Journal of Criminal Law and Criminology

Volume 83
Issue 1 *Spring*

Article 5

Spring 1992

# The Variable Effects of Arrest on Criminal Careers: The Milwaukee Domestic Violence Experiment

Lawrence W. Sherman

Janell D. Schmidt

Dennis P. Rogan

Douglas A. Smith



## Recommended Citation

Lawrence W. Sherman, Janell D. Schmidt, Dennis P. Rogan, Douglas A. Smith, The Variable Effects of Arrest on Criminal Careers: The Milwaukee Domestic Violence Experiment, 83 J. Crim. L. & Criminology 137 (1992-1993)

0091-4169/92/8301-0137
THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY
Vol. 83, No. 1


# THE VARIABLE EFFECTS OF ARREST ON CRIMINAL CAREERS: THE MILWAUKEE DOMESTIC VIOLENCE EXPERIMENT

LAWRENCE W. SHERMAN, JANELL D. SCHMIDT, DENNIS P. ROGAN, DOUGLAS A. SMITH, PATRICK R. GARTIN, ELLEN G. COHN, DEAN J. COLLINS, and ANTHONY R. BACICH*

## I. INTRODUCTION

The jurisprudence of the criminal sanction has long recognized diverse objectives: deterrence, justly deserved punishment, incapac-

This research was supported by grant 86IJCXKO43 to the Crime Control Institute from the National Institute of Justice. The points of view or opinions stated herein are those of the authors and do not necessarily represent the official views of the U.S. Department of Justice or the Milwaukee Police Department. We are deeply indebted to former Milwaukee Police Chief Robert J. Ziarnik for his support of this research and to Chief Philip Arreola for continuing that support. We also thank Kathleen Stolpman and the staff of the Sojourner Truth House for their maintenance of hotline records of repeat violence. This article also reflects the work of many Crime Control Institute staff members and interviewers and the advice of Drs. Joel Garner, Albert J. Reiss, Jr., Robert Boruch, and Kinley Larntz, as well as Lucy Friedman and Allen Andrews. Most important were the officers who carried out the experiment, all of whom entered eligible cases into it: Joseph Vukovich, Zygmunt Lipski, Kenneth Jones, Michael Dubis, Lawrence Roberts, Thomas Skovera, Alan Singer, Frederick Birts, Thomas Bohl, Daniel Halbur, Peter Panasiuk, Scott Rinderle, Michael Braunreiter, Timothy Koceja, John Bogues, Edgar Bullock, Kim Stack, "Mick" Heinrich, Jerome Sims, John Wallace, Wayne Armon, Rosalie Gallegos, Edward Prah, Dean Schubert, Richard Thompson, Debra Glass, Cheryl Switzer, Robert Eckert, Daniel Kent, Tracy Becker, Steven Fyfe, Jeffery Watts, Gregory Blumenberg, Mark Hilt, and Kathleen Borkowski.

* Lawrence W. Sherman is Professor of Criminology, University of Maryland and President, Crime Control Institute. Ph.D. Yale University, 1976. Janell D. Schmidt is Director, Milwaukee Office, Crime Control Institute. M.S., University of Wisconsin at Milwaukee, 1985. Dennis P. Rogan is Vice President, Crime Control Institute. Ph.D., University of Maryland, 1988. Douglas A. Smith is Professor of Criminology, University of Maryland. Ph.D., Indiana University, 1982. Patrick R. Gartin is Associate in Criminology, University of Florida. Ph.D., University of Maryland, 1992. Ellen G. Cohn is Visiting Assistant Professor of Criminal Justice, Indiana University at Indianapolis. Ph.D., Cambridge University, 1991. Dean J. Collins is Deputy Inspector, Milwaukee Police Department. M.S., University of Wisconsin at Milwaukee, 1973. Anthony R. Bacich is Captain, Milwaukee Police Department. M.S., University of Wisconsin at Milwaukee, 1986.

itation and perhaps rehabilitation.[1] Yet it has rarely recognized arrest as a form of sanctioning, despite the widely acknowledged use of arrest for that purpose.[2] While the Supreme Court has held that pre-trial detention does not legally constitute punishment,[3] the jurisprudence of arrest must nonetheless confront problems of potential conflict among the diverse objectives of arrest as a sanction. Perhaps the most perplexing problem involves empirical evidence of conditions under which arrests *increase*, rather than reduce, the frequency of repeat offending by arrested individuals. This problem is particularly challenging for misdemeanor offenses that rarely result in prosecution and for which arrest may be the only criminal sanction ever applied.

Mandatory arrest laws for misdemeanor domestic battery have become the leading example of this problem in the jurisprudence of arrest. Enacted by some fifteen state legislatures[4] despite implicit knowledge that few arrests are ever prosecuted,[5] mandatory arrest was widely viewed as a criminal sanction that produced a specific deterrent effect.[6] This view was supported by the findings of the pioneering Minneapolis, Minnesota domestic violence arrest experiment (also called the Minneapolis Spouse Abuse Experiment), the first controlled experiment in the use of arrest for any offense, which found a substantial specific deterrent effect in a sample of 314 cases.[7] But as the authors of that experiment pointed out, the sam-

---

1 *See* HERBERT L. PACKER, LIMITS OF THE CRIMINAL SANCTION (1968).

2 WAYNE R. LAFAVE, ARREST: THE DECISION TO TAKE A SUSPECT INTO CUSTODY 437 (1965).

3 Bell v. Wolfish, 441 U.S. 520 (1979).

4 NAT'L CENTER ON WOMEN & FAMILY LAW, INC., MANDATORY ARREST SUMMARY CHART (1991). *See* ARIZ. REV. STAT. ANN. § 13-3601B (1991); CONN. GEN. STAT. ANN. § 466-386(a) (West 1990); D.C. CODE ANN. § 16-1031(a) (1991); HAW. REV. STAT. § 709-906(4) (1991); 1991 IOWA ACTS 2160; ME. REV. STAT. ANN. tit. 19, § 770 (West 1991); MO. REV. STAT. § 455.085 (1990); NEV. REV. STAT. § 171.137 (1991); N.J. REV. STAT. § 2C:25-5a (1991); OR. REV. STAT. § 133.055 (1989); R.I. GEN. LAWS § 12-29-3(B) (1991); S.D. CODIFIED LAWS ANN. § 23A-3-21 (1989); UTAH CODE ANN. § 30-6-8(2) (1991); WASH. REV. CODE ANN. § 10.31.100(2) (West 1991); WIS. STAT. ANN. § 968.075(2) (West 1990).

5 In Milwaukee, Wisconsin, for example, the prosecution rate for misdemeanor domestic battery was about ten percent at the time the Wisconsin State legislature enacted mandatory arrest for probable cause cases of that offense. In the Milwaukee experiment reported in this article, the prosecution rate was under five percent of all arrests.

6 U.S. ATTORNEY GENERAL'S TASK FORCE ON FAMILY VIOLENCE, REPORT (1984); Lawrence W. Sherman & Ellen G. Cohn, *The Impact of Research on Legal Policy: The Minneapolis Domestic Violence Experiment*, 23 LAW & SOC'Y. REV. 117 (1989) [hereinafter Sherman & Cohn, *The Impact of Research*].

7 *See* Richard A. Berk & Lawrence W. Sherman, *Police Responses to Family Violence Incidents: An Analysis of An Experimental Design With Incomplete Randomization*, 83 J. AM. STAT. ASS'N 70 (1988); Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 AM. SOC. REV. 261 (1984).

ple size precluded thorough testing of an important possibility: "that for some kinds of people, arrest may only make matters worse." They went on to recommend that "until subsequent research addresses that issue more thoroughly, it would be premature for state legislatures to pass laws requiring arrests in *all* misdemeanor domestic assaults."[8]

This article reports subsequent research that has now addressed the issue more thoroughly. Just as the Minneapolis study's authors feared, the Milwaukee, Wisconsin domestic violence arrest experiment provides substantial evidence that arrest makes some kinds of people more frequently violent against their cohabitants. This evidence creates a philosophical conflict between the objectives of punishment and deterrence, a problem with little previous commentary in the jurisprudence of sanctions. The evidence shows that, while arrest deters repeat domestic violence in the short run, arrests with brief custody increase the frequency of domestic violence in the long run among offenders in general. The evidence also shows that, among cases predominantly reported from Milwaukee's black urban poverty ghetto, different kinds of offenders react differently to arrest: some become much more frequently violent, while others become somewhat less frequently violent.

These variable effects of arrests on criminal careers raise important questions about whether crime prevention or just deserts is to be the paramount goal of the criminal sanction.[9] The longstanding jurisprudential premise that punishment always deters, or at least never backfires,[10] can no longer be accepted. Such a serious claim requires substantial documentation. This article expands upon find-

---

[8] LAWRENCE W. SHERMAN & RICHARD A. BERK, THE MINNEAPOLIS DOMESTIC VIOLENCE EXPERIMENT 7 (1984). *See also* Lawrence W. Sherman, *Experiments in Police Discretion: Scientific Boon or Dangerous Knowledge?* 47 LAW & CONTEMP. PROBS. 61 (1984) [hereinafter Sherman, *Experiments in Police Discretion*].

[9] This question is not new, even though it lacks systematic treatment in modern jurisprudence. In 1764, Cesare Beccaria argued that when the infliction of punishment produces no effect, then punishment is not morally justified and violates the social contract. CESARE BECCARIA, ON CRIMES AND PUNISHMENTS 14 (Henry Paolucci trans., 1963). A century later, Sir Arthur Conan Doyle answered the question this way: "To revenge crime is important, but to prevent it is more so." 2 THE ANNOTATED SHERLOCK HOLMES 672 (William S. Baring-Gould ed., 1967).

[10] von Hirsch, for example, has observed that

> When one seeks to justify the criminal sanction by reference to its deterrent utility, desert is called for to explain why that utility may justly be pursued at the offenders' expense. When one seeks to justify punishment as deserved, deterrence is needed to deal with the countervailing concern about the suffering inflicted. The interdependence of these two concepts suggests that the criminal sanction rests, ultimately, on *both*.

ANDREW VON HIRSCH, DOING JUSTICE: CHOICE OF PUNISHMENTS 55 (1976).

ings presented elsewhere, providing considerably more detail than has been previously reported[11] about the results of the Milwaukee domestic violence experiment.

## II. The Criminology and Jurisprudence of Police Discretion

The factual premise of mandatory arrest advocates has been that police discriminate against victims of domestic violence, largely because most police officers are men.[12] The indisputable evidence cited in support of this premise is that police often fail to make arrests in cases of misdemeanor domestic battery, a claim supported by repeated field observation studies of police decisionmaking.[13] On occasion, police have even failed to make arrests for domestic violence felonies committed in their presence. For example, in 1983, Torrington, Connecticut police officer Frederick Petrovits stood and watched as Charles Thurman, holding a bloody knife, kicked his wife Tracey in the head.[14] She was already bleeding from knife wounds in the chest, neck and throat. Petrovits did nothing as Mr. Thurman went into the house, grabbed his three-year old son, came back out and kicked his wife in the head again. Three other officers arrived and also did nothing but call for an ambulance while Mr. Thurman wandered around, continuing to threaten his wife. Only when he approached his wife again as she was lying on a stretcher did the police finally arrest Mr. Thurman, a short-order cook at a cafe frequented by local police officers.[15]

The evidence for police discrimination against women domestic battery victims is bolstered by incidents of police officers committing battery against their own wives. In the City of Chicago in 1988,

---

[11] *See* LAWRENCE W. SHERMAN, POLICING DOMESTIC VIOLENCE: EXPERIMENTS AND DILEMMAS (1992) [hereinafter SHERMAN, POLICING DOMESTIC VIOLENCE]; Lawrence W. Sherman et al., *From Initial Deterrence to Long-Term Escalation: Short-Custody Arrest for Poverty Ghetto Domestic Violence*, 29 CRIMINOLOGY 821 (1991) [hereinafter Sherman, *From Initial Deterrence*]; Lawrence W. Sherman & Douglas A. Smith, *Crime, Punishment and Stake in Conformity: Legal and Extralegal Control of Domestic Violence* (forthcoming 1992 in AM. SOC. REV.) [hereinafter Sherman & Smith, *Crime, Punishment and Stake in Conformity*].

[12] SHERMAN, POLICING DOMESTIC VIOLENCE, *supra* note 11, ch. 2.

[13] DONALD J. BLACK, THE MANNERS AND CUSTOMS OF THE POLICE 94 (1980), reports that in a 1966 study of high-crime area policing in Boston, Washington and Chicago, arrests were made in only forty-seven percent of all misdemeanors involving family members. Nan Oppenlander, *Coping or Copping Out*, 20 CRIMINOLOGY 449, 455 (1982), reports similar results from a 1977 observation study of policing in twenty-four agencies in three metropolitan areas (Tampa, Rochester, N.Y., and St. Louis): arrests were made in only twenty-two percent of all family assault cases. *See also* Delbert S. Elliott, *Criminal Justice Procedures in Family Violence Crimes, in* FAMILY VIOLENCE 427 (Lloyd F. Ohlin & Michael H. Tonry, eds. 1989).

[14] Thurman v. City of Torrington, 595 F. Supp. 1521, 1526 (1984).

[15] *Id.*

for example, at least four of the city's 12,000 police officers killed their wives and then killed themselves.[16] A 1991 lawsuit filed against the Chicago Police Department claimed it had a continuing pattern of covering up police violence against spouses.[17] The plaintiff, a police officer's wife, claimed to have been beaten for years, with no help from police supervisors to whom she complained or from officers who responded after beating incidents. After the plaintiff obtained a court order of protection, her husband stopped her on the street while she was driving her son in her car. Her husband was in uniform, in a marked squad car, with his uniformed partner sitting in the car. The husband beat his wife in full view of both his partner and his son. The partner did nothing to intervene, even though he knew there was a valid order of protection being violated. The officer was later tried and convicted on battery charges but was not immediately dismissed from the police force.[18]

The problem with the use of these facts as evidence of *discrimination* against women victims of domestic violence is that they are silent about *disparity*. If one assumes full enforcement of laws against other offenses, then the evidence of under-enforcement of this offense is sufficient. But if full enforcement is only a myth, then the question becomes how much difference there is between the probability of arrest (given an opportunity) for domestic violence and that for other offenses. That this is the appropriate question is clear. Full enforcement is indeed a myth, and American police practice "aggressive" under-enforcement of most offenses.[19] One study found that even with the suspect present and with legally sufficient evidence, police made arrests in only forty-four percent of all reported misdemeanors and fifty-eight percent of all reported felonies.[20] Other studies reach similar findings.[21] For a wide variety of reasons, police ignore most opportunities to make arrests.[22]

Criminological study of police discretion has established little

16 Jacob R. Clark, *Policing's Dirty Little Secret?*, LAW ENFORCEMENT NEWS, April 15, 1991, at 1, 10.

17 *Id.* at 1.

18 *Id.*

19 *See* Harold E. Pepinsky, *Better Living Through Police Discretion*, 47 LAW & CONTEMP. PROBS. 249 (1984).

20 BLACK, *supra* note 13, at 90.

21 *See* Douglas A. Smith & Christy A. Visher, *Street-Level Justice: Situational Determinants of Police Arrest Decisions*, 29 SOC. PROBS. 167 (1981).

22 This fact has stimulated extensive social science theorizing and commentary. *See, e.g*, MICHAEL P. BANTON, THE POLICEMAN IN THE COMMUNITY (1964); BLACK, *supra* note 13; MICHAEL K. BROWN, WORKING THE STREET (1981); ALBERT J. REISS, THE POLICE AND THE PUBLIC (1971); JEROME H. SKOLNICK, JUSTICE WITHOUT TRIAL (1966); JAMES Q. WILSON, VARIETIES OF POLICE BEHAVIOR (1968).

consistent explanation of the causes of police behavior,[23] but one nearly universal finding is that police attend to the demeanor or overall "moral worth" of the suspect and victim. If police are clearly not blind "ministerial" agents automatically carrying out the legislature's commands, a more accurate description seems to be that they are judicial officials administering their own conceptions of just deserts.[24] As sociologist William Westley observed in the first systematic field study of an American police department (Gary, Indiana in 1949), police do not enforce the law so much as their own morality.[25] Police routinely speak of suspects who "fail the attitude test," or who are guilty of "contempt of cop," or who are just plain bad people, denoted by the widespread police use of the label "asshole."[26] The importance of police "gut" reactions to people and situations has so shaped our understanding of what police do that one scholar makes it part of the very definition of policing: "a mechanism for the distribution of non-negotiably coercive force employed in accordance with the dictates of an *intuitive* grasp of situational exigencies."[27] Much of this intuition goes beyond the "craft" of how to accomplish a goal in a particular situation[28] to a moral judgment about what that goal should be.

The "police justice" model of discretion has a clear consequence for domestic violence: leading police to arrest the "unemployed, unmarried, nonchurchgoing riffraff," while letting the more respectable (and deferential) suspects they encounter go free.[29] This practice is clearly supported by a just deserts view of police as judicial officials and a free will conception of human behavior. It falls down, however, on a premise of deterrence and determinism. Criminological theory for the past half century has suggested that persons most likely to be arrested for domestic violence are the persons least likely to be deterred by an arrest.[30] That was one reason why a controlled experiment was necessary to test the effects of arrest—even on people whom police would normally not arrest. The more important reason, though, was to determine how police

23 *See* Lawrence W. Sherman, *Cause of Police Behavior: The Current State of Quantitative Research*, 17 J. RES. CRIME & DELINQ. 69 (1980).

24 *See* Sherman, *Experiments in Police Discretion*, *supra* note 8.

25 *See* WILLIAM A. WESTLEY, VIOLENCE AND THE POLICE (1970).

26 *See* John Van Maanen, *The Asshole*, *in* POLICING: THE VIEW FROM THE STREET (Peter K. Manning & John Van Maanen eds. 1978).

27 EGON BITTNER, THE FUNCTIONS OF THE POLICE IN MODERN SOCIETY 46 (1970).

28 *See* WILSON, *supra* note 22.

29 *See* Sherman, *Experiments in Police Discretion*, *supra* note 8, at 78.

30 *See, generally*, TRAVIS HIRSCHI, CAUSES OF DELINQUENCY (1969); JOHN LOFLAND, DEVIANCE AND IDENTITY (1969).

could best prevent future domestic violence, regardless of any "intuitive" grasp of the justice of the situation.

The results of the pioneering Minneapolis experiment helped proponents of mandatory arrest to try to eliminate the police justice model and restore the ministerial model, for that one offense.[31] To our knowledge, no other type of offense has ever been subjected to an offense-specific mandatory arrest statute by any state legislature. While field research suggests that police may easily evade such mandates,[32] the laws have at least increased substantially the chances of suspects' being arrested for domestic violence.[33] They may even have created the closest approximation of full enforcement ever achieved by American police. Whether or not this approach can ever eliminate discretion is less important than the content of the mandate: to arrest everyone, regardless of the likely effects of the arrest on future violence.

An alternative to the ministerial approach is to take the likely consequences of arrest into account in exercising police discretion. The key criterion for deciding to arrest in any specific case would be the probable effect of the arrest on the suspect's future conduct, based on predictions derived from controlled experiments in arrest. This "professional crime control" model poses enormous difficulties in finding legally and ethically acceptable guidelines for when arrests should and should not be made.[34] Yet the difficulties may be no greater than the inequitable consequences resulting from a mandatory arrest policy. Equal protection for suspects may produce unequal protection for victims.

The choice between "justice" and "crime control" models of police discretion, up to now, has been moot. As long as criminology merely raised questions of justice by documenting the inequities of police discretion, the choice was limited to legislative versus police conceptions of justice. This choice attracted relatively little public concern outside the scholarly community of criminal law and criminology, allowing the jurisprudence of arrest to lie dormant in recent years. But if the evidence presented below is at all persuasive, it demonstrates the need for a new approach to police discretionary rule-making: one that confronts the variable effects of arrest on criminal careers.

---

31 *See* Sherman & Cohn, *The Impact of Research*, *supra* note 6. *See also* James W. Meeker & Arnold Binder, *Experiments as Reforms: The Impact of the 'Minneapolis Experiment' on Police Policy* , 17 J. POLICE SCI. & ADMIN. 147 (1990).

32 *See* Kathleen J. Ferraro, *Policing Woman Battering*, 36 SOC. PROBS. 61 (1989).

33 *See* SHERMAN, POLICING DOMESTIC VIOLENCE, *supra* note 11.

34 Sherman, *Experiments in Police Discretion*, *supra* note 8, at 76.

## III. Research Design

From April 7, 1987 to August 8, 1988, the Milwaukee Police Department conducted a controlled experiment in the use of arrest for misdemeanor domestic battery.[35] A controlled experiment is a research design which attempts to isolate a cause and effect relationship between two variables;[36] in this case, police decisions to arrest or not to arrest and subsequent domestic violence by the suspects. The essential logic of a controlled experiment is to make two or more groups virtually identical in all respects except one: the treatment to be evaluated (in this case, arrest). The elimination of rival hypotheses allows a very strong inference of cause and effect to be made about differences in the groups observed.

The method by which pre-existing differences between the groups are minimized or almost eliminated is called *random assignment*, a lottery method giving each suspect an equal probability of receiving each treatment.[37] Thus, whether a suspect is arrested or not is purely a matter of chance, regardless of police officers' intuitive grasp of the circumstances. This method of evaluating legal practices has been endorsed by an advisory committee of the Chief Justice of the United States,[38] and it has not been subject to legal challenge in the arrest experiments conducted to date. The equal probability of arrest and no arrest in the Milwaukee experiment was produced by a computer-generated sequence of police responses (treatments) in advance of the experiment. This sequence was sealed and kept secret from all participants in the experiment until the actual occurence of each of the 1200 cases eligible for entry into the experiment.

Unlike the earlier Minneapolis experiment (and all of its other replications), the Milwaukee experiment was conducted well after the May 1, 1986 implementation of a citywide policy of mandatory arrest. Thus it had the effect of reducing the severity of police response in the control group, rather than increasing it in the experimental group. While that effect improved the ethical posture of the experiment,[39] it is unclear what effect it may have had on the results. The effect of giving a "break" to the control group may be different from "cracking down" on the experimental group. The low level of

---

35 *See* SHERMAN, POLICING DOMESTIC VIOLENCE, *supra* note 11, at app. 2; Sherman, *From Initial Deterrence*, *supra* note 11, at 826.

36 *See* SOCIAL EXPERIMENTATION (Henry W. Riecken & Robert F. Boruch, eds., 1974).

37 *See* STUART J. POCOCK, CLINICAL TRIALS: A PRACTICAL APPROACH (1983).

38 ADVISORY COMMITTEE ON EXPERIMENTATION IN THE LAW, FED. JUDICIAL CTR., REPORT (1981).

39 *See* Norval Morris, *Impediments to Penal Reform*, 33 U. CHI. L. REV. 627 (1966).

awareness of the mandatory arrest policy among the victims and suspects in the sample, however, suggests that the prior existence of mandatory arrest had little effect on the results.[40]

### A. SAMPLE

The experiment was conducted in four of the six police patrol districts in Milwaukee. While the districts were racially and economically diverse, most of the cases in the experiment came from poor black neighborhoods. This is consistent with the often-observed pattern of greater frequency of requests for police intervention in domestic disturbances in such areas than in predominantly white working class and middle-class neighborhoods.[41] The resulting sample of suspects was ninety-one percent male, seventy-six percent black, sixty-four percent never married to the victim, fifty-five percent unemployed, thirty-one percent high school graduates, forty-two percent intoxicated at the time police arrived, and fifty percent with a prior arrest record, consisting of thirty-two percent with a prior arrest for domestic battery against anyone and twenty-six percent with a prior arrest for a battery against the same victim as in the presenting case. These characteristics of the 1200 eligible cases were not very different from the 854 ineligible cases encountered by the thirty-five specially selected officers who participated in the experiment; the most frequent reason for ineligibility was the absence of the offender from the scene (Table 1).

---

[40] Twenty-four percent of the victims and nineteen percent of the suspects interviewed correctly identified the city's policy of mandatory arrest. Sherman, *From Initial Deterrence*, *supra* note 11, at 845.

[41] *See* BLACK, *supra* note 13, ch. 6; M. P. Baumgartner, *Law and the Middle Class: Evidence From a Suburban Town*, 9 LAW & HUM. BEHAV. 3 (1985).

**TABLE 1**
**CASE ACTIVITY AND INELIGIBILITY REASONS BY DISTRICT**

| | District 2 | | District 3 | | District 5 | | District 7 | | Total |
|---|---|---|---|---|---|---|---|---|---|
| Number of Ineligible Cases | 119 | | 363 | | 157 | | 215 | | 854 |
| Primary Ineligible Reason | N | % | N | % | N | % | N | % | |
| Suspect Not On Scene | 63 | 53 | 211 | 58 | 63 | 40 | 144 | 67 | 481 |
| Open Warrants, Commitments | 12 | 10 | 36 | 10 | 24 | 15 | 11 | 5 | 83 |
| Imminent Danger To Victim | 12 | 10 | 24 | 7 | 9 | 6 | 8 | 4 | 53 |
| Serious Injury To Victim | 6 | 5 | 19 | 5 | 10 | 6 | 11 | 5 | 46 |
| Both Parties Arrested | 2 | 2 | 13 | 4 | 5 | 3 | 5 | 2 | 25 |
| Officer Decision | 6 | 5 | 6 | 2 | 7 | 4 | 4 | 2 | 23 |
| Valid Restraining Order | 1 | 1 | 5 | 1 | 6 | 4 | 4 | 2 | 16 |
| Victim Insists On Arrest | | | | | | | | | 13 |
| Officer Assaulted | | | | | | | | | 9 |
| Victim Assaulted At Scene | | | | | | | | | 7 |
| Other | | | | | | | | | 98 |

The 1200 eligible cases encountered by the thirty-five experimenting officers constituted twenty-five percent of all domestic violence incidents reported by all police in those four districts during the eight-hour shift (7:00 p.m to 3:00 a.m.) in which the experiment was conducted. There is good reason to believe that the experimental cases were typical of all cases citywide, since all officers in those four districts in those eight hours produced forty percent of all domestic batteries citywide, twenty-four hours a day. Dispatchers were instructed to refer cases to the experimental officers whenever they were available, regardless of the area of the district in which the case was located. Other officers also frequently referred cases to the experimental officers, especially when they judged the cases to be eligible: suspect and victim both present; probable cause to arrest; victim and suspect currently or formerly married, cohabiting, or parents of a child in common; no valid restraining order in effect; no outstanding arrest warrants against either party; one party only eligible for arrest; no serious injury; no apparent threat of immediate violence after police leave; and a victim who did not insist upon an arrest being made.

These restrictions created some limitations on the generalizability of the results but apparently allowed about half of all mandatory arrest situations into the experiment (with fifty-eight percent of the cases the experimenting officers encountered). Moreover, inspection of the cases deemed ineligible in each of the four districts shows a fairly high level of consistency (Table 1).

### B. RANDOM ASSIGNMENT AND TREATMENTS

If the case was deemed eligible, participating officers agreed to radio headquarters for a warrant check. If no warrants were outstanding, they were to radio or phone the Crime Control Institute (CCI) office with the names and dates of birth of the suspect and victim, as well as the officer's payroll number. The CCI staff would then open a wax-sealed envelope (prepared in Washington, D.C.) in a pre-arranged sequence, containing a piece of paper marked "1," "2," or "3." The numbers were codes for police actions (treatments):

Code 1: Standard arrest under mandatory arrest policy; suspect eligible for release on $250 bail, cash or credit card.

Code 2: Suspect to be arrested in the same way, but to be released on personal recognizance as quickly as possible after arrival at central booking, preferably within two hours.

Code 3: Suspect not to be arrested, but police to read a standard warning of arrest if police had to return that evening.

The labels "1", "2" and "3" as well as "Full Arrest," "Short Arrest" and "Warning" are used below in both text and tables as shorthand for these three treatments.

The purpose of comparing two lengths of time in custody was to determine whether differences across police agencies in average dosage of custody time affected the results of arrest. The earlier Minneapolis experiment had been conducted with a night in jail as the minimum dosage, while other agencies around the country were reportedly releasing arrested suspects within two hours.[42]

This screening process was to be undertaken regardless of prior contact with the experiment, just as the earlier Minneapolis experiment had done. The one exception was for prior Code 3 cases on the same night. If the officers had to return again, they were instructed to abort the random assignment and make an arrest, consistent with the warning delivered on the first encounter. Handling each *event* as the unit of analysis for separate randomization—rather than consistent application of the same treatment once a *suspect* had been randomized as the unit of analysis—was a major difference between Milwaukee and several of the other replications of the pioneering Minneapolis experiment, such as in Omaha, Nebraska.[43]

An even greater difference between the Milwaukee and Minneapolis experiments was the high degree of compliance with the randomized design achieved by the Milwaukee officers. As Table 2 shows, in over ninety-eight percent of the cases, the treatments actu-

[42] Our own survey of fifteen Wisconsin police departments found that eight of them released domestic violence suspects in less than three hours. Sherman, *From Initial Deterrence*, *supra* note 11, at 824.

[43] *See* Franklyn W. Dunford et al., *The Role of Arrest in Domestic Assault: The Omaha Police Experiment*, 28 CRIMINOLOGY 183 (1990).

**TABLE 2**
**TREATMENTS AS RANDOMIZED AND DELIVERED**

| | Treatments as Randomized | | | |
|---|---|---|---|---|
| Treatments as Delivered | Arrest/Hold | Arrest/Release | Warn | Total |
| Arrest/Hold | 400 | 13 | 1 | 414 |
| Arrest/Release | 1 | 384 | 1 | 386 |
| Warn | 3 | 1 | 396 | 400 |
| Total | 404 | 398 | 398 | 1200 |

ally delivered were the same as the randomly assigned treatments contained in the envelopes. This includes repeat randomization of some couples, for a total of 1,112 couples across the 1200 cases.[44]

Most of the twenty "treatment failures," as we trained police to think of them, were cases randomly assigned to arrest and release which had to be misassigned to arrest and hold. Most of those, in turn, were due to failures of information systems supporting police in the field. The most common problem (6 of the 20) was incorrect field information about whether the suspect was wanted on a warrant. When the arrest/release suspects were brought to headquarters for booking, they were subjected to a second warrant check. Three of those suspects were found to have given false names in the field, and three were found to have had a warrant that the original, radio-transmitted warrant check had not found. A seventh case was barred for early release by the booking officers because of outstanding municipal warrants, in violation of the official orders for the experiment.

The remaining thirteen reasons for misassignments reveal the human limitations on random assignment in these circumstances. Seven of those cases were caused by unpredictable events after the envelope was opened. Two of those cases were changed from arrest/release to arrest/hold after the suspect became violent in the booking area. Two cases were changed because the suspects were hospitalized and could be neither booked nor released on recognizance. Two cases were changed to arrest and hold after evidence of additional crimes was discovered at the scene (theft in one case, drug possession in another). One case was changed to arrest and hold due to an escalation of danger at the scene after the envelope was opened. The last six misassignments were due to simple officer error.

The three treatments produced substantially different exper-

44 This means that 7.3 % (88 of 1200) of the randomized cases were repeat couples, almost identical to the 7.5% (25 of 330) in the earlier Minneapolis study.

iences for both victims and suspects.[45] Perhaps the most important difference was the special processing needed to get the short arrest suspects out of custody within the two-hour goal. The result of their being taken to the head of the line at most stages of the booking process was an average time in custody of about three hours, compared to an estimated eleven hours or more for the suspects randomly assigned to full arrest.[46] Whether this experience is comparable to speedy booking for everyone in smallèr police agencies remains an unanswered question.

### C. OUTCOME MEASURES

Four outcome measures were used to estimate the prevalence and frequency of repeat violence by the sample suspects. The most comprehensive and precise was the "hotline" reports called in by all police citywide to the battered women's shelter whenever they encountered a case of domestic battery, whether or not they could make an arrest. These reports encompassed most, but not all, of the second and third data sources: arrests of the suspects for repeat violence (against any victim, including the same one as in the presenting incident), and offense reports of repeat violence by the same suspect against the same victim. All three of these "official" sources were available for 100% of the cases.

The fourth data source was up to two face-to-face interviews conducted with the victim in each randomized case. One interview was attempted shortly after the presenting incident, for the first 900 of the 1200 cases. A separate interview was attempted in all 1200 cases six to twelve months after the presenting incident. The initial interviews were suspended after 900 cases to test for any possible influence of the interviews on the rate of repeat violence.[47] Response rates for both interviews were fairly high, at seventy-eight percent for the initial interviews and seventy-seven percent for the long-term follow-ups.

## IV. MAIN EFFECTS

The analysis of the Milwaukee experiment proceeded in two stages. The first stage was the analysis of the "main effects" of the randomized experiments, or the differences (or lack of them) in outcome measures *between* the three treatment groups. The second

---

45 Sherman, *From Initial Deterrence*, *supra* note 11, tbl. 2, at 831.

46 *Id.*

47 No differences in effects of arrest were found between the last 300 cases and the first 900 cases.

stage analyzed differences in treatment effects *within* various subgroups of the sample. Of the two, the main effects are more statistically powerful and more straightforwardly interpretable. Their analysis begins with an examination of the effects of the treatments on the amount of time each couple spent together during the follow-up period. Answering this question is a necessary first step in determining whether any differential incapacitation effect has occurred which might obscure or falsely portray any deterrent effects.[48]

### A. TIME-AT-RISK

One possibility is that making an arrest might be more likely than failing to make an arrest to break up a couple; the arrested suspect may simply never return home after the arrest, whereas the warned suspect was never taken away. This did not happen very often, however. Among the Milwaukee arrest group couples, seventy-four percent had been together again by the time of the initial interview. By the time of the six month interview, forty-one percent of all victims said they were living with the suspect then, and another thirty-one percent said they had lived with the suspect for at least part of the time since the randomized police response. Among those who were living together, seventy-two percent had cohabited all of the time since the randomized response.

The key question for our analysis is whether time-at-risk varied by treatment group. One way to answer that question is by analyzing the set of interviews that were done consistently near the six month anniversary of the randomized response, namely the 563 follow-up interviews completed between case 473 and case 1200. These data have the least amount of error in estimating time-at-risk due to variations in the amount of time since the presenting incident. They show that there were only slight differences in the extent of cohabitation across the three treatment groups.

Among those interviewed close to six months in all three treatment groups (N=563, see *supra*), the majority of couples were no longer cohabiting: only forty-five percent of the full arrest cases, forty-four percent of the short arrest cases, and thirty-eight percent of the warning cases were cohabiting at the time of the interview. Of those still cohabiting at six months, the proportions who had cohabited the entire time since the presenting incident were seventy

48 *See* Albert J. Reiss, *Some Failures in Designing Data Collection That Distort Results*, *in* COLLECTING EVALUATION DATA: PROBLEMS AND SOLUTIONS 161 (Leigh Burstein et al., eds., 1985).

TABLE 3
COHABITATION DAYS TO FOLLOW-UP INTERVIEW
BY TREATMENT GROUP

| Treatment Group | N of Interviews | Mean Days to Interview | Mean Days Cohabitation | Standard Deviation | Cohabitation Ratio |
|---|---|---|---|---|---|
| Full Arrest | 315 | 292 | 136 | 151 | .47 |
| Short Arrest | 280 | 279 | 115 | 139 | .41 |
| Warning | 287 | 295 | 121 | 147 | .41 |

| t tests | | | |
|---|---|---|---|
| Full arrest vs. Short arrest | t = 1.79, | df = 593, | p = .074 |
| Full arrest vs. Warning | t = 1.27, | df = 600, | p = .205 |
| Short arrest vs. Warning | t = − .49, | df = 565, | p = .623 |

percent, seventy-one percent, and eighty-five percent, respectively. Among couples not cohabiting at the time of the interview, thirty-one percent, thirty-two percent and twenty percent, respectively, had cohabited for some portion of the six month follow-up period. Note that the differences between the arrest and warning groups are not always consistent in direction, although they do show lower prevalence of any cohabitation in the warning group compared to the arrest group.

Another test for differences in time-at-risk is to estimate the total number of days of cohabitation reported by the victims at all follow-up interviews, regardless of when the interviews were done (N= 882, Table 3). This procedure required distinguishing four categories from among the interview data: (1) those who cohabited continuously; (2) those who had not cohabited continuously but were cohabiting on the date of the interview; (3) those who were not cohabiting on the date of the interview but had cohabited some of the time since the presenting incident; and (4) those who had not cohabited at all since the presenting incident. Precise estimates of the number of days of cohabitation were available for the first and fourth categories from the dates of the presenting incident and the interview. The two middle categories, however, provide only victim recall, in days, weeks or months, to estimate the days of cohabitation.[49]

Table 3 presents the results of our estimates (expressed as Mean Days Cohabitation) of actual days at risk, for each treatment

49 Victim recall of the days of cohabitation is far from perfect. In five cases, for example, the victim's estimates for groups 2 and 3 were in excess of the time between the presenting incident and the interview. In twenty-two other cases, the victim said they had lived together some of the time but provided no estimate for how much time. For reasons like this, we treated 39 of the 921 interviews as missing data, without examining their treatment groups. That left 882 interviews across all three treatment groups.

group. It shows that there were no greater differences in time-at-risk than we would expect by chance variation (p =.05). It also shows that, on average, all three treatment groups were cohabiting less than half the time from the presenting incident to the interview. We do not know whether this represents a before-after decrease in the cohabitation ratio (days cohabiting divided by total days). The relationships could have been just as intermittent and variable in level of cohabitation in the period before the presenting incident as in the period after. We do know, however, that ninety percent of the 1200 police reports and seventy-four percent of the 900 initial victim interviews reported that the couples were cohabiting on the date of the presenting incident. This compares to only forty-one percent of the total follow-up interviews reporting cohabitation since the presenting incident. Moreover, thirty-six percent (114) of the full arrest group's victims, forty-two percent (119) of the short arrest victims, and forty-seven percent (135) of the warning group's vitim's reported zero days of cohabitation since the presenting incident. The evidence suggests, then, that there was a reduction in the prevalence of cohabitation (as a percentage of all couples), even if there might not have been an overall reduction in couple-days at risk.

We conclude two things from these findings. First, the differences in contact across treatment groups are not great enough to affect the findings presented below about differences in repeat violence between the groups. Whatever differences in recidivism we find are more likely attributable to deterrence or escalation than to differential time-at-risk. Second, the potential differences in deterrence *within* the relationship are greatly attenuated by the low time-at-risk overall and by reduced prevalence of cohabitation. This is significant from a policy standpoint, since it suggests that no matter which of the three responses police provided, one major sequel was a tendency for the couple to split up. This fact alone may help explain, for example, the finding that women who called police about domestic violence in the late 1970s (when police did not usually make arrests) were half as likely to suffer repeat violence as those who did not call police—an effect possibly due entirely to reduced time-at-risk.[50]

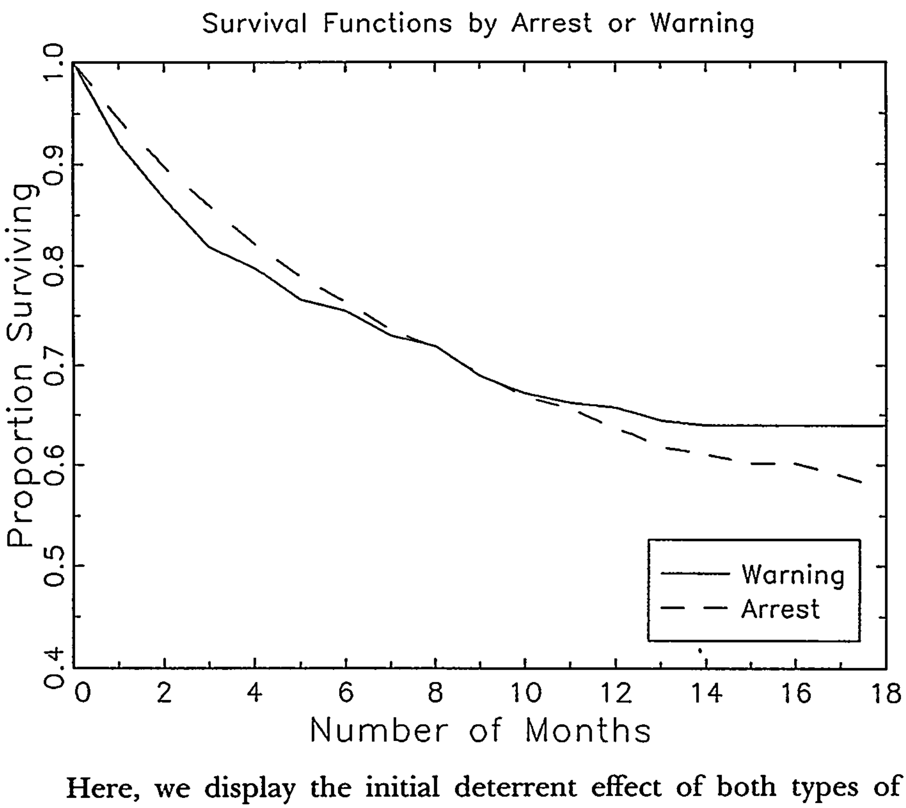
### B. INITIAL DETERRENCE

We have reported elsewhere a clear initial deterrent effect of

[50] *See* PATRICK A. LANGAN & CHRISTOPHER A. INNES, BUREAU OF JUSTICE STATISTICS, PREVENTING DOMESTIC VIOLENCE AGAINST WOMEN (1986).

both short and full arrest treatments in comparison to the warning treatment.[51] For thirty days or more after the presenting incidents, the *prevalence* (proportion of cases with one or more instances of) of repeat violence reported in victim interviews is substantially lower for the arrest groups. For short arrest only, the *frequency* (average number of instances per case) of violence reported to the hotline is significantly lower than for the warning group. Other official measures (arrest and offense reports) show no evidence of initial deterrence, either in frequency or prevalence.

FIGURE 1




Here, we display the initial deterrent effect of both types of arrest combined, a procedure recommended by some commentators.[52] Figure 1 shows the "survival" trend in the prevalence of repeat violence over time, with an obviously clear advantage for the arrested suspects in the early days. At about seven to nine months after the presenting incidents, however, the arrest and non-arrest

---

[51] Sherman, *From Initial Deterrence*, *supra* note 11, at 836.

[52] *See, e.g.*, Arnold Binder & James W. Meeker, *Experiments as Reforms*, 16 J. CRIM. JUST. 347 (1988).

**TABLE 4**
**LONG TERM PREVALENCE OF SAME-VICTIM REPEAT VIOLENCE**
(during period up to follow-up interview date)

| Measure | Treatment | | | P Value of Pair Differences* | | |
|---|---|---|---|---|---|---|
| | Full Arrest | Short Arrest | Warning | 1&2 | 1&3 | 2&3 |
| Sample N = 921 | 324 | 300 | 297 | | | |
| All Victim Interviews | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 113 | 89 | 92 | | | |
| Prevalence Ratio | 35% | 30% | 31% | | | |
| Hotlines to Interview Date | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 88 | 80 | 78 | | | |
| Prevalence | 27% | 27% | 26% | | | |
| Arrests to Interview Date | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 66 | 62 | 69 | | | |
| Prevalence Ratio | 20% | 21% | 23% | | | |
| Offenses to Interview Date | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 86 | 75 | 75 | | | |
| Prevalence Ratio | 27% | 25% | 25% | | | |
| Any Measure to Interview Date | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 148 | 131 | 131 | | | |
| Prevalence Ratio | 46% | 44% | 44% | | | |

* $P \leq .05$, two tailed tests. n.s. means non-significant.

curves cross over, and from there on out the arrest group does worse.[53]

C. LONG-TERM ESCALATION

Whatever the initial effects may be, there is clearly no long-term deterrence from arrest in the Milwaukee experiment. Tables 4 and 5 show no reductions in either the prevalence of same-victim violence or the frequency of any-victim violence in the arrest groups compared to the non-arrest (warning) group. The only significant differences, in fact, are those showing arrest *increasing* the risk of violence. These differences are not consistent enough across measures for us to draw a conclusion that arrest backfired, and the magnitude of the increased risk from arrest is generally small. But the direction of the difference is fairly consistent across measures in favor of warnings yielding lower long-term risks of repeat violence.

The problem with Tables 4 and 5 is that they suffer from "truncation," as statisticians call it, in their long term effects. The follow-up period is cut off arbitrarily, and the truncation is inconsistent across cases. This raises various problems of interpretation, and

[53] In order to make them comparable, Figures 1 and 2 are limited to the 1,133 cases for which employment data are available.

TABLE 5
LONG-TERM FREQUENCY OF ANY-VICTIM REPEAT VIOLENCE
(unrestricted follow-up period)

| Measure | Treatment | | | P Value of Pair Differences* | | |
|---|---|---|---|---|---|---|
| | Full Arrest | Short Arrest | Warning | 1&2 | 1&3 | 2&3 |
| Sample N = 1200 | 404 | 398 | 398 | | | |
| Hotlines | | | | n.s. | .02 | .00 |
| Repeat Violence N | 296 | 301 | 261 | | | |
| Mean Events Per Suspect | .73 | .76 | .66 | | | |
| Arrests | | | | n.s. | n.s. | n.s. |
| Repeat Violence N | 146 | 157 | 151 | | | |
| Mean Events Per Suspect | .36 | .39 | .38 | | | |
| Offenses | | | | .04 | n.s. | n.s. |
| Repeat Violence N | 200 | 168 | 179 | | | |
| Mean Events Per Suspect | .49 | .42 | .45 | | | |
| Offenses Without Arrests | | | | .00 | .02 | n.s. |
| Repeat Violence N | 134 | 84 | 101 | | | |
| Mean Events Per Suspect | .33 | .21 | .25 | | | |

* $P \leq .05$, two tailed tests. n.s. means non-significant.

also violates important assumptions necessary to use the tests of statistical significance we have employed here. In order to deal with the truncation problem and to take full advantage of the maximum period of observation completed after each randomized case, we computed the mean number of days to the first repeat incident of domestic violence among the thirty-six percent of all cases with any repeat violence at any time, during a period of up to twenty-two months after the randomized police response.[54] This comparison of arrest and warning yields a statistically significant escalation effect for the arrest treatment. At a mean of 124 days to first repeat violence, the combined arrest group recidivated twenty-three percent sooner than the warning group, which averaged 160 days to first failure.

The time to failure measures, however, have great limitations for policy research on violence. Originally designed to analyze the permanent "failures" of light bulbs burning out or medical patients dying, the models lose the important information on what happens after the initial failure. The question of total repeat violence, and not just whether there has been any, is also an important one to the police officers who conducted the experiment. As they told us in our last meeting of the experiment, their primary concern was the reduction of calls to police about domestic violence citywide. This

54 This computation is also restricted to the 1,133 cases used in Figures 1 and 2.

concern requires that effects on high-rate offenders be weighted more heavily than effects on low rate offenders, with analyses that take total numbers of violent events into account.

If that is the case, then the victim interview data must be cast aside, given the difficulty of obtaining a precise count of events in the victim interviews. (They were also set aside, of course, from the time-to-failure analyses in Minneapolis and Omaha because victims also have difficulty in giving a precise date for even one offense.) The hotline data, however, are ideally suited to the task of providing exact counts. And as reported elsewhere, the hotline data also show a statistically significant long-term escalation effect from arrest.[55] The effect is limited to the short arrest treatment only, but that fact may have broad policy significance for the many police agencies releasing domestic violence arrestees within three hours of arrest.

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. While the large number of tests showing statistical nonsignificance may make some readers suspect that some of the effects occurred by chance, there is little doubt that the main effects of the Milwaukee experiment fail to replicate the strong specific deterrence showing of the earlier Minneapolis experiment.

## V. Variable Effects

The large sample size of the Milwaukee experiment was explicitly designed to go beyond the main effects and to explore the possibility that arrest may have different effects on different kinds of people. Four years before the experiment began, the central hypothesis was described at a Duke Law School conference on police discretion: that more socially marginal people, as indicated by such characteristics as unemployment and unmarried cohabitation, would be less deterrable than less marginal people.[56] The experiment collected data on both those indicators of marginality, as well as several others, including high school graduation, length of prior cohabitation, and race (because of its effect on employment rates). The hypothesis was not necessarily that arrest would backfire for the more marginal groups, although that is generally what was found with respect to frequency of repeat violence and less so with respect to its prevalence.

[55] Sherman, *From Initial Deterrence*, *supra* note 11, at 837.

[56] Sherman, *Experiments in Police Discretion*, *supra* note 8, at 78.

A. TWO CAUTIONS

### 1. *Experimental vs. Correlational Results*

The whole purpose of doing experiments is to reduce the uncertainty associated with correlational analysis. The endless number of possible correlations to test always leaves researchers uncertain whether the correlations found are true "causes" or mere coattails to a hidden truly causal factor. By randomizing, experimenters virtually eliminate such unknown rival hypotheses. That is why the main effects analysis is more straightforward.

A problem arises when one begins to explore how different subgroups within an experiment react to the experimental treatment. The strongest way to examine that question is to plan those explorations in advance, building them into the design. By "blocking," or assigning police responses under a separate random schedule for each subgroup, one would still eliminate rival hypotheses for the apparent effects of the treatments within each group.[57] If we had done that in Milwaukee, for example, we would have had a separate set of pre-randomized envelopes for black and white suspects, or for employed and unemployed suspects. We could even have used separate sets of envelopes for some combinations of such factors (called factorial designs), such as employed unmarried suspects, unemployed married suspects, employed married suspects, etc.

Just contemplating such a design, however, shows how complicated it can become. It can also raise major political problems in the selection of factors for blocking randomization within the separate lists. When this experiment was negotiated in 1986, the use of randomized experiments in arrest was still a very fragile idea, with only one precedent. Blocking randomization in advance on individual suspect or victim factors could have caused enough controversy to kill the whole venture and so was eschewed.

Many analysts have advocated examining the underlying structure of main effects in randomized experiments, much as surveys analyze demographic patterns in attitudes and reported behavior. These "post-hoc" analyses of experiments can strongly suggest causal relationships for some kinds of people. But *what post-hoc analysis cannot do is prove that there is a causal interaction effect between a randomized treatment and a correlated characteristic.* The second stage analysis results of differences in treatment effects within subgroups, which are reported in this article, are couched in strong language because we believe the findings to be theoretically coherent and

57 *See* POCOCK, *supra* note 37.

very likely to represent truly causal relationships. But without a randomized design within each of the subgroups, we cannot be nearly as certain of the interaction effects as we are of the "main effects" of no difference across treatments.

### 2. *Replicated vs. Unreplicated Results*

A second caution is also in order. The earlier Minneapolis experiment had a broad-ranging policy impact long before any attempt was made to replicate it. This fact has been the subject of considerable discussion[58] and criticism.[59] While it is arguably better to make policy on an unreplicated finding than on no finding at all, it is important to know the difference.

The Milwaukee findings are not unreplicated. At the time of this writing, they have been replicated on two out of two attempts, as reported below. But we must remind the reader that these three experiments are just snapshots of three cities at three times. Not enough is yet known about how social experiments generalize to other times and places to be certain the thrice-observed effects will hold true. This is true no matter how often a finding is replicated. Nonetheless, the replication of the findings increases our confidence about their generalizability.

## B. ESCALATION AMONG MARGINAL SUSPECTS

### 1. *Prevalence of Repeat Violence*

Table 6 presents the differences in the prevalence rates of each treatment group within a uniform six month (183-day) period following the presenting randomized incident, controlling for various indicators of individual characteristics. These rates show, in effect, the odds of any given *individual* suspect committing at least one new act of domestic violence. The *relative* (as distinct from absolute) percentage differences in those odds, calculated using the warning group rate as the base of one hundred percent, all show that arrest versus non-arrest treatment has very different effects for different kinds of people.

The most consistent *prevalence* effect is that those with high stakes in social conformity, experience a deterrent effect from both versions of arrest, while those with low stakes in conformity show no such effect. Those who are employed, high school graduates, white, or married and those who have cohabitated for over two years all

58 *See* Sherman & Cohn, *The Impact of Research*, *supra* note 6.

59 *See* Richard O. Lempert, *Humility is a Virtue: On the Publicization of Policy-Relevant Research*, 23 LAW & SOC'Y REV. 145 (1989).

TABLE 6
ANY-VICTIM PREVALENCE OF REPEAT HOTLINE REPORTS PER 10,000 SUSPECTS
(during a six-month period)

| Individual Characteristic | Full Arrest | Short Arrest | Warning | % Difference 1 & 3 | % Difference 2 & 3 |
|---|---|---|---|---|---|
| Prior | 3,341<br>(137) | 3,950<br>(119) | 3,846<br>(130) | −13.3 | 02.7 |
| No Prior | 1,873<br>(267) | 1,828<br>(279) | 2,089<br>(268) | −10.3 | −12.5 |
| Blacks | 2,656<br>(305) | 2,721<br>(305) | 2,633<br>(300) | 00.8 | 03.3 |
| Whites | 1,481<br>( 81) | 1,538<br>( 78) | 2,436<br>( 78) | −39.2 | −36.8 |
| Employed | 2,011<br>(189) | 1,702<br>(188) | 2,766<br>(141) | −27.3 | −38.5 |
| Unemployed | 2,775<br>(209) | 3,140<br>(207) | 2,629<br>(251) | 05.5 | 19.4 |
| High School | 2,278<br>(158) | 2,958<br>(142) | 3,235<br>(102) | −29.6 | −08.6 |
| Less than H.S. | 2,327<br>(202) | 2,000<br>(220) | 2,466<br>(296) | −05.6 | −18.9 |
| Married | 1,700<br>(147) | 1,509<br>(106) | 2,564<br>(117) | −33.7 | −41.1 |
| Not Married | 2,813<br>(256) | 2,808<br>(292) | 2,734<br>(278) | 02.9 | 02.7 |
| Yrs. Cohabit $> 2$ | 2,438<br>(201) | 2,795<br>(161) | 2,895<br>(152) | −15.8 | −03.5 |
| Yrs. Cohabit $\leq 2$ | 2,456<br>(114) | 2,185<br>(119) | 2,444<br>(135) | 00.5 | −10.6 |

N for each prevalence rate is shown in parentheses.

show substantially lower prevalence rates of repeat violence when randomly assigned to arrest than when warned. Yet their opposites (unemployed, dropouts, etc.) show little difference in prevalence of recidivism between being arrested or warned.

Looking *solely* on rates of prevalence of repeat violence, there is apparently good reason to adopt a policy of mandatory arrest. Arrest has strong deterrent effects for some groups, with up to one-third fewer suspects repeating their violence in the next six months. Its failure to deter others does not, at least, cause any harm. This apparent conclusion, however, demonstrates the importance of looking beyond prevalence to a robust examination of the frequency of repeat violence. In the Milwaukee experiment, where frequency was well measured, prevalence alone as an outcome measure would be a *very misleading* basis for policy implications.

### 2. *Frequency of Repeat Violence*

More important from a policy standpoint are the frequency rates, which show the effects of a mandatory arrest policy on the total incidence of violence in the community. The group (not individual) frequency results per days at risk in the Milwaukee experiment are shown in Table 7. This table shows that over an unrestricted follow-up period, arrest not only deters some groups; it also escalates other groups into far higher frequency of domestic violence. The magnitude of the percentage differences (again using the warning group as the base of one hundred percent) in effects across subgroups is quite large by the normal standards of social research and statistics. The table consistently shows arrest to make those with less stake in conformity more violent, and those with more stake in conformity less violent.

The difference in reaction to full arrest between blacks and whites is startling. The fact that 10,000 arrested whites produce 2,504 (=5,212−2,708) fewer acts of domestic violence a year than warned whites, while 10,000 arrested blacks produce 1,803 (= 7,296−5,493) more acts of violence per year than warned blacks, is a far larger magnitude than we ever expected. If three times as many blacks as whites are arrested in a city like Milwaukee, which is a fair approximation, then an across-the-board policy of mandatory arrests prevents 2,504 acts of violence against primarily white women at the price of 5,409 acts of violence against primarily black women. While one explanation is that this effect is mostly due to racial differences in unemployment rates, the differential impact by race is just as morally troublesome whatever the underlying cause.

There is even less reduced-violence benefit due to full arrest by employed suspects at the price of increased violence by unemployed suspects. With 958 (= 5,991-5,033) fewer acts of violence committed against victims of 10,000 employed suspects who had been arrested than of those who had been warned, the price equals 2,274 (= 7,504-5,230) more acts of violence per 10,000 unemployed suspects who had been arrested than if they had only been warned. Some might reason that since most people are employed, this policy seems to be reasonable as a utilitarian tradeoff. But wherever the majority of the domestic violence incidents police respond to involve unemployed suspects—as they do in Milwaukee—then mandatory arrest fails to produce the greatest good for the greatest number. The fact that this is not evident in the main effects reflects the relatively even splits of most of the three treatment groups on most of the characteristics presented in the table. Figure 2 displays

TABLE 7
AFTER-ONLY MEAN FREQUENCY OF HOTLINE REPORTS PER ANNUM PER 10,000 SUSPECTS
(unrestricted follow-up period)

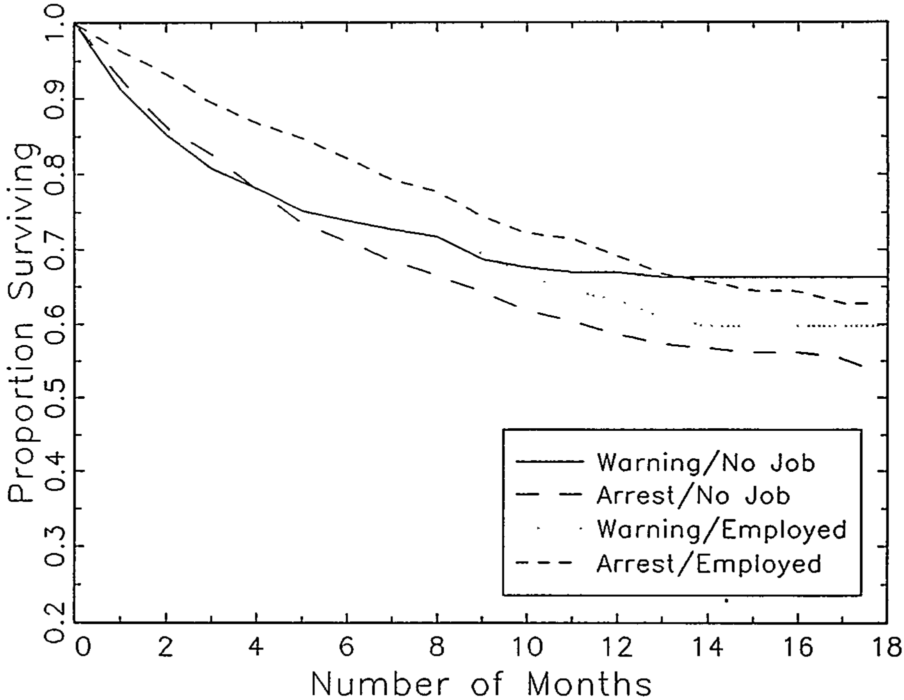
| Individual Characteristic | Full Arrest | Short Arrest | Warning | % Difference 1 & 3 | % Difference 2 & 3 |
|---|---|---|---|---|---|
| Prior | 10,771 (137) | 11,318 (119) | 8,403 (130) | 28.2 | 34.7 |
| No Prior | 4,204 (267) | 4,899 (279) | 4,179 (268) | 00.5 | 17.2 |
| Blacks | 7,296 (305) | 7,410 (305) | 5,493 (300) | 32.8 | 34.9 |
| Whites | 2,708 ( 81) | 4,942 ( 78) | 5,212 ( 78) | −48.0 | −05.2 |
| Employed | 5,033 (189) | 4,842 (188) | 5,991 (141) | −16.0 | −19.2 |
| Unemployed | 7,504 (209) | 8,428 (207) | 5,230 (251) | 43.5 | 61.1 |
| High School | 5,869 (158) | 7,355 (142) | 6,367 (102) | −07.8 | 15.5 |
| Less than H.S. | 6,360 (202) | 6,211 (220) | 5,106 (296) | 24.6 | 21.6 |
| Married | 4,720 (147) | 4,774 (106) | 5,386 (117) | −12.3 | −11.3 |
| Not Married | 7,222 (256) | 7,441 (292) | 5,545 (278) | 30.2 | 34.2 |
| Yrs. Cohabit > 2 | 7,048 (201) | 4,774 (161) | 5,386 (152) | 30.9 | −11.4 |
| Yrs. Cohabit ≤ 2 | 6,195 (114) | 5,666 (119) | 5,661 (135) | 09.4 | 00.0 |

N for each mean is shown in parentheses.

the differences in survival curves over time among the four groups divided by arrest and employment status.

FIGURE 2

Survival Functions by Arrest and Employment Status




It is particularly interesting that the worst escalation effect in Table 7 is found among unemployed suspects who received the short arrest treatment. While the unemployed have largest percentage increase in violent acts per 10,000 of any group (61.1%), the employed have one of the largest deterrent effects from short arrest (−19.2%)—even slightly larger than the effects of full arrest (−16.0%). This may suggest that the employed react to getting a break, or to getting out early enough to go to work, by avoiding a second chance to lose their job.

An important point about the employment data in an industrial town like Milwaukee is that the suspects' jobs were primarily at lower levels of occupational prestige. The first ten most often listed occupations in the sample, for example, were general assistance (a part-time workfare program), factory maintenance, security guard, retail stock handler, grocery store meat wrapper, grocery cashier, car wash attendant, and valet shop clerk. A review of all of the suspects' listed occupations shows a total of six mid-level prestige jobs: one teacher, one child's counselor, one editor, one retail sales manager, one insurance salesman, and one bank executive. The difference between the working class and the underclass is often forgotten by middle-class. The middle class is concerned with having any great job to lose or career to ruin, as opposed to the underclass, which is concerned with working at all.

A high school education predicts a fairly weak deterrent effect of arrest, but lack of a high school education predicts a fairly strong criminogenic effect of arrest. Marriage is more powerful than education, with a marriage license enhancing the deterrent effect and its absence aggravating the adverse reaction to arrest. Contrary to our expectations, length of cohabitation goes the other way, although it is not inversely correlated with marriage. Arrest appears to make suspects more violent if they have lived with the victim for over two years than if they have not.

### 3. *Are the Interactions Significant?*

The next question is whether these differences are due to chance. Table 8 presents the results of a Poisson regression model for both main effects and two-way interactions. Only prior domestic hotline reports show a main effect that is not due to chance, and strongly predicts the number of after-treatment incidents. But the interaction effects are generally significant and are all in the theoretically predicted direction. Our interpretation of the model is that being black rather than white increases the recidivism frequency rate for the arrest group by sixty-two percent, while having a job reduces it by fifty-eight percent and being a high school graduate reduces it by forty-three percent. The interaction effects with marriage and length of cohabitation are weaker and not significant.

The key indicators of marginal social status, then, fairly consistently show that arrest increases the frequency of violence among marginal suspects. The deterrent effects of arrest on persons with higher stakes in conformity are not consistently strong, possibly because there are so few middle class persons in the sample. The deterrent effects could, for example, become stronger as the value of the suspect's stake in conformity increased. The most startling interaction effect is the strongly opposite directions of the effects of arrest for whites and blacks.

### 4. *Why Do Prevalence and Frequency Results Conflict?*

Accounting for the differences between Tables 6 and 7 is an important policy question. They differ both in follow-up period covered and in treatment effects. In order to ensure that the differ-

TABLE 8
POISSON REGRESSION COEFFICIENTS MAIN EFFECTS AND TWO-WAY INTERACTIONS (CONTROLLING PRIOR OFFENDING AND DAYS AT RISK) FOR ANY VICTIM AFTER-ONLY FREQUENCY OF HOTLINE REPORTS

| Variable | Coefficient | Std. Error | T-ratio | Prob $\geq x^t$ |
|---|---|---|---|---|
| Main Effects | | | | |
| Intercept | −8.34405 | 1.02519 | −8.139 | .00000 |
| Arrest | −.113133 | .347519 | −.326 | .74477 |
| Black | .066061 | .189910 | .348 | .72795 |
| Employed | .248127 | .156097 | 1.590 | .11193 |
| High School | .226241 | .152576 | 1.483 | .13812 |
| Married | −.185799 | .171293 | −1.085 | .27806 |
| Cohabit. > 2 | .114660 | .154494 | .742 | .45799 |
| Prior | .842037 | .102621 | 8.205 | .00000* |
| LogADAYS | 1.21703 | .165801 | 7.340 | .00000 |
| Two-Way Interactions | | | | |
| Arrest & Black | .624443 | .297136 | 2.102 | .03559* |
| Arrest & Employed | −.584834 | .214761 | −2.723 | .00647* |
| Arrest & High School | −.430113 | .209724 | −2.051 | .04028* |
| Arrest & Married | −.070437 | .230941 | −.305 | .76037 |
| Arrest & Cohabit. | .177514 | .217008 | .818 | .41336 |

* $P \leq .05$
T − ratio Prob $|t| \geq x$

ence in treatment effects was not due simply to the difference between the six-month restriction in Table 6 and the individually varying, longer-term follow-up (up to twenty-two months) in Table 7, we recalculated Table 7 with a six-month (183-day) restriction.[60] The results reveal little difference. We can therefore disregard an assertion that prevalence and frequency results differ due to differences in follow-up periods.

Given a clear difference between frequency and prevalence rates, the next question is how the higher frequency rates are distributed across individuals. The two extreme alternatives would be:

1) similar frequency rates for most offenders, with a small number of extremely arrest-reactive offenders driving the overall group frequency rate upward, and
2) generally higher frequency rates among all subgrouping of the arrest-reactive offenders, such as greater percentages of suspects with two events regardless of subgroup membership.

These alternatives have major policy significance, since a mandatory arrest policy with a few exceptions might become an alternative to a

60 The results are available on request; ask the first author for Table 8-12.

more generally discretionary policy. Table 9 examines the frequency distribution of repeat events by treatment group and the three key individual predictors of treatment effects: employment, education, and race. It shows that between two-thirds and three-quarters of all recidivist events in the first six months are concentrated among offenders who had only one repeat event. Only whites randomized to full arrest pose an exception to that pattern, with ninety-two percent of the recidivism committed by offenders with only one repeat event. Similarly, the concentrations of offenders with three or more subsequent events in six months vary little, with the exception of whites.

These data falsify the hypothesis that the difference in frequency rates by subgroup characteristics is due to a small number of highly arrest "allergic" offenders. Rather, the data are consistent with the conclusion that arrest produces generally higher frequency rates among more socially marginal persons.

### C. REPLICATIONS

The generalizability of these findings is greatly enhanced by their having been replicated in experiments in two other cities: Omaha, Nebraska and Colorado Springs, Colorado.[61] Statistically significant interaction effects might be obtained within any one experiment just by chance, if enough tests are performed. It is highly unlikely, however, that similar interaction effects in three independent experiments are due merely to chance.

The most complete replication is found in Omaha. Our own reanalysis of the Omaha data allowed us to examine both the prevalence and frequency data among different subgroups within that sample. Both results are consistent with the Milwaukee results. Consistent with the less severe concentration of urban problems in Omaha, the benefits of mandatory arrest there appear to outweigh the risks. The frequency data, reported elsewhere,[62] show a stronger deterrent effect among the employed, and a weaker escalation effect among the unemployed, than in Milwaukee. Nonetheless, the directions of the interactions are consistent with the Milwaukee results.

The same is generally true for the Omaha prevalence data re-

61 The interactions for Colorado Springs are reported in this issue. Richard A. Berk et al., *A Bayesian Analysis of the Colorado Springs Spouse Abuse Experiment*, 83 J. CRIM. L. & CRIMINOLOGY 170 (1992). The interactions for Omaha are reported in SHERMAN, POLICING DOMESTIC VIOLENCE, *supra* note 11, and in Sherman & Smith, *Crime, Punishment and Stake in Conformity*, *supra* note 11.

62 Sherman & Smith, *Crime, Punishment and Stake in Conformity*, *supra* note 11.

**TABLE 9**
**FREQUENCY DISTRIBUTIONS OF REPEAT HOTLINE REPORTS BY TREATMENT GROUP BY EMPLOYMENT STATUS, EDUCATION AND RACE**

| Number of Repeat Hotline Reports | Employment | | | | Education | | | | Race | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Employed | | Not Employed | | High School | | Drop Out | | White | | Black | |
| | N | (*) | N | (*) | N | (*) | N | (*) | N | (*) | N | (*) |
| Full Arrest | | | | | | | | | | | | |
| total | 189 | | 209 | | 158 | | 202 | | 81 | | 305 | |
| 0 | 151 | | 151 | | 122 | | 155 | | 69 | | 224 | |
| 1 | 25 | 66% | 37 | 64% | 25 | 69% | 29 | 62% | 11 | 92% | 50 | 62% |
| 2 | 8 | 21% | 10 | 17% | 7 | 19% | 8 | 17% | 1 | 8% | 16 | 20% |
| 3+ | 5 | 13% | 11 | 19% | 4 | 11% | 10 | 21% | — | | 15 | 19% |
| Short Arrest | | | | | | | | | | | | |
| total | 188 | | 207 | | 142 | | 220 | | 78 | | 305 | |
| 0 | 156 | | 142 | | 100 | | 176 | | 66 | | 222 | |
| 1 | 23 | 72% | 43 | 66% | 31 | 74% | 27 | 61% | 9 | 75% | 54 | 65% |
| 2 | 7 | 22% | 10 | 15% | 5 | 12% | 9 | 20% | 1 | 8% | 17 | 20% |
| 3+ | 2 | 6% | 12 | 18% | 6 | 14% | 8 | 18% | 2 | 17% | 12 | 14% |
| Warning | | | | | | | | | | | | |
| total | 141 | | 251 | | 102 | | 296 | | 78 | | 300 | |
| 0 | 102 | | 185 | | 69 | | 223 | | 59 | | 221 | |
| 1 | 29 | 74% | 44 | 67% | 24 | 73% | 50 | 68% | 13 | 68% | 56 | 71% |
| 2 | 9 | 23% | 16 | 24% | 8 | 24% | 17 | 23% | 6 | 32% | 17 | 22% |
| 3+ | 1 | 3% | 6 | 9% | 1 | 3% | 6 | 8% | — | | 6 | 8% |

N = Number of Subjects.
(*) = % of Suspects with one or more repeat incidents.

TABLE 10
OMAHA, NEBRASKA
PREVALENCE OF REPEAT OFFICIAL VIOLENCE BY RANDOMIZED POLICE TREATMENT AND SOCIAL STATUS
(401 day maximum follow-up period)

| Social Status | Arrested | Not Arrested |
|---|---|---|
| Employed | 19% | 28% |
| Unemployed | 57% | 53% |
| Married | 29% | 18% |
| Unmarried | 35% | 48% |
| High School Graduate | 24% | 34% |
| High School Dropout | 48% | 32% |
| Whites | 17% | 27% |
| Blacks | 55% | 47% |

ported here in Table 10. With the exception of marriage, the differences in prevalence of officially measured repeat violence (any rearrest or new complaint, combined) go in the same directions as in Milwaukee. Three out of four indicators of marginality are associated with less deterrence and generally with some escalation.

The Bayesian analysis of the Colorado Springs experiment is confined to prevalence only. Absent a mandatory arrest policy and a strong custom of reporting all domestic battery misdemeanors, the Colorado Springs experiment does not offer a very robust test of differences in frequency. The prevalence results, however, are consistent with the Milwaukee results, showing clear deterrence of persons with higher stakes in conformity and much weaker evidence of escalation effects of arrest for less marginal people. As the Milwaukee results suggest—but with only Omaha as a replication—the analysis of prevalence as the only outcome may obscure important consequences of mandatory arrest policies on the total amount of domestic violence in a community. Frequently rates more clearly show the escalation effects of arrest.

## VI. CONCLUSIONS AND IMPLICATIONS

The Milwaukee domestic violence experiment finds no evidence of an overall long-term deterrent effect of arrest. The initial deterrent effects observed for up to thirty days quickly disappear. By one year later, short arrest alone, and short and full arrest combined, produce an escalation effect. The first reported act of repeat violence following combined arrest treatments occurs an average of twenty percent sooner than it does following the warning treatment.

The Milwaukee experiment does find strong evidence that arrest has different effects on different kinds of people. Employed, married, high school graduate and white suspects are all less likely to have any incident of repeat violence reported to the domestic violence hotline if they are arrested than if they are not. Unemployed, unmarried, high school dropouts and black suspects, on average, are reported much more frequently to the domestic violence hotline if they are arrested than if they are not. The magnitudes of the increased domestic violence associated with arrest of the latter groups are substantial, ranging up to sixty percent. The Milwaukee findings are replicated clearly in Omaha, as well as by a more limited data set in Colorado Springs.

These results strongly suggest that arrest has variable effects on criminal careers, depending upon the social marginality of the offenders. At least for the offense of misdemeanor domestic battery—or harassment, in the case of Colorado Springs—arrest appears to deter less marginal persons and to escalate the frequency of violence among more marginal persons. Whether this pattern applies to other types of offenses is still unknown, but it is certainly plausible. As one leading labeling theorist observed two decades ago, arrest probably serves to keep the large majority of people in line, even while it causes a small group of social outcasts to become more criminal.[63]

The accumulating empirical support for the proposition of variable effects of arrest on criminal careers raises major questions for criminology, jurisprudence, and public policy. The question for criminology is how theory can clearly account for these differential reactions to a relatively minor application of the criminal sanction. Competing theoretical perspectives, such as shaming, control, and power, might all account for these facts. Future experiments can now be more closely focused on comparative tests of competing theoretical perspectives.

The question for jurisprudence is whether the ministerial approach to police discretion is proper, especially with a mandatory arrest statute. Previous jurisprudence has rejected the proposition that punishment should be made more severe than just deserts allow in order to increase a deterrent effect. But it has never before considered the proposition that punishment should be made less severe in order to reduce an escalation effect. Indeed, jurisprudence seems to have hardly considered the problem of escalation effects at all. Andrew von Hirsch, for example, suggests that

> the disposition of convicted offenders should be commensurate with

[63] LOFLAND, *supra* note 30.

> the seriousness of their offenses, even if greater or lesser severity would promote other goals. For the principle, we have argued, is a requirement of justice, whereas deterrence, incapacitation and rehabilitation are essentially strategies for controlling crime. The priority of the principle follows from the assumptions we stated at the outset: the requirement of justice ought to constrain the pursuit of crime prevention.[64]

Yet, he concedes elsewhere that all punishment depends upon the assumption of deterrence for its moral justification.[65] How punishment can be justified when it escalates violence is not at all clear. Yet how the punishment of some, and the failure to punish others, could be justified is equally unclear. The conflict between justice and crime control seems never to have been framed so baldly.

The short-term implications of this dilemma for public policy are daunting. At the least, it suggests a need for other approaches to the control of domestic violence among marginal persons, such as greater investment in battered women's shelters. At best, it suggests a serious and thoughtful debate about the effects of domestic violence on our society, as well as the current inequities in police discretion that have been tolerated for years. The price of reduced violence may be changing the nature of the inequities and making the fact of inequity explicit. Whether we are willing to pay that price is a matter for every citizen to consider.

---

64 VON HIRSCH, *supra* note 10, at 74-75.
65 *Id.* at 55.