WHEN RANDOM ASSIGNMENT FAILS: SOME LESSONS FROM THE MINNEAPOLIS SPOUSE ABUSE EXPERIMENT

Richard A. Berk Gordon K. Smyth
University of California, Santa Barbara

Lawrence W. Sherman
University of Maryland

1 Introduction

The well-known assets of random assignment to treatment and control groups have lead social scientists and statisticians increasingly to advocate the use of field experiments for estimating the impact of social programs (Reicken and Boruch 1974; Cook and Campbell 1979; Tanur 1983; Fienberg et al. 1985; Berk et al. 1985). These exhortations have apparently had an impact. Beginning over a decade ago, large scale field experiments have either been proposed or initiated by a number of federal agencies, including the Department of Justice, the Department of Labor, the Department of Health and Human Services, the Internal Revenue Service, and others. Progress at the state and local level has been slower, but the recent developments are promising (e.g., California Attorney General's Commission on the Prevention of Drug and Alcohol Abuse 1986: Chapter 8).

Unfortunately, many instances of random assignment implemented in the field have been some distance from ideal. Sometimes, random assignment has been aborted for particular subjects because of ethical or practical concerns. For example, in an experiment underway at a major hospital in Los Angeles, crime victims who come to the emergency room are assigned at random to one of two intensities of counseling. However, subjects who, in the course of counseling, evidence particularly aberrant or self-destruction behavior, are dropped from the study and given special help.

Sometimes subjects choose not to cooperate, or in the natural course of events, may reject the intervention assigned. Thus, in an experiment on job training for single parents being undertaken in San Jose, subjects may drop out of the program for a variety of reasons: poor health, lack of child care, unreliable transportation, and the like.

And sometimes, the randomization mechanism is misunderstood or partially subverted. For example, in an experiment on police responses to family violence, recently begun in Colorado Springs, dispatchers have intermittently failed to follow the randomization protocol.[1]

Violations of random assignment create difficult problems for the design and analysis of field experiments. Clearly, one risks reporting seriously biased estimates of treatment impact (Maddala, 1983: 257-290). In this paper, we build on the experience of the Minneapolis Spouse Abuse Experiment (Sherman and Berk 1984; Berk and Sherman, 1987) to extract some general lessons about implementing random assignment in the field. Section 2 briefly describes the Minneapolis Experiment. Section 3 presents strategies developed to minimize violations of random assignment. Section 4 addresses the need to collect proper data on how the assignment was *actually* undertaken and shows how these data may be used to improve the quality of impact estimates produced. The analysis produces estimates of treatment effects that approximately *double* those reported in earlier publications. Finally, section 5 draws some general conclusions.

2 The Minneapolis Experiment

Police departments across the country long have been unclear about how best to respond to incidents of wife battery. Law enforcement officials certainly realize that wife battery is not only a serious felony, but often a precursor to homicide. Indeed, the report of Attorney General's Task Force on Family Violence (1984: 11) observed, "Battery is a major cause of injury to women in America. Nearly a third of female homicide victims are killed by their husbands and boyfriends. Almost 20 percent of all murders involve family relationships." However, several factors have mitigated against the routine use of arrest.

First, police in most jurisdictions know that prosecutors traditionally have been reluctant to pursue wife battery cases and when the rare conviction is obtained, judges have been reluctant to apply serious penalties. Whatever the reasons for such actions,[2] police officers have understood that arrested offenders were often back home before the arresting officers completed their shift (Berk et al. 1982).

Second, particularly in the 1970's, clinical psychologists (Bard 1970; Potter 1978; Fagin 1978: 123-124) were arguing that mediation was the best police strategy. While the efficacy of mediation was never demonstrated under rigorous conditions, "crisis intervention counseling" was offered as a viable alternative to arrest.

Third, there may well have been among some police a reluctance to intervene aggressively in wife battery incidents because of beliefs in the sanctity of the home and

[1] The senior author is a consultant for all three experiments, but there is no written material yet available for dissemination.

[2] While it is true that wives sometimes are reluctant witnesses, threats from the assailant are part of the explanation. In recent years, Witness/Victim Assistance Programs, often housed with the offices of District Attorneys, have attempted to help wife battery victims cope with each step in process from indictment to sentencing (Goolkasian 1986).

in male prerogatives. It is also certain that at least some police officers capitalized on these beliefs in their own households. In short, there was by the early 1980's no scientific evidence on which to base police policy. Nevertheless, business as usual would probably have continued had not pressure for reform begun to grow. In particular, widely publicized lawsuits in New York and Oakland sought to compel mandatory arrests in wife battery incidents, while successful lobbying efforts in several states reduced the evidentiary requirements necessary for an arrest in misdemeanor domestic assault incidents.

Partly in response to the turmoil and partly as a result of progressive views of Minneapolis Police Chief Anthony V. Bouza, the Minneapolis Spouse Abuse Experiment was designed and implemented. In brief, (see Sherman and Berk 1984, and Berk and Sherman 1987, for more details), for instances of misdemeanor of spouse abuse,[3] there were to be three interventions: arrest, ordering the offender form the premises for eight hours, and some form of advice that might include informal mediation. All three treatments were to be assigned randomly, but allowance had to be made for certain anticipated violations. For example, police officers had to be permitted to make an arrest if the assailant refused to leave the house as ordered for eight hours. Under such circumstances, random assignment was clearly going to be violated.

The target sample size was about 300 cases, with an unsuccessful outcome defined as new violence (between the same victim and offender) coming to the attention of the police, or reported by the victim during one of 13 followup interviews (conducted approximately every two weeks for six months). A variety of statistical analyses for the 313 households included in the experiment all led to the same conclusion: the arrest intervention was most effective in reducing post-treatment violence (Sherman and Berk 1984; Berk and Sherman 1987). Partly as a result of these findings, many police departments made arrest at least presumptory (Sherman and Hamilton 1984).

3 Quality Control for the Random Assignment

Clearly, the best way to minimize the biases than can result from faulty random assignment is to minimize the likelihood that the random assignment will be compromised to begin with.[4] In Minneapolis, five strategies were employed. First, an effort was made to recruit police officers for the experiment who were capable of understanding the experiment's goals and procedures, and who were committed to a fair test of the three interventions. In the end, the participating officers were volunteers who were no doubt among the best officers in the Department. While this raises important questions about the generalizability of the experiment, it also improved the chances that the experiment would be undertaken properly.

Second, all of the officers were given extensive training (as paid overtime) on the problem of wife battery, the reasons why an experiment was desirable, and how the experiment should be implemented. For example, there were a number of role playing sessions in which police offices had the opportunity to practice the experimental procedures, including the random assignment. By the last training session, the police officers were apparently highly motivated and seemingly in control of the tasks to be done.

Third, frequently during the course of the experiment, meetings were held between the police officers and researchers to exchange experiences and ideas. In addition to reinforcing the commitment of all parties, potential difficulties were typically resolved before experimental procedures were seriously affected.

Fourth, research staff maintained regular contact with the participating officers, often through "ride-alongs" when the officers were on patrol. Much like the meetings, these activities helped bolster morale and facilitated collective problem solving. Finally, the devices used to randomly assign treatments were constructed to make "cheating" difficult. As described elsewhere (Berk and Sherman 1987):

> Police officers who participated in the experiment were asked to carry a special pad of report forms, color-coded for the three treatments. Each time officers encountered a situation meeting the experimental criteria, they were to apply the treatment indicated by the color of the report form on the top of the pad. The report, which asked for a few observations about the setting and the participants (e.g. was the offender drunk, was a gun involved), was to be filled out as soon as possible after the encounter. It was then to be removed from the pad and forwarded to the research staff.
>
> All of the color-coded forms were arranged in random order, stapled together in sets of 25, and numbered sequentially. The colors were meant to assist officers who might have to apply a random treatment rapidly under difficult circumstances. The stapling and numbering were meant to discourage even well-intentioned effort by officers to match particular treatments to particular incidents. The numbering also provided us with one check on whether the treatments were being implemented as designed.

In addition to attempts to insure that random assignment was properly implemented, were efforts to clearly define the situations in which random assignment could be voided. "Upgrading" to an arrest from the separation or advice interventions was to be permitted if (1) the offender would not leave the premises when orders, (2) the officers were assaulted, (3) a restraining order was violated, or (4) the victim persistently demanded that the

[3] For ethical and legal reasons, incidents of felony spouse abuse, for which an arrest was legally required, were excluded from the experiment.

[4] Virtually all of the work necessary for instituting the experiment and maintaining its integrity was undertaken by Lawrence Sherman, with the help of a dedicated research staff.

offender be arrested. Clearly, opening this loophole left lots of room for police discretion, but during training and monitoring, we stressed repeatedly that random assignment should not be discarded unless absolutely necessary and then only under the specified conditions. For example, one of the frequent role playing situations involved an uncooperative assailant.

4 Analysis of the Assignment Process

4.1 Data Collection

Despite efforts to insure that the random assignment would be implemented properly, it was clear that random assignment would sometimes not be employed. By design, upgrading was permitted under specified circumstances. In addition, some police would occasionally make errors, or place law enforcement concerns ahead of the research design. For example, it was easy to imagine police officers being tempted to discard a case when the preferred action did not correspond to the treatment assigned. If an arrest was desired, for instance, the offense could be redefined from a misdemeanor to a felony, for which an arrest was by law required.

It was necessary, therefore, to collect data on factors that might lead police officers to abort random assignment. Obvious candidates included measures of situations in which "upgrading" was explicitly permitted. In addition, data were needed on variables that might affect police decisions despite training: whether the assailant was rude, whether there were weapons in the house, whether the assailant was intoxicated, and so on. Note that this information had to be collected *whether or not random assignment was employed, and not overtly in response to why the random assignment was discarded.* We anticipated the need to construct statistical models of the assignment process in which explanatory variables had to available under both random and non-random assignment, and not *in response* to a decision about random assignment that had already been made.[5] For example, it would be a mistake to ask officers directly to explain their actions solely when random assignment was not applied.

Ideally, information on the assignment process should have been collected by observers who accompanied police. However, this was not practical. As a result, we were forced to rely on information that the police could record either as part of their own official forms, or as part of the color-coded forms they were carrying for the experiment. In short, for each incident that was eligible, we made arrangements to collect information on the treatment that was randomly assigned, the treatment that was actually delivered, and a number of factors that might affect whether random assignment was applied.[6]

[5]One would risk simultaneous equation bias, since the treatment assigned might well affect the rationales reported.

[6]We also asked victims during the initial, post-treatment interview about what was happening when the police arrived and what the police did when they were there. By and large, this material was consistent with what the police reported, but was necessarily far less specific and complete. How could respondents know, for example, whether the randomly assigned treatment was applied rather than some other treatment?

Table 1: Implementation of random assignment. (Table entries are numbers of cases.)

	Designed Treatment			
Delivered Treatment	Arrest	Advice	Separate	
Arrest	91	18	26	135
Advice	0	84	5	89
Separate	1	6	82	89
	92	108	113	313

4.2 Data analysis

Table 1 cross-tabulates the designed (randomly assigned) treatment with the delivered treatment.[7] It is clear from the main diagonal that the vast majority (82 percent) of subjects got the randomly assigned treatment. Of those that did not, the most common pattern was, just as the experiment allowed, an "upgrade" to arrest from advice or separate. Not surprisingly, therefore, a multinomial logit analysis of Table 1 reveals that the randomly assigned treatment was the most important cause of the treatment delivered (details in Berk and Sherman 1987). In addition, for the randomly assigned advice and separation treatments, the expected situational variables predict upgrading. Particularly potent are factors for which the loopholes had been explicitly designed (e.g. if an officer was assaulted). Finally, once the set of explanatory variables include the treatments randomly assigned, and for advice and separation, the "upgrading" variables, the model's implied cell frequencies and the actual cell frequencies differ by no more than chance at conventional levels.[8] In short, all of our statistical evidence supports the conclusion that by and large, the assignment of treatments was implemented as planned.

However, the fact that some of the treatments were not assigned at random raises potential difficulties for the usual analysis of randomized experiments. Suppose, for example, that the offenders who were "upgraded" from advice or separation were particularly violent individuals who were likely to commit new assaults in the fu-

[7]The original sample size was 330. Seventeen cases were dropped because they did not fall within the definitions on which the experiment was premised. For example, any assaults between mothers and daughters were dropped. Also note that the numbers along the bottom margin reflect the randomization and that the 1/3–1/3–1/3 null hypothesis is not rejected at conventional levels.

[8]The upgrading was captured as interaction effects via product variables. In one instance, we used a single variable coded "1" if the random treatment was advice or separate and if one or more of the upgrading factors was present, and coded "0" otherwise. When this was added to a model already including two binary (1, 0) variables for randomly assigned advice and randomly assigned separation, we failed to reject the null hypothesis that all of the systematic variation in the table was accounted for (Feinberg 1980: 40–43). That is, the model "fit" the data.

Table 2: Logit analyses of treatment effects. Sample size was 313 for the first two analyses, 301 for the third.

Logit Coefficients for Treatments			
Variable	As Assigned	As Delivered	Both
Intercept	−1.21*	−1.05*	−2.08*
Random Arrest	−0.90*	−0.82*	−1.81*
Random Advice	−0.21	−0.46	−0.14
Loophole	—	—	0.84**
Logit Coefficients for Assignment			
Intercept	—	—	2.91*
Upgrade	—	—	−2.19*

* Statistically significant at the .05 level for a one-tailed test for a null hypothesis of zero.

** Statistically significant at the .10 level for a one-tailed test for a null hypothesis of zero.

ture. This would mean that the randomly assigned advice and separation groups were losing some of their high risk members, while the group actually experiencing arrest was gaining some high risk members. Then, an analysis comparing the three randomly assigned treatments or the three treatments actually received would be biased against the relative effectiveness of arrest.

Table 2 shows three sets of estimates of the treatment effects using official police data. The results from the self-report data are much the same and need not concern us here (see, however, Sherman and Berk 1984 and Berk and Sherman 1987). In a similar spirit, all the results in Table 2 are based on a logit formulation. Reported elsewhere are findings within a time-to-failure framework using Cox proportional hazard regression, but again, the story is much the same and is probably not usefully introduced into this discussion.[9]

The outcome variable is simply whether there was new violence between the same offender and victim recorded on either police offense reports or police arrest reports. An offense report is completed by police officers after they arrive at a location where a crime has been committed. An arrest report is completed by police officers if an arrest is actually made. To underscore the distinction, officers arriving at the scene of a spousal assault incident will fill out an offense report, but not an arrest report, if alleged assailant had left the premises before they arrived.

The first analysis (starting from the left) compares the three randomly assigned treatments, ignoring that some of the randomly assigned treatments were not delivered. It is clear that an arrest is more effective than separation by itself and other analyses show arrest to be more effective than separation and advice combined (Berk and Sherman 1987). Arresting offenders seems to cuts the odds of new violence by a multiplicative factor of about .40.

The second analysis compares the three treatments actually delivered. The findings are almost identical, which should not be too surprising. Even if the upgrading process does bias the results, the number of cases upgraded is modest, and more important, the biases are probably in the same direction whether the treatments considered are as assigned or as delivered.[10]

The final analysis reported in Table 2 attempts to adjust for potential biases that could result from the non-random assignment of arrest. Since the estimates are produced through a likelihood function designed specifically for the Minneapolis experience, some background exposition is required.

Consider the joint probability distribution of two random variables: whether or not an experimental subject fails and whether or not an experimental subject is assigned randomly to treatments. This joint probability distribution can be expressed, of course, as the product of an appropriate marginal and conditional probability distribution.

For those who fail, the probability of failing can be expressed as the marginal probability of being randomly assigned times the conditional probability of failing given random assignment, plus the marginal probability of not being randomly assigned times the conditional probability of failing given non-random assignment. Likewise, for those who do not fail, the probability of not failing can be expressed as the marginal probability of being randomly assigned times the conditional probability of not failing given random assignment, plus the marginal probability of not being randomly assigned times the conditional probability of not failing given non-random assignment. From this, one can write a likelihood function for *a given individual* assigned to *separation or advice* as follows:

$$\begin{aligned} L \;=\; & \{NRA \times P(NRA) \times P(NotFail|NRA) + \\ & RA \times P(RA) \times P(NotFail|RA)\}^{NotFail} \times \\ & \{NRA \times P(NRA) \times P(Fail|NRA) + \\ & RA \times P(RA) \times P(Fail|RA)\}^{Fail} \qquad (1) \end{aligned}$$

where RA is a binary variable coded "1" if the case is randomly assigned and "0" otherwise, NRA is a binary variable coded in exactly the opposite way $(1 - RA)$, *Fail* is a binary variable coded "1"if the case fails and

[9] Cox's proportional hazard regression in continuous time is approximated by a discrete time representation with logistic regressions estimated for each time period. Each regression uses as its data those cases still at risk to failure (i.e. cases not yet lost from censoring or "death"), and constrains the logit regression coefficients to be the same across time periods, except for the intercepts, which are free to vary. As the duration of the discrete time periods shrink, the discrete form estimates increasingly approximate the Cox continuous time estimates (Lawless 1982: 372–377). Hence, the major gain from the continuous time approach is some statistical efficiency.

[10] Of course, one can define the problem away by claiming interest only in the assigned treatments. The first analysis then provides unbiased estimates of the nominal treatment. However, since the point of the experiment was to test a hypothetical set of policy options, it was important to obtain good estimates of what was actually deliverd. More discussion of this point can be found in Berk and Sherman (1987).

"0" otherwise, *NotFail* is coded in exactly the opposite way ($1 - Fail$), $P(RA)$ is the probability of being randomly assigned, $P(NRA)$ is the probability of not being randomly assigned ($1 - P(RA)$). Given random assignment, $P(Fail|RA)$ is the probability of failing, and $P(NotFail|RA)$ is the probability of not failing ($1-P(Fail|RA)$). $P(Fail|NRA)$ and $P(NotFail|NRA)$ are the corresponding probabilities given non-random assignment. In addition, we make each of the probabilities in equation 1 a function of set of explanatory variables. Hence, the probability of being randomly assigned can be written as,

$$P(RA) = \frac{1}{1+e^{z'\alpha}}, \tag{2}$$

where z is a vector of variables thought to affect the chances of being randomly assigned, and α is a conformable vector of regression coefficients. At this point, we are allowing for a single set of explanatory variables and associated parameters, although this is not a necessary restriction in principle. For example, there could be different logistic regressions for different assigned treatments. The probability of failure can be written as,

$$P(Fail|RA) = \frac{1}{1+e^{x'_1\beta_1}}$$

and

$$P(Fail|NRA) = \frac{1}{1+e^{x'_2\beta_2}}, \tag{3}$$

where x_1 and x_2 are vectors of variables thought to affect the chances of failure, especially binary variables for the treatments received, and β_1 and β_2 are conformable vectors of regression coefficients. Recall that Table 1 shows that by far the most common violation of random assignment was, consistent with the experiment's guidelines, an upgrade from advice or separation to arrest. For this paper, we ignore all other violations[11] and simply drop those cases. Then, we can define z as a single binary variable coded "1" if least one of situational factors surfaced that could properly lead to an upgrade (e.g., an uncooperative offender) and "0" otherwise. Table 1 also suggests that there were effectively four kinds of delivered treatments: (1) randomly assigned arrest, (2) randomly assigned advice, (3) randomly assigned separation, and (4) non-randomly assigned arrest. All other possibilities are too few to analyze separately, and are deleted. Thus, there are four treatments defined as a set of binary variables (coded "1" or "0"), with the first three of them included within x_1 and the fourth in x_2.

Two additional complications remain. First, it was not possible to observe for each the assigned treatments, cases which were randomly assigned and cases which were not. For cases assigned randomly to arrest, we observe only a delivered arrest. For such cases, therefore, the probability of random assignment must be, in effect, imputed. If we assume that for cases assigned to arrest at random, equation 2 holds, then for each case randomly assigned to arrest, the probability of non-random assignment may be imputed from cases assigned at random to the other treatments. Second, among those randomly assigned to arrest, a subset of difficult subjects would have been upgraded had they been randomly assigned to separation or advice. Therefore, it was necessary distinguish the impact of arrest for those who in principle would have been randomly assigned from the impact of arrests for those who in principle would not have been randomly assigned.

The two complications associated with the set of subjects randomly assigned to arrest are played out in the details of the likelihood function. The likelihood for a given individual assigned to arrest can be written as follows, with complete details for the full likelihood function provided in the appendix to this paper:

$$\begin{aligned} L = {} & \{P(NRA) \times P(NotFail|NRA) + \\ & P(RA) \times P(NotFail|RA)\}^{NotFail} \times \\ & \{P(NRA) \times P(Fail|NRA) + \\ & P(RA) \times P(Fail|RA)\}^{Fail}. \end{aligned} \tag{4}$$

The final column in Table 2 shows the maximum likelihood estimates. Each of the coefficients are fully consistent with expectations.[12] The estimate of the impact of a randomly assigned arrest is about twice as large as for the earlier results, while the estimated impact of arrest for those not assigned to arrest at random (via the "loophole" given to police officers) is actually ***positive***. That is, the upgraded subjects are very poor risks indeed. Finally, the negative coefficient for "upgrade" indicates that consistent with the experiment's guidelines, the probability of random assignment was decreased when any of the designated situational factors materialized.

Yet, we do not want to make too much of the more intricate analysis. The data are very thin in places so that the movement of but a few cases from "failure" to "success" (or the reverse) could change some of the results. Our major substantive point is that the general conclusions from earlier work are confirmed using a far more sophisticated statistical procedure. Our major technical point is that the likelihood framework provided is generally applicable for instances of "failed" randomization, not just those that parallel the Minneapolis experience. In other words, the the likelihood approach we propose should be more widely useful. Moreover, equations 1 and 4 can be applied in situations where the outcome or assignment process is characterized by densities we have not used.[13] Hence, there is an second important dimension along which generalizations are possible.

[11] The others kinds of violations are too few to treat as a separate problem. The data matrix will turn out to be rather sparse as it is.

[12] Because of the sparse data, estimates of standard errors obtained for the estimated information matrix were suspect. Consequently, likelihood ratio tests were used instead.

[13] For example, the outcome might be the number of "failures" represented by a Poisson distribution. Readers interested in more conventional econometric approaches should consult Maddala's text (1983: 257–290; see also Amemiya, 1985: 360–408)). Perhaps the major difference is that our strategy responds to the particular information available in principle when some experimental subjects are assigned at random and some are not.

5 Conclusions

Researchers undertaking randomized field experiments should routinely anticipate implementation difficulties. In particular, random assignment will often be imperfect. At least four actions should be taken to minimize the impact of flawed random assignment. First, a wide range of potential obstacles to random assignment should be discussed with the personnel who will implement randomization. Where possible, appropriate responses should be articulated and practiced. Second, during the course of the experiment, the assignment of cases to treatments should be carefully monitored. Documented problems should be immediately addressed. Third, data should be collected on the implementation of random assignment in a form that can be used in later data analyses. Finally, statistical procedures should be used that give one some purchase on the impact of any faulty random assignment. However, prevention is always the best strategy, and if the first two recommendations are effectively employed, the last two will be unnecessary.

Acknowledgement

An earlier version of this paper was presented at the 1987 meetings of the American Statistical Association. Discussants Stephen Fienberg and Michael Dennis made a number of useful observations from which this draft has benefited.

References

Attorney General's Task Force on Family Violence, (1984), Washington, D.C.: U.S. Department of Justice.

Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard Press.

Bard, M. (1970), "Training Police as Specialists in Family Crisis Intervention," Washington, D.C.: U.S. Department of Justice.

Berk, R.A., Boruch, R.F., Chambers, D.L., Rossi, P.H., Witte, A.D., (1985), "Social Policy Experimentation: A Position Paper," *Evaluation Review*, 9: 387–429.

Berk, R.A., and Sherman, L.W. (1987), "Police Responses to Family Violence Incidents: An Analysis of an Experimental Design with Incomplete Randomization," *J. Amer. Statist. Ass.*, forthcoming 1987.

Berk, R.A., Rauma, D., Loseke, D.R., Berk, S.F. (1982), "Throwing the Cops Back Out: The Decline of a Local Program to Make the Criminal Justice System More Responsive to Incidents of Domestic Violence," *Social Science Research*, 11: 245–279.

Commission on the Prevention of Drug and Alcohol Abuse, (1986), Sacramento, California: Office of the Attorney General.

Cook, T.D., and Campbell, D.T. (1979), *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago: Rand McNally.

Fagin, J.A. (1978), "The Effects of Police Interpersonal Communications Skills on Conflict Resolution," Ph.D. Dissertation, Southern Illinois University, Ann Arbor: University Microfilms.

Fienberg, S.E. (1980), *The Analysis of Cross-Classified Categorical Data*, Cambridge, Mass: MIT Press.

Fienberg, S.E., Singer, B. and Tanur, J.M. (1985), "Large-Scale Social Experiments in the United States," in A.C. Atkinson and S.E. Feinberg (eds.), *A Celebration of Statistics: The ISI Centenary Volume*, New York: Springer-Verlag.

Goolkasian, G.A. (1986), "Confronting Domestic Violence: The Role of Criminal Court Judges," Washington, D.C.: National Institute of Justice,

Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, New York: Cambridge Press.

Parnas, R.I. (1972), "The Police Response to Domestic Disturbance," in L. Radzinowitz and M.E. Wolfgang (eds.), *The Criminal in the Arms of the Law*, New York: Basic Books.

Reicken, H.W., and Boruch, R.F. (1974), *Social Experimentation: A Method for Planning and Evaluating Social Intervention*, New York: Academic Press.

Sherman, L.W., and Berk, R.A. (1984), "The Specific Deterrent Effects of Arrest for Domestic Assault," *American Sociological Review*, 49: 261–271.

Sherman, L.W., and Hamilton, E. (1984), "The Impact of the Minneapolis Domestic Violence Experiment: Wave I Findings," Washington, D.C.: Police Foundation.

Tanur, J.M. (1983), "Methods for Large-Scale Surveys and Experiments," in S. Leinhardt (ed.), *Sociological Methodology*, 1983–1984, San Francisco: Jossey Bass.

Appendix: The Likelihood Function

The likelihood function shown in equation 1 has eight parts, depending on binary variables for whether random assignment was applied and whether the subject failed.

(1) For those randomly assigned to separation who failed

$$L = \frac{1}{(1 + e^{z'\alpha})(1 + e^{\beta_0})}$$

(2) For those randomly assigned to separation who did not fail

$$L = \frac{e^{\beta_0}}{(1 + e^{z'\alpha})(1 + e^{\beta_0})}$$

(3) For those randomly assigned to advice who failed

$$L = \frac{1}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_1})}$$

(4) For those randomly assigned to advice who did not fail

$$L = \frac{e^{\beta_0+\beta_1}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_1})}$$

(5) For those non-randomly assigned arrest who failed

$$L = \frac{e^{z'\alpha}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_2})}$$

(6) For those non-randomly assigned to arrest who did not fail

$$L = \frac{e^{z'\alpha}e^{\beta_0+\beta_2}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_2})}$$

(7) For those randomly assigned arrest who failed

$$L = \frac{1}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_3})} + \frac{e^{z'\alpha}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_2})}$$

(8) For those randomly assigned to arrest who did not fail

$$L = \frac{e^{\beta_0+\beta_3}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_3})} + \frac{e^{z'\alpha}e^{\beta_0+\beta_2}}{(1 + e^{z'\alpha})(1 + e^{\beta_0+\beta_2})}$$

where:

z is a vector of variables affecting the treatment delivered;

α is a vector of parameters for variables affecting the treatment delivered;

β_0 is the intercept representing the failure rate of individuals randomly assigned to separation;

β_1 is the increment or decrement in the failure rate for individuals randomly assigned to advice;

β_2 is the increment or decrement in the failure rate for individuals arrested, but not randomly assigned to arrest;

β_3 is the increment or decrement in the failure rate for individuals randomly assigned to arrest.

One wrinkle is found in parts 5, 6, 7, and 8. In parts 7 and 8, there is a parameter for the impact of arrest for those assigned at random (β_3) not found in parts 5 and 6. In parts 7 and 8, there is another parameter for the impact of arrest for those *not* assigned at random (β_2) *found* in parts 5 and 6. That is, in parts 5, 6, 7 and 8, β_2 captures the impact of arrest for those not assigned at random.

Another wrinkle is that for the subjects assigned to arrest, we cannot observe what would have happened had they been troublesome; they got arrested regardless. Nevertheless, we can use information in the other parts of the likelihood function to help estimate for those randomly assigned to arrest, the impact of arrest separately for those who in principle would have been assigned at random and those who in principle would not have been assigned at random. This is the source of larger treatment effect compared to earlier published analyses.

The likelihood was maximized using the general likelihood function procedure in GAUSS.