Journal of Quantitative Criminology, Vol. 11, No. 1, 1995

Published Findings from the Spouse Assault Replication Program: A Critical Review

Joel Garner,[1,3] Jeffrey Fagan,[1] and Christopher Maxwell[2]

Published reports from seven jointly developed experiments have addressed whether or not arrest is an effective deterrent to misdemeanor spouse assault. Findings supporting a deterrent effect, no effect, and an escalation effect have been reported by the original authors and in interpretations of the published findings by other authors. This review found many methodologically defensible approaches used in these reports but not one of these approaches was used consistently in all published reports. Tables reporting the raw data on the prevalence and incidence of repeat incidents are presented to provide a more consistent comparison across all seven experiments. This review concludes that the available information is incomplete and inadequate for a definitive statement about the results of these experiments. Researchers and policy makers are urged to use caution in interpreting the findings available to date.

KEY WORDS: violence; deterrence; spouse assault; experiment.

1. MINNEAPOLIS DOMESTIC VIOLENCE EXPERIMENT

Although American social institutions have increased the range of formal and informal controls available to address perceived limitations and inadequacies in legal responses to victims of spouse assault, social control through law dominated theory and policy on domestic violence in the 1980s (Fagan and Browne, 1994). The Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984a) was a critical event in changing public and scholarly perceptions of spouse assault from a "family problem" amenable to mediation and other informal, nonlegal interventions (Bard and Zacker, 1971) to a law violation requiring a formal criminal justice sanction.

In that experiment, street-level police officers' selections of the most appropriate response to misdemeanor spouse assault were determined by an

[1]School of Criminal Justice, Rutgers University, Newark, New Jersey 07102.
[2]Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109.
[3]To whom correspondence should be addressed.

0748-4518/95/0100-0003$07.50/0

experimental design, i.e., random assignment to one of three treatments: (1) arresting the suspect, (2) ordering one of the parties out of the residence, and (3) advising the couple. Using victim interviews and official records of subsequent police contact, Sherman and Berk (1984a, p. 267) reported that the prevalence of subsequent offending—assault, attempted assault, and property damage—was reduced by nearly 50% when the suspect was arrested. On the basis of the results from what they emphasized was the "first scientifically controlled test of the effect of arrest on any crime" Sherman and Berk (1984b, p. 1) concluded that

> these findings, standing alone as the result of one experiment, do not necessarily imply that all suspected assailants in domestic violence incidents should be arrested. Other experiments in other settings are needed to learn more. But the preponderance of evidence in the Minneapolis study strongly suggests that the police should use arrest in most domestic violence cases.

In the decade since the preliminary results were announced in the "Science" section of the *New York Times* (Boffey, 1983, p. L1), the study's findings were reported in over 300 newspapers in the United States, three major television networks broadcast the study's results in prime time news programs or documentaries, and numerous nationally syndicated columnists and editorials featured the study and its findings (Sherman and Cohn, 1989). The Attorney General's Task Force on Family Violence endorsed the study's findings and recommended that state and local agencies adopt a pro-arrest policy toward spouse assault (U.S. Attorney General, 1984). Following the attention given to this study's results, a dramatic change in formal policy consistent with the study's proarrest findings has been reported by police departments in both large and small U.S. cities (Cohn and Sherman, 1987).

The Minneapolis experiment was designed, funded, and implemented as a test of specific deterrence theory (Sherman, 1980) and was a direct response to the call for such tests by the National Academy of Sciences (Blumstein *et al.*, 1978). The Minneapolis experiment is atypical for its innovative experimental design, its test of theory, its extensive visibility, its focus on a controversial issue, and its apparent impact on public policy. This experiment is atypical for another reason: it was replicated [National Institute of Justice (NIJ), 1985].

2. THE SPOUSE ASSAULT REPLICATION PROGRAM

Between 1985 and 1991, teams of police departments and researchers implemented in six jurisdictions similar experiments designed to provide independent and complementary tests of the theories that informed the Minneapolis experiment: the specific deterrent effects of arrest on subsequent criminality. These experiments, which collectively came to be known under

the acronym of SARP[5] (Spouse Assault Replication Program), were implemented with the following requirements: Cases must be eligible for arrest for misdemeanor spouse assault, alternative treatments must be determined by randomization after eligibility was established, one of the treatments must be arrest, and the primary outcome of interest was subsequent criminality as measured by official police records and victim interviews (NIJ, 1985).

Common data elements for measuring treatments, outcomes, and the characteristics of suspects (NIJ, 1987) and a set of "core analyses" were established for the program (NIJ, 1988). The common data elements identified a variety of offense types, victim types, and other considerations that could be used in determining what constituted a reoffense; the core analysis specified the use of three dimensions of criminal careers (prevalence, incidence, and time to failure) and the nature of the comparisons (e.g., posttreatment contrasts of randomized groups) common to all SARP experiments. These were reporting requirements; beyond these core elements, each investigator was free to pursue other analyses and there was considerable variability in the timing, jurisdiction, and nature of the experiments (Garner, 1989).

3. PUBLISHED CONCLUSIONS

The initial published findings from these tests of deterrence theory are available for six SARP experiments in five sites. The following is a presentation (alphabetically by city) of the original investigators' narrative conclusion from their research.

Charlotte

Hirschel *et al.* (1992a, p. 29) reported that, for Charlotte,

> based on a thorough analysis of data from official police records of rearrest, as well as from intensive interviews with victims of abuse, there is no evidence that arrest is a more effective deterrent to subsequent assault. This conclusion remains regardless of the measure of recidivism used—prevalence, incidence, or time to failure.

Colorado Springs

A Bayesian analysis of data from Colorado Springs (Berk *et al.*, 1992a, p. 200) concluded that

[5]Three of the six jurisdictions implemented two experiments. In addition to the replication of Minneapolis, Omaha experimented with the use of an arrest warrant in those situations where the suspect was not present when the police arrived. The two studies are referred to here as Omaha—Offender Present and Omaha—Offender Absent. Atlanta and Dade County included a factorial design, where both arrested and not arrested cases were randomly assigned to follow-up counseling. These two experiments are not considered here.

> it is clear no treatment effect is apparent when the treatment variable is defined as the treatment that was randomly assigned and when the outcome variable is constructed from official data . . . . However, when the Colorado Springs outcome is constructed from victim reports rather than official data, a strong [deterrent] treatment effect surfaces . . . .

Dade County

The report by Pate and Hamilton (1992, p. 7–2) on Dade County states,

> Based on the results of the second-wave (6 month) victim interview, significant [deterrent] effects were found attributable to arrest with respect to both prevalence and time to failure of attacks against the original victim; significance level of the effect on incidence was one decimal point short of the standard 0.05 level.

and

> No significant arrest treatment effect [for prevalence or incidence] was found with respect to subsequent offense reports.

Milwaukee

The assessment of the Milwaukee data by Sherman *et al.* (1992b, p. 156) was that

> in sum, the main effects analysis shows some evidence of initial deterrent effects, no evidence of long term deterrent effects, and some evidence of long term escalation in both the timing and frequency of violence against any victim.

Omaha—Offender Absent

In their report on the offender absent experiment with arrest warrants, Dunford *et al.* (1990b, p. 642) concluded that

> suspects assigned to the warrant treatment were *always* found to have lower prevalence and frequency rates of repeated conflict than suspects assigned to the no warrant treatment.

Omaha–Offender Present

For the offender present experiment, Dunford *et al.* (1990a, p. 204) concluded that

> . . . arrest in Omaha, by itself, did not appear to deter subsequent domestic conflict any more than did separating or mediating those in conflict.

On the same page, they go on to report that

> arrest did not appear to place victims in greater danger of increased conflict than did separating or mediation. It would appear that what the police did in Omaha . . . neither helped nor hurt victims in terms of subsequent conflict.

Thus, where there was once one experiment with two consistent findings (prevalence and time to failure), there are now seven similar experiments

where the findings on the specific deterrent effect of arrest on the prevalence of reoffending—the central finding of the Minneapolis experiment—differ internally by data source and externally by site.

Once made public, the findings from the SARP experiments have been interpreted in ways that on occasion conflict with the original authors. For instance, Sherman (1992, pp. 16–17) summarize SARP findings in the following manner:

> The best way to compare the findings across experiments is still to focus on the effects of arrest compared to nonarrest, the central policy issue for state legislatures and police agencies. The most important finding for them is that arrest increased domestic violence in Omaha, Charlotte and Milwaukee.

He also reports (p. 17),

> There is evidence that arrest had a deterrent effect in Minneapolis, Colorado Springs, and Metro-Dade, but the cumulative evidence is somewhat mixed.

Notably, these statements about an escalation effect in Omaha and Charlotte are not the conclusions of the original authors of the Omaha and Charlotte experiments cited above. This fact warrants emphasis since the Sherman interpretation has become a commonly accepted representation of the published results (Mastrofski and Uchida, 1993; McCord, 1993; Jones, 1992; Blumstein and Petersilia, 1994; Zorza and Woods, 1994). Thus, to the diverse set of findings from the SARP experiments, the derivative publications have added distinct interpretations of each other's results.

4. INTERACTION EFFECTS

The diverse findings reported for the SARP experiments have led to some speculation and research on possible causes. Sherman *et al.* (1992a) and Berk *et al.* (1992b), in multisite analyses, and Pate and Hamilton (1992), in a single-site analysis, report on the interaction of arrest with two measures of "stake-in-conformity" (Toby, 1957), the marital and employment status of the suspect. This line of analysis suggests that the arrest *increases* violence for unmarried and also for unemployed suspects and *deters* it for married and for employed suspects. Sherman *et al.* (1992a, p. 687) evaluation of the published findings from Colorado Springs, Dade County, Milwaukee, and Omaha is that

> all four experiments that have examined this hypothesis report on interaction with unemployment consistent with the stake-in-conformity hypothesis, at least in the official data.

Upon close examination (Table VI), it is clearer that the evidence supporting this conclusion stems from reports on a small subset of SARP data and, for the data reported, the evidence is less consistent than Sherman and his

colleagues suggest. In addition, as Sherman *et al.* (1992a) recognize, the "stakes in conformity" explanation does not have the rigor of an experimental finding and has been applied *post hoc* to the SARP findings.

The quotations from the SARP publications cited above hint at a general failure in these studies to replicate the Minneapolis findings of a deterrent effect for arrest. Our review of the published findings from SARP led us to believe that that conclusion is not warranted, at least not yet, given the evidence published to date. What is available is a series of inconsistent individual-site reports and a few incomplete and highly selective cross-site comparisons. There is no published assessment that accumulates in a consistent and systematic fashion all the evidence from each of the SARP experiments. Our purpose here is to make substantial progress toward that goal.

5. COMPARING AND ASSESSING THE SARP RESULTS

While we think that it is important to restate accurately the conclusions reported by the original investigators, such narrative accounts are limited in the amount of information they can convey about the substantive and methodological basis for each of these conclusions. In addition, the design and reporting of the SARP experiments as "replications" (NIJ, 1985; Berk *et al.*, 1992a; Dunford *et al.*, 1990a; Garner, 1990; Hirschel *et al.*, 1991; Pate *et al.*, 1991; Sherman *et al.*, 1991) could easily leave the impression of more similarity in implementation, data analysis, and reporting than a more detailed review of the SARP publications reveals. For instance, the SARP conclusions reached above are based primarily, though not exclusively, on a small set of analytical comparisons: the prevalence, incidence, and time to failure for (1) violent offenses, (2) against the same victim, (3) over a 6-month period, (4) from official records, and (5) contrasting cases randomly assigned to arrest with all other cases.

SARP was explicitly designed to permit multiple analytical comparisons across sites (NIJ, 1985, 1987, 1988) and a complete understanding of the available information about recidivism in SARP would include cross-site comparisons with (1) different offense types (threats, property damage, etc.), (2) different victim types (other family members, any other victim), (3) different treatment comparisons (arrest versus mediation, mediation versus separation), (4) different periods at risk (1, 3, 6, or 12 months), and (5) different groups of cases (cases as assigned versus cases as treated). These nonexhaustive choices we have listed would result in 162 analytical comparisons (3 offense types × 3 victim types × 3 treatment groups × 3 risk periods × 2 case groups) just for the *prevalence* measures. As we shall see, there are analytically distinct methods for analyzing frequency and time to failure measures.

There are three issues of note about the availability of so many reasonably plausible comparisons. First, none of the "replications" report the same analytical comparisons reported in the original Minneapolis experiment. In the Minneapolis experiment, Sherman and Berk (1984a, b) included threats of violence and property damage as failures equivalent to actual violence.[6] In addition, the original Minneapolis publications did not report the contrast between treatments as assigned; the findings reported were "corrected" statistically to account for misapplication of some of the nonarrest cases to arrest (Berk and Sherman, 1988). None of the publications from the SARP experiments uses these outcome criteria or this analytical approach.

The second notable issue arising from the availability of multiple comparisons in the SARP experiments is that there was no a priori consensus about the most appropriate of these possible analytical comparisons. While one report (e.g., Sherman *et al.*, 1992) provides an explicit, albeit *post hoc*, rationale for emphasizing one measure (frequency) over another (prevalence), neither our theories nor our policy preferences are sufficiently well developed to specify which outcome should change within what period of time if arrest (or any treatment) is to be considered an effective police response to spouse assault. Each of the SARP reports includes several analytical comparisons (see Dunford *et al.*, 1990a; Pate *et al.*, 1991); none reports more than a handful of possible comparisons or acknowledges the potential variety of possible comparisons. The authors of these articles have published what they think (and their editors agree) are the most interesting comparisons, but as we shall see, the published works do not agree among themselves as to which comparisons are the most interesting or appropriate for "replications."

The design of the SARP experiments to permit multiple comparisons raises a third concern; the comparisons are not independent of each other. The presence of hundreds of possible comparisons already complicates the interpretation of the few statistical tests that have been reported. One of the reports (e.g., Hirschel *et al.*, 1992a) recognizes some of the problems nonindependent comparisons generate; others do not. If, in fact, there is no effect from the alternative police reponses to spouse assault tested in SARP, a 0.05 level of statistical significance means that (if certain underlying assumptions are met), of 100 independent comparisons, by chance 5 will show significant differences. When the comparisons are not independent of each other (such as prevalence and frequency measures) and when not all

[6]The original Minneapolis report defines threats of violence and property damage as violence; we use information on outcomes excluding threats and property damage from the Gartin (1991) reanalysis of Minneapolis data for comparability. The prevalence rates of Gartin are about half of that reported by Sherman and Berk (1984b).

comparisons are reported, the formal interpretation of tests of statistical significance is muddled (Robertson, 1992; Toothaker, 1993). Finally, despite the fact that (1) the most commonly reported finding in the SARP experiments is that there are no statistically significant differences between treatments and (2) assessments of statistical power were explicit considerations in the design and funding decisions within SARP (NIJ, 1988), virtually all of the SARP publications fail to report the power of their statistical comparisons. Thus, when a finding of no effect is reported, the readers have no formal way to assess whether the failure to find an effect was due to the absence of an effect or to the likelihood that the research design would not find an effect if it did exist.

6. STANDARDIZED COMPARISONS ACROSS SARP EXPERIMENTS

These concerns would be misplaced if it were possible from the published works to extract standardized or even comparable comparisons across SARP experiments. After several readings and rereadings of the published articles, books, and final reports on the SARP experiments, we were unable to do this. In fact, we could not find one comparison that could be extracted precisely for all SARP experiments.

In order to produce a consistent set of findings across the SARP experiments, we have identified six central comparisons commonly reported in published reports. The first two comparisons are the prevalence in official records and victim reports of a violent offense by the same suspect against the same victim within 6 months from the presenting incident. We examined the findings comparing all cases randomized to arrest versus those randomized to any other treatment. These are the simplest and most common comparisons among SARP sites and the original Minneapolis experiment. We also attempted to extract from official records and from victim interviews two other sets of comparisons: (1) the frequency rates of reoffending for the same offense type, victim type, risk period, and treatment comparison and (2) measures of the time to first failure.

The prevalence, incidence, and time to failure comparisons represent the dimensions of career behavior measurements that have become the *lingua franca* of deterrence and criminal career research (Blumstein *et al.*, 1978, 1986). Our selection of offense type, victim type, risk period, and treatment comparisons is determined in great part by similarity with the original Minneapolis experiment and the availability of information in published reports. We make no claim here that these comparisons are more relevant theoretically or to policy than any of the others possible in the SARP experiments (NIJ, 1988); we do believe that they constitute the pedestrian essentials of

the SARP experiments, without which any understanding of the findings from SARP is unlikely.

6.1. Prevalence of Reoffending

Table I compares cases assigned to arrest versus those assigned to other treatments for seven SARP experiments. It includes the reported number of experimental cases and the number and percentage of cases with at least one reported incident of violence by the same suspect against the same victim within 6 months recorded in either official records or victim interviews. Table I also lists the direction of the effect. If in the posttreatment period the group of cases randomly assigned to arrest had higher failure rates, we designated this escalation; if the rates were lower, as deterrence.[7]

If the comparisons were reported to be statistically significant at the level of 0.05 or better, the cell with the direction of the effect is shaded. We use a P value of 0.05 as a consistent standard; Dunford *et al.* (1990a, b) report actual P values. In the offender absent experiment (1990b, p. 642) they invoke (in footnote 12) a P value standard of 0.1.[8]

Table I exhibits effects in the direction of deterrence using *both* interviews and official records in four of the six experiments—Minneapolis, Colorado Springs, Dade County, and Omaha—Offender Absent. There are consistent escalation effects in only one experiment—Milwaukee. In two experiments, Charlotte and the Omaha—Offender Present, official records show escalation effects but victim interviews show effects in the direction of deterrence. Of the 14 possible comparisons in Table I, only 3 (21%) are statistically significant: Significant deterrent effects are reported from both official records and victim reports in Minneapolis and from victim interviews in Dade.

The footnotes to Table I clarify the extent to which the data reported there are not based on precisely the same measures, cases or comparisons. For instance, Hirschel *et al.* (1992a) do not report subsequent *offenses* in official police records, but only subsequent *arrests*; Pate and Hamilton

[7]We are grateful to an anonymous reviewer for emphasizing that these are posttreatment-only comparisons and that, with individual-level data, the terms escalation and deterrence might imply pre-post comparisons. In the context of this research, we believe that our use of these terms is justified, as are the reviewer's caveats.

[8]We are indebted to an anonymous reviewer for reminding us of the importance of the difference between using one- and using two-tailed tests in a replication, an issue unexplored in SARP publications. Readers might consider which test is most appropriate for comparisons made before and after the publication of the Omaha results or for comparisons not reported in Minneapolis (i.e., frequency rates) before writing us with their conclusions. We employ the two-tailed test to be consistent across studies.

Table I. Prevalence of Reoffending After 6 Months

	SOURCES	Charlotte[1]		Col. Springs[2]		Dade County[3]		Milwaukee[4]		Minneapolis[5]		Omaha-OA[6]		Omaha-OP[7]	
		RANDOMIZED TREATMENT GROUPS: A - ARRESTED (And Booked) NA - NOT ARRESTED SHADED CELLS DENOTES STATISTICALLY SIGNIFICANT EFFECT REPORTED													
	MEASURES	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA
RECORDS	Cases	214	436	421	1,158	465	442	624	297	93	237	111	136	109	221
	Failed Cases	39	68	81	224	89	91	161	75	6	35	16	30	19	36
	Prevalence	18.2	15.6	19.2	19.3	19.1	20.6	25.8	25.2	6.5	14.8	14.4	22.1	17.4	16.3
	Direction	Escalation		Deterrent		Deterrent		Escalation		Deterrent		Deterrent		Escalation	
SURVEYS	Cases Surveyed	112	226	Not Reported		199	182	624	297	52	115	84	112	77	165
	Failed Cases	66	142			50	71	202	92	9	35	29	53	29	69
	Prevalence(Cases)	30.8	32.6			10.8	16.1	25.2	23.1	9.7	14.8	26.1	39.0	26.6	31 2
	Prevalence(Surveyed)	58.9	62.8			25.1	39.0	32.4	31.0	17.3	30.4	34.5	47.3	37.7	41.8
	Direction	Deterrent		Deterrent		Deterrent		Escalation		Deterrent		Deterrent		Deterrent	

[1]Hirschel *et al.* (1992a): Collapses citation and advise/separate treatments into not arrested treatment. Six hundred fifty cases excludes 28 couples who entered the experiment twice and 4 couples who entered the experiment three times. Police records: Findings on page 19, Table 2. Victim interviews: Findings on page 24, Table 4.

[2]Berk *et al.* (1992a): Total reported cases = 1658; 79 cases not used in reported analyses, all from the not arrested treatment. Police records: Findings extrapolated from page 183, Table 3, by multiplying failure proportion by sample size for two arrest and two no-arrest categories. Victim interviews: Data not reported. Direction and size of effect based on statement on page 197, "The odds multipler for arrest is approximately 0.65 and the Bayesian ninety percent confidence region no longer includes 1.0."

[3]Pate *et al.* (1991): Violence-only offenses not reported; substituted all subsequent offenses. Police records: Findings on page 6-55, Table 6-8c. Victim interviews: Findings on pages 6-32 and 6-40, Table 6-4E.

[4]Sherman *et al.* (1992b): Police records: Findings on page 154, Table 4, limited to 921 (of 1200) cases for which interviews were obtained. Victim interviews: Findings on page 154, Table 4, not limited to 6-month posttreatment interval.

[5]Gartin (1991): Police records, findings on pages 140–141, Tables 5.06 and 5.07; Victim interviews: findings on page 120, Table 5.01.

[6]Dunford *et al.* (1990b): "Complaint recidivism." Police records: Findings on page 639, Table 3. Victim interviews: findings on page 640, Table 4.

[7]Dunford (1990a): Complaint recidivism. Police records: Findings on page 200, Table 6. Victim interviews: Findings on page 201, Table 7, column 2.

(1992) do not report prevalence rates separately for violent offenses; Sherman (1992) does not limit his analyses to a 6-month time at risk; and Berk *et al.* (1992a) couch prevalence measures within a Bayesian discussion of interaction effects. In addition, the raw data for all experimental cases

have been published for only two of the experiments—Minneapolis[9] and Dade County. The published data from Charlotte and both Omaha experiments exclude repeat cases[10] (about 5% of the sample in each case); Colorado Springs excludes, without comment, 79 cases, all from the nonarrested group; and the Milwaukee reports include official record recidivism only for the 927 cases (of 1200) with initial interviews.

Table I suggests one possible basis for the different conclusions about the SARP experiments reported by the original SARP authors and by Sherman (1992). Where Dunford *et al.* (1990a) and Hirschel *et al.* (1992a) report no statistically significant difference for either a deterrent or an escalation effect, Sherman (1992) reports that the findings for these studies are in the *direction* of escalation, not deterrence. Both statements are derived from the reported data, but both allow for misunderstanding of the actual results. The display of raw data in Table I conveys, we believe, more precisely what was reported in each experiment and provides a firmer basis for understanding SARP results than the narrative interpretations of the previously cited authors. However, given the lack of common reporting formats, the variety of measures used, the variety of standards for including or excluding randomized cases, and the variety of results from different experiments and data sources, Table I is still an inadequate basis for understanding the prevalence findings in SARP. We recommend caution in interpreting the published results on the prevalence of reoffending until more complete and consistent analyses are available.

6.2. Power Analysis

Three of the 14 comparisons recorded in Table I generated "statistically significant" differences. "Statistically significant differences" is short-hand jargon for assessing results obtained when testing a null hypothesis; in Table I, the null hypothesis is that suspects assigned to arrest will have the same recidivism as suspects not assigned to arrest. Given certain well-specified assumptions, a traditional frequentist approach interprets the results from Minneapolis and from the survey data in Dade County as likely to occur by chance in fewer than 5 of 100 tests. Thus, the error of rejecting the null hypothesis when it is true has, in these three tests, a probability of 0.05. By convention, differences this great (an α level of 0.05) are considered statistically significant.

[9]Sixteen cases given randomly assigned treatments in the Minneapolis experiment were not included in any of the reports on that study until the recent Gartin (1991) dissertation.

[10]Both studies report that the inclusion of the repeat cases does not substantively alter the direction or size of the reported effects, but they do not report the results with the repeat cases included.

Eleven (78%) of the 14 comparisons recorded in Table I did not generate statistically significant differences. It would be a grave but common error to interpret these 11 results by themselves as meaning that the null hypothesis (arrest and nonarrest are equally effective) is true. Under these conditions, the null hypothesis may be true *or* the research designs used may be unlikely to detect the expected effect. Statistical power is concerned with this second type of error (β), failing to reject the null hypothesis when it is false; power is defined as the complement of β, that is, $(1-\beta)$. The smaller the likelihood of type II error, the greater the power of the test.

Conventions have been developed to assess the power of a particular test (Cohen, 1988) and they were used by the NIJ when it was deciding on the design and planning of the SARP experiments (Garner, 1987); unfortunately, the SARP publications typically do not address the issue of statistical power.[11] Thus, readers are unable to discern what these studies say about the likelihood of their design detecting an effect similar to the Minneapolis experiment. However, since statistical power is a function of the sample size, the expected effect size, and the significance criterion (Cohen, 1988), it is possible to compute the statistical power of the tests in Table I. We can use the reported sample sizes and the conventional significance criterion of 0.05. Although determining an expected effect size is somewhat arbitrary, Cohen (1988) suggests a range of effects sizes of 0.2, 0.5, and 0.8, which he labels small, medium, and large.

In addition, we have the advantage of prior research (the Minneapolis Domestic Violence Experiment) which can provide an expected effect size directly relevant to its replications (Lipsey, 1983, 1990). When using differences in proportions (as in prevalence measures), examples of "small" effect sizes are differences between a 5% failure rate and a 10% failure rate or differences between a 40% failure rate and a 50% failure rate (Cohen, 1988, p. 181). Examples of a "medium" effect size are 5 versus 21% or 40 versus 65%. Differences in recidivism between arrest and not arrest extracted from the original Minneapolis publications (Sherman and Berk, 1984a) are 10 versus 21% (for official records) and 19 versus 35% (for victim interviews). Both of these comparisons compute to an effect size of approximately 0.3, or somewhere between Cohen's small and medium effects.

In Table II, we have used the conventional 0.05 significance-level criterion, the small, medium, and Minneapolis effect sizes, and the sample sizes from the SARP experiments to assess the power of the tests reported in Table I.[12] The findings reported in Table II reveal that for official records,

[11]Dunford and co-workers' (1989) unpublished report to NIJ and Gartin's (1991) reanalysis of the Minneapolis data do discuss the issue of statistical power.

[12]Power statistics computations assisted by Borenstein and Cohen (1988).

Table II. Statistical Power of Prevalence Measures in 7 SARP Experiments

EXPERIMENTAL SITE	Charlotte	Colorado Springs	Dade County	Milwaukee	Minneapolis	Omaha-OA	Omaha-OP
EFFECT SIZE	OFFICIAL RECORDS						
Small (.2)	0.67	0.93	0.85	0.99	0.37	0.34	0.4
Minneapolis (.3)	0.95	0.99	0.99	0.99	0.68	0.64	0.72
Medium (.5)	0.99	0.99	0.99	0.99	0.99	0.99	0.95
*Mean Sample Size	287	628	453	531	133	122	146
EFFECT SIZE	VICTIM INTERVIEWS						
Small (.2)	0.41	Sample Sizes Not Reported	0.5	0.81	0.22	0.28	0.3
Minneapolis (.3)	0.74		0.83	0.99	0.44	0.55	0.58
Medium (.5)	0.99		0.99	0.99	0.85	0.93	0.95
*Mean Sample Size	150		190	402	72	96	105

Harmonic Mean for Unequal Cell Sizes

haded Cells Exceed Cohen's Suggested Convention of Power > .80

four out of the seven SARP experiments had power exceeding 0.80,[13] and on average, the statistical tests had more power than a majority of the published studies in criminology (Brown, 1989). Thus, tests made using official records from these four experiments, all of which reported evidence that did not reject the null hypothesis, have a low probability of failing to reject (accepting) the null hypothesis when it is false. Under these conditions, the absence of statistically significant effects in Table I cannot be interpreted easily as the result of a weak research design. However, this is not the situation with the Omaha experiments or the tests using victim interviews. The Omaha—Offender Present experiment has a high probability (0.60) of failing to reject the null hypothesis (for a small effect) when that null hypothesis is false; for an effect similar in size to that found in Minneapolis, the Omaha—Offender Present experiment has a smaller but still generally unacceptable probability (0.28) of failing to reject the null hypothesis when that null hypothesis is false. Only in testing for an effect of arrest that Cohen classifies as medium or large does the Omaha—Offender Present experiment have "acceptable" levels of statistical power.[14]

The victim interviews tell a different story. In these series of tests, only the Milwaukee experiment exceeds the recommended power level of 0.80 for a small effect and only Dade and Milwaukee exceed this power level for an

[13]This is the level recommended by Cohen (1988, p. 56) as a convention for power, much as 0.05 is the convention for statistical significance.

[14]Dunford *et al.* (1989, pp. 44–46) use a *P* value of 0.1 (rather than the 0.05 used here) and reach a different conclusion concerning the statistical power of the Omaha experiments.

effect size similar to the Minneapolis experiment. Thus, based on Cohen's criteria, we cannot reject the hypothesis that the failure of the tests using interview data in other SARP experiments to reject the null hypothesis is due to limitations in the design and implementation of the experiments.[15]

The addition of statistical power calculations clarifies the extent to which the nonsignificant results in Table I can be attributed to the absence of an effect; unfortunately, the SARP reports did not include this calculation and readers could not discern readily the potential contributions of weak designs and weak effects. These power estimates are directly relevant only to prevalence measures. We have eschewed reporting power estimates for other SARP analyses because (1) such calculations are more complicated to explain, (2) the information necessary for such tests is not available in published reports, and (3) we have made our point—power analyses are necessary to make a formal interpretation of the reported findings of no significant differences.

6.3. Frequency Rates

The design of SARP called for the new experiments to go beyond the simple prevalence measures reported in the original Minneapolis study and include measures of the frequency of reoffending and the time to failure. The use of these independent dimensions of criminality had been encouraged by the National Academy of Science's panels on deterrence (Blumstein *et al.* (1978) and criminal careers (Blumstein *et al.*, 1986) and they were adopted as part of SARP's core analysis (NIJ, 1988). The repetitive nature of domestic violence and its high participation rate distinguish family violence from stranger violence (Fagan and Browne, 1994) and may have important implications for our ability to predict and prevent future domestic violence (Petrie and Garner, 1990). The use of these measures in the SARP provides an additional basis for assessing the effect of alternative police treatments on subsequent criminal behavior and, alas, another source of variation in the reported findings.

The original reports on the Minneapolis experiment did not report frequency measures for either official records or victim interviews. Most (but not all) of the published reports on the replication findings included not only the number of cases that fail at least once but also the total number of repeat incidents. Table III displays the findings on frequency levels[16] published as of September 1994. While the evidence on the direction of the

[15]Statistical significance is a characteristic of the results of research; statistical power is a characteristic of a design for research. Powerful designs may or may not find statistically significant effects. Ironically, the design with the least power (Minneapolis) and the design with the second-most power (Dade) found statistically significant deterrent effects.

[16]Table III is based on the same criterion for determining failures as Table I.

Table III. Reported Frequency of Reoffending After 6 Months

	SOURCES	Charlotte[1]		Col. Springs[2]		Dade County[3]		Milwaukee[4]		Minneapolis[5]		Omaha-OA[6]		Omaha-OP[7]	
	MEASURES	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA
RECORDS	Cases	214	436	431	1227	465	442	802	398	93	237	111	136	109	221
	# of Failures	43	84	Not Reported		135	167	304	155	Not Available		22	46	29	43
	Failures/Cases	0.20	0.19			0.29	0.38	0.38	0.39			0.20	0.34	0.27	0.19
	Direction	Escalation				Deterrent		Deterrent				Deterrent		Escalation	
SURVEYS	Cases Surveyed	112	226	Not Reported		199	182	624	297	52	115	84	112	77	165
	# of Failures	241	488			56+	96+	Not Reported		28	102	227	383	162	266
	Failures/Cases	1.13	1.12			0.12	0.22			0.30	0.43	2.05	2.82	1.49	1.20
	Failures/Surveyed	2.15	2.16			0.28	0.53			0.54	0.89	2.70	3.42	2.10	1.61
	Direction	Deterrent				Deterrent				Deterrent		Deterrent		Escalation	

A - ARRESTED NA - NOT ARRESTED NO STATISTICALLY SIGNIFICANT EFFECTS REPORTED

[1]Hirschel *et al.* (1992a): Collapses citation and advise/separate treatments into not arrested treatment. Police records findings from page 19, Table 2. Victim interview findings from page 24, Table 4.

[2]Frequency data not reported.

[3]Pate *et al.* (1991): Separate violence-only data not reported; all subsequent offenses included. Police records findings from page 6-55, Table 6-8C, and page 6-62, Table 6-10A. Victim interview findings from page 6-28. Used "hit, slapped or hurt" measure.

[4]Sherman (1992): Includes all victims. Police records findings from page 354, Table A2.20. See also Sherman *et al.* (1992b, p. 155, Table 5) for slightly higher frequency rates and an effect in the direction of escalation from an unrestricted follow-up period.

[5]Gartin (1991): Police records data no longer available. Victim interview findings from page 120, Table 5.01.

[6]Dunford *et al.* (1990b): Complaint recidivism. Police records findings from page 639, Table 3. Victim interview findings from page 640, Table 4. (Effect statistically significant at 12 months or when using "Victim physically injured" measure.)

[7]Dunford *et al.* (1990a): Complaint recidivism. Police records findings from page 200, Table 6. Victim interview findings extrapolated from frequencies on page 201, Table 7.

effect of arrest on the frequency of violent reoffending is as varied as that on the prevalence of violent reoffending, none of the comparisons in Table III rejected the hypothesis that the difference between arrested and non-arrested groups is equal to zero. Of course, frequency rates are subject to all of the same limitations—multiple measures, multiple comparisons, lack of theory and policy guidance—noted above for prevalence rates and for which we suggest caution in interpreting the published results.

Table IV. Computed Incident Rates of Reoffending After 6 Months

	SOURCES	Charlotte		Col. Springs		Dade County		Milwaukee		Minneapolis		Omaha-OA		Omaha-OP	
	MEASURES	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA	A	NA
RECORDS	Cases	214	436	431	1227	465	442	802	398	93	237	111	136	109	221
	Failed Cases	39	68	Not Reported		89	91	208	99	6	35	16	30	19	36
	# of Failures	43	84			135	167	304	155	Not Available		22	46	29	43
	Failures/ Failed Cases	1.10	1.24			1.52	1.84	1.46	1.57			1.38	1.53	1.53	1.19
	Direction	Deterrent				Deterrent		Deterrent				Deterrent		Escalation	
SURVEYS	Cases Surveyed	112	226	Not Reported		99	182	624	297	52	115	84	112	77	165
	Failed Cases	66	142			29	49	202	92	9	35	29	53	29	69
	# of Failures	241	488			56+	96+	Not Reported		28	102	227	383	162	266
	Failures/ Failed Cases	3.65	3.44			1.93	1.96			3.11	2.91	7.83	7.23	5.59	3.86
	Direction	Escalation				Deterrent				Escalation		Escalation		Escalation	

A-ARRESTED NA-NOT ARRESTED

6.4. Incident Rates

There is at least one other complication that affects the reported measures of frequency rates. Each of the SARP sites defined the rate of offending as the ratio of the number of offenses to the number of total cases. This measure, while not without some *ad hoc* rationale as policy relevant, does not conform to the understanding of a dimension of criminality that is independent of the measure of prevalence. As formulated by the National Academy of Sciences (Blumstein *et al.*, 1986), incidence measures are defined as the ratio of the number of offenses to the number of active cases. The Academy's rationale is based on the notion that simple frequency rates confound being active with the rate of activity.

This reformulation is not merely definitional. As the results displayed in Table IV reveal, the computation of the incidence measures (the number of failures/cases with at least one repeat incident) is different from that of the frequency measures (number of failures/all treated cases) in either the direction or the size of the effects in several experiments. For instance, the frequency measures reported for interview data from Minneapolis (Gartin, 1991) show a deterrent effect for arrest; incidence measures for the same site and data source show an escalation effect.[17] The results in Charlotte are

[17]Gartin's 1991 (p. 120) reanalysis computes both frequency and incidence rates.

reversed for both victim interviews and official records. None of these reversals change a significant effect, but they do reveal that there can be real differences depending on how frequency and incidence rates are computed.

Nor is this reformulation definitive. The central concept in the career criminal paradigm is λ, the annualized rate of offending among active offenders *while free*. Given the generally low rate of prosecution, conviction, and incarceration among misdemeanor spouse assault cases (Ford, 1991; Elliott, 1989), an assumption of 100% time free to commit new offenses may not be unreasonable. However, given the substantial minority of victims who report no subsequent contact with the offender during the period at risk and the prospects that these offenders might be incarcerated for other offenses, a definition of an offending rate among active offenders annualized on the time of possible contact with the victim may be more interesting theoretically or policy relevant in the context of spouse assault. In addition, incidence rates exclude some members (nonactives) of a randomized group, removing the patina of equivalence from the remaining groups.

This review cannot, by itself, settle on the appropriate measures but it can establish the strengths and the limitations of the frequency measures in SARP published reports. The strength of using multiple measures is that they can capture salient differences in the nature of the effects observed. The justification for the use of prevalence and incidence rates (Blumstein *et al.*, 1986) includes arguments that these parameters are independent of each other and we should not be surprised if alternative police responses to misdemeanor spouse assault do not have effects of the same size, direction, or statistical significance on incident rates as they do on prevalence rates. Thus, the divergent findings between prevalence and incidence rates contribute to the program's ability to assess the range of effects of alternative police responses to misdemeanor spouse assault. However, frequency rates do not necessarily provide the same independent dimension of criminality and the two measures should not be confused when testing theory or evaluating policy.

6.5. Time to Failure

The third approach to analyzing the outcomes of the SARP experiments considered here is the analysis of the time to first failure. Originally reported in the first Minneapolis reports (Sherman and Berk, 1984b), this measure was included as part of the program's core analysis (NIJ, 1988) and has been reported in several of the available SARP publications (Dunford *et al.*, 1990a, b; Sherman *et al.*, 1992a; Hirschel *et al.*, 1992b; Pate *et al.*, 1991). Analyses of the probability of failure within a temporal space, or hazard functions, have been prominent in assessing the effects of sanctions or inter-

Table V. Time to Failure Models of Reoffending

Characteristics of Analysis	Experimental Sites						
	Charlotte[1]	Col. Springs[2]	Dade County[3]	Milwaukee[4]	Minneapolis[5]	Omaha-OA[6]	Omaha-OP[7]
Treatments	Arrest Citation Separate /Advise	Arrest Restore Order Protection Order Counseling	Arrest No Arrest	Short Arrest Warning	Arrest Separate Advise	Warrant Advise	Arrest Separate Advise
Time to Failure Model	Not Reported	Not Reported	Not Reported	Mean Daily Rate	Cox regression	Kaplan-Meier	Kaplan-Meier
Statistical Tests	Lee-Desu		Lee-Desu Log-Rank Wilcoxon	t - test	t test	Mantel-Cox	Mantel-Cox
Length of Observations	180 Days	200 Days	180	180	180	360	180
Experimental Cases	650	1589	907	796	314	247	330
Direction of Effect	Escalation	Deterrent	Deterrent	Escalation	Deterrent	Deterrent	Escalation
Cases Surveyed	Not	Not	381	Not	161	196	242
Direction of Effect	Reported	Reported	Deterrent	Reported	Deterrent	Deterrent	Deterrent

SHADED CELLS DENOTE STATISTICALLY SIGNIFICANT EFFECT REPORTED

[1]Hirschel *et al.* (1992b, p. 106–107), using subsequent arrest for offense against the same victim.
[2]Berk *et al.* (1991, p. 123).
[3]Pate *et al.* (1991, pp. 6-47 to 6-51).
[4]Sherman (1992, pp. 188–205).
[5]Berk and Sherman (1988, p. 75).
[6]Dunford *et al.* (1990b, pp. 645–647).
[7]Dunford *et al.* (1990a, pp. 200–203).

ventions (see Visher *et al.*, 1991; Maltz, 1984). Table V displays some of the characteristics of these analyses and the reported findings.

The comparisons in Table V are even more disparate than the prevalence and frequency rates reported earlier. Some compare arrest and nonarrest (Dade); others compare up to four different treatments. Some are limited to 6 months (180) days; others use longer periods. Some are explicit about the underlying model; others are silent on this issue. Finally, the statistical tests reported in Table V vary with the model and the comparisons used. Given this variability, Table V is less a set of comparable comparisons and more like a census of time to failure findings from SARP publications. Eight of the 11 reported findings in Table V show an effect in the direction of deterrence, and 5 of these were reported to be statistically significant. None of the escalation findings are statistically significant. Three comparisons were not reported.

Typically, time to first failure measures are not independent of measures of prevalence; they are measures of prevalence that incorporate additional information about the time it takes to fail. Thus, when the definition of failure and the time at risk are constant, the direction of the effects in time to failure models cannot vary from those effects obtained in prevalence measures. The use of timing information, however, provides a more sensitive test of effect sizes and this is one of the primary justifications for the use of this type of analysis (Maltz, 1984; Schmidt and Witte, 1984). A comparison of Tables I and V reveals this to be the case: the size and direction of the reported prevalence and time to failure effects are the same within each site and data source. What is different about these summary findings are the tests of significance, best illustrated by the Omaha—Offender Absent Experiment. In Table I these effects are not statistically significant; in Table V they are.[18] This is not trickery or happenstance but the result of using more sensitive tests.

A second strength of the analysis of time to failure is not revealed in Table V or, in our opinion, in the reports in SARP findings. While each site reports a graph of the cumulative proportion failing over time, the evaluation of these graphs is, with one possible exception (Sherman, 1992), heavily qualitative, underdeveloped, and atheoretical. This is unfortunate. At the start of SARP, there was (and still is) no accepted theory or conventional wisdom as to when the deterrent (or escalation) effects of arrest, if they exist at all, would become apparent or when they would decay. The collection and reporting of the time to failure were incorporated into the SARP design to help provide some empirical basis for future theory, policy, and research. If we are to understand the dynamics of domestic violence and the mechanisms by which police action affects criminal behavior, we need to make systematic use of the information we have on how the distribution of time to failure varies from treatment to treatment, if at all. The published analyses of time to failure from the SARP experiments have not contributed much to that goal.

6.6. Interaction Effects

We are not the first to note the variation in the results of the SARP experiments. Others have sought to explain the dissimilarity in findings from one experiment to the next on the basis of differences in the preexisting characteristics of suspects (Sherman *et al.*, 1992a, b; Sherman, 1992; Pate and Hamilton, 1992; Berk *et al.*, 1992b). They argue that while randomiza-

[18]Tables I and V are not directly comparable since Table I uses offenses and Table IV uses arrests. However, the prevalence findings for arrest in Omaha—Offender Absent are at 0.07 and 0.08, respectively.

tion does a good job equalizing preexisting differences between treatment groups in any one experiment, the kinds of cases that became eligible for the different experiments varied from site to site. For instance, the Dade County experiment (Pate *et al.*, 1991) included only married or formerly married couples until the last few months of the experiment; other sites had more cohabitating than married couples. Most suspects in Colorado Springs were employed; most suspects in Milwaukee were not. These kinds of differences could arguably explain the apparent differences from site to site if the effectiveness of police actions vary from one type of person to another.

Three companion studies have reported analyses that explore the differences in preexisting conditions among some of the SARP sites. These analyses report multivariate models that incorporate not only the direct effects of arrest but also the interaction of arrest with two measures of stakes in conformity (Toby, 1957). If, as this line of research suggests, arrest is a more effective deterrent with married or employed suspects, some or all of the differences in the reported direct effects of arrest in SARP publications would be accounted for by differences in the proportion of married or employed suspects in each site.

Table VI displays the characteristics of these studies and their findings. The figure also shows the limitations of these published analyses. First, data from only four of the seven experiments have been analyzed[19] and the Milwaukee and Omaha analyses are not based on a complete sample of experimental cases. The analyses for Dade County, Colorado Springs, and the Omaha—Offender Present experiment are based on official police records of subsequent offenses; the original Milwaukee interaction analysis was based on a data source unique to that site—police records of calls concerning a new offense to a shelter hotline. It is not clear from the available publications which official record data source from Milwaukee was used by Berk *et al.* (1992b).

Perhaps the most striking characteristic of these studies is that they do not use the same outcome measure. The *frequency of any violence* is used in the Milwaukee analysis, the *frequency of any crime* in the Omaha analysis, the *prevalence of any repeat incident* in Colorado Springs (and the associated four-site analysis), and the *prevalence of assault* in the Dade County analysis. No rationale is provided for these varied selections. Figure 6 summarizes the direction and statistical significance of the reported findings from these studies. For only one variable is there a consistent finding: the interaction of arrest with suspect employment. Only one (Milwaukee) of the four findings for marriage is in the predicted direction; two of the marriage interaction findings (Dade and the combined sites) are in the direction opposite

[19]At the time these analyses were conducted, data from Charlotte may not have been publicly available.

Table VI. Interaction of Arrest with Stakes in Conformity in Four SARP Experiments

Characteristics	Study Authors			
	Sherman, et al., 1992a		Berk, et al., 1992b	Pate and Hamilton, 1992
Sites	Milwaukee	Omaha	Col. Springs Dade Milwaukee Omaha	Dade
Cases	1,133	239	1,658 3,937 (4 sites)	907
Data Source	Hot Line	Official Records	Official Records	Official Records
Measure	Frequency of Violence	Frequency of any crime	Prevalence of Any Repeat Incident	Prevalence of Assault
Period of Risk	6 to 22 months	13 to 31 months	6 Months	6 Months
Analysis	Negative Binomial Regression	Negative Binomial Regression	Logistic Regression	Logistic Regression
Variables in Model				
Prior Violence	Positive	Not Reported	Positive	Positive
Race	Positive	Not Reported	Positive	Negative
Education	Negative	Not Reported	Not Reported	Not Reported
Interaction Terms				
Marriage	Negative	Not Reported	Positive	Positive
Employment	Negative	Negative	Negative	Negative
Marriage and Employment	Negative	Negative	Not Reported	Negative

of that predicted by the stakes in conformity hypothesis. Analysis of this variable is not reported for Omaha. The interaction of a stakes in conformity variable (a combination of marriage and employment) is consistent in the three studies where it is reported.

While comparisons of similar studies do not always consistently report the effect for particular variables, these findings were published in companion articles in the same issue of the *American Sociological Review* and are referred to as "replications" by the authors. Moreover, none of the published articles on the interaction effects use the outcome measures from victim interviews. The inconsistencies in the published findings and the availability of analytic reports for only 4 of the 28 main comparisons[20] suggest that the published reports provide only fragmentary information from which to draw conclusions on the interaction effects. As with prevalence, incidence, and time to failure analyses, our current understanding of the interaction effects may not warrant revision after the completion of a more thorough exploration of the SARP data. However, our current knowledge relies on substantially incomplete and inconsistent reporting and analysis.

We would add that the "stakes in conformity" argument, while certainly interesting, is only one of several rival explanations for the diverse findings across SARP experiments. For instance, it is very plausible but as yet untested that the variation in both the interview procedures and the official record keeping systems could account for some of the cross-site variation in findings. Another rival and not fully explored hypothesis is that the arrest or other treatments were implemented with such dissimilarity to preclude an expectation of similarity in result. Until these and other plausible alternative explanations are examined, until more comprehensive tests of the "stakes in conformity" are completed, and, as the bulk of this article points out, until consistent measures of recidivism are employed across sites, a complete understanding of SARP findings will remain elusive.

7. CONCLUSION

The conclusions of the original SARP investigators are varied; so are the analytical approaches they used to reach those conclusions. More importantly, it is difficult, if not impossible, to assess from the published works how sensitive their conclusions are to the selection of comparisons reported. These considerations suggest caution in comparing and contrasting the published results of the SARP experiments at this time.

While some real contributions to understanding of the effectiveness of alternative police responses to spouse assault has been produced, there is much unexplored. The published articles, books, and reports on the SARP, individually and collectively, do not provide sufficient information to assess the generalizability of the original Minneapolis results; the available findings

[20]That is, 7 sites × 2 data sources (official records, victim interviews) × 2 outcome measures of recidivism (prevalence and incidence).

do not offer a single replication of the measures and analysis used in the Minneapolis study. In addition, the studies lack (a) comparable measures of prevalence for any single criterion variable, (b) either a complete census of alternative analytical comparisons or explicit rationale for the selective reporting of these comparisons, (c) measures of frequency that are independent of measures of prevalence, (d) analyses of statistical power for analytical comparisons, or (e) insights into the history of events following from the original police intervention that will allow for more detailed theoretical explanation of the mechanisms by which arrest influences subsequent violence. Finally, the published research on stakes in conformity does not include consistent tests of the interaction of arrest with marriage and employment, nor does it include a consistent set of findings for any particular model of interaction effects.

The gaps in empirical knowledge about the effectiveness of alternative police responses to spouse assault limit our understanding of deterrence theory and its relevance for spouse assault policy. These criticisms of SARP publications are not, and should not be seen to be, damning[21] of the quality of the research in SARP or the individual authors. It is because SARP was so well implemented and the publications so explicit about their methods and findings that our detailed literature review was possible. Much of the information we have identified as important and inconsistently reported could not have been provided by any of the original authors alone. The final reports to NIJ were constructed independently and in sequence over several years. Nor is it reasonable that any of the journals could be expected to publish all of the comparisons and tests we have identified.

Neither are our criticisms unfair. The information we have identified as missing does not arise from a new standard or expectation of social science. The information that was not reported or reported inconsistently was identified as an integral part of the SARP design (NIJ, 1988) long before any reports were published. These criticisms and the diverse findings from the SARP may engender less optimism about prospects for linking research and policy than the original Minneapolis experiment aroused (Lempert, 1984).

It is also evident that the full story has not yet been told. In most, if not all, instances, the limitations of the separate analyses of the SARP experiments can be redressed. Using the common data elements from these experiments, comparable measures can be computed, alternative measures can be reported for each experiment, analyses of time to failure can be conducted, consistent tests of the interaction hypotheses can be performed,

[21]Any of the authors of this article would be proud to have any of the SARP publications on our résumé; the quality of the science evident in each of them is exemplary, but not perfect.

and power analyses can be computed. Moreover, the data sets can be pooled for hypothesis tests that avoid the limitations and selection biases inherent in the individual experiments and that enhance the external validity of their outcomes. Until the kinds of common data analysis originally anticipated for SARP are completed and thorough and rigorous reanalyses of the archived data by independent investigators have established the empirical soundness of SARP findings, the fragmentary evidence and incomplete records provide a less than perfect foundation for understanding alternative police responses to spouse assault.

Replication programs such as SARP are to be recommended as essential mechanisms for organizing knowledge about the external validity of single site social experiments such as the Minneapolis Domestic Violence Experiment. The full value of such programs, however, is rarely achieved by uncoordinated individual project reports and the very strength of the replication model can be dissipated if the reporting of findings is neither comprehensive, or standardized. These steps are necessary not only to unravel the substantive puzzles in the SARP data but also to enhance the process of accumulating knowledge, advancing theory, and improving policy.

ACKNOWLEDGMENTS

This research was supported in part by grants from the Harry Frank Guggenheim Foundation and the National Institute of Justice, 93-IJ-CX-0021. All opinions are those of the authors and do not necessarily reflect the views of the Guggenheim Foundation or the U.S. Department of Justice.

REFERENCES

Bard, M. and Zacker, J. (1971). The prevention of family violence: Dilemmas of community interaction. *J. Marriage Family* 33: 677–682.

Berk, R. A., and Sherman, L. W. (1988), Police responses to family violence incidences: An analysis of an experimental design with incomplete randomization, *J. Am. Stat. Assoc.* 83: 70–76.

Berk, R. A., Black, H., Lilly, J., and Rikoski, G. (1991). Colorado Springs Spouse Assault Replication Project: Final Report. Final Report to the National Institute of Justice, National Institute of Justice, Washington, DC.

Berk, R. A., Campbell, A., Klap, R., and Western, B. (1992a). Bayesian analysis of the Colorado Springs Spouse Abuse Experiment. *J. Crim. Law Criminol.* 83: 170–200.

Berk, R. A., Campbell, A., Klap, R., and Western, B. (1992b). The deterrent effect of arrest in incidents of domestic violence: A Bayesian analysis of four field experiments. *Am. Sociol. Rev.* 57: 698–708.

Blumstein, A., Cohen, J., and Nagin, D. (eds.) (1978). *Estimating the Effects of Criminal Sanctions on Crime Rates*, National Academy Press, Washington, DC.

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

Blumstein, A. and Petersilia, J. (1994). NIJ and its research program, In National Institute of Science, 25 Years of Criminal Justice Research, National Institute of Justice, Washington, DC.

Boffey, P. M. (1983). Domestic violence: Study favors arrest. *New York Times* April 5.

Borenstein, M. and Cohen, J. (1988). *Statistical Power Analysis: A Computer Program,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Brown, E. (1989). Statistical power and criminal justice research. *J. Crim. Just.* 17: 115–122.

Cohen, J. (1988). *Statistical Power for the Behavioral Sciences, 2nd ed.,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Cohn, E. and Sherman, L. (1987). *Police Policy on Domestic Violence, 1986: A National Survey* (Report 5). Crime Control Institute, Washington, DC.

Dunford, F. W., Huizinga, D., and Elliott, D. S. (1989). *The Omaha Domestic Violence Police Experiment,* Final Report, Grant 85-IJ-CX-K435, National Institute of Justice, U.S. Department of Justice, Washington, DC.

Dunford, F. W., Huizinga, D., and Elliott, D. S. (1990a). The role of arrest in domestic assault: The Omaha experiment. *Criminology* 28: 183–206.

Dunford, F. W., Huizinga, D., and Elliott, D. S. (1990b). Victim initiated warrants for suspects of misdemeanor domestic assault: A pilot study. *Just. Q.* 7: 631–653.

Elliott, D. S. (1989). Criminal justice procedures in family violence crimes. In Ohlin, L., and Tonry, M. (eds.), *Family Violence, Volume 11 of Crime and Justice: An Annual Review of Research,* University of Chicago Press, Chicago, pp. 427–480.

Fagan, J. and Browne, A. (1994). Violence against spouses and intimates. In Reiss, A. J. and Roth, J. (eds.), *Understanding and Controlling Violence, Vol. 3,* National Academy Press, Washington, DC.

Ford, D. (1991). Prosecution as a victim power resource: A note on empowering women in violent conjugal relationships, *Law Society Rev.* 25: 313–334.

Garner, J. H. (1987). Second phase funding Milwaukee replication, Unpublished memorandum, National Institute of Justice, Washington, DC, Nov. 2.

Garner, J. H. (1989). Replicating the Minneapolis Domestic Violence Experiment, Paper presented at the British Criminology Conference, Bristol, England.

Garner, J. H. (1990). Two, three . . . many experiments: The use and meaning of replication in social science research, paper presented at the Annual Meeting of the American Society of Criminology, Baltimore, Nov.

Gartin, P. (1991). *The Individual Effects of Arrest in Domestic Violence Cases: A Reanalysis of the Minneapolis Domestic Violence Experiment,* Final Report submitted to the National Institute of Justice.

Hirschel, J. D., Hutchison, I. W., III, Dean, C. W., Kelley, J. J., and Pesackis, C. E. (1991). Charlotte Spouse Assault Replication Project: Final Report, Grant No. 87-IJ-CK-K004, National Institute of Justice, Washington, DC.

Hirschel, J. D., Hutchison, I. W., III, and Dean, C. W. (1992a). The failure of arrest to deter spouse abuse. *J. Res. Crime Delinq.* 29: 7–33.

Hirschel, J. D., Hutchison, I. W., III, and Dean, C. W. (1992b). Female spouse abuse and the police response: The Charlotte, North Carolina Experiment. *J. Crim. Law Criminol.* 83: 73–119.

Jones, A. (1992). *Next Time She'll Be Dead: Battering and How to Stop It,* Beacon Press, Boston.

Lempert, R. (1984). From the Editor. *Law Society Rev.* 18: 505–513.

Lipsey, M. W. (1983). Treatment implementation, statistical power, and internal validity. *Eval. Rev.* 4: 543–549.

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research,* Sage, Beverly Hills, CA.

Maltz, M. (1984). *Recidivism*, Academic Press, New York.
Mastrofski, S. D., and Uchida, C. D. (1993). Transforming the police. *Journal Res. Crime Delinq.* 30: 330–358.
McCord, J. (1993). Deterrence of domestic violence: A critical view of research. *J. Res. Crime Delinq.* 29: 229–239.
National Institute of Justice (1985). *Replicating an Experiment in Specific Deterrence: Alternative Police Responses to Spouse Assault,* National Institute of Justice, Washington, DC.
National Institute of Justice (1987). Common data elements. Spouse Assault Replication Program, unpublished memorandum, National Institute of Justice, Washington, DC.
National Institute of Justice (1988). Core analysis, Unpublished memorandum, National Institute of Justice, Washington, DC.
Pate, A., Hamilton, E. and Annan, S. (1991). *Metro-Dade Spouse Assault Replication Project: Draft Final Report,* The Police Foundation, Washington, DC.
Pate, A. and Hamilton, E. E. (1992). Formal and informal deterrents to domestic violence: The Dade County Spouse Assault Experiment. *Am. Sociol. Rev.* 57: 691–697.
Petrie, C. and Garner, J. H. (1990). Is violence preventable? In Besharov, D. J. (ed.), *Family Violence*, AEI Press, Washington, DC pp. 164–184.
Robertson, L. S. (1992). *Injury Epidemiology*, Oxford University Press, Oxford.
Schmidt, P., and Witte, D. (1984). *Survival Analysis*, Academic Press, New York.
Sherman, L. W. (1980). Specific deterrent effect of spouse assault, Proposal submitted to the National Institute of Justice Crime Control Theory Program, U.S. Department of Justice, Washington, DC.
Sherman, L. W. (with J. D. Schmidt and D. P. Rogan) (1992). *Policing Domestic Violence: Experiments and Dilemmas,* Free Press, New York.
Sherman, L. W., and Berk, R. A. (1984a). The specific deterrent effects of arrest for domestic assault. *Am. Sociol. Rev.* 49: 261–272.
Sherman, L. W., and Berk, R. A. (1984b). *The Minneapolis Domestic Violence Experiment,* Police Foundation Reports, No. 1, Washington, DC.
Sherman, L. W., and Cohn, E. (1989). The impact of research on legal policy: The Minneapolis Domestic Violence Experiment. *Law Society Rev.* 23: 117–144.
Sherman, L. W., Schmidt, J. D., Rogan, D. P., Gartin, P., Cohen, E. G., Collins, D. J., and Bacich, A. R. (1991). From initial deterrence to long-term escalation: Short custody arrest for poverty ghetto domestic violence. *Criminology* 29: 821–850.
Sherman, L. W., Smith, D. A., Schmidt, J. D. and Rogan, D. P. (1992a). Crime, punishment, and stake in conformity: Legal and informal control of domestic violence. *Am. Sociol. Rev.* 57: 680–690.
Sherman, L. W., Schmidt, J. D., Rogan, D. P., Smith, D. A., Gartin, P. R., Cohn, E. G., Collins, D. J. and Bacich, A. R. (1992b). The variable effects of arrest on crime control: The Milwaukee Domestic Violence Experiment. *J. Crim. Law Criminol.* 83: 137–169.
Toby, J. (1957). Social disorganization and stakes in conformity. *J. Crim. Law Criminol. Police Sci.* 48: 12–17.
Toothaker, L. (1993). *Multiple Comparison Procedures*, Sage, Newbury Park.
U.S. Attorney General's Task Force on Family Violence (1984). Final Report, Government Printing Office, Washington, DC.
Visher, C., Lattimore. P., and Linster, R. (1991). Predicting recidivism of serious youthful offenders using survival models. *Criminology.* 29: 329–366.
Zorza, J. and Woods, L. (1994). *Analysis and Policy Implications of the New Police Domestic Violence Studies*, National Center on Women and Family Law.