.58	−.05	.95	.32	1.38	−.54	.58
Age	.01	1.01	−.01	.99	.02	1.02	.01	1.01
Ethnicity (African-American)								
Hispanic	.06	1.07	−.15	.86	−.81	.44	−.67	.51
West Indian/Caribbean	−.02	.98	.48	1.62	.58	1.78	2.48	11.98
Other race	.42	1.52	−.60	.55	−.54	.58	−1.02	.36
Married	−1.07	.34**	−.85	.43	−.60	.55	−.71	.49
Employed	−.34	.71	−1.09	.34*	−.60	.55	−2.05	.13
Prior arrest	−.70	.50	−.48	.62	−.57	.56	−.63	.53
Alpha					8.56***		13.67**	
Intercept	−1.48		−.52	.00	.06		.49	
Initial Log-Likelihood	−90.56		−63.82		−307.20		−242.60	
Final Log-Likelihood	−84.11		−57.78		−161.18		−104.98	
P value	.17		.15		.00		.00	

[a]Logistic regression
[b]Negative binomial regression
*p < .05; **p < .01; ***p < .001

estimated effect for treatment was negative (although the results were nonsignificant); collectively these 15 negative effects suggest that persons assigned to treatment may report 44 percent fewer than the expected incidents of victimization. Even though none of the 16 tests achieved statistical significance, the empirical regularity of the results suggests that treatment is working.

DISCUSSION

A number of reviewers (Hamberger and Hastings 1993; Rosenfeld 1992; Tolman and Bennett 1990) have observed that after years of asking "Are batterer treatment programs effective?" we still do not have a conclusive answer. Yet a new generation of evaluations (based on experimental models) promises to provide such an answer. Results from our experiment are consistent with those obtained by Palmer et al. (1992). Both experiments found that the rate of new incidents reported to criminal justice authorities was reduced significantly among batterers assigned to treatment.

As in the Palmer et al. (1992) study, our treatment effects were relatively large. In the analyses of the prevalence and frequency rates, we found that men assigned to a therapeutic treatment group

for batterers were less likely to be subject to future criminal complaints and arrests involving the same victim than were men assigned to a control group: the rate was 59 percent smaller at the six-month follow-up and 45 percent smaller at 12 months. Also, the 12-month follow-up showed that men assigned to batterer treatment, on average, registered a longer period of nonviolence or longer time to first failure than men assigned to the control group. On the basis of the victim survey data, the size of the treatment effect was similar to that found in the criminal justice data: the average violence rate was 44 percent smaller as a result of treatment. The effects were nearly always in the same direction. In 15 of 16 models, victims of men assigned to the treatment group reported fewer violent acts than victims of men in the control group.

We are more cautious about the outcomes of victim surveys, however, because in contrast to the results based on the criminal justice records) none of the differences between the two groups were statistically significant at the standard .05 alpha level. This nonsignificance may be due in part to the small sample sizes ($n = 171$ for the six-month data; $n = 189$ for the 12-month data). Alternatively, as one anonymous peer reviewer suggested, the men assigned to batterer treatment may have become more skillful than their control group counterparts at avoiding calls to the police when they are violent.

Some observers are likely to question the generalizability of our positive treatment findings because our sample represented only a small proportion of the spouse abuse cases adjudicated by the Brooklyn Criminal Court during our intake period: 3.4 percent, or 376 of nearly 11,000 adjudicated cases, met our study criteria. The sample size was limited because the prosecutor, the defendant, and the judge had to agree to treatment in order to make a case eligible for the lottery process. Therefore the treatment effects we observed cannot be generalized even to the population of adjudicated spouse abuse cases in the Brooklyn Criminal Court. Similarly, many of the earlier quasi-experimental evaluations of batterer treatment programs can be criticized on the grounds that participants are only a small and probably unrepresentative proportion of batterers processed in the courts from which their samples were drawn. Thus the Palmer et al. (1992) experiment is probably the most readily generalizable because it used a more representative sample of all batterers placed on probation by the court where the study was conducted. Their sample size, however, was only 59 cases, and not all of those batterers were necessarily motivated to undergo treatment.

Which of the two sampling approaches is better for evaluation? The answer depends on public policies defining who should participate in treatment programs. Many program administrators argue that treatment is appropriate only for batterers who demonstrate a willingness to change their behavior. We accept that argument; accordingly our sampling approach makes more sense because it filtered potentially unmotivated batterers. The Palmer et al. (1992) study, however, provides a test of cases more typically referred to court-mandated batterer groups.

The issue of generalizability crops up in another way as well. In many treatment studies that have relied on batterer or victim surveys to assess violence, interview response rates have been poor, some as low as 30 percent. Low response rates are a problem because the cases in which follow-up data are available may be different from those in which data are not available. For example, Edleson and Syers (1990) reported higher levels of education and income for batterers who completed follow-up surveys than for those who did not. Therefore it is unclear whether their analysis of treatment effects applies to the low-SES batterers who did not complete the survey as well as the higher-SES batterers who did so.

Although we succeeded in interviewing only about half of the victims in our sample, we were able to demonstrate the statistical equivalence of completers and noncompleters across a range of participant characteristics. Moreover, we had complete official recidivism data on virtually the entire sample, and we found that rates of victimization reported to the police did not differ significantly between victims who completed the surveys and those who did not.

Another limitation of our study is the problem of treatment misassignment: cases assigned to the control group that a judge reassigned to the treatment group. Our study includes 53 misassigned cases; this raises some questions about the precision of our treatment estimates. Consequently our results may reflect more clearly the potential for failures under a policy of assigning people to treatment than the theoretical issue of how well treatment works for those who receive it. Probably this misassignment tends to suppress the effectiveness of the ATV treatment program which we estimated. That is, if the 53 batterers in our control group who benefited from treatment because of the misassignment had been denied treatment, the effect of treatment would have been increased. Therefore it is likely that our reported treatment estimates are conservative.

Studies of batterer treatment programs have grown increasingly sophisticated: They use designs with a high degree of internal validity. As recognized by Fagan (1996), randomized experiments

are the preferred evaluation design. The few experimental evaluations that have been completed provide useful information on statistical power and sample sizes for designers of future studies. As more studies are completed, we may learn with some confidence how much and under what conditions treatment programs reduce violence.

If that proves to be the case, attention will turn to policy issues regarding how these programs are conducted. For example, the trend has been to lengthen treatment; yet in contrast to findings in the therapeutic drug treatment literature (see S.R. Maxwell 1994), virtually no research supports one treatment length over another. Also, practitioners in the field disagree about the content of treatment sessions. Many programs have adopted a feminist orientation which assumes that eliminating violent behavior requires changing the participant's perception of men's and women's roles in society. Other programs take a different approach, emphasizing anger control, stress management, and better communication techniques. Currently there is no empirical basis for deciding which approach is best. One thing seems certain, however: group treatment programs for batterers will continue in some form, if only because no good sentencing alternatives exist for spouse abusers.

REFERENCES

Armitage, P. 1996. "The Design and Analysis for Clinical Trials." Pp. 1-30 in *Design and Analysis of Experiments: Handbook of Statistics*, vol. 13, edited by S. Ghosh and C. R. Rao. Amsterdam: Elsevier.

Berk, R.A., H. Black, J. Lilly, and G. Rikoski. 1991. *Evaluating Alternative Police Response to Spouse Assault in Colorado Springs: An Enhanced Replication of the Minneapolis Experiment, 1987-1989*. Colorado Springs, CO: Colorado Springs Police Department.

Berk, R.A., G.K. Smyth, and L.W. Sherman. 1988. "When Random Assignment Fails: Some Lessons from the Minneapolis Spouse Abuse Experiment." *Journal of Quantitative Criminology* 4:209-33.

Blumstein, A., J. Cohen, J. Roth, and C. Visher, eds. 1986. *Criminal Careers and Career Criminals*. Washington, D.C.: National Academy Press.

Buzawa, E.S. and C.G. Buzawa. 1996. *Domestic Violence: The Criminal Justice Response*. 2nd ed. Newbury Park, CA: Sage.

Chen, H.T., C. Bersani, S. Myers, and R. Denton. 1989. "Evaluating the Effectiveness of a Court Sponsored Abuser Treatment Program." *Journal of Family Violence* 4:309-22.

Crowell, N.A. and A.W. Burgess, eds. 1996. *Understanding Violence Against Women*. Washington, D.C.: National Academy Press.

Davis, R.C., B.E. Smith, and L. Nickles. 1997. *Prosecuting Domestic Violence Cases With Reluctant Victims: Assessing Two Novel Approaches*. Washington, DC: American Bar Association.

Davis, R.C. and B.G. Taylor. 1997. "A Proactive Response to Family Violence: The Results of a Randomized Experiment." *Criminology* 35:307-33.

———. 1999. "Does Batterer Treatment Reduce Violence? A Synthesis of the Literature." *Women and Criminal Justice* 10(2):69-93.

Dobash, R., R. Dobash, K. Cavenagh, and R. Lewis. 1995. "Evaluating Criminal Justice Programmes for Violent Men." Pp. 88-116 in *Gender and Crime*, edited by R. Dobash and R. Dobash. Cardiff: University of Wales Press.

———. 1996. *Research Evaluation of Programmes for Violent Men*. Edinburgh: Centre Research, Scottish Office.

Dubin, J.A. and D. Rivers. 1989. "Selection Bias in Linear Regression, Logit and Probit Model." *Sociological Methods and Research* 18(2):360-90.

Dunford, F. 1999. "The San Diego Navy Experiment: An Assessment of Interventions for Men Who Assault Their Wives." Unpublished Manuscript.

Dutton, D. 1986. "The Outcome of Court-Mandated Treatment for Wife Assault: A Quasi-Experimental Evaluation." *Violence and Victims* 1(3):163-75.

———. 1988. "Profiling of Wife Assaulters: Preliminary Evidence for a Trimodal Analysis." *Violence and Victims* 3(3):5-29.

Edleson, J. 1996. "Controversy and Change in Batterers' Programs." Pp. 154-69 in *Future Interventions with Battered Women and Their Families*, edited by J. Edleson and Z. Eisikovits. Thousand Oaks, CA: Sage.

Edleson, J. and M. Syers. 1990. "Relative Effectiveness of Group Treatments for Men Who Batter." *Social Work Research and Abstracts* 26(2):10-17.

Eisikovits, Z. and J. Edleson. 1989. "Intervening with Men Who Batter: A Critical Review of the Literature." *Social Service Review* 37:84-414.

Fagan, J.A. 1996. "The Criminalization of Domestic Violence: Promises and Limits." *NIJ Research Report*. Washington, DC: National Institute of Justice, U.S. Department of Justice.

Fagan, J.A. and A. Browne. 1994. "Violence Against Spouses and Intimates." Pp. 112-51 in *Understanding and Controlling Violence*, vol. 3., edited by A.J. Reiss Jr. and J.A. Roth. Washington, DC: National Academy Press.

Fagan, J., E. Friedman, S. Wexler, and V.L. Lewis. 1984. *National Family Violence Evaluation*. San Francisco, CA: URSA Institute.

Fagan, J.A., J. Garner, and C.D. Maxwell. 1997. *Reducing Injuries to Women in Domestic Assaults*. Atlanta, GA: Centers for Disease Control and Prevention.

Feder, L. 1996. *A Test of the Efficacy of Court-Mandated Counseling for Domestic Violence: A Broward County Experiment*. Proposal submitted to the National Institute of Justice. Florida Atlantic University, Boca Raton, Florida.

———. 1999. "The Efficacy of Court-Mandated Counseling for Convicted Misdemeanor Domestic Violence Offenders: The Broward Experiment." Washington, DC: American Society of Criminology.

Garner, J., J.A. Fagan, and C.D. Maxwell. 1995. "Published Findings From the Spouse Assault Replication Program: A Critical Review." *Journal of Quantitative Criminology* 11:3-28.

Gartin, P.R. 1995. "Dealing with Design Failures in Randomized Field Experiments: Analytic Issues Regarding the Evaluation of Treatment Effects." *Journal of Research in Crime and Delinquency* 32:425-45.

Gelber, R.D. and M. Zelen. 1986. "Planning and Reporting of Clinical Trials." Pp. 406-25 in *Medical Oncology*, edited by P. Calabresi, P.S. Schein, and S.A. Rosenberg. New York: Macmillan.

Gondolf, E. 1995. *Batterer Intervention: What We Know and Need to Know*. Washington, DC: National Institute of Justice.

———. 1997. "Expanding Batterer Program Evaluations." Pp. 208-18 in *Out of Darkness: Contemporary Research Perspectives on Family Violence*, edited by G. K. Kaufman and J. Jasinski. Thousand Oaks, CA: Sage.

Greenfeld, L.A., M.R. Rand, D. Craven, P. Klaus, C.A. Perkins, C. Ringel, G. Warchol, C. Maston, and J. Fox. 1998. *Violence by Intimates: Analysis of Data on Crimes by Current or Former Spouses, Boyfriends, and Girlfriends*. Washington, DC: U.S. Department of Justice.

Hamberger, K. and J. Hastings. 1993. "Court-Mandated Treatment of Men Who Assault Their Partner: Issues, Controversies and Outcomes." Pp. 188-232 in *Legal Responses to Wife Assault: Current Trends and Evaluation*, edited by Z. Hilton. Newbury Park, CA: Sage.

Hanna, C. 1996. "No Right to Choose: Mandated Victim Participation in Domestic Violence Prosecutions." *Harvard Law Review* 109:1849-1910.

Harrell, A. 1991. *Evaluation of Court-Ordered Treatment of Domestic Violence Offenders*. Washington, DC: Urban Institute.

Healey, K., C. Smith, and C. O'Sullivan. 1998. *Batterer Intervention: Program Approaches and Criminal Justice Strategies*. Washington, DC: U.S. Department of Justice.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153-61.

Hotaling, G.T. and D. Sugarman. 1986. "An Analysis of Risk Markers in Husband to Wife Violence: The Current State of Knowledge." *Violence and Victims* 1(2):101-24.

Little, R.J.A. and N. Schenker. 1995. "Missing Data." Pp. 39-76 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C.C. Clogg, and M.E. Sobel. New York: Plenum.

Maxwell, C.D. 1998. "The Specific Deterrent Effect of Arrest on Aggression Between Intimates and Spouses." PhD dissertation, Graduate School of Newark, School of Criminal Justice, Rutgers University.

Maxwell, S.R. 1994. "Formal and Informal Social Controls of Drug Treatment Retention of Offenders." PhD dissertation, Graduate School of Newark, School of Criminal Justice, Rutgers University.

Palmer, S., R. Brown, and M. Barrera. 1992. "Group Treatment Program for Abusive Husbands: Long-Term Evaluation." *American Journal of Orthopsychiatry* 62: 276-83.

Patel, H.I. 1996. "Clinical Trials in Drug Development: Some Statistical Issues." Pp. 31-57 in *Design and Analysis of Experiments: Handbook of Statistics*, vol. 13, edited by S. Ghosh and C.R. Rao. Amsterdam: Elsevier.

Rebovich, D.J. 1996. "Prosecution Response to Domestic Violence: Results for a Survey of Large Jurisdictions." Pp. 175-76 in *Do Arrest and Restraining Orders Work?*, edited by E.S. Buzawa and C.G. Buzawa. Thousand Oaks, CA: Sage.

Rosenbaum, A. and K.D. O'Leary. 1986. "The Treatment of Marital Violence." Pp. 44-69 in *Clinical Handbook of Marital Therapy*, edited by N.S. Jacoben and A.S. Gurmand. New York: Guilford.

Rosenfeld, B.D. 1992. "Court Ordered Treatment of Spouse Abuse." *Clinical Psychological Review* 12:205-26.

Saunders, D. 1996. "Feminist-Cognitive-Behavioral and Process-Psychodynamic Treatments for Men Who Batter: Interaction of Abuser Traits and Treatment Models." *Violence and Victims* 11:393-414.

Saunders, D.G. and S.T. Azar. 1989. "Treatment Programs for Family Violence." Pp. 481-546 in *Family Violence*. vol. 11, edited by L. Ohlin and M. Tonry. Chicago, IL: University of Chicago Press.

Sherman, L.W. 1992. *Policing Domestic Violence: Experiments and Dilemmas*. New York: Free Press.

Straus, M.A. 1979. "Measuring Family Conflict and Violence: The Conflict Tactics Scale." *Journal of Marriage and the Family* 41:75-88.

Tjaden, P. and N. Thoennes. 1998. *Prevalence, Incidence and Consequences of Violence Against Women: Findings from the National Violence Against Women Survey*. Washington, DC: National Institute of Justice.

Tolman, R.M. and L. Bennett. 1990. "A Review of Quantitative Research on Men Who Batter." *Journal of Interpersonal Violence* 5(1):87-118.

Tolman, R.M. and J.L. Edelson. 1995. "Interventions for Men Who Batter: A Review of Research." Pp. 211-55 in *Understanding Partner Violence: Prevalence, Causes, Consequences, and Solutions*, edited by S.M. Stith and M.A. Straus. Minneapolis, MN: National Council on Family Relations.

Weisberg, S. 1985. *Applied Linear Regression*. New York: Wiley.

Winship, C. and R.D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 11(1):111-44.