A Bayesian Analysis of the Colorado Springs Spouse Abuse Experiment

Richard A. Berk

Alec Campbell

Ruth Klap

Bruce Western

A BAYESIAN ANALYSIS OF THE COLORADO SPRINGS SPOUSE ABUSE EXPERIMENT

RICHARD A. BERK, ALEC CAMPBELL, RUTH KLAP, AND BRUCE WESTERN*

ABSTRACT

This Article analyzes data from the Colorado Springs Spouse Abuse Experiment. In that experiment, suspects apprehended for misdemeanor spouse abuse were assigned at random to one of four treatments: (1) an emergency order of protection for the victim coupled with arrest of the suspect; (2) an emergency order of protection for the victim coupled with immediate crisis counseling for the suspect; (3) an emergency order of protection only; or (4) restoring order at the scene with no emergency order of protection. Outcome measures are taken from official police data and from follow-up interviews with victims. Using Bayesian procedures to take previous experiments into account, the balance of evidence supports a deterrent effect for arrest among "good risk" offenders, who presumably have a lot to lose by being arrested. The balance of evidence is far more equivocal for a "labeling effect" in which an arrest increases the likelihood of new violence.

I. INTRODUCTION

The pioneering Minneapolis Spouse Abuse Experiment was arguably one of the most influential and controversial criminal justice studies conducted over the past several decades. In 1981 and 1982, the Minneapolis Police department, in cooperation with the Police Foundation and with the support of the National Institute of Justice, conducted a randomized field experiment on the impact of three

different intervention strategies in cases of misdemeanor spousal violence: (1) arresting the suspect; (2) ordering the suspect from the premises for twenty-four hours; or (3) "simply" trying to restore order. Outcome measures were obtained from new offenses reported to the police and from follow-up interviews with victims. The study concluded that arresting the suspect was the most effective treatment in reducing the likelihood of renewed violence.[1]

The findings of the experiment were immediately introduced into the policy area. Soon after, a number of police departments across the country made an arrest the presumptive intervention in incidents of spousal violence. A number of states also revised their penal codes to more seriously address spousal violence and the importance of sanctioning suspects.[2] Yet, there was also substantial controversy. Policy makers' uncritical acceptance of the findings from a single study caused major concern.[3]

The researchers responsible for the reports of the Minneapolis experiment stated clearly the limitations of a single study. For example, commenting on the impact of the arrest treatment, Berk and Sherman note that "the effectiveness of arrest is relative to the other treatments applied. If arrest were pitted against other interventions, or the same interventions implemented somewhat differently, the results could change."[4]

In response to these and other concerns, the National Institute of Justice funded follow-up experiments at six new sites. An effort was made to coordinate the follow-up studies and to replicate the Minneapolis protocols as much as possible. To date, there are published results from only one of the replications. In Omaha, Nebraska, arrest did not stand out as the most effective intervention.[5] We will have more to say about that study shortly. Working papers from at least one other site are being circulated. Here, we report the first findings from the Colorado Springs, Colorado replication,

---

[1] Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 AM. SOC. REV. 261, 261 (1984); Richard A. Berk & Lawrence W. Sherman, *Police Responses To Family Violence Incidents*, 83 J. AM. STAT. ASS'N 70, 74 (1988) [hereinafter Berk & Sherman, *Police Responses*]; Richard A. Berk et al., *When Random Assignment Fails: Some Lessons from the Minneapolis Spouse Abuse Experiment*, 4 J. QUANTITATIVE CRIMINOLOGY 209, 209 (1988) [hereinafter Berk et al., *When Random Assignment Fails*].

[2] Lawrence W. Sherman & Ellen G. Cohn, *The Impact of Research on Legal Policy: The Minneapolis Domestic Violence Experiment*, 23 LAW & SOC'Y REV. 117, 126 (1989).

[3] Richard Lempert, *Humility Is a Virtue: On the Publicization of Policy-Relevant Research*, 23 LAW & SOC'Y REV. 145, 146 (1989).

[4] Berk & Sherman, *Police Responses, supra* note 1, at 76.

[5] Franklyn W. Dunford et al., *The Omaha Domestic Violence Police Experiments: Final Report to the National Institute of Justice, Technical Report*, INST. OF BEHAVIORAL SCIENCE, U. OF COLORADO (1989).

building on unpublished findings from the Milwaukee, Wisconsin replication, and also weaving in new tabulations from the Omaha study. We focus especially on the Milwaukee experiment because of its new and important findings, the credibility of which depends in part on similar results in at least one other site.[6] To this end, we employ a Bayesian framework in which priors based on the Milwaukee study can be explicitly introduced into the analysis of data from Colorado Springs. In other words, we undertake a form of "meta-analysis" across three randomized field experiments.[7] A full "stand-alone" analysis of the Colorado Springs experiment will be reported elsewhere.

II. The Omaha Replication

Beginning in early 1986, the Omaha replication began comparing their versions of "arrest," "separation," and "mediation" for cases of misdemeanor spousal violence. The Omaha authors note, however, that little effort was made to standardize treatments, and the information on treatment content is sketchy.[8] Three hundred and twenty seven suspects were assigned to one of the three treatments. Of that total, ninety-seven percent received the treatment to which they were randomly assigned. Random assignment, therefore, was successfully implemented.[9]

Outcome measures were of two types: (1) "official" recidivism measured by new arrests or officially reported complaints for any crimes committed by suspects against their original victims, and (2) victim reports of new incidents including fear of injury, pushing/hitting, and physical injury. The victim report data were collected through face-to-face interviews six months into the study. The response rate for these interviews was seventy-three percent, with only slight differences in response rates across treatments.[10]

Using a variety of outcome measures, the Omaha authors claim that no treatment proved more successful than any other. That is, in no case could the hypothesis of equal "failure" rates across the three treatments be rejected at conventional statistical confidence levels. Yet, the Omaha results appear to be within the Minneapolis

---

6 *See* Lawrence W. Sherman et al., Crime, Punishment and Stake in Conformity: Legal and Extra Legal Control of Domestic Violence (1991) (unpublished manuscript, on file with the J. Crim. L. & Criminology) [hereinafter Sherman, Crime, Punishment, and Stake in Conformity].

7 *See* Larry V. Hedges & Ingram Olkin, Statistical Methods for Meta-Analysis (1985).

8 *See* Dunford, *supra* note 5, at 189-91.

9 *Id.*

10 *Id.*

confidence limits. Just as in the earlier Minneapolis experiment, it is a bit difficult to confidently interpret the results. For example, if victims in the separation or mediation treatments were frustrated by the failure of police officers to arrest their assailants, they might well be less inclined to report future assaults. After taking the considerable risk of serious retaliation from the suspect by calling the police, "nothing" happened. This would downwardly bias reports of new violence in official police records and artificially improve the apparent performance of the mediation and separation treatments. Moreover, there is no discussion of the treatments' possible interaction effects, the importance of which the following makes clear.

III. THE MILWAUKEE REPLICATION

As described by Sherman and his colleagues,[11] between April 1987 and August 1988, the Milwaukee police department randomly assigned 1,200 suspects charged with misdemeanor spousal battery to one of three conditions: (1) standard "full" arrest in which the suspect was held until morning, unless he could post bail; 2) "short" arrest in which the suspect was released on recognizance within three hours; and 3) no arrest at all coupled with a scripted warning of an arrest if the police had to return. Ninety-eight percent of the suspects received the treatment to which they were randomly assigned.

While both victim report data and official data were used to construct outcome measures, the Milwaukee authors have (to date) focused on the official data, which they felt was far more reliable than the data extracted from the interviews with victims. The overall failure rates from these data were very similar across all three treatments, but the two arrest treatments seemed to delay the onset of new violence by a little more than a month compared to the warning treatment.

Far more important for our purposes were findings that the arrest treatments reduced both the likelihood (prevalence) and the number (frequency) of failures (repeat violence) for suspects who were employed and increased the likelihood and number of failures for suspects who were not employed. The Milwaukee researchers anticipated the former result before the data analysis began. Employed individuals would presumably have more to lose through an arrest, perhaps even their employment itself. According to Social Control Theory, employed individuals have stronger ties to the local community and, therefore, are more likely to feel shame after an

---

[11] Sherman, Crime, Punishment, and Stake in Conformity, *supra* note 6.

arrest for domestic violence. The latter result was explained in a *post hoc* fashion. According to Labeling Theory, suspects with little to lose will not be deterred by an arrest; and if they blame the victim for it, they may be more likely to retaliate.

However, some grounds exist for caution about the arrest treatments' interaction effects as shown in the researchers' general analysis of the date. A number of interaction effects were examined by the Milwaukee researchers, and the prospect of capitalizing on chance exists. Moreover, whatever the average effect across all suspects, if a pool of suspects is divided into two groups, and one group does better than average, the other group necessarily must do worse than average. And since for the Milwaukee experiment, the average effect across all suspects is effectively zero, finding a beneficial treatment effect for employed suspects requires finding a harmful effect for unemployed suspects. Put another way, any split of the Milwaukee subject pool in which arrest appears to have a beneficial effect for one of the two groups necessarily requires finding a harmful effect for the other group. However, in conventional frequentist terms, one or both effects might not be "statistically significant" because, in part, of small sample sizes.

In summary, in Milwaukee, as in Omaha, there are at most only weak hints of a main treatment effect for arrest. However, claims were made of rather substantial interaction effects for arrest depending upon whether the suspect is employed.

IV. THE COLORADO SPRINGS REPLICATION

Over a period of about two years beginning in June of 1987, the Colorado Springs Police Department randomly assigned 1,658 suspects of misdemeanor spousal violence to one of four treatments: (1) an emergency order of protection for the victim coupled with arrest of the suspect; (2) an emergency order of protection for the victim coupled with immediate crisis counseling for the suspect; (3) an emergency order of protection only; or (4) restoring order at the scene with no emergency order of protection. After police arrived at the scene and determined that a case was eligible for the experiment, they called the dispatcher to learn what treatment should be applied. The dispatcher, in turn, consulted a computer which generated the assignment at random. Eighty-two percent of the suspects received the treatment to which they were randomly assigned. The eighteen percent rate of "misassignment" is comparable to the rate in Minneapolis and is addressed below. Some of the misassignments had systematic explanations. For example, occasionally a sus-

pect assigned to the restoring-order intervention refused to cooperate and had to be arrested.

Follow-up data again were of two kinds: reports of new offenses in police records and interviews with victims. For the latter, the response rate for the last interview, which took place six months after entering the study, and from which the victim reported outcome measures are most easily constructed, was sixty-four percent. The implications of this is addressed below.

Table 1 shows some descriptive statistics for suspects and victims subject to random assignment. The data in Table 1 were taken from "implementation forms" filled out at the scene by the police for every "household" included in the experiment. These data, therefore, have the fewest number of missing observations of any data available to us. Subsequent data collection efforts lost cases through attrition (*e.g.*, the victim and suspect moved from Colorado Springs).

As anticipated by the research design, virtually all of the victims and suspects are either married or otherwise "romantically" involved, with roughly two-thirds married. The average age for both victims and suspects is nearly thirty. Roughly sixty percent of the victims are white, which is comparable to the mix of victims in Minneapolis and Omaha, but more than double the white percentage in Milwaukee. The vast majority of victims are female. For thirty percent of the victims, the police recorded no occupation. In addition, eleven percent are unemployed. The remaining victims are spread fairly evenly across a number of skilled and unskilled occupations. The relatively large fraction (seven percent) of victims working in the military is unusual. There is no significant military representation at any of the other sites, while in and around Colorado Springs are several military installations, including Fort Carson (Army), and the U.S. Air Force Academy. Not surprisingly, the suspects have rather similar backgrounds. Perhaps most distinctive is the very large percentage (twenty-four percent) of suspects in the military. No other sites have comparable figures, raising issues to which we will shortly return.

Table 2 provides information on how the experiment was implemented. For nearly twenty percent of the suspects, the instant offense was a repeat offense. Consistent with the research design, when the offense was recorded, it was a misdemeanor. Felonies were excluded from the research because of ethical and legal issues; for example, for these more serious offenses, the law required that an arrest be made. About a quarter were assigned at random to each of the four treatments. As noted earlier, however, eighteen

Table 1
BACKGROUND INFORMATION FOR COLORADO SPRINGS DATA
(N=1658)

Variable	Proportion or Mean
Victim White	.59
Victim Black	.25
Victim Latina/o	.12
Suspect White	.53
Suspect Black	.31
Suspect Latina/o	.14
Victim Female	.89
Suspect Male	.89
Victim Occup. Unskilled	.20
Victim Occup. Skilled	.16
Victim Occup. Professional	.04
Victim in Military	.07
Suspect Occup. Unskilled	.19
Suspect Occup. Skilled	.19
Suspect Occup. Professional	.03
Suspect in Military	.24
Victim's Age	28.7
Suspect's Age	29.7
Couple Married/Living Together	.64
Couple Married/Not Living Together	.04
Couple Friends/Lovers	.28
Prior Domestic Violence by Suspect	.19

percent of the suspects did not receive the treatment to which they were assigned. The marginals for the imposed treatment indicate that the major deviation from the design involved the counseling treatment. Counseling was apparently less frequently delivered than called for by design, with the other treatments, save emergency orders of protection alone, being given more often. "Other" included such things as transporting the suspect to an emergency room for immediate medical care.

Table 2
IMPLEMENTATION OF THE EXPERIMENT (N=1658)

Offense Committed	Proportion
Menacing	.03
Harassment	.54
3rd Degree Assault	.38
False Imprisonment	.03
Other	.02
Treatment Assigned	Proportion
EPO/Arrest	.26
EPO/Counseling	.24
EPO alone	.27
Restore Order	.23
Treatment Imposed	Proportion
EPO/Arrest	.28
EPO/Counseling	.19
EPO alone	.27
Restore Order	.24
Other	.02

EPO = Emergency protection order.

We do not know precisely why there was a shortfall for the counseling treatment. The difficulty of implementing this treatment no doubt contributed. It required the police officers to transport the suspect to another location at which counseling could occur. If the suspect was seriously intoxicated or impaired by drugs, counseling could not be undertaken.

We have completed some analyses of the fit between the treatment assigned and the treatment imposed, consistent with research on the Minneapolis experiment.[12] Just as in the Minneapolis experiment, in Colorado Springs there is some evidence to explain "upgrading" from the counseling, the emergency orders of protection alone, and the restore order treatments to the arrest treatment.

---

[12] Berk et al., *When Random Assignment Fails, supra* note 1.

This could occur, for example, if the suspect refused to cooperate with efforts to "restore order" and, as a result, was arrested.

Since the suspects who were "upgraded" were sometimes just those suspects who had a history of spouse abuse, the shift of some suspects to the arrest treatment from the other treatments could bias the study against finding any beneficial effects for arrests. We empirically address these concerns below in the data analysis.

V. ESTIMATION

This article focuses primarily on the possibility of different treatment effects for different classes of suspects and, in particular, on whether the impact of arrest depends on whether the offender is employed or in the military. We use Bayesian procedures to take explicit account of two replication studies (Omaha and Milwaukee) done before the Colorado Springs experiment. The results of those earlier studies are translated into "prior distributions" of interaction effects. When combined with the results from the Colorado Springs experiment, that produces "posterior distributions" of interaction effects, which take account of all three replication experiments. In principle, one could simply pool the necessary data, since we have access to the requisite tabulations for each site. However, for reasons that will be made more clear below, simple pooling would, in our view, give too much weight to the Milwaukee results.

We also considered pooling the Minneapolis data, but the necessary tabulations could not be computed from the information available. Yet, because the Minneapolis sample is small compared to the sample size for the three replications explored here, including the Minneapolis data would make little difference in the results. We also favor Bayesian inference more generally for policy-related work for reasons that are beyond the scope of this article.[13]

We do not focus on main effects. By "main effects" we mean average effects across all suspects. Given the potential saliency of interaction effects, main effects are, in one important sense, not well defined. They depend in part on the particular mix of the relevant offender classes. Given certain interaction effects, varying proportions of offenders in the different offender classes could produce varying main effects. Such variation is of more than academic concern, since the mix of suspects varies greatly across the three sites.

To maximize comparability to past spouse abuse experiments, we collapse the treatments into two categories: arrest versus every-

---

13 *But see* VIC BARNETT, COMPARATIVE STATISTICAL INFERENCE (1982); WILLIAM E. POLLARD, BAYESIAN STATISTICS FOR EVALUATION RESEARCH: AN INTRODUCTION (1986).

thing else. In addition, we also define a failure in ways that maximize comparability. A failure (via the data collected by police) is defined as a new reported offense by the suspect involving the same victim; and a failure (via the victim report data) is defined as an incident in which the same suspect either struck or caused injury to the same victim. In both instances, therefore, a failure is defined as an incident that would be classified as a crime in most locales.

There are, of course, a number of other ways failure could be defined. But in order to meaningfully combine the results across sites, the definitions across sites must be as similar as possible. For example, Colorado's spousal violence statutes proscribe a wider range of behavior than Wisconsin's statutes. If one relied solely on such local definitions, different findings across sites could result simply from different offense content.

There may also be concern that the use of a yes/no (binary) definition of failure throws out too much information compared to the use of survival time. In conventional frequentist terms, however, the only price is a bit of efficiency, and our sample sizes are already very large. Finally, since very few suspects in Colorado Springs committed more than one repeat offense during the follow-up period, using a count of the number of new incidents instead of a binary variable effectively makes no difference in the results.

Our estimation procedures are taken from a recent article by Clogg and his colleagues.[14] We have a sample of 1,658 randomly assigned suspects, and the goal is to determine how failure (Y) varies with the treatment (T). In addition, we have a covariate, employment status (E), which can be used to separate the suspects into groups so that different effects for the treatment can be examined for each group. In particular, the E can in principle distinguish between five mutually exclusive and exhaustive categories: (1) employed in the civilian labor force; (2) military personnel; (3) unemployed; (4) other (e.g., student, retired); and (5) employment status unknown. Building on the Milwaukee experiment, we expect that an arrest will work better than the other treatments for suspects who are employed in the civilian labor force or are in the military. We expect that an arrest will work worse than the other treatments for suspects who are unemployed or whose occupation is unreported. The prior distributions and their rationales are described below.

More formally, let Y_{ij} denote the value of the (binary) outcome

---

14 Clifford C. Clogg et al., *Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression*, 86 J. AM. STAT. ASS'N 68, 68-77 (1991).

variable (1=failure, 0=success) for the jth suspect in group i (i = 1,. . ., I; j = 1,. . ., n_i), where n_i is the number of suspects in the ith employment status group, and $\Sigma_i n_i$ = N. In our application, N is 1,658; and I the number of categories of **E**, can be up to 5, at least in principle. In an analogous fashion, let T_{ij} denote the value of the (binary) treatment variable (1=arrest, 0=not arrest) for the jth suspect in the ith group. Then, the logistic regression model specifies that for given π, a vector of probabilities of failure, the Y_{ij} are independent with $Pr(Y_{ij} = 1 \mid \pi) = \pi_i$, and that

$$\phi_i = \chi_i\beta \tag{5.1}$$

where $\phi_i \equiv logit(\pi_i) = log[\pi_i/(1 - \pi_i)]$ is the logit transformation of π_i, X_i is the variable T_i appended to a vector of 1's (of the same length), and β is a 2 by 1 vector of logistic regression coefficients.

Continuing in the fashion of Clogg and his colleagues,[15] the likelihood function can be expressed as

$$L\left(\beta \mid f_{1i}, f_{0i}; i = 1, ..., I\right) \propto \prod_{i=1}^{I} [\pi_i(\beta)]^{f_{1i}}[1 - \pi_i(\beta)]^{f_{0i}} \tag{5.2}$$

where $\pi_i(\beta)$ is the inverse logit transformation, and f_{1i} is the number of 1's in group i and f_{0i} is the number of 0's in group i. The conjugate prior is of the form

$$p(\beta) \propto \prod_{i=1}^{I} [\pi_i(\beta)^{g_{1i}}] [1 - \pi_i(\beta)]^{g_{0i}} \tag{5.3}$$

where g_{1i} and g_{0i} are positive constants to be specified. Then, the posterior is

$$p(\beta \mid f_{1i}, f_{0i}; i = 1, ..., I) \propto \prod_{i=1}^{I} [\pi_i(\beta)]^{f_{1i}+g_{1i}}[1 - \pi_i(\beta)]^{f_{0i}+g_{0i}} \tag{5.4}$$

Equation 5.4 implies that the posterior is essentially the "likelihood" of a dataset that includes not just the original observations but additional observations representing one's priors. In 2 × 2 table form, for example, new counts are simply added to the counts in each cell computed initially from the sample data alone.

Clogg and his colleagues incremented their data with cell counts constructed from a single prior empirical distribution. The

---

[15] *Id.* at 75.

goal was to "shrink" the group means toward the overall (observed) mean of the sample. Our goal is different; we want to include prior information about the differential effectiveness of arrest for different groups of suspects. That is, we need to specify a prior distribution for each group, represented by appropriate increments to the cell counts. Moreover, our priors are not drawn from the data on hand but rather from past research.

A. CHOICE OF PRIORS

As a practical matter, constructing prior distributions for the interaction effects first required defining groups of suspects for whom we anticipated different treatment impacts. Building directly on the work of Sherman and his colleagues,[16] we divided the total suspect pool into two groups: good risks and bad risks. The good risks included suspects who were employed in the civilian labor force or were in the military. We expected them to be deterred by an arrest.

The bad risks included people who were unemployed or people for whom the police recorded no employment information. We surmised that one of the reasons why no employment information was recorded on the experiment's implementation forms for a substantial proportion of those cases was that police officers considered such information beside the point; namely, for example, from what else was apparent about the suspect, a job was very unlikely. Moreover, some preliminary analyses undertaken to learn why these data were missing suggested that the suspects listed as unemployed and the suspects with no occupations reported looked a lot alike. In any case, taking the findings of the Milwaukee experiment seriously, we anticipated that for bad risks, an arrest would actually increase the likelihood of new family violence because deterrence would fail and bad risks who were arrested would perhaps be more likely to retaliate against the victim.

Given the good risk/bad risk distinctions, specification and use of the prior distributions as Clogg and his colleagues suggest meant that data from Milwaukee and Omaha had to be pooled with the data from Colorado Springs. To maximize comparability across sites, the intervention was considered to be the treatment randomly assigned, and the outcome was the presence or absence of a new family violence incident recorded in official documents. All of the sites employed nominally sound randomization procedures, but the success of random assignment varied across sites. In a similar fash-

---

16 Sherman, Crime, Punishment, and Stake in Conformity, *supra* note 6.

ion, the official data were more comparable across sites than were the interviews with victims, which varied in sampling design, content and response rates.

Table 3 shows the interaction effects for each of the three sites, with treatment as assigned and failure as measured through the official data. The tabulations for Omaha and Milwaukee were kindly provided by Lawrence Sherman. Note that in all three sites, there is a tendency for good risks to do better when arrested and bad risks to do worse. However, some of these treatment effects are very small. Moreover, in the Colorado Springs data, the interaction effects are determined primarily by the distinction between suspects in the military and suspects who were otherwise either employed or unemployed; "good risk" really means being in the military compared to all other employment statuses. Clearly, some overall assessment is necessary.

Proper pooling, however, required two statistical adjustments. First, we did not take the Milwaukee interaction effects at face value. Since they were, in part, the result of a data analysis in which several different interaction effects were considered, we were concerned about overfitting and capitalizing on chance. Capitalizing on chance may also occur when several different forms of the response variable are tried or when various recodings of key variables are examined. Accordingly, we shifted the estimates of treatment effects towards zero by cutting them in half. Since our comparison is relative among treatments rather than absolute, this shift has no effect on the outcome. No such adjustments were applied to the Omaha data, since these were data we requested and were the only data we requested.

Second, we are interested in comparisons of the failure rate for suspects arrested and suspects not arrested. Consequently, the site-specific overall rate of failure is effectively a nuisance parameter. By representing our priors with a constructed dataset based on the Milwaukee and Omaha experiments, the overall failure rate becomes a very serious nuisance indeed. Since the overall failure rates in Milwaukee and Omaha differed from the overall failure rate in Colorado Springs, and since the fractions of cases assigned to arrest also differed, combining a Milwaukee-driven 2 × 2 table and an Omaha-driven 2 × 2 table with the Colorado Springs 2 × 2 table confounded the failure rate for the different sites with comparisons of failure rates for the experimentals and controls. In other words, the pooled data led to an unbalanced design. Suppose, for example, that Milwaukee had applied only the arrest intervention; there was no control group. And suppose that Colorado Springs had applied

Table 3
INTERACTION EFFECTS BETWEEM RISK AND ARREST BY SITE

Omaha: Risk/Treatment	Failure Proportion	Sample Size
Good Risk/Arrest	.19	53
Good Risk/No Arrest	.28	128
Bad Risk/Arrest	.56	30
Bad Risk/No Arrest	.53	47
Milwaukee: Risk/Treatment	Failure Proportion	Sample Size
Good Risk/Arrest	.20	334
Good Risk/No Arrest	.28	163
Bad Risk/Arrest	.28	427
Bad Risk/No Arrest	.26	209
Colorado Springs: Risk Treatment	Failure Proportion	Sample Size
Good Risk/Arrest	.19	279
Good Risk/No Arrest	.20	801
Bad Risk/Arrest	.20	142
Bad Risk/NO Arrest	.18	357

only the control condition of no arrest; there was no experimental group. Clearly, any comparisons between the experimentals and controls would confound site differences in the likelihood of failure with any treatment effects. Our situation is not so extreme, but the difficulties are of the same kind.

In order to separate site effects from treatment effects, we constructed the prior 2 × 2 tables forcing the Milwaukee and Omaha overall failure rates to be the same as the overall failure rate in Colorado Springs. We used the marginal distribution for failures in Colorado Springs when constructing the prior 2 × 2 tables for the Milwaukee and Omaha experiments. In summary, the prior 2 × 2 tables based on the Milwaukee and Omaha experiments were easily constructed with four pieces of information: (1) the marginal proportions of successes and failures for the Colorado Springs experiment; (2) the site specific, marginal proportions assigned to the treatment and control conditions; (3) the overall sample size; and (4) the difference in failure proportions for the experimentals versus

the controls for both the good and the bad risks. As an alternative to applying the marginal distribution of failures (overall) from Colorado Springs to the Omaha and Milwaukee data, one could have simply included dummy variables for site in pooled analysis. The results would be virtually identical.

Since some readers may be uneasy with our Bayesian use of prior information, and since our decision to reduce by one-half the Milwaukee treatment effects was clearly a judgement call, it is important to explore how sensitive our posterior distributions are to the Omaha and Milwaukee data. Therefore, we undertook a number of analyses in which the combined Omaha and Milwaukee sample size was varied as these data were pooled with the data from Colorado Springs. Zero was the smallest same size used for the combined Omaha and Milwaukee data. With that value (N=0), the results would then depend solely on the Colorado Springs data. The actual complete combination of the full Omaha data and the full Milwaukee data was the largest sample size used. With that value (N=670), the results would then derive from a "full" meta-analysis across sites. Various sample sizes in between were also used. As shown below, this process permitted us to graph how the posterior distributions change as the amount of prior information varies. It also allows readers to evaluate for themselves how our results vary depending on how much relevant information they believe is contained in the Omaha and Milwaukee experiments. In all cases, however, the Milwaukee treatment effects (*i.e.*, the difference between the proportions failing under the experimental and control conditions) are halved.

VI. RESULTS

The results are presented in a series of eight graphs. Each graph plots on the vertical axis the mean odds multiplier of the posterior distribution for the treatment effect of arrest against, on the horizontal axis, the number of observations of "prior data" added to the Colorado Springs data. For example, the mean odds multiplier in Figure 1 equals .92 when the Omaha and Milwaukee data on the good risk suspects are ignored (N=0) and .85 when all the Omaha and Milwaukee data on the good risk suspects are included (N=670). Also shown on each graph is the Bayesian 90% confidence region and a "no effect" value for the odds multiplier of 1.0.

We use the mean odds multiplier as the point estimate from the posterior distribution. The treatment is coded so that arrested = 1 and not arrested = 0. Failure is coded so that a new offense = 1 and

no new offense $= 0$. Hence, a treatment odds multiplier of less than 1.0 represents a reduction in new violence for the suspects who were arrested compared to the suspects who were not arrested. For good risk suspects, a treatment odds multiplier of less than 1.0 was anticipated. For the bad risk suspects, a treatment odds multiplier of greater than 1.0 was anticipated.
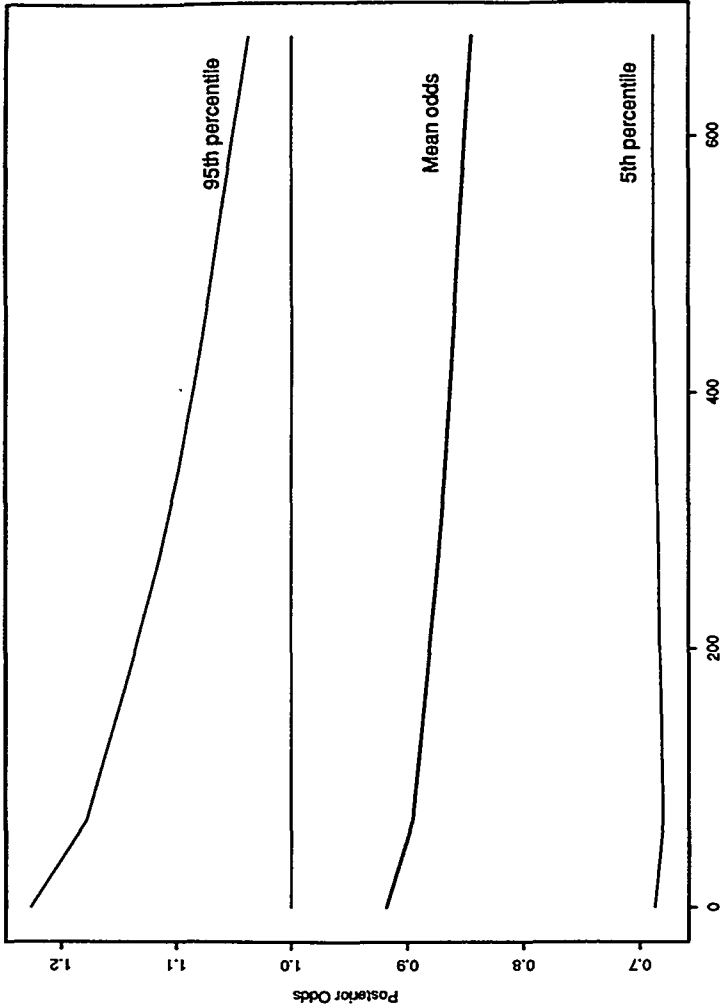
Figure 1. Posterior odds of failure for good risks assigned to the arrest treatment for increasing prior information (official data).
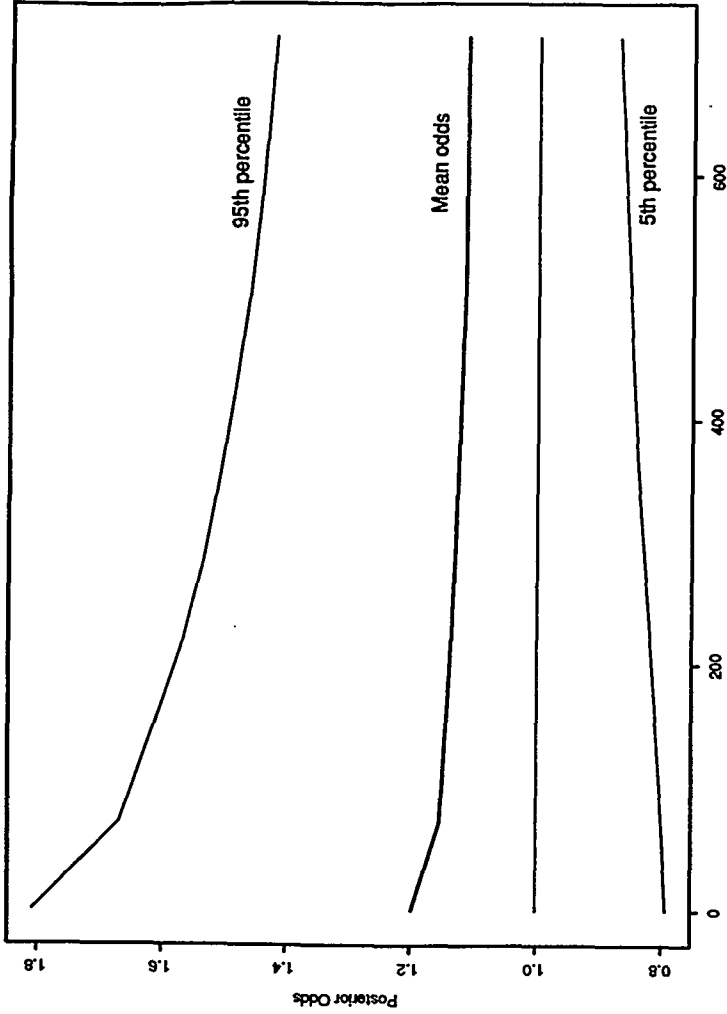
Bad Risk/Arrest Assigned
(Official Data)

95th percentile

Mean odds

5th percentile

Posterior Odds

Figure 2. Posterior odds of failure for bad risks assigned to the arrest treatment for increasing prior information (official data).
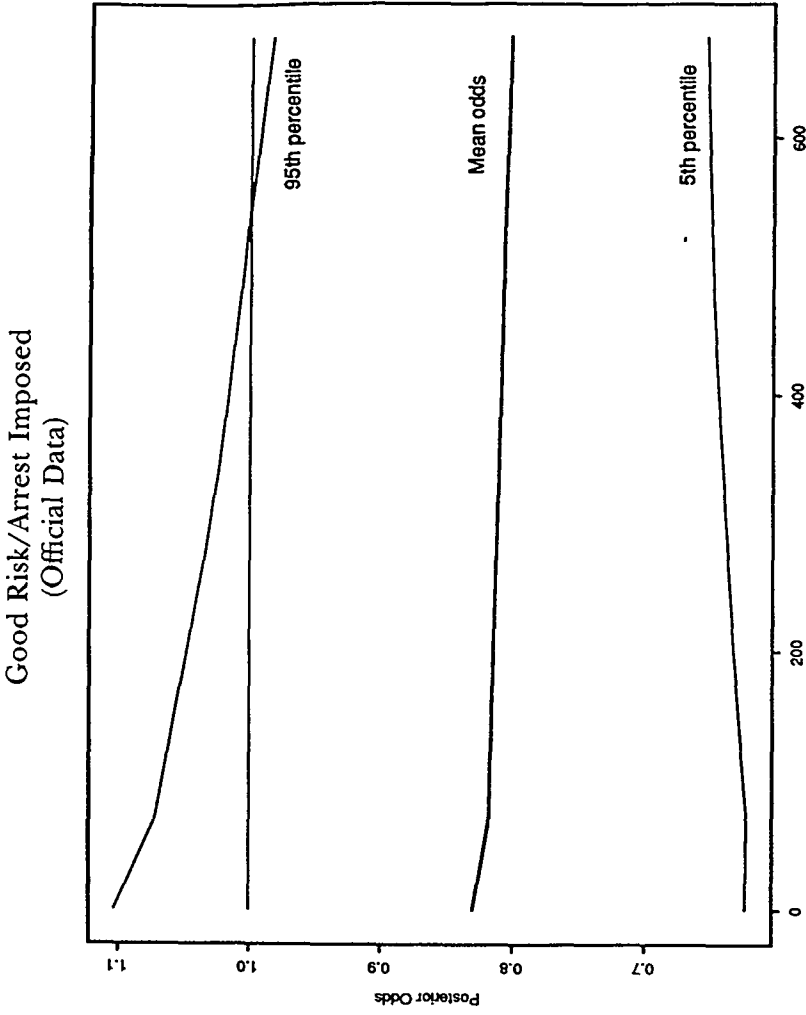
Good Risk/Arrest Imposed
(Official Data)

Figure 3. Posterior odds of failure for good risks with arrest as the imposed treatment for increasing prior information (official data).
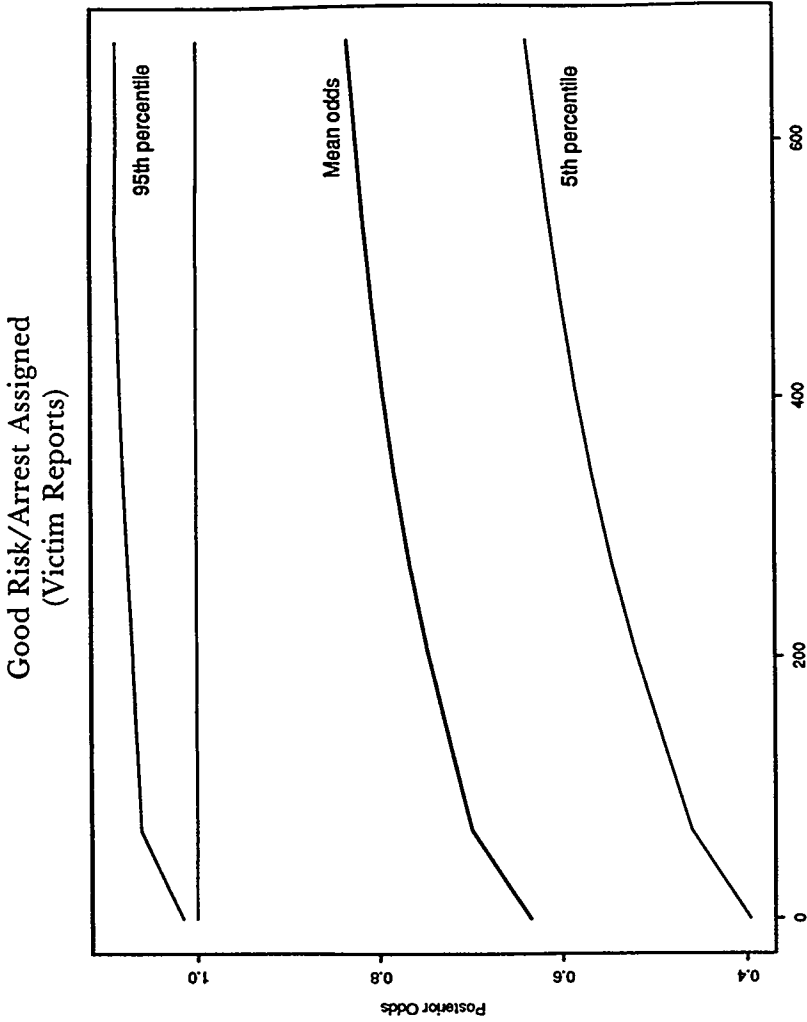
Good Risk/Arrest Assigned
(Victim Reports)

Figure 4. Posterior odds of failure for good risks assigned to the arrest treatment for increasing prior information (victims' reports).

Figure 5. Posterior odds of failure for good risks with arrest as the imposed treatment for increasing prior information (victims' reports).

Figure 6. Posterior odds of failure for bad risks with arrest as the imposed treatment for increasing prior information (official data).

Figure 7. Posterior odds of failure for bad risks assigned to the arrest treatment for increasing prior information (victims' reports).
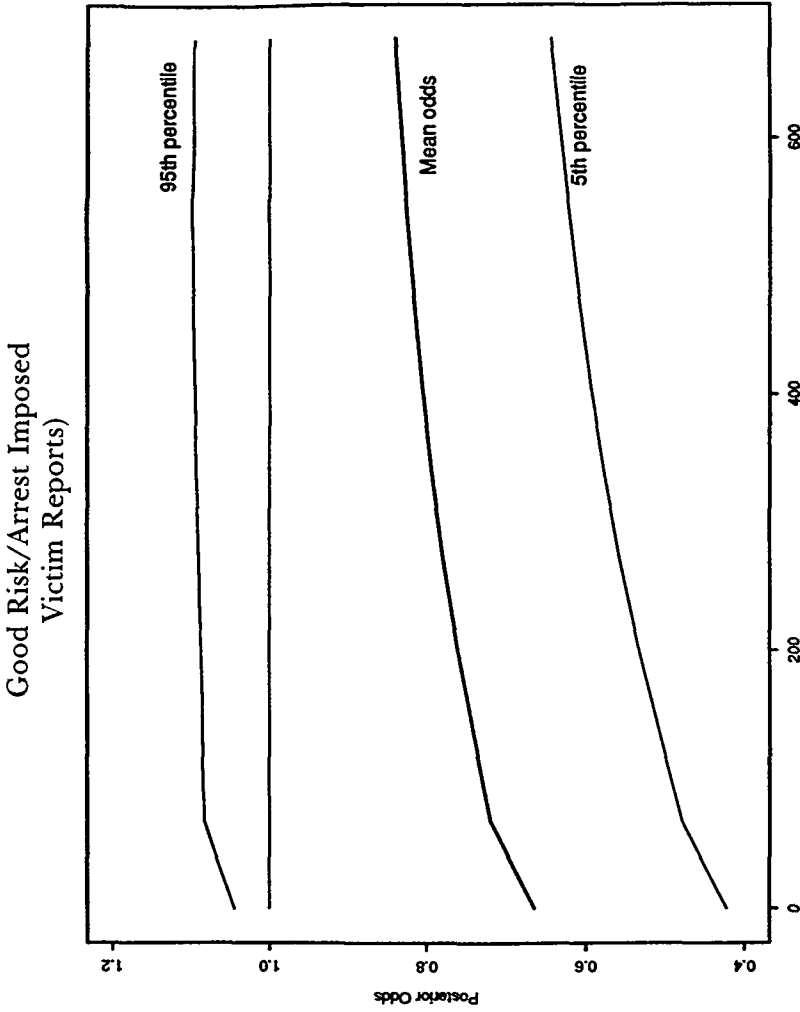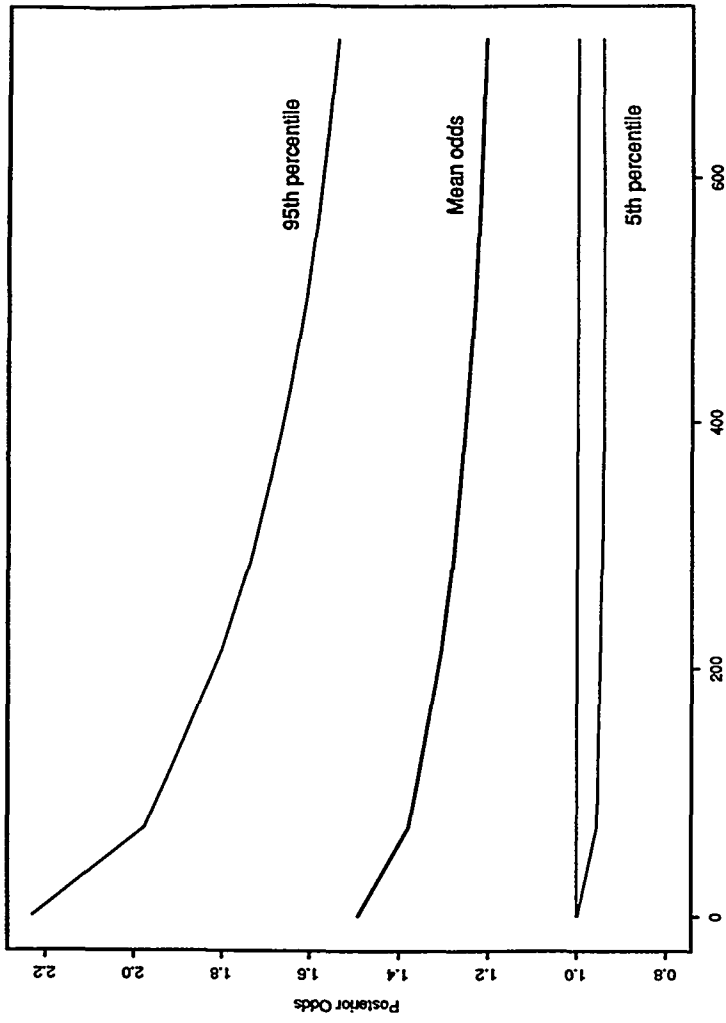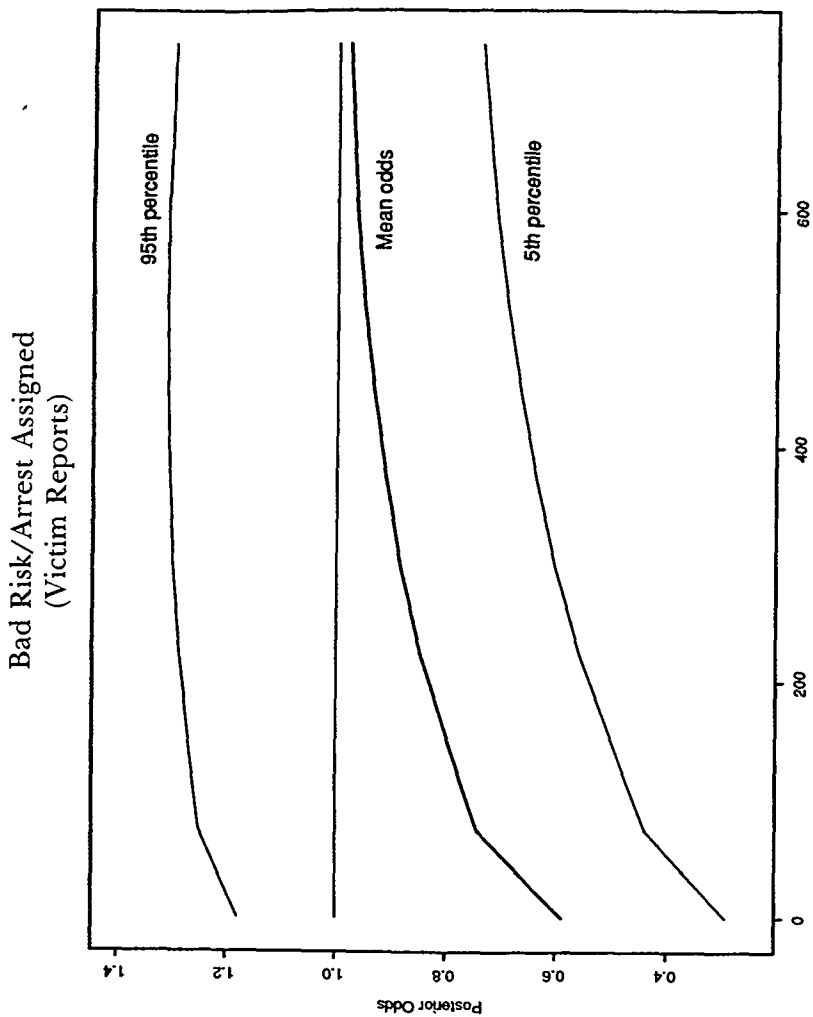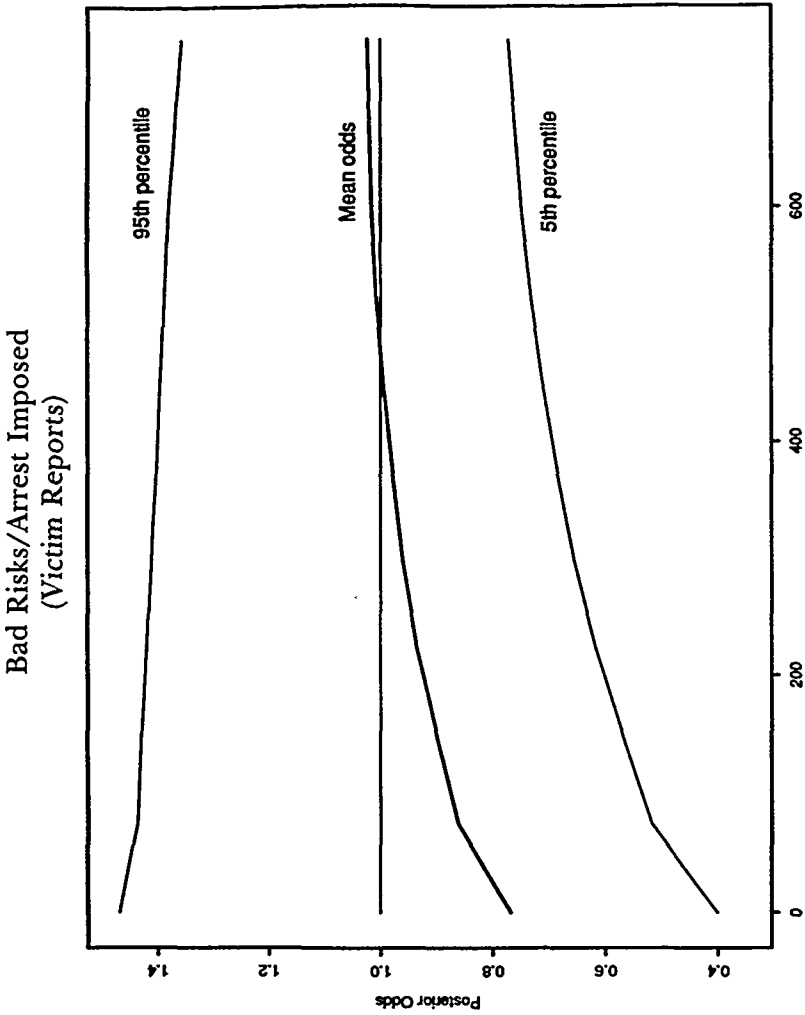
Figure 8. Posterior odds of failure for bad risks with arrest as the imposed treatment for increasing prior information (victims' reports).

Figures 1 and 2 show the graphed results for the assigned treat-ment and the official data outcome variable. The far left side of both graphs shows the mean odds multiplier and ninety percent Bayesian confidence region without the combined Omaha and Mil-waukee prior distribution taken into account; the Colorado Springs data are analyzed alone. The far right side of both graphs shows the mean odds multiplier and ninety percent Bayesian confidence re-gion when the combined Omaha and Milwaukee prior distribution is fully added (but with the interaction effects for Milwaukee still halved). As one moves from left to right, one can see how the mean odds multiplier and ninety percent Bayesian confidence region change as a function of an increasing number of cases from the Omaha and Milwaukee data are added. In short, the graphs provide a sensitivity analysis.

From Figure 1, it is clear that the mean odds multiplier is always less than 1.0; an arrest leads to a reduction in violence. As more of the Omaha and Milwaukee data are added, the mean odds multiplier drops from .92 to .85, and the ninety percent Bayesian confidence region shrinks substantially. A mean odds multiplier of .85 indi-cates that the odds of new violence are reduced by a multiplier of .85; that is, the odds of new violence are eighty-five percent of what they would have been without an arrest. Without the Omaha and Milwaukee data, the balance of evidence favors arrest, but only mod-erately. With the Omaha and Milwaukee data fully included, the balance of evidence strongly favors arrest. In fact, the odds that the odds multiplier is less than 1.0 are about 10 to 1. This is the rough equivalent of a significance test in frequentist inference.

From Figure 2, it is clear that the mean odds multiplier is always greater than 1.0; an arrest leads to an increase in violence. As more of the Omaha and Milwaukee data are included, however, the mean odds declines from about 1.2 to about 1.1. A mean odds multiplier of 1.1 indicates that the odds of new violence are increased by a multiplier of 1.1; that is, the odds of new violence are one hundred and ten percent of what they would have been without an arrest, even as the ninety percent confidence region shrinks. Overall, here the weight of the evidence for the labeling effect is far from compel-ling. Indeed, the odds that the odds multiplier is greater than 1.0 are only about 3 to 1. It is also worth mentioning that had we taken the Milwaukee priors at face value and not discounted them for the possible capitalization on chance, our results in Figures 1 and 2 would have been somewhat stronger.

Recall that, in order to maximize comparability across sites, we built our prior Omaha and Milwaukee distributions from the treat-

ment as assigned and the outcome based on official data. In Figures 1 and 2, those priors were applied to parallel data collected in Colorado Springs. But once we have our "best" priors in hand, they can be used more broadly. In Figures 3 through 8, we apply those priors to the Colorado Springs data varying whether the treatment analyzed is assigned or imposed and whether the outcome analyzed is from the official data or the victim-report data. In our judgement, the "cleanest" story is told from the treatment assigned and the official data; hence the construction of our "best" priors. Nevertheless, there may be some interest seeing how well our findings hold up in the weaker ("less clean") data. Such analyses also provide an opportunity to see how the results can differ as a result of problems with the implementation of the assignment process and the collection of the interview data.

We have already addressed potential problems from the failure to impose the assigned treatment in eighteen percent of the Colorado cases.[17] Reliance on the follow-up interviews and/or questionnaires can also generate significant sample bias. Since the response rate was sixty-four percent for the interviews from which an outcome measure could properly be constructed, bias could result if the attrition from the sample is associated with the treatment assigned and with the outcome variable. We found little evidence of such an association, but our analyses were necessarily limited to the characteristics of suspects and victims reported on the experiment's implementation forms. Recall that these were filled out by the police at the scene, and the forms contain information on virtually all of the suspects and victims in the experiment. However, the amount of information collected on each suspect and victim was necessarily rather limited. It is perhaps important to distinguish between two kinds of artifacts. First, the estimates of any treatment effects may be biased because attrition is related to the treatment assigned and to the outcome. This is conventional sample selection bias.[18] We have no evidence of any sample selection bias as best we could explore it with the data on hand. Second, attrition is unrelated to the treatment assigned but is related to the outcome. Therefore, estimates of any treatment effects are unbiased, but one must be very careful in how those treatment effects are generalized (since the subject pool is not the pool originally selected). We have more to say about this issue below.

---

17 See *supra*, at section IV.

18 Richard A. Berk, *An Introduction to Sample Selection Bias in Sociological Data*, 48 AM. SOC. REV. 386, 396 (1983).

Equally important, the events captured in official data do not fully overlap with the events captured in the victim reports. For example, an assault reported by the victim may not be reported to the police. Conversely, an arrest triggered by a call to the police by a neighbor might not be properly recalled by the victim. Again, the key question is whether the errors are related to the treatment assigned. For example, police officers may be less inclined to arrest an employed suspect. They may define the incident as aberrant or accidental, or they may not want to risk the suspect's job or standing in the local community. Consequently, in the follow-up data, the number of new incidents for employed suspects would be reduced compared to the number of new incidents for unemployed suspects. But unless the underreporting for employed suspects was more common when an arrest intervention had previously been applied, the estimated treatment effects of arrest would be unaffected. In short, it would not be surprising if the posterior distributions for the interaction effects differed depending on the data source, but it is difficult to anticipate the direction in which any biases would go.

Figures 3 through 5 show results for the good risk suspects. By and large, the findings are about the same as those seen in Figure 1. Perhaps the strongest evidence of a deterrent effect comes from Figure 3: imposed arrest and official data. But the basic conclusions seem to hold despite potential problems with the random assignment and the interview data.

Figures 6 through 8 show the results for the bad risk suspects. The strongest evidence of the labeling effect comes from Figure 6: imposed arrest and official data. But Figures 7 and 8 show almost no support for a labeling effect, once the Milwaukee and Omaha data are fully included. For the labeling effect, therefore, method artifacts are very real indeed.

To summarize, the weight of the statistical evidence supports a beneficial treatment effect for good risk suspects. Because they have much to lose, arrest may serve as a deterrent. The evidence on bad risk suspects depends substantially on whether the assigned treatment or the imposed treatment is used and also on the data source for the outcome variable. One's conclusions are sensitive to variations in how the data are analyzed. While the balance of evidence is still in the direction of a labeling effect, it is difficult to draw any firm conclusions, especially conclusions on which one might choose to make public policy.

Finally, we turn briefly the main effect of arrest pitted against all other interventions. This is nothing more than the average treatment effect across sites with the treatment once again defined as the

treatment randomly assigned and the outcome once again constructed from the official data. In contrast to our earlier analyses, we are able to include the pioneering Minneapolis results. Note that there is no need to discount (*e.g.*, in our prior analysis, above) the Milwaukee data for this main effects analysis. In short, we are doing nothing more than pooling the data across all four sites. The only complication is that, once again, one must control for site differences in the average failure rate across treatments.

Table 4 shows the main effect results. 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 the official data. The odds multiplier is virtually 1.0, and the ninety percent Bayesian confidence region is balanced around 1.0. The baseline site is Colorado Springs and the baseline intervention is everything but arrest. Note the large multiplier for Milwaukee. Had we not controlled for site differences in the average failure rate across treatments, a spurious harmful impact for arrest would have been found. When the Colorado Springs data are considered alone, the story is much the same. However, when the Colorado Springs outcome is constructed from the victim reports rather than from the official data, a strong treatment effect surfaces; the odds multiplier for arrest is approximately .65, and the Bayesian ninety percent confidence region no longer includes 1.0.

Table 4
IMPACT OF ARREST ACROSS SITES

Variable	Odds Multiplier	Lower 90% Bound	Upper 90% Bound
Minneapolis	0.92	0.71	1.20
Omaha	0.83	0.64	1.08
Milwaukee	2.29	1.96	2.67
Arrest	0.99	0.85	1.14

Why the results based on the victim report data differ from the results based on official data is currently being explored. One suspicion, consistent with the data analyses done to date, is that victims linked to bad risk suspects are more difficult to recontact for follow-up interviews. For example, they may be more likely to move. Hence, households with bad risk suspects were lost at a higher rate than households with good risk suspects. If the arrest is really an effective deterrent only for the good risk suspects, the greater attrition of the bad risk suspects will lead to an overestimate of the bene-

ficial impact of arrest, averaged over all suspects for whom interview follow-up data are available. This serves to underscore the point that main effects, averaged over all suspects, are not well defined when there are interaction effects present. In brief, we are not taking the apparent treatment effect seriously.

VII. DISCUSSION AND CONCLUSIONS

The balance of statistical evidence from Omaha, Milwaukee, and Colorado Springs suggests that arresting suspects in incidents of spousal violence has a deterrent effect for at least a large and identifiable subset of "good risk" suspects. There is also a hint that for the group of "bad risk" suspects there may be a labeling effect: an arrest can sometimes make things worse. Whether there are deterrent effects on the average across all suspects depends on the relative sizes of the deterrent and labeling effects and on the relative sizes of the two groups of suspects. These will vary from site to site and will, therefore, produce different effects averaged over all suspects.

It is, of course, unreasonable to expect that any spousal violence intervention will produce exactly the same outcome for all suspects, just as it is unreasonable to expect that a given kind of medication will produce exactly the same outcome in all people who have a particular disease. Just as there will often be side effects with medication, there is good reason to expect side effects from interventions in spousal violence. Therefore, one must consider the balance of beneficial and harmful effects, not just upon particular individuals, but upon the set of individuals for whom a prospective intervention is relevant. That balance is not addressed in this article or in any analyses of the Minneapolis replications we have seen to date.

However, if there really are on balance harmful effects for an identifiable subset of offenders and their victims, serious ethical questions are raised. Can one legitimately recommend police practices, even if beneficial in the aggregate, which place identifiable victims at additional risk? While we have no suggestions on the ethical issues as posed, it may be possible to broaden somewhat the scope of possible policies to minimize the ethical problems. One might couple further constraints on high risk offenders who are arrested that would reduce their likelihood of new violence. For example, bail could be made much higher, so that far fewer suspects would be free while awaiting trial. One could seek to make the victim less vulnerable. For example, victims could be strongly encouraged to

make use of local shelters for battered women, which might also re-
duce the risk of retaliation. In short, perhaps a better policy than
simply mandatory arrest for all offenders, regardless of risk cate-
gory, would be to couple an arrest for high risk offenders with addi-
tional measures to protect victims.

Given the many misunderstandings surrounding the interpreta-
tion and use of the findings from the Minneapolis Experiment, four
points should be stressed. First, the good risk/bad risk distinction
needs far better conceptual underpinnings and far more direct
measures. If good risks, for example, are suspects who have a lot to
lose from an arrest, we need to understand precisely what those
possible losses are. If the key is psychic costs such as shame or guilt
as opposed to economic costs, such as the possible loss of a job,
these too need to be measured directly. If the mechanisms do not
involve rational choice, we need to measure the manner in which
social integration *per se* helps establish a psychological environment
in which deterrence can work. In any case, the simple presence or
absence of a job is clearly only a very rough proxy. We have begun
to explore these issues further with the Colorado Springs dataset.

Second, for the good risk/bad risk interactions to be taken seri-
ously, the same story should surface across a number of empirical
measures. In this article, we have only considered a single measure:
employment status. Work is underway that will include a far larger
number of good risk/bad risk indicators.

Third, even if an arrest is more effective than a range of practi-
cal alternatives for particular suspects at a given site, it is hardly a
panacea. It is not a cure for domestic problems, violent or other-
wise, and it is not the only policy instrument that could be brought
to bear. There is plenty of room for many different kinds of inter-
ventions from many different sources.

Fourth, there is never any once-and-for-all answer to any policy
question. As opportunities occur in which social policies may be re-
examined, the current balance of scientific evidence needs to be
scrutinized. That balance can (and likely will) change over time with
improvements in scientific understanding and the changing environ-
ment into which policies are introduced. With the virtual informa-
tion vacuum in the middle 1980s about what police should do in
cases of spousal violence, the Minneapolis Experiment was appro-
priately an important input into the development of useful social
policy. To ignore the Minneapolis results would have meant basing
policy on little more than the habitual practices, with which many
police departments were already deeply dissatisfied. In a similar
fashion, the new findings coming from the replications of the Min-

neapolis Experiment should also be seriously factored into the evolving policy discourse. However, no one should expect the findings to be definitive. Definitive findings do not exist in the policy world.

Is there some sort of policy bottom line? Perhaps most important is that in none of the three replication sites was arrest shown to be less effective overall than any of the other interventions. Therefore, with monetary costs roughly equal, interventions can be picked based on legal and moral concerns. We are not legal scholars or ethicists, but to us, an assault is an assault no matter what the relationships between the parties.

We are more cautious about the interaction effects. The balance of evidence clearly supports a statistical interaction effect between employment status and arrest, at least for employed offenders. The policy implications of this effect, however, are unclear. Even if there are conceptual classes of "good risks" and "bad risks," there remains the tricky problem of how police officers can practically determine which risk class a suspect fits into. Perhaps once all of the replication data are fully analyzed, more definitive recommendations will be forthcoming.