88IJCX 0007

129728

When Bigger is Not Better: Design Sensitivity

In a Sample of Criminal Justice Experiments*

David Weisburd

with

Anthony Petrosino
and
Gail Mason

88-IJ-CX-007

*Research for this article was supported by the National Institute of Justice (Grant #88IJCX-0007) and by the School of Criminal Justice, Rutgers University. I wish to express special thanks to Lawrence Sherman, who developed the idea for a study of randomized experiments in sanctions, Christopher Maxwell and Ana Lopes who assisted in preparation of the manuscript, and Martha J. Smith who helped in developing data for analysis.

Only experimental designs allow researchers to make an unambiguous link between causes and their effects (Sechrest and Rosenblatt 1987; Campbell and Stanley 1966). Random assignment of subjects into treatment and "control"[1] groups makes it possible to assume that the only systematic difference between experimental and control subjects is found in the interventions that are studied. In contrast, correlational or quasi-experimental designs are always plagued by the possibility that some important confounding factor has not been taken into account by investigators (Farrington, Ohlin and Wilson 1986; Brody 1978). While criminal justice researchers have long recognized the advantages of randomized experiments, the difficulties of developing experimental designs in real life criminal justice settings and the ethical questions raised by random allocation of criminal justice sanctions have generally led researchers to other less controversial and more easily developed research methods (see Clarke and Cornish 1972).[2]

Whatever the disadvantages of randomized experiments, in recent years there has been a growing interest in experimental

---

[1] In criminal justice experiments there is seldom a group that receives no treatment. More commonly, as is illustrated later, offenders are given different types of treatments, for example, intensive versus traditional probation.

[2] For an interesting and persuasive rebuttal to many of these criticisms, see Boruch, 1976.

methods in the social sciences (Fagan 1990; Berk et al. 1985). This interest has been mirrored in criminological circles, where a number of influential scholars have sought to encourage and expand the use of randomized experiments (e.g. see Farrington, Ohlin and Wilson 1986). The success of their efforts is evidenced in the large number of experimental studies ongoing by the late 1980s (see Garner and Visher 1988), and the fact that the annual program plan of the National Institute of Justice in 1990 was "strongly encouraging" experimental as compared with non-experimental research designs (1990, p. 6).

This growing interest in experimentation has been accompanied by a concern with the adequacy of experimental methods in criminal justice. Petersilia, for example, argues that little attention has been paid to the special difficulties of designing and managing field experiments in the justice system, or the potential strategies that might be used to overcome such problems (1989; see also Dennis 1988). The fact that most experimental studies in criminal justice have failed to show significant research findings adds support to such concerns (e.g. see Farrington 1983; Weisburd, Sherman and Petrosino 1990), though the link between experimental design and study outcomes has not been explicitly tested. In this essay we examine this question in the context of a review of experimental studies in criminal justice sanctions conducted by Weisburd, Sherman and Petrosino (1990). Focussing upon the problem of statistical power, we challenge traditional assumptions about the relationship between research design and experimental results.

In section I statistical power is described with particular attention focussed on conventional means of increasing the power of research designs. Section II presents a brief description of the experimental studies identified by Weisburd et al., and section III looks at general methodological characteristics of those experiments. In section IV we examine the relationship between sample size, usually seen as a primary determinant of statistical power, and the actual outcomes of experiments. Our finding in that section, that larger studies which should lead to more powerful research designs do not, is discussed in section V in terms of various components of the design and management of field experiments. In Section VI we suggest ways of overcoming the weaknesses of large sample designs and present some concluding comments on the implications of our findings for future experimental research.

I. Statistical Power

In contrast to statistical significance--which identifies for the researcher the risk of stating that factors are related when they are not-- statistical power provides an estimate of how often one would fail to identify a relationship that in fact existed. In statistical terms, power is defined as "1 - type II (beta) error," or one minus the probability of accepting the null hypothesis (usually "no difference") when it is false. Its relationship to the proposition that many experiments are designed for failure is straightforward. Statistical power can identify when a research enterprise is likely to fail to provide support for the existence of an effect when, in fact, it is present in the population.

The importance of statistical power is often not fully understood by researchers, who are generally much more concerned with the concept of statistical significance. It has become virtually impossible to present research findings without attention to the statistical significance of research results, and norms concerning significance criteria are strongly established. While little if any attention is paid to statistical power in the design of criminal justice studies (Brown 1989), researchers carefully set significance (or alpha) levels at the outset of an experiment in order to avoid accusations of bias later on. Generally, a .05 level of significance is set. In other words, it is assumed that taking a risk of rejecting the null hypothesis five in a hundred times, when it is correct, is acceptable. Such clear standards for significance thresholds have allowed researchers to guard against the problem of biasing results to the research hypothesis.[3]

The notion that researchers may be biasing results against the research hypothesis (or for a finding of "no effect") has appeared less troubling to criminal justice scholars. Especially in experimental studies, which are often developed to test the effectiveness of expensive government interventions, the possibility that a study would be designed in a way that made it

[3] When significance thresholds that make it easier to reject the null hypothesis (e.g. alpha=.10) are used, the researcher is generally expected to carefully explain his or her departure from established convention.

difficult to identify program success has appeared unlikely. Nonetheless, evidence from primarily non-experimental criminological research (Brown 1989) suggests that criminal justice studies are often severely underpowered.[4] This means that research is often designed in such a way that even if the effect the researcher posits is present in the population it is unlikely to be detected in the sample under study.

At this point it will help to clarify our discussion if we take a concrete example. Suppose a researcher wanted to examine the effects of methadone treatment on the six-month recidivism rate of drug addicts. Following the experimental method, he or she would randomly allocate addicts into control and treatment groups. The statistical power of this experiment is the probability that the statistical test employed would lead to a significant finding. Clearly, the researcher would not want to design a study that would make it highly unlikely to establish a relationship between methadone treatment and reduced recidivism if one existed. But, importantly if a test is not powerful, than the risk of such an error (a Type II error) is very high. How then can one design a powerful study? More simply, what are the components that make up statistical power?

---

[4] There is also a substantial literature that documents the fact that research in the social sciences is generally underpowered (e.g. Chase and Chase 1976; Orme and Tolman 1986).

Most obvious of these components is the significance criterion employed in a statistical test.[5] Clearly, the simplest way to decrease the likelihood of failing to reject the null hypothesis is to adjust the test statistic used as a threshold for statistical significance. One way to do this is to change the risks of type I error employed in an experiment. Because statistical power and statistical significance are directly related, when a less stringent level of significance is chosen (for example .10 as opposed to .05) it makes it easier to reject the null hypothesis and achieve statistical significance, and thus the experiment becomes more powerful. While this method for increasing statistical power is direct, it is usually not a practical suggestion since, as already discussed, norms concerning levels of significance are fairly well established.

A more practical method for changing the value of the test statistic needed to reject the null hypothesis is to limit the direction of the research hypothesis. A "one-tailed test" provides greater power than a "two-tailed test" for the same reason that a less stringent level of significance provides more power than a more stringent one. By choosing a one-tailed test of significance the researcher reduces the value of the test statistic needed to reject the null hypothesis. This occurs because the critical

---

[5] The type of statistical test used in an experiment can also affect its statistical power. Some tests are more appropriate in particular situations and provide more powerful tests of research hypotheses. Nonetheless, the differences between the power of different tests (equally appropriate to the problem at hand) is usually relatively small.

region (the part of a sampling distribution which defines the area of rejection of the null hypothesis) is shifted to test only one of the two potential outcomes in an experiment. For example, in the methadone experiment discussed above, the researcher could rule out the possibility at the outset that treatment might backfire and increase drug use. While there are many cases in which a directional research hypothesis is appropriate, once a one-directional test is posited, a surprising finding in the opposite direction cannot be touted as a major result.[6]

A second component of statistical power is what statisticians define as effect size. Effect size measures the influence of the intervention that is being assessed by looking both at the magnitude of the differences between treatment and control groups and the stability of those differences (see Cohen 1988).[7] The relationship between statistical power and effect size is a straightforward one. When an effect in a population is larger it is harder to miss in any particular sample. Since statistical power asks what the likelihood is of detecting a particular effect (i.e., achieving statistical significance) in a given sample, when effect size is larger the experiment is more powerful. Where effects are hypothesized to be relatively small, other aspects of design must

---

[6] This is a case where you cannot have your cake and eat it too. If the researcher chooses to gain the advantage of a one-tailed test, than he or she must sacrifice any finding in a direction opposite to that originally predicted. To do otherwise brings into question the integrity of the researcher's statistical design.

[7] We will speak more about the computation of effect size coefficients in Section IV.

be maximized in order to achieve an acceptable level of statistical power.

Effect size is generally seen as the characteristic of statistical power which is most difficult to manipulate. A test is ordinarily conducted in order to determine the influence of an intervention on subjects. In experimental field research the intervention itself is usually arrived at through a complex series of negotiations between researchers and practitioners. Though, as we will point out later, effect size can be manipulated in ways that do not adversely affect the theoretical or practical goals of an experiment, there has been relatively little consideration of effect size in efforts to increase the power of experimental designs.[8]

The final component of statistical power, and the one most often used to manipulate power in social science research, is sample size (see Kraemer and Thiemann 1987). Larger samples, all else being equal, provide more stable and reliable results than do smaller samples. Though later on we will illustrate this fact in the context of experiments in sanctions, the statistical logic here is not complex. Larger samples are more "trustworthy" than smaller ones. For example, one would not be surprised to get two or three heads in a row from a toss of an honest coin. However, if the coin produced only heads in a sample of 25 tosses, we would be much more suspicious. Getting one hundred heads in a hundred coin tosses would lead even the most trusting person to doubt the fairness of

[8] For an important exception, see Lipsey 1990.

the coin. In this same sense, larger samples are more powerful, since they are more likely to be able to identify an effect, if it exists in a population, than are smaller studies. Conversely, as Kraemer and Thiemann note, "the smaller the sample size the smaller the power" (1987, p. 27). Because sample size provides a method for increasing statistical power that is straightforward and does not involve manipulations in either the significance levels employed or the treatments administered, it has played a central role in power analyses in the social sciences.

Returning to the methadone example, the researcher would, as David Farrington has suggested, "assess the size of effect (e.g. percentage difference) that would have practical significance and then calculate the sample size that would be needed to obtain statistical significance with this size of effect" (1983, p. 206). Put differently, the researcher's central problem is to identify the sample size needed to provide a powerful experiment based on the significance criteria and the effect size hypothesized. At a minimum, it is recommended that a statistical test have a power level greater than .50--indicating that the test is more likely to show a significant result than not (e.g. see Gelber and Zelen 1985, p. 413). But it is generally accepted that the most powerful experiments seek a power level of .80 or above (e.g. see Cohen 1973; Gelber and Zelen 1985). Such experiments are, given the assumptions about significance and effect size outlined by the researcher, highly likely to evidence a significant finding. The problem for our methadone researcher, simply stated, is to collect enough cases to achieve this threshold of power.

This is the process generally followed in developing powerful research designs. It is on its face a way of ensuring that a particular study is not designed for failure. While it makes assumptions about significance and effect size, it is primarily reliant on sample size to achieve a desired level of statistical power. It is based on the assumption that all else being equal, larger samples provide for a more powerful research design. But as we will see shortly, the simple assumption that effect size is a fixed parameter, staying basically constant across samples of different sizes, is a flawed assumption.

II. The Sample: Experiments in Sanctions

Our analysis of experimental design is drawn from a review carried out by Weisburd, Sherman, and Petrosino (1990). They attempted to identify all randomized studies reported in English which were conducted in criminal justice settings and utilized coercive "treatment" or "control" conditions. They used five specific criteria for inclusion of studies in their review:

1) That individuals were used as the primary unit of analysis;

2) That those individuals were randomly allocated into multiple treatment groups or treatment and control groups;[9]

---

[9] In seven cases Weisburd et al. include studies which randomized according to alternative allocation schedules. For example, in the Denver Drunk Driving Experiment (Ross and Blumenthal 1974; 1975) investigators allocated subjects based on alternative months. In the Hamilton Juvenile Services Project

3) That at least one outcome variable (whether self-report or drawn from a criminal justice agency) measured crime related activities;

4) That the intervention or treatment (or the control condition) be coercively applied by a criminal justice agency in response to or anticipation of a criminal act;

5) That there be a minimum of 15 cases included in at least two of the groups examined.[10]

Weisburd et al. were able to identify some 76 experiments that fit their specific criteria after a careful search of both computerized

---

Experiment (Byles and Maurice 1979) and the California Juvenile Behavior Modification and Transactional Analysis Experiment (Jesness, DeRisi, McCormick, and Wedge 1972; Jesness 1975) the investigators used an odd/even system for placing offenders in treatment and control groups. In the Police Foundation Shoplifting Arrest Experiment investigators noted that offenders were "alternatively assigned to an arrest or release category" (Williams et al., 1987). Such allocation procedures are random in the sense that there were not systematic biases in the choice of subjects who would be placed in each of the allocation sequences. However, because such studies might be seen as violating components of a classical experimental design, we replicated our basic analyses without them. The results do not differ substantively from those reported here.

[10] Farrington, Ohlin and Wilson (198, p. 66) argue that a "randomized experiment can control for all extraneous variables... only if a reasonably large number of people (at least 50) are assigned to each condition." We could find no statistical reason for using this particular threshold, and Farrington et al. do not detail their thinking on this question. Our understanding is that the disadvantages of smaller samples are already taken into account by statistical tests in estimates of the standard errors of sampling distributions. The relatively low threshold used by Weisburd et al. reflects their desire to include as broad a sample as possible. A sample of thirty cases total in a study, is generally seen as the minimum N that will allow researchers to use parametric statistical tests (e.g. see Hays 1981, p. 218). The usefulness of the Weisburd et al. approach is illustrated in our discussion of the relationship between sample size and statistical power in Section IV.

and non-computerized criminal justice and general social science bibliographic indexes (see Appendix I).[11] Once identified, each experiment was described in a registry and included in a computerized data base that detailed specifics of the subjects, sanctions, methods and outcomes of the experiment.

Some mention should be made at the outset of the limitations created by identifying a sample of experiments through published studies and reports. A sample of what is reported is not the same as the universe of all studies. We might expect, for example, that studies that show a significant effect for criminal justice interventions would be more likely to be disseminated and published. And accordingly, there may be a bias to "successes" in this review as in others (see Coleman 1989). Moreover, we suspect that studies conducted in criminal justice settings by agency researchers, as compared with studies supervised by university researchers, are also more likely to escape inclusion in a review

---

[11] The search for studies to include in the data base began with a review of Farrington (1983) and Farrington, Ohlin and Wilson (1986). From the references and studies included there, additional references and studies were reviewed, including bibliographies, qualitative works on the topic of randomized field experiments, and elaborations of studies already included in the sample. A search of the Criminal Justice Abstracts data base was also conducted. At the same time, additional narrative review articles on experimentation, deterrence, rehabilitation, sentencing, and corrections were examined. In addition, a search of the National Criminal Justice Reference System (NJCRS) was completed in June 1989 for 1973-88 using the following keywords: a) randomization; b) controlled study; c) random assignment; d) randomly assigned; e) random allocation; f) field experiment; g) randomized experiment; and h) controlled trial.

Almost 70% of the experiments were reported in scholarly journals or books. Twenty-eight percent were discussed in government publications and 3% were identified only in non-governmental research reports.

of published materials. We are certain there are other biases as well that relate to the dissemination of research findings. Nevertheless, we do not want to over-emphasize such limitations. Most of the studies included by Weisburd et al. did not report any statistically significant results,[12] and many were conducted without any substantial university (or research institute) involvement.[13] Moreover, it is likely that the major studies conducted with significant research resources would have been disseminated regardless of their results.

The criteria employed by Weisburd et al. laid a fairly wide net for the identification of experimental criminal justice studies. There is, for example, tremendous diversity in the sanctions evaluated by researchers. While such penalties as probation, parole and imprisonment occur most often in the studies examined, there are also examples of studies evaluating police interventions, such as arrests (e.g. see Sherman and Berk 1984; Williams, Forst and Hamilton 1987), prison tours, like the Scared Straight experiment in New Jersey (Finckenauer 1982), and restitution (e.g. see Schneider and Schneider 1983; Schneider 1986).

Most often the experiments tested the influence of alternative criminal justice sanctions or the application of differing dosages of a particular sanction. For example, Ross and Blumenthal (1974;

---

[12] In seven out of ten studies no significant differences were found between groups included in the experiment.

[13] Thirty-seven of the experiments were conducted without major support from a university or research institute.

1975) randomly assigned drunk drivers to three groups: a group that received a fine; one that received regular probation; and one that received therapeutic probation. In the Sacramento 601 Diversion project (Baron, Feeney, and Thornton 1972; 1973), one of thirteen diversion studies in the review, juvenile delinquents were randomly assigned to an experimental group receiving family and individual counseling or to a control condition that went before the juvenile court. A number of parole studies varied the intensity of caseloads or supervision services. This was the case, for example, in the California Special Intensive Parole Experiment (Reimer and Warren 1957) conducted in the early 1950s.

In one unusual probation study, Illinois parolees were randomly assigned to regular probation supervision, or probation supervision carried out by volunteer lawyers (Berman 1975; 1978).

There are relatively few experiments where the experimental or control group was able to avoid criminal justice intervention altogether,[14] though some of these are particularly well known. For example, in the Minneapolis Domestic Violence Experiment (Sherman and Berk 1984; 1984a; Berk and Sherman 1985; 1988) suspects were randomly allocated either to an arrest group, or to a group that received discretionary mediation, or to one in which suspects were ordered to stay away from home for eight hours. In

[14] It could be argued that the diversion experiments did this as well. But when offenders were diverted from traditional criminal justice processing they usually received a fairly intrusive regimen of counseling or supervision.

the Police Foundation Shoplifting Arrest Experiment (Glick, Hamilton, and Forst 1986; Sherman and Gartin 1986; Williams, Forst and Hamilton 1987) those in the experimental group were arrested after being identified as shoplifters. Members of the control group were released. While a few prison experiments contrasted continued incarceration with some type of work release or halfway house supervision (e.g. Lamb and Goertzel 1974; 1974a), only one contrasted imprisonment with release. In the California Reduced Prison Sentence Experiment (Berecochea and Jaman 1973; 1981), inmates were randomly assigned to six-month-early release or a group which finished out their full sentences.

Eight of the studies tested the effects of group assignment to different institutional "wards," "regimes," or "communities." For example, in the Fricot Ranch Experiment (Jesness 1965; 1971), male delinquents were randomly assigned to an experimental 20-bed dormitory, or to the more traditional 50-bed unit. In the English Borstal Allocation Experiment (Williams 1970; 1975) youths were assigned to three types of borstal institutions: one that emphasized therapeutic treatment, one that included group counseling, or one that emphasized hard work and paternalistic control.

More than a quarter of the experiments involved treatments that are added onto traditional criminal justice sanctions, often in the context of a prison or jail stay. Many of these studies would not have been seen by the original investigators as sanctioning experiments but rather as attempts at arriving at effective rehabilitative treatments. For example, in the Copenhagen Short Term Offender Experiment (Bernsten and Christiansen 1965) adult male prisoners were randomly assigned to an experimental group receiving psychological examination, interviews with social workers, or some form of individualized treatment geared toward resocialization. Members of the control group received services available through routine custody. In the California Juvenile Probation and Group Counseling Experiment (Adams 1965) juvenile male probationers were randomly assigned to an experimental group that received counseling sessions each week over six months and a control group that received normal probation services. Such experiments were included by Weisburd et al. when inmates were coerced into participating. In cases where participation in the experiment was voluntary, the study was excluded (e.g. see Annis 1979).

The experiments reviewed included a substantial degree of diversity in the types of offenders examined. Nonetheless, most of the studies had predominantly male samples, and a majority of the subjects in most of the experiments were white. Half of the studies reviewed were conducted only with juveniles and most included offenders prosecuted for relatively minor offenses. Indeed a number of the experiments specifically excluded high-risk offenders. Though, as already discussed, there are difficulties in making inferences from a sample of published materials, we suspect that the controversy surrounding random allocation of criminal justice interventions makes it more difficult to include offenders convicted of serious crimes and perhaps easier to conduct studies with juveniles. Still, almost half the experiments included some adult offenders, and a few randomly allocated persons convicted of more serious crimes.[15]

The sample includes experiments from 18 states as well as the District of Columbia. Fourteen studies were conducted outside the United States, with twelve carried out in England. Perhaps it is not surprising, given the tradition of support for empirical research in California, that more than forty percent of the studies came from that state. Overall, most of studies were carried out across institutions within a state or local jurisdiction. Nonetheless, two studies were carried out across institutions in the federal justice system. Nineteen of the experiments were carried out in only one institution.

Weisburd, Sherman and Petrosino thus identify a broad spectrum of experimental studies for our analysis. Nonetheless, their inclusion criteria led them to exclude a number of better-known experiments in criminal justice. For example, the Kansas City Preventive Patrol Experiment (Kelling, Pate, Dieckman and Brown 1974), which randomly allocated varying amounts of police patrol, was excluded because it involved random allocation of geographic areas (beats) rather than people. Similarly, Tornudd's (1968) study of the effects of differential prosecutions on drunkenness randomly allocated towns rather than offenders.

---

[15] For example, see the North Carolina Butner Correctional Facility Experiment (Love, Allgood & Samples, 1986) and the English Prison Intensive Social Work Experiment (Shaw, 1974).

The sanctioning criteria employed by Weisburd et al. also led to the exclusion of a number of well-known studies. For example, the Living Insurance for Ex-Prisoners (LIFE) and the Transitional Aid Research Project (Tarp) experiments, both often thought of as criminal justice studies, were excluded from the sample (see Berk, Lenihan, and Rossi 1980; Rossi, Berk and Lenihan 1980). These experiments randomly assigned subjects released from prison to groups which received weekly stipends and a control group which did not. The study did not meet the requirements for inclusion in the sample because payments were not administered by criminal justice agents. The classic Cambridge-Somervile Youth Study (Powers and Witmer 1951) was excluded for similar reasons. It involved a social work response that could be refused by the subjects or their families.

The criterion that the experiment include crime-related outcome measures meant that studies like the Manhattan Bail Project (Ares, Rankin, and Sturz 1963), an often-cited experiment, also do not appear in this sample. There it was the success of pre-trial recommendations for release or bail rather than the influence of sanctions upon recidivism that was assessed. Similarly, Taylor's (1967) study of the effects of psychotherapy on borstal girls was excluded because only psychological outcome measures were examined.

III. Experiments in Sanctions: Methodological Characteristics

Comparatively few experimental studies in criminal justice provide very great detail about the methods employed in designing and carrying out research (Lipsey 1990). This is due, in part, to the norms of report writing and publishing. There is just not the same demand for discussion of methodological details of research as there is for elaboration about outcomes or theoretical perspectives that led up to the studies. Nonetheless, it is possible to examine in a general way a number of characteristics of the experimental research reviewed by Weisburd et al. Before turning specifically to the relationship between effect size and sample size, we examine below a series of other design questions that are related to the power of experimental studies.

As described earlier, the size of a sample is directly related to the statistical power of an experiment. All else being equal, larger experiments are more powerful, and for this reason sample size has become the primary design characteristic manipulated by experimental researchers in order to increase the power of their research. Interestingly, while focussing on larger samples, few researchers have taken advantage of the fact that studies in which the size of the groups examined are relatively similar are more powerful than those in which the size of the groups is markedly different. While the benefit here is usually small, it can be large when the Ns per group differ widely. And this is the case for a number of the experimental studies examined by Weisburd, Sherman and Petrosino.

The problem is illustrated by a formula for standardizing "N" sizes in experimentation used by Cohen (1988) in developing statistical power computations:

$$\frac{2(N_1)(N_2)}{N_1 + N_2}$$

For example, if there is a total N of 500, but 400 in one group and 100 in another, the weighted N used in power (and significance) calculations is only 160, while the weighted N for a two-group study equally divided between experimental and control groups is 250. Though the overall size of both studies is the same, the design of the latter is more powerful.

Often it is impossible to identify why the sizes of experimental and control groups are unequal in the experiments.[16] We suspect that the reason is usually linked to randomization itself. Many studies randomly allocated subjects in ways that limited their control over the number of individuals that fell in each group. For example, in the Sacramento Juvenile 601 Diversion Experiment (Baron, Feeney, and Thornton 1972; 1973) offenders were allocated to treatment and control groups based on randomly chosen days. Five of the experiments used a toss of a coin or die to randomly allocate subjects. In eighteen of the forty-four experiments that described randomization procedures, researchers reported the use of random numbers tables. Though one might expect relatively equal groups using this technique, this was not always true. For instance, in the National Restitution Experiment in Washington D.C. (Schneider and Schneider 1983; Schneider 1986) the control group had 137 subjects and the treatment group 274.

[16] It should be noted that four in ten of the studies reviewed did not describe how randomization was carried out.

Table 1 illustrates the direct relationship between statistical power and sample size. The experiments are divided into five groups based on their standardized N levels: "15-50" "51-100" "101-200" "201-400" "over 400." Across the table are the average levels of power for the experiments in each group given assumptions of "small" "moderate" and "large" effects (see Cohen 1988).[17] Looking at Table 1 it is clear that there is a substantial amount of diversity in the sizes of the samples chosen. For example, 12 of the studies include 50 or fewer standardized cases per group, and 11 include more than 400 standardized cases per group. As is to be expected, as the average sample size gets larger the power levels associated with each hypothesized effect size also increase. For the smallest experiments, very large effect sizes would be needed for the researcher to be confident of identifying a statistically significant effect. For the very largest experiments, even a very small effect would, on average, achieve a power level of .90.

The experiments overall do not support the notion that criminal justice experiments are designed for failure, at least in terms of the number of cases studied by investigators. On average, experiments we examine allow a very high likelihood of detecting a

---

[17] Cohen's estimates are commonly used, but like other conventions are fairly arbitrary. As he notes in his widely cited text on statistical power: "Although arbitrary, the proposed conventions will be found to be reasonable by reasonable people. An effort was made in selecting these operational criteria to use levels of ES [effect size] which accord with a subjective average of effect sizes such as are encountered in behavioral science" (1988:13).

moderate effect and are almost certain to detect a large effect (see Table 2). While the power level achieved for a small effect is less than .40, here criminal justice experiments in sanctions are not very much different from research in other social sciences. When we compare experiments in sanctions with other reviews of statistical power in other disciplines, we find that criminal justice experiments are, on average, using these standardized criteria, fairly powerful (see Table 2). In most areas where power has been assessed, studies have not been designed for detection of small effects, and in this regard criminal justice experiments in sanctions are more powerful than research in areas such as social work, applied and abnormal psychology, education and speech pathology.

When experimenters are unable to ensure the integrity of the randomization process, the power of experimental research is also affected. Breakdowns in randomization bring into question the computed significance levels reported by investigators. Such levels are dependent on certain assumptions, fair randomization being one of them. While slight violations of this assumption, like others, is unlikely to seriously bias study results, in a number of cases randomization breakdowns were serious. For example, in the Denver Drunk Driving Sentencing Experiment (Ross and Blumenthal 1974; 1975) judges circumvented the randomization process in more than half the cases, mostly in response to defense attorney pleas to have their clients receive fines rather than the probation conditions. In 16 of the studies reviewed by Weisburd, Sherman, and Petrosino, randomization failures were reported by investigators.[18]

"Randomization overrides" present similar problems, though the fact that they are planned allows investigators to more carefully measure their influence upon experimental results. For example, in the Minneapolis Domestic Violence Experiment (Sherman and Berk 1984; 1984a; 1985; Berk and Sherman 1985; 1988), overrides were allowed if the offender attempted to assault the police, the victim demanded an arrest, a restraining order was violated, or offenders would not leave the premises when ordered to do so by the police. Though such overrides occurred in 18% of the cases, the investigators had carefully documented overrides and were able to analyze their occurrence and their influence on the experimental results. In 11 of the experiments, researchers reported allowing practitioners to override the randomization process.[19]

Treatment breakdowns have a clear effect on the statistical outcomes of experiments. When the investigator cannot ensure that a "treatment" has been administered, or that it has been

---

[18] We suspect that such failures are underreported in published studies.

[19] A somewhat similar problem is evidenced in 8 studies where offenders were allowed to opt out of the less punitive sanction condition. For example, in the Ellsworth House study (Lamb and Goertzel 1974; 1974a) offenders could choose to remain in prison rather than be assigned to a half way house. (As noted earlier, Weisburd et al. did not include studies that allowed subjects to refuse the more punitive "sanction" condition.) Breakdowns in assignment, like those reported in the Ellsworth House Experiment, were generally small, for the obvious reason that offenders were likely to want to take advantage of the more lenient randomized condition.

administered in the dosage planned, the statistical power of a study is reduced. This happens because the "effect" of an experiment is directly related to the intensity of treatments administered. For example, in the California Special Intensive Parole Experiment (Reimer and Warren 1957) parole officers in the control group increased their contacts with parolees (and thus simulated the treatments found in the experimental condition). Reimer and Warren offer this as one potential explanation for the small and insignificant differences between the treatment and control groups studied. Similarly, in the California Parole Research Project Experiment (Johnson 1962; 1962a), control subjects often received more contact with their officers than did those in the experimental group, a factor which Johnson argues led to a finding of no difference between the experimental and control groups. In this case a non-experimental reanalysis of the study showed that when supervision was classified by actual intensity (rather than experimental allocation) a strong relationship existed between parole success and increased contact (Johnson 1962a).

IV. Statistical Power and Sample Size: A Reevaluation of Common Assumptions

As we noted earlier, sample size is generally viewed as the most straightforward method for affecting the power of experiments.[20] All else being equal, the power of a study grows with each increment in the number of cases included. This fact was

[20] Though Lipsey argues that affecting change in "effect size" is more cost effective (1990:169).

illustrated well in Table 1, where we estimated the expected power of the Weisburd et al. experiments under assumptions of small, moderate and large "effects." If we assumed a small effect, the average power of the studies grew from .12 for those with fifty or fewer standardized cases (per group) to over .90 for those with over four hundred cases. While the expected design benefits of larger samples decrease as assumed effect size grows (see Lipsey 1990), we still found an average difference of .50 in estimated power for the largest and smallest experiments under assumptions of moderate effects and a .17 difference if large effects were assumed.

These results help explain why researchers concerned with statistical power try to gather as many cases for inclusion in their samples as possible. In criminological research, which often tackles very serious public policy problems that are very difficult to effect, the benefits of larger samples are particularly attractive. For example, it might not be expected that a particular prison program would have a very large influence on subsequent violence by offenders. Nonetheless, even if a relatively small group were deterred from committing future murders or rapes the benefits for the community would be great. It is precisely in such studies, where researchers seek to design a test sensitive to even relatively small changes, that sample size has its largest influence on statistical power.

But the benefits associated with larger samples are based on the assumption that there is little relationship between the number of cases in a study and the effect of treatments on the subjects examined.[21] If for example, we assumed that the effect of a study declined the larger it became, the gain in statistical power associated with larger samples would be offset by the smaller effect coefficients found in such studies. This fact can be illustrated by turning to measures of statistical significance and their relationship to the standardized effect coefficients and sample size estimates used in statistical power.

Generally, significance tests in experimental research are derived by taking the ratio of the size of the differences between an experimental and control group to the estimated sampling error for the study (see Equation 1):

$$\text{Equation 1:} \quad \frac{\text{Size of Difference}}{\text{Standard Error}}$$

The size of the difference between the two samples is simply the magnitude of the difference in the dependent measures employed (usually means or percentages). The standard error or sampling error associated with these estimates, the denominator of the equation, is a ratio of the variability of the sample estimates to the size of the study. Taking a commonly used test, the t test, we

[21] It also assumes that there is little relationship between sample size and significance criteria and the type of statistical tests employed, other components of statistical power we discussed earlier. But in the case of these characteristics of power, an assumption of no relationship is not problematic. The size of a study does not alter the substance of significance criteria, nor does it influence the basic characteristics of a statistical test--except to the extent that it has an impact upon the choices made by researchers themselves.

can see why larger studies are more powerful (see Equation 2). As the number of cases grows, the standard error (the denominator of the t test) will get smaller. This increases the overall size of the t statistic and thus the likelihood of rejection of the null hypothesis:

Equation 2: $$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

The t test also illustrates why tests with larger "effects" are more powerful. In power analysis, effect size is generally computed by taking the ratio of the difference between sample estimates to the variability of those estimates (see Cohen 1988; Lipsey 1990). In Equation 2 effect size would be expressed as "d" (Cohen 1986, p. 20), the ratio of the difference of means to the pooled within group standard deviation of those means (i.e. $\bar{X}_1 - \bar{X}_2/s$). As d grows in size--either through a growth in the absolute difference in the means ($\bar{X}_1 - \bar{X}_2$) or a decline in the amount of variability of the estimates (s)--the t statistic also gets larger and again rejection of the null hypothesis is more likely. Returning to our earlier concern, if "d" were to get smaller as the number of cases in a study increased, then the benefits of a larger sample might be offset. Given the strong reliance on sample size as a means of increasing statistical power, we set out to examine this relationship directly.

Our primary empirical problem was to develop estimates of effect size for each of the experiments examined. We were aided in this process by the fact that many of the experiments included only one outcome measure, usually assessed at only one time period. Nonetheless, about six in ten of the studies reviewed included either multiple outcome measures or multiple follow-up periods or both. Our problem was to decide which of these estimates, or which combination of them, to use for identifying the "observed" effect size for each particular study.

One solution used by others who have reviewed effect size across studies (e.g. Levernson 1980; Chase & Chase 1976) is to take the mean of all of the outcome measures included by investigators. This solution has the benefit of not focusing upon a "deviant" effect in a study. Because we wanted to get some degree of consistency across the experiments studied, we operationalized this "average effect size" (AES) measure by taking the mean of all the effect coefficients at the follow-up period closest to one year.[22] While AES provides one overall view of the influence of the studies on their subjects, it does not take into account the fact that investigators often thought of their studies as tests of a series of research hypotheses, often linked to different outcome measures. In order to get some sensitivity to this problem we developed an additional measure--maximum effect size (MES)--that provides an

---

[22] To calculate "average effect size" we took the follow-up period closest to 12 months for each experiment (thirty-six of the studies used a one-year follow-up period; most of the others had a follow-up period somewhere between 6 and 18 months). For experiments with more than one outcome measure, we took the mean of the effect size for each outcome measure. In experiments where the subjects were divided into more than two groups, the effect size was calculated by taking the difference between each of the groups and then calculating the mean of those differences.

upper range of effect for the experiments.[23] MES identifies the largest effect for the 12-month (or closest) follow-up period.[24]

Assessing effect size from these measures, our first conclusion relates not to the relationship between sample size and effect size but to the magnitude of the effects found in criminal justice experiments concerned with sanctions (see Table 3). Of the 74 studies in which effect size estimates could be computed,[25] less than four in ten have standardized effects above .20 using our average effect size measure. Using the less conservative MES estimates, still only half of the studies have effects of this magnitude. This means that most of the studies did not even achieve what is generally defined as the threshold for a small effect (see Cohen 1988). Only 1 experiment evidences what Cohen describes as a large effect (a standardized effect coefficient above .80) using either measure of effect size.

Following these results, we might conclude that adjustments in sample size are likely to have a large yield in criminal justice studies. As we noted earlier, it is precisely in the case where the investigator desires to detect small effects that the influence

---

[23] To calculate "maximum effect size" we took the follow-up period closest to 12 months for each experiment. For experiments with more than one outcome measure, we took the measure with the largest effect size. If an experiment had only one outcome measure, the effect size for that measure was used.

[24] We also developed another measure which examines the largest standardized effect size for any outcome measure at any follow-up period. The results using this measure were similar to those reported in Tables 3 and 4.

[25] In two cases there was insufficient information provided by investigators to develop effect size coefficients.

of the number of cases on statistical power is greatest. But this conclusion does not seem to hold when we turn to the relationship between sample size and effect size. If we look at the mean standardized N for studies that fall in each of the effect size categories reviewed in table 3,[26] we can see that there is a generally inverse relationship between sample size and effect size. Indeed, the mean standardized N for the studies with the largest effects (.61-1.00) is between one quarter and one fifth of that for the studies with the smallest effects (0-.20) whether we use the MES or AES coefficients. Only in the case of the comparison between studies with effects of .41-.60 and .61-.80 does the number of standardized cases increase, and here the change is relatively small.

What this means substantively, is that estimates of statistical power arrived at by manipulating sample size, while assuming a constant effect size, are misleading. While there is clearly a gain to be had from increasing the size of a sample, the negative relationship between sample size and effect size offsets, at least in part, the design benefits of increasing the number of cases studied. How much of a loss is illustrated in Table 4. Here we calculate the statistical power of the experiments based on the sample average effect size measure.[27]

---

[26] In experiments with more than one outcome measure the standardized N was calculated by taking the mean of the standardized Ns for each outcome measure.

[27] While statistical power relates to the population characteristics of a study, we use these measures as a "best guess" of the true parameters under the assumptions made by investigators.

Quite surprisingly, given the general trend of using sample size as a method to increase statistical power, we do not find that larger studies have a power advantage. Indeed, the largest studies (those with more than 400 standardized cases per group) are less powerful, under these assumptions, than the smallest ones. They are also less powerful than studies with between 201 and 400 standardized cases, and cnly marginally more powerful than those with 51-200 standardized cases per group. There is a slight increase in power between the second (51-100) and third (101-200) sample size categories, and an increase in average power of .12 between the third and fourth (201-400) sample size groupings (though the latter estimate is based on only 5 cases in the larger group).

In the face of a result so at odds with conventional assumptions about statistical power, we were concerned that specific characteristics of our sample, rather than a more general process inherent to experimentation in criminal justice sanctions, might be responsible for our results. If, for example, it was the case that a particular type of experimental research was more likely to include fewer subjects and such experiments were also more effective, this would explain in part our basic finding. Our efforts to examine this problem were hampered by the fact that the experiments varied so greatly. But we were able to look at the basic relationship between sample size and a series of specific characteristics that cut broadly across the studies. In the case of type of outcome measure (e.g. % arrested, % violating parole), type of investigator (practitioner versus university researcher), type of sanction (e.g. parole or probation), and gender of subjects, we found little evidence that would lead us to challenge our conclusion that larger studies, regardless of their type, yielded generally smaller effects. However, we did find that the smallest studies were more likely to include only juvenile offenders, or to involve treatments added onto conventional sanctions.[28]

Those experiments that were conducted primarily with juveniles are much more likely than others to fall in the smallest sample-size categories (see Table 5). Indeed, the larger the study the less likely it is to involve primarily juvenile offenders. Nine of twelve of the studies including less than 50 standardized cases per group were "juvenile" studies. This was true for only three of the eleven studies in the largest sample-size grouping. Nonetheless, when we examine the relationship between sample size and effect size within the experiments including only juveniles, our results are generally consistent with the earlier findings (see Table 5).

Experiments that involve treatments added onto conventional sanctions (e.g. coercive group counseling programs in a prison), accounted for less than one in three of the studies reviewed by

---

[28] We also found that studies with a six-month follow-up period or shorter were more likely to include fewer cases. We do not include this question in our discussion because the number of studies involved here is small (only 11 overall, with 5 in the 15-50 sample size category) and makes a substantive analysis suspect. However, when we do examine the AES estimates across sample size categories for these eleven experiments, we find a similar pattern to that evidenced in our overall analysis.

Weisburd et al. But they make up half of the experiments in the smallest sample grouping and none in the two largest categories. Accordingly, it might be argued that our basic finding reflects the relationship between sample size and experiment type. While it is the case that treatment experiments overall have larger effects than other experiments we reviewed, the relationship of sample size and effect size for treatment experiments follows the general pattern of our results (see Table 6). There is a very large drop in effect size between the smallest studies and those with 51-200 standardized cases.

V. Why Larger Studies are Not More Powerful

The simple assumption that statistical power can be increased merely by adding cases to a study is not supported by our data. The largest studies are not necessarily more powerful than smaller ones, indeed using sample estimates as a guide, the very largest investigations are no more likely to lead to rejection of the null hypothesis than are the very smallest. This fact challenges common wisdom in experimental research (e.g. see Kolata 1990). Nonetheless, we believe this result is consistent with the experiences of those who have approached the very difficult task of designing and implementing randomized experiments in the real world of criminal justice.

It is generally easier to keep track of 100 or 200 subjects than 800 or a thousand. Similarly, 3 or 4 administrators are easier to monitor than 25 or 50. As the scale of experimental research grows so do the difficulties of implementation and management. But, even when criminal justice researchers set out with an awareness of the potential problems that large field studies entail, they are often surprised at the special difficulties they encounter. For example, Joan Petersilia, in describing her experience as an evaluator of a large Bureau of Justice Assistance probation study, provides a good example of how even experienced researchers are likely to underestimate the complexities of large-scale experimental research:

> The author anticipated that monitoring a field experiment of these dimensions would require tremendous effort. However, the extra burdens imposed by high turnover and loss of motivation among the projects' staff and administrators was not anticipated. Nor did we realize how difficult it would be to get adequate data from the sites, which were responsible for collecting and forwarding the data to RAND. (Petersilia 1989, p. 452).

The fact that larger studies are more difficult to monitor and control than are smaller ones has two important implications for the statistical power of experimental research in sanctions. First, the management and monitoring problems associated with larger studies often lead to treatments being administered less effectively or less consistently. Second, the need to gather large numbers of cases for study often leads to a great deal of heterogeneity in the nature of the samples studied. Because these characteristics of larger studies influence the magnitude of differences between groups in an experiment (the numerator of the effect size coefficient) and the variability of those differences (the denominator of effect size), they also influence the statistical power of experimental studies.

Problems in administering treatments in larger studies are illustrated in a number of the experiments we examined. In some cases, treatment failures result from the difficulty of keeping track of a very large number of subjects. For example in the California Reduced Prison Experiment (Berecochea and Jaman 1981; 1973), which included more than a 1000 inmates, the experimental subjects who were supposed to serve longer prison terms, sometimes served less prison time than the control group. But, the difficulties in managing large numbers of criminal justice practitioners also led to treatments not being administered in the dosages proposed by experimenters. In the Vera Institute Pretrial Adult Felony Offender Diversion Experiment (Baker and Sadd 1979; 1981; Baker and Rodrigues 1979), for example, some forty percent of the diversion group (N=410) never received the experimental condition.[29]

[29] We believe that treatment failures are more likely to occur in larger studies, and when they do occur, are likely to be more serious. Nonetheless, using evidence of any treatment breakdowns as described by investigators, we do not find a clear linear relationship between sample size and treatment failure. There is comparatively little difference between the smallest studies and the sample groupings ranging up to 400 standardized cases per group. Among studies that fall into these categories treatment failures noted by investigators average between 15 and 20 percent. The largest studies, have a somewhat higher rate of failure, about a third, though the absolute difference here in the number of cases that have treatment failures (as contrasted with the smallest studies) is not large. It is important to note that these results reflect not only actual difficulties in administering treatment, but also the attention paid by investigators to reporting such failures. It is likely that the greater attention given to detail in the smaller studies, also led to more careful identification and reporting of problems encountered.

These cases illustrate how an inability to ensure the implementation of treatments can have an impact upon the outcomes of experiments by minimizing the differences between the experiences of treatment and control group members. But breakdowns in treatment integrity may also affect the variability of outcome measures. In the Memphis Juvenile Diversion Experiment, for example, the principal investigators note that the 785 youths in the experimental group received somewhere between 11% and 140% of their projected treatments (Whitaker and Severy 1984). Because different offenders received different treatments, we would expect that the overall effects of the study would vary tremendously from subject to subject. While heterogeneity in the administration of treatment is common in both large and small experiments, our readings of the cases suggest that such variability is likely to be much greater in larger studies.

Variability is also increased by the heterogeneity of subjects studied in larger experiments. In planning such investigations it is often necessary to establish very broad eligibility requirements in order to gain the number of cases that investigators desire. For example in the California Special Intensive Parole Experiment (Reimer and Warren 1957) described earlier, some eighty percent of the prison population qualified for inclusion in the study.[30] Many times investigators in larger studies are forced to relax eligibility requirements even further once the project is ongoing.

---

[30] Interestingly, even though the eligibility requirements were so broad, in the second Phase of this project it was necessary to include subjects that were not eligible in the first phase.

In the English Intensive Probation Experiments (Folkard et al. 1974; Folkard, Smith and Smith 1976) for example, the original design, which called for high-risk male probationers, was changed when researchers saw that they were unable to fulfill project quotas. In the Rand study described by Petersilia, overestimation of the number of eligible offenders also led the sites involved to relax eligibility requirements in the midst of the experiment. Indeed, Petersilia argues that it eventually became "unclear who was participating" (1989, p.450).

The effects of this heterogeneity in the subjects examined upon the statistical power of experiments is illustrated in two studies that analyzed sub-groups of offenders separately after the original project design had failed to yield significant results. The Police Foundation Shoplifting Arrest Experiment (Williams, Forst and Hamilton 1987) examined shoplifters six years of age and older. Looking at the entire sample no significant deterrent effects of arrest were noted. But when subjects were categorized into those under 16 years, and those 17 years of age and older, significant results were found for the juvenile group. In the Memphis Juvenile Diversion Experiment (Severy and Whitaker 1982; 1984; 1984a; Whitaker, Severy, and Morton 1984), investigators also found no significant differences when the entire sample was examined. But within the experimental group those youths needing social adjustment or education assistance were more likely to have a successful experience when compared to those needing family or individual counselling.

Though in both these cases the statistical design of the experiments was violated by a post-facto division of the experimental and treatment groups, they follow a developing consensus among criminologists (see Farrington Ohlin and Wilson 1986) that different types of offenders will respond very differently to different types of sanctions. Where an experiment includes a heterogeneous population, effects of sanctions on one sub-group of offenders may be hidden by a different effect on another, as appears to be the case, for example, in the Police Foundation Shoplifting Arrest Experiment. Where there is less systematic variation in the study, but still great diversity in the types of subjects included, the variability of the estimates gained will grow, again leading to a smaller effect coefficient and thus a less powerful study.

Our observations on the relationship between sample size and problems of implementing and monitoring experimental research are based on a relatively small group of studies. Nonetheless, they are consistent with findings that develop out of a very large review of correctional treatment programs conducted by Lipton, Martinson, and Wilks (1975). Though Lipton et al. did not look specifically at the relationship between sample size and the quality of the 231 studies they examined, they did rate the studies in terms of the strength of the overall research design and the success of investigators in carrying out the studies.[31] In a

---

[31] Studies were selected for inclusion in the survey by Lipton et al. on the basis of the following criteria: the study must represent an evaluation of a treatment method applied to criminal

reanalysis of these data, conducted by Palmer (1978), he found "a strong inverse relationship between both quality and strength [of the studies], on the one hand, and sample size, on the other" (1978, p. 160). Among the better designed and implemented studies ("A" studies) the average sample size was 459. Among lower quality studies ("B" studies), the average sample size was 900. While Palmer relegated these findings to an Appendix, they suggest to us that our observations concerning the difficulties of developing and managing larger studies are not limited to our sample of criminal justice experiments in sanctions.

VI. Implications and Conclusions

In examining the statistical power of experiments in sanctions we are lead to an ironic conclusion about the relationship between experimental design and study outcomes. Had more attention been paid to statistical power in developing sanctioning studies in criminal justice, the power of the studies themselves would

---

offenders; it must have been completed after January 1, 1945; it must include empirical data resulting from a comparison of a treatment group and control group(s); these data must be measures of improvement in performance on some relevant dependent variables. Studies specifically excluded were after-only studies without comparison groups, prediction studies, studies that only describe and subjectively evaluate treatment programs, and clinical speculations about feasible treatment methods.

Following assessment by a professional researcher each study was reviewed by a committee and allocated to one of three categories: "A" studies, acceptable for the survey with no more than minimal research shortcomings; "B" studies, acceptable for the survey with research shortcomings that place reservations on interpretation of the findings; and "Other Studies". Under "Other Studies" were reports and articles excluded because two or more of a possible 11 conditions existed. See Lipton et al. (1975), pages 6-7, for a list of these conditions.

probably not have increased significantly. The naive assumption behind much power analysis, that sample size is unrelated to effect size, is not consistent with our findings. Investigators of larger studies are likely to encounter more serious problems in implementing treatments than smaller studies. They are often forced as well to draw more heterogeneous samples. Both these factors influence the outcome measures of experiments, and thus the power advantages of larger samples are often offset.

Our results would seem to suggest that smaller studies are to be preferred over larger investigations. Nonetheless, there are significant difficulties in generalizing from small and restricted samples, and the design advantages they seem to offer do not offset the power disadvantages inherent to studies of such a small size. Using sample estimates as a guide, we found that the very smallest studies examined were more powerful than the very largest. Yet, such studies did not offer even an equal chance of finding a statistically significant difference between treatment and control groups. Just as the small effects of large investigations offset the advantages of increasing the number of cases examined, the small samples in smaller investigations offset the advantages gained from larger effects. The task accordingly is not to focus on smaller studies, but rather on strategies that will allow researchers to increase sample size while maintaining the integrity of treatments and minimizing variability.

Petersilia (1989), provides one lesson in this regard from the RAND Intensive Supervision Demonstration Project. Experimenters cannot allow practitioners to control the implementation of important aspects of study design, even though this is often one way of conserving much-needed research funds. As is the case for many other large scale investigations, economic and practical constraints forced RAND to utilize practitioners to carry out many research tasks that would have been more directly controlled by researchers had the investigation been smaller. For RAND these decisions did not turn out to be cost effective in the long run. They created both greater variability in treatments and in the pool of offenders examined than had been proposed in the original project design (Petersillia and Turner 1991; Turner 1991). More generally, the RAND experience illustrates the importance of maintaining researcher control over each stage of an experiment's design and implementation.

One example of a method for monitoring the implementation of treatments when they are controlled by practitioners is provided by the Minneapolis Hot Spots Patrol Experiment (Sherman and Weisburd, 1990). Sherman and Weisburd randomly allocated increased police patrol to 55 of 110 high crime locations, called hot spots. While the number of cases in that study was relatively small, the number of practitioners involved was very large. Indeed, the entire patrol force in Minneapolis was used in increasing the patrol dosage in the experimental locations. In trying to avoid a problem encountered in the earlier Kansas City Preventive Patrol Experiment (Kelling et al., 1974), in which there was some doubt as to whether the treatments were successfully administered, Sherman and Weisburd conducted 7500 hours of random observations of the experimental and control sites. While the observations were intended primarily as a means of documenting dosage, they also were seen by investigators as a method for keeping practitioners "honest."

Variability in the larger studies we examined often developed from over-estimation of the number of cases that fit the original eligibility requirements of investigators. This problem has become widely recognized in recent experimental studies (see Petersilia 1989) and has led a number of investigators to conduct what have been termed "case flow" studies. In the National Institute of Justice's Domestic Violence Replication Program, for example, researchers in each of the five sites involved in the program conducted a careful analysis before the study began of the potential universe of cases available for randomization (Uchida 1991). This process allowed investigators to avoid midstream changes in eligibility requirements. More generally, case flow studies provide an effective method for preventing the "watering down" of the experimental pool in order to achieve quotas set in the original research design.

Even when following the original project design, investigators often include a great deal of diversity in the types of subjects examined. As we observed earlier, larger studies are likely to be more variable than smaller ones, a fact that explains in part the reduction in effect size that is found in the largest investigations. Statisticians offer one solution to this problem--randomization within blocks (e.g. see Lipsey 1990)--that has generally been ignored by criminal justice researchers. Block designs, which randomly allocate subjects within groups, minimize the effects of variability in an experiment by making sure that like subjects will be compared one to another. A commonly used method randomly allocates subjects within pairs, for example by random allocation of twins in psychological studies. While randomization of matched pairs is unlikely to be practical in criminological field experiments, blocking within larger groups does provide an effective method for minimizing the effects of variability in a study. Sherman and Weisburd (1989), for example, randomly allocated police patrol within five independent blocks based on prior crime activity. While blocking demands more complex statistical analyses than traditional experimental designs, it provides a relatively inexpensive method for dealing with the diversity of subjects found in most large studies.

These examples provide some evidence of recent attempts to manage the design difficulties that are likely to be encountered in large experiments. But such efforts have not been joined systematically, nor linked directly to the issues we have raised. The nature of the problems criminal justice researchers examine demand that they design for relatively large studies. Our findings suggest that there will be few gains in increasing sample size until the design difficulties that larger samples pose are directly addressed. For the future, this demands much greater attention to problems of method and design in experimentation than has been given by investigators to date. For the present, it suggests that commonly used approximations of statistical power which do not take into account the relationship between sample size and effect size, provide a very misleading view of the design advantages of larger studies.

References

Adams, S. 1965. "An Experimental Assessment of Group Counseling with Juvenile Probationers." Journal of the California Probation, Parole and Correctional Association 2: 19-25.

Adams, S. 1970. "The Pico Project." In The Sociology of Punishment and Correction, edited by N.B. Johnston, L. Savitz and M.E. Wolfgang. New York: John Wiley and Sons.

Annis, H.M. 1979. "Group Treatment of Incarcerated Offenders with Alcohol and Drug Problems: A Controlled Evaluation." Canadian Journal of Criminology 21: 3-15.

Ares, C.E., A. Rankin and H. Sturz. 1963. "The Manhattan Bail Project: An Interim Report on the Use of Pre-Trial Parole." New York University Law Review 38: 67-93.

Austin, J.F. 1980. Instead of Justice: Diversion. Doctoral dissertation, University of California. Ann Arbor: University Microfilms International.

Baker, S.H. and O. Rodrigues. 1979. "Random Time Quota Selection: An Alternative to Random Selection in Experimental Evaluation." In Evaluation Studies Review Annual. Volume 4, edited by L. Sechrest. Beverley Hills, CA: Sage.

Baker, S.H. and S. Sadd. 1979. Court Employment Project: Evaluation. Final Report. New York: Vera Institute of Justice.

Baker, S.H. and S. Sadd. 1981. Diversion of Felony Arrests; An Experiment in Pretrial Intervention: Evaluation of the Court Employment Project. Summary Report. Washington, D.C.: National Institute of Justice.

Barkwell, L.J. 1976. "Differential Treatment of Juveniles on Probation: An Evaluative Study." Canadian Journal of Criminology and Corrections 18: 363-378.

Baron, R., F. Feeney and W.E. Thornton. 1972. Preventing Delinquency through Diversion. The Sacramento County Probation Department 601 Diversion Project. A First Year Report. Sacramento, CA: Sacramento County Probation Department.

Baron, R., F. Feeney and W. Thornton. 1973. "Preventing Delinquency through Diversion." Federal Prison 37: 13-18.

Baron, R. and F. Feeney. 1976. Juvenile Diversion through Family Counseling: A Program for the Diversion of Status Offenders in Sacramento County, California. Washington, D.C.: National Institute of Law Enforcement and Criminal Justice.

Berecochea, J.E. and D.R. Jaman, W.A. Jones. 1973. Time Served in Prison and Parole Outcome: An Experimental Study. Report Number 1. Sacramento: California Department of Corrections Research Division.

Berecochea, J.E. and D.R. Jaman. 1981. Time Served in Prison and Parole Outcome: An Experimental Study. Report Number 2. Sacramento: California Department of Corrections Research Division.

Berg, I., M. Consterdine, R. Hullin, R. McGuire and S. Tryer. 1978. "The Effect of Two Randomly Allocated Court Procedures on Truancy." British Journal of Criminology 18: 232-244.

Berg, I., R. Hullin, R. McGuire, and E.S. Tryer. 1979. "Truancy and the Courts: Research Note." Journal of Child Psychiatry and Psychology 18: 359-365.

Berg, I., R. Hullin and R. McGuire. 1979. "A Randomly Controlled Trial of Two Court Procedures in Truancy." In Psychology, Law and Legal Processes, edited by D.P. Farrington, K. Hawkins and S.M. Lloyd-Bostock. NJ: Humanities Press.

Berk, R.A., K.J., Lenihan and P.H. Rossi. 1980. "Crime and Poverty: Some Experimental Evidence from Ex-Offenders." American Sociological Review 45: 766-786.

Berk, R.A., R.F. Boruch, D.L. Chambers, P.H. Rossi and A.D. Witte. 1985. "Social Policy Experimentation. A Position Paper." Evaluation Review 9: 387-430.

Berk, R.A. and L.W. Sherman. 1985. "Data Collection Strategies in the Minneapolis Domestic Assault Experiment." In Collecting Evaluation Data: Problems and Solutions, edited by L. Burnstein, H.E. Freeman and P.H. Rossi. Beverly Hills, CA: Sage.

Berk, R.A. and L.W. Sherman. 1988. "Police Responses to Family Violence Incidents: An Analysis of an Experimental Design with Incomplete Randomization." Journal of the American Statistical Association 83: 70-76.

Berman, J.J. 1975. "The Volunteer in Parole Program." Criminology 13: 111-113.

Berman, J.J. 1978. "An Experiment in Parole Supervision." Evaluation Quarterly 2: 71-90.

Bernsten, K. and K.O. Christiansen. 1965. "A Resocialization Experiment with Short-Term Offenders. In Scandanavian Studies in Criminology, Volume I, edited by K.O. Christiansen. London: Tavistock.

Boruch, R. 1976. "On Common Contentions About Randomized Field Experiments." Evaluations Studies Review Annual. Volume I. Edited by Jean V. Glass. Beverly Hills: Sage.

Brewer, J.K. and Owen, P.W. 1973. "A Note on the Power of Statistical Tests." The Journal of Educational Measurement 10: 71-74.

Burkhart, W. 1969. "The Parole Work Unit Programme: An Evaluation." British Journal of Criminology 9: 125-147.

Brody, S. 1978. "Research into the Aims and Effectiveness of Sentencing." Howard Journal of Penology and Crime Prevention 17: 133-148. Edinburgh, Scotland.

Brown, S.E. 1989. "Statistical Power and Criminal Justice Research." Journal of Criminal Justice 17: 115-122.

Byles, J.A. and A. Maurice. 1979. "The Juvenile Services Project: An Experiment in Delinquency Control." Canadian Journal of Criminology 21: 155-165.

Campbell, D.T. and J.C. Stanley. 1966. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Chase, L.J. and R.B. Chase. 1976. "A Statistical Power Analysis of Applied Psychological Research." Journal of Applied Psychology 61: 234-237.

Clarke, R.V.G. and D.B. Cornish. 1972. The Controlled Trial in Institutional Research-Paradigm or Pitfall for Penal Evaluators? London: Her Majesty's Stationery Office.

Cohen, J. 1962. "The Statistical Power of Abnormal-Social Psychological Research: A Review." Journal of Abnormal Social Psychology 65: 145-153.

Cohen, J. 1973. "Statistical Power Analysis and Research Results." American Educational Research Journal 10: 225-230.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Coleman, D. "Charge Dropped on Bogus Work." The New York Times, 4, April, 1989.

Cornish, D.B. and R.V.G. Clarke. 1975. Residential Treatment and It's Effects on Delinquency. London: Her Majesty's Stationery Office.

Cornish, D.B. 1987. "Evaluating Residential Treatments for Delinquents: A Cautionary Tale." In Social Intervention: Potential and Constraints, edited by K. Hurrelmann, F. Kaufmann and F. Losel. Berlin: Gruyter.

Craft, M., G. Stephenson and C. Granger. 1964. "A Controlled Trial of Authoritarian and Self-Governing Regimes With Adolescent Psychopaths." American Journal of Orthopsychiatry 34: 543-554.

Dennis, M.D. 1988. Implementing Randomized Field Experiments: An Analysis of Criminal and Civil Justice Research. Doctoral dissertation, Northwestern University.

Ditman, K.S., G.G. Crawford, E.W. Forgy, H. Moskowitz and C. Macandrew. 1967. "A Controlled Experiment on the Use of Court Probation for Drunk Arrests." American Journal of Orthopsychiatry 124: 160-163.

Dunford, F.W., D.W. Osgood and H.F. Weichselbaum. 1982. National Evaluation of Diversion Projects. Washington, D.C.: National Institute of Juvenile Justice and Delinquency Prevention.

Empey, L.T. and J. Rabow. 1961. "The Provo Experiment in Delinquency Rehabilitation." American Sociological Review 26: 679-696.

Empey, L.T. and S.G. Lubeck. 1971. The Silverlake Experiment. Chicago: Aldine.

Empey, L.T. and M.L. Erickson. 1972. The Provo Experiment. Lexington, MA: D.C. Heath and Co.

Fagan, J.A. 1990. "Natural Experiments in Criminal Justice." In Measurement Issues in Criminology, edited by Kimberley L. Kempf. New York: Springer-Verlag.

Farrington, D.P. 1983. "Randomized Experiments on Crime and Justice." In Crime and Justice: An Annual Review of Research, edited by M. Tonry and N. Morris. Chicago: Universty of Chicago Press.

Farrington, D.P., L.E. Ohlin and J.Q. Wilson. 1986. Understanding and Controlling Crime. New York: Springer-Verlag.

Finckenauer, J.O. 1982. Scared Straight. Englewood Cliffs, NJ: Prentice Hall.

Folkard, M.S., A.J. Fowles, B.C. McWilliams, D.D. Smith, D.E. Smith and G.R. Walmsley. 1974. IMPACT: Intensive Matched Probation and After-Care Treatment. Volume I. The Design of the Probation Experiment and an Interim Evaluation. London: Her Majesty's Stationery Office.

Folkard, M.S., D.E. Smith and D.D. Smith. 1976. IMPACT: Intensive Matched Probation and After-Care Treatment. Volume II. The Results of the Experiment. London: Her Majesty's Stationery Office.

Fowles, A.J. 1978. Prison Welfare. London: Her Majesty's Stationery Office.

Garner, J. and C.A. Visher. 1988. "Experiments Help Shape New Policies." NIJ Reports. September/October, No. 211.

Gelber, R.D. and M. Zelen. 1985. "Planning and Reporting Clinical Trials." In Basic Principles and Clinical Management of Cancer, edited by P. Calabrese, P.S. Schein and S.A. Rosenberg. Medical Oncology. New York: MacMillan.

Glick, B., E. Hamilton and B. Forst. 1986. Shoplifting: An Experiment in Lesser Crimes and Punishments. Draft Final Report. Washington, D.C.: Police Foundation.

Greater Egypt Regional Planning and Devolpment Commission. 1979. Menard Correctional Center Juvenile Tours Impact Study. Carbondale, IL: Greater Egypt Regional Planning and Development Commission.

Guttman, E. 1963. Effects of Short-Term Psychiatric Treatment on Boys in Two California Youth Authority Institutions. Research Report No. 36. Sacramento, CA: California Department of Youth Authority.

Holden, R.T., L.T. Stewart, J.N. Rice and E. Manker. 1981. Tennessee DUI Probation Follow-Up Demonstration Project. Final Report. (Dept. of Transportation Contract No, DOT Hs-5-01199). Springfield, VA.

Holden, R.T., 1982. Legal Reactions to Drunk Driving. Doctoral Dissertation, Vanderbilt University, 1981. University Microfilms International 8200766.

Holden, R.T. 1983. "Rehabilitative Sanctions for Drunk Driving: An Experimental Evaluation." Journal of Research and Crime and Delinquency 20: 55-72.

Hudson, C.H. 1973. An Experimental Study of the Differential Effects of Parole Supervision for a Group of Adolescent Boys and Girls. Summary Report. Minneapolis: Minnesota Department of Corrections.

Hudson, J. and C.D. Hollister. 1976. "An Experimental Study of Parole Supervision of Juveniles and Social Service Utilization." Iowa Journal of Social Work 4: 80-89.

Jackson, P.C. 1978. The Bay Area Parole Study. Sacramento, CA: Department of the Youth Authority.

Jackson, P.C. 1983. "Some Effects of Parole Supervision on Recidivism." British Journal of Criminology 23: 17-34.

Jesness, C.F. 1965. The Fricot Ranch Study. Sacramento, CA: California Youth Study.

Jesness, C.F. 1971. "Comparative Effectiveness of Two Institutional Treatment Programs for Delinquents." Child Care Quarterly 1: 119-130.

Jesness, C.F. 1971. "The Preston Typology Study." Journal of Research in Crime and Delinquency 8: 38-52.

Jesness, C.F., W.J. DeRisi, P.M. McCormick and R.F. Wedge. 1972. The Youth Center Research Project. Sacramento, CA: California Youth Authority.

Jesness, C.F. 1975. "Comparative Effectiveness of Behavior Modification Transactional Analysis Programs for Delinquents." Journal of Consulting and Clinical Psychology 43: 758-779.

Jesness, C.F. No Date. "The Youth Center Project: Transactional Analysis and Behavior Modification Programs for Delinquents." Behavioral Disorders 1: 27-36.

Johnson, B.M. 1962. Parole Performance of the First Year's Releases: Parole Research Project: Evaluation of Reduced Caseloads. Research Report No. 27. Sacramento, CA: California Youth Authority.

Johnson, B.M. 1962a. An Analysis of Predictions of Parole Performance and of Judgments of Supervision in the Parole Research Project. Research Report No. 32. Sacramento: California Youth Authority.

Kassebaum, G., D. Ward and D. Wilner. 1971. Prison Treatment and Parole Survival. New York: Wiley.

Kelling, G.L., T. Pate, D. Dieckman and C.E. Brown. 1974. The Kansas City Patrol Experiment. A Technical Report. Washington, D.C.: Police Foundation.

Kirby, B.C. 1969. "Crofton House: An Experiment With a County Halfway House." Federal Probation 33: 53-58.

Klein, M.W. 1986. "Labeling Theory and Delinquency Policy: An Experimental Test." Criminal Justice and Behavior 13: 47-79.

Kolata, G. "In Clinical Trials, Some Contend, Big is Beautiful." The New York Times, 15, April, 1990.

Kraemer, H.C. and S. Thiemann. 1987. How Many Subjects? Statistical Power Analysis in Research. Newbury Park, CA: Sage.

Kroll, R.M. and Chase, L.J. 1975. "Community Disorders. A Power Analytic Assessment of Recent Research." Journal of Communication Disorders 8: 237-247.

Lamb, H.R. and V. Goertzel. 1974. "Ellsworth House: A Community Alternative to Jail." The American Journal of Psychiatry 131: 64-68.

Lamb, H.R. and V. Goertzel. 1974. "A Community Alternative to County Jail: The Hopes and the Realities." Federal Probation 38: 33-39.

Lee, R. and N.M. Hayes. 1978. "Counseling Juvenile Offenders: An Experimental Evaluation of Project Crest." Community Mental Health Journal 14: 267-271.

Lee, R. and N.M. Hayes. 1980. "Project Crest and the Dual-Treatment Approach to Delinquency: Methods and Research Summarized." In Effective Correctional Treatment, edited by R.R. Ross and P. Gendreau. Toronto: Butterworths.

Levernson, JR., R.L. 1980. "Statistical Power Analysis: Implications for Researchers, Planners and Practitioners in Gerontology." The Gerontologist 20: 494-498.

Lewis, R.V. 1979. The Squires of San Quentin. Preliminary Findings on an Experimental Study of Juvenile Visitation at San Quentin Prison. Sacramento, CA: Department of the Youth Authority. Division of Research.

Lewis, R.V. 1981. The Squires of San Quentin. An Evaluation of a Juvenile Awareness Program. Sacramento, CA: Department of the Youth Authority. Division of Research.

Lewis, R.V. 1983. "Scared Straight-California Style." Criminal Justice and Behavior 10: 209-226.

Lichtman, G.M. and S.M. Smock. 1981. "The Effects of Social Services on Probational Recidivism." Journal of Research in Crime and Delinquency 18: 81-100.

Lincoln, C.M., M.W. Klein, K.S. Teilmann and S. Labin. No Date. Control Organizations and Labeling Theory: Official Versus Self-Reported Delinquency. Unpublished manuscript, University of Southern California.

Lipsey, M.W. 1990. Design Sensitivity: Statistical Power for Experimental Research. Newbury, CA: Sage.

Lipton, D., R. Martinson and J. Wilks. 1975. The Effectiveness of Correctional Treatment. New York: Praeger.

Love, C.T., J.G. Allgood and F.P.S. Samples. 1986. "The Butner Research Projects." Federal Probation 50: 32-39.

National Institute of Justice. 1990. Fiscal Year 1990 Program Plan. Office of Justice Programs. U.S. Department of Justice: Washington, D.C.

Orme, J.G. and T.D. Combs-Orme. 1986. "Statistical Power and Type II Errors in Social Work." Social Research and Abstracts 22: 3-10.

Orme, J.G. and R.M. Tolman. 1986. "The Statistical Power of a Decade of Social Work Education Research." Social Service Review 60: 619-632.

Ostrom, T.M., C.M. Steele, L.K. Rosenblood and H.L. Mirels. 1971. "Modification of Delinquent Behavior." Journal of Applied Social Psychology 1: 118-136.

Owen, G. and P.W. Mattesich. 1987. Community Assistance Program: Results of a Control Study of the Effects of Non-Residential Corrections on Adult Offenders in Ramsey County. St. Paul: Wilder Foundation.

Palmer, T.B. 1971. "California's Community Treatment Program for Delinquent Adolescents." The Journal of Research in Crime and Delinquency 8: 74-92.

Palmer, T.B. 1974. "The Youth Authority's Community Treatment Project." Federal Probation 38: 3-14.

Palmer, T.B. 1978. Correctional Intervention and Research. Toronto: Lexington Books.

Persons, R.W. 1966. "Psychological and Behavioral Change in Delinquents Following Psychotherapy." Journal of Clinical Psychology 22: 337-340.

Persons, R.W. 1967. "Relationship Between Psychotherapy With Institutionalized Boys and Subsequent Community Adjustment." Journal of Consulting Psychology 31: 137-141.

Petersilia, J. 1989. "Implementing Randomized Experiments: Lessons from BJA's Intensive Supervising Project." Evaluation Review 13: 435-458.

Petersilia, J. and S. Turner. 1990. Intensive Supervision for High-Risk Probationers: Findings from Three California Experiments. Santa Monica, CA: Rand.

Powers, E. and H. Witner. 1951. An Experiment in the Prevention of Delinquency. The Cambridge-Somerville Youth Study. New York: Columbia University Press.

Quay, H.C. and C.T. Love. 1977. "The Effects of a Juvenile Diversion Program on Rearrests." Criminal Justice and Behavior 4: 377-396.

Reimer, E. and M. Warren. 1957. "Special Intensive Parole Unit: Relationship between Violation Rate and Initially Small Caseload." National Probation and Parole Association Journal 3: 222-229.

Reimer, E. and M. Warren. 1958. Special Intensive Parole Unit. Phase II. Thirty-Man Caseload Study. Sacramento, CA: California Department of Corrections.

Rose, G. and R.A. Hamilton. 1970. "Effects of a Juvenile Liaison Scheme." British Journal of Criminology 10: 2-20.

Ross, H.L. and M. Blumenthal. 1974. "Sanctions for the Drinking Driver: An Experimental Study." The Journal of Legal Studies 3: 53-61.

Ross, H.L. and M. Blumenthal. 1975. "Some Problems in Experimentation in a Legal Setting." The American Sociologist 10: 150-155.

Rossi, P.H., R.A. Berk and K.J. Lenihan. 1980. Money, Work and Crime. New York: Academic Press.

Sarason, I.G. and V.J. Ganzer. 1973. "Modeling and Group Discussion in the Rehabilitation of Juvenile Delinquents." Journal of Counseling Psychology 20: 442-449.

Sarason, I.G. 1978. "A Cognitive Social Learning Approach to Juvenile Delinquency." In Psychopathic Behavior: Approaches to Research, edited by R.D. Hare and D. Schalling. Chichester: Wiley.

Schneider, A.L. 1980. "Effects of a Status Offender Deinstitutionalization: A Case Study." In Evaluation and Criminal Justice Policy, edited by R. Roesch and R.R. Corrado. Beverly Hills, CA: Sage.

Schneider, A.L. 1986. "Restitution and Recidivism Rates of Juvenile Offenders: Results from Four Experimental Studies." Criminology 24: 533-552.

Schneider, P.R. and A.L. Schneider. 1983. An Analysis of Recidivism Rates in Six Federally-Funded Restitution Projects in Juvenile Courts. A Statistical Summary. Washington D.C.: National Institute of Justice.

Schuster, D.H. 1974. "The Effectiveness of Official Action Taken Against Problem Drivers: A Five-Year Follow-Up." Journal of Safety Research 6: 171-176.

Sechrest, L. and A. Rosenblatt. 1987. "Research Methods." In Handbook of Juvenile Delinquency, edited by H.C. Quay. New York: Wiley and Sons.

Seckel, J.P. 1965. Experiments in Group Counseling at Youth Authority Institutions. Sacramento, CA: Youth Authority Division of Research.

Seckel, J.P. 1967. The Fremont Experiment: Assessment of Residential Treatment at a Youth Authority Reception Center. Sacramento, CA: Youth Authority Division of Research.

Sedleimer, P. and G. Gigerenzer. 1989. "Do Studies of Statistical Power Have an Effect on the Power of Studies?" Psychological Bulletin 105: 309-316.

Severy, L.J. and J.M. Whitaker. 1982. "Juvenile Diversion: An Experimental Analysis of Effectiveness." Evaluation Review 6: 753-774.

Severy, L.J. and J.M. Whitaker. 1984. "Memphis-Metro Youth Diversion Project: Final Report." Child Welfare 63: 269-277.

Shaw, M. 1974. Social Work in Prison. London: Her Majesty's Stationery Office.

Sherman, L.W. and R.A. Berk. 1984. "The Deterrent Effects of Arrest for Domestic Assault." American Sociological Review 49: 261-272.

Sherman, L.W. and R.A. Berk. 1984. The Minneapolis Domestic Violence Experiment. Washington, D.C.: Police Foundation Reports.

Sherman, L.W. and R.A. Berk. 1985. "The Randomization of Arrest." In Randomization and Field Experimentation, New Directions for Program Evaluation, Number 28, edited by R.F. Boruch and W. Wothke. San Francisco: Jossey-Bass.

Sherman, L.W. and P.R. Gartin. 1986. "Differential Recidivism: A Field Experiment of the Specific Sanction Effects of Arrest for Shoplifting." Unpublished paper presented at the American Society of Criminology Conference, Atlanta, GA.

Sherman, L.W. and D. Weisburd. 1989. Policing the Hotspots of Crime: A Redesign of the Kansas City Preventive Patrol Experiment. Washington, D.C.: Crime Control Institute.

Shivratt, J.L. 1988. "Social Interactional Training and Incarcerated Juvenile Delinquents." Canadian Journal of Criminology 30: 145-163.

Star, D. 1978. Summary Parole: A Six and Twelve Month Follow-Up. Research Report No. 60. Sacramento, CA: California Department of Corrections.

Stark, H.G. 1963. "A Substitute for Institutionalization of Serious Delinquents. A California Youth Study Experiment." Crime and Delinquency 9: 242-248

Stratton, J.G. 1975. "Effects of Crisis Intervention Counseling on Predelinquent and Misdemeanor Juvenile Offenders." Juvenile Justice 26: 7-18.

Taylor, A.J.W. 1967. "An Evaluation of Group Psychotherapy in a Girl's Borstal." International Journal of Psychotherapy 17: 168-177.

Tornudd, P. 1968. "The Preventive Effect of Fines for Drunkeness: A Conrolled Experiment." Scandanavian Studies in Criminology 2: 109-124.

Turner, Susan. 1991. Personal communication with author, February.

Traux, C.B., D.G. Wargo and L.D. Silber. 1966. "Effects of Group Psychotherapy With High Accurate Empathy and Non-Possesive Warmth Upon Female Institutionalized Delinquents." Journal of Abnormal Psychology 71: 267-274.

Uchida, Craig. 1991. Personal communication with author, February.

Venezia, P.S. 1972. "Unofficial Probation: An Evaluation of its Effectiveness." Journal of Research in Crime and Delinquency 9: 149-170.

Waldo, G.P. and T.G. Chiricos. 1977. "Work Release and Recidivism: An Empirical Evaluation of a Social Policy." Evaluation Quarterly 1: 87-108.

Warren, M.Q. 1967. "The Community Treatment Project: History and Prospects." In Law Enforcement Science and Technology, edited by S.A. Yefsky. Washington, D.C.: Thompson Book Company.

Weisburd, D., L. Sherman and A.J. Petrosino. 1990. Registry of Randomized Criminal Justice Experiments in Sanction. Rutgers University and Crime Control Institute. Los Altos: Sociometrics Corporation, Data Resources Program of the National Institute of Justice.

Welsh, J.D. 1978. "Is Pretrial Performance Affected by Supervision?" In Pretrial Services Annual Journal, 1978. edited by D.A. Henry. Washington, D.C.: Pretrial Services Resource Center.

Whitaker, J.M. and L.J. Severy. 1984. "Service Accountability and Recidivism for Diverted Youth." Criminal Justice and Behavior 11: 47-73.

Whitaker, J.M., L.J. Severy and D.S. Morton. 1984. "A Comprehensive Community-Based Youth Diversion Program. Child Welfare 63: 175-181.

Williams, H., B. Forst and E.E. Hamilton. 1987. "Stop! Should You Arrest that Person?" Security Management 31: 52-58.

Williams, M. 1970. A Study of Some Aspects of Borstal Allocation. London: Home Office Prison Department. Office of the Chief Psychologist.

Williams, M. 1975. "Aspects of the Psychology of Imprisonment." In The Use of Imprisonment: Essays in the Changing State of English Penal Policy, edited by S. McConville. London: Routledge.

Woolley, T.W. 1983. "A Comprehensive Power-Analytic Investigation of Research in Medical Education." Journal of Medical Education 58: 710-715.

Yarbough, J.C. 1979. Evaluation of JOLT as a Deterrence Program. Lansing, MI: Michigan Department of Corrections.

TABLE 1.

Statistical Power Under Assumptions of Small, Moderate, and Large Effect Size

		Assumed Effect Size		
Sample Size	(N)	Small	Moderate	Large
15 - 50	(12)	.12	.49	.82
51 - 100	(25)	.26	.87	.99
101 - 200	(21)	.37	.98	.99
201 - 400	(5)	.60	.99	.99
Over 400	(11)	.91	.99	.99

TABLE 2.

Statistical Power in Various Fields (under assumptions of small, moderate, and large effect).

	Effect Size		
Field	Small	Moderate	Large
Criminal Justice Experiments in Sanctions			
Weisburd et al. (1990)	.39	.86	.96
Gerontology			
Levernson (1980)	.37	.88	.96
Social Work			
Orme & Combs-Orme (1986)	.35	.76	.91
Applied Psychology Research			
Chase & Chase (1976)	.25	.67	.86
Abnormal and Social Psychology			
Sedlmeier & Gigerenzer (1989)	.21	.50	.84
Education, General			
Brewer & Owen (1973)	.28	.79	.91
Speech Pathology			
Kroll & Chase (1975)	.16	.44	.73

TABLE 3.

Effect Size and Sample Size

Average			Maximun		
Effect Size	N (%)	Mean Stnd. N	Effect Size	N (%)	Mean Stnd. N
0.00 - 0.20	45 (61)	235	0.00 - 0.20	37 (50)	253
0.21 - 0.40	20 (27)	118	0.21 - 0.40	22 (30)	136
0.41 - 0.60	3 (4)	37	0.41 - 0.60	5 (7)	56
0.61 - 0.80	5 (7)	51	0.61 - 0.80	9 (12)	66
0.81 - 1.00	1 (1)	32	0.81 - 1.00	1 (1)	32

Table 4.

Average Effect Size and Statistical Power (by Sample Size)

Sample Size	(N)	Mean Effect	Mean Power *
15 - 50	(12)	.42	.46
51 - 100	(25)	.23	.29
101 - 200	(21)	.17	.33
201 - 400	(5)	.18	.45
Over 400	(11)	.08	.35

* Power estimates are derived by taking the mean power of all outcomes measures examined.

Table 5.

Average Effect Size for Experiments including Only Juveniles

Sample Size (N)	Mean Average Effect
15 - 50 (9)	.52
51 - 100 (16)	.22
101 - 200 (7)	.21
201 - 400 (2)	.28
over 400 (3)	.09

Table 6.

Average Effect Size for "treatment" Experiments

Sample Size	Effect Size (N)	
15 - 50	.59	(6)
51 - 100	.19	(11)
101 - 200	.18	(4)
201 - 400	--	--
over 400	--	--

Appendix One

Chronological List of the Experiments
(by year experiment began)

(1951) Copenhagen Short-Term Offender Experiment
(1953) California Special Intensive Parole Experiment-Phase I
(1955) California Pico Experiment
(1956) California Special Intensive Parole Experiment-Phase II
(1957) Fricot Ranch Delinquent Dormitory Experiment
(1959) California Short-Term Psychiatric Treatment Experiments--two experiments
(1959) California Parole Research Project Experiment
(1959) English Psychopathic Delinquent Experiment
(1959) Utah Provo Experiment
(1960) California Paso Robles and Youth Training Center Group Counseling Experiments
(1961) California Juvenile CTP Phase I Experiments
(1961) California Group Counseling Prison Experiment
(1961) California Fremont Program Experiment
(1963) California Juvenile Probation and Group Counseling Experiment
(1963) English Police Cautioning Experiment
(1964) English Borstal Allocation Experiment
(1964) San Diego (CA) Chronic Drunk Offender Experiment
(1964) Kentucky Village Psychotherapy Experiment
(1964) Fairfield School for Boys Experiment
(1965) California Crofton House Experiment
(1965) California Parole Work Unit Experiment
(1965) English Juvenile Therapeutic Community Experiment
(1965) Los Angeles Silverlake Experiment
(1966) California Preston School Typology Experiment
(1966) Los Angeles Community Delinquency Control Project Experiment
(1968) California Juvenile Behavior Modification and Transactional Analysis Experiment
(1968) English Prison Intensive Social Work Experiment
(1969) Denver Drunk Driving Sentencing Experiment
(1969) Florida Inmate Work Release Experiment
(1969) Ohio Juvenile Probationer Behavior Modification Experiment
(1970) California Reduced Prison Sentence Experiment
(1970) California Unofficial Probation Experiment
(1970) Minneapolis Informal Parole Experiment
(1970) Sacramento (CA) Juvenile 601 Diversion Experiment
(1971) California Ellsworth House Experiment
(1971) Illinois Volunteer Lawyer Parole Supervision Experiment
(1971) English Intensive Probation Experiments--four experiments
(1971) Tacoma Juvenile Inmate Modeling and Group Discussion Experiment
(1972) Sacramento (CA) Juvenile 602 Diversion Experiment

(1973) Canadian I-Level Maturity Probation Experiment
(1973) English Intensive Welfare Experiment
(1973) San Fernando (CA) Juvenile Crisis Intervention Experiment
(1974) Juvenile Diversion and Labeling Paradigm Experiment
(1975) Leeds (UK) Truancy Experiment
(1975) San Pablo (CA) Adult Diversion Experiment
(1975) Pinellas County (FL) Juvenile Services Program Experiment
(1975) Washington, D.C., Pretrial Supervision Experiment
(1976) California Early Parole Discharge Experiment
(1976) California Summary Parole Experiment
(1976) Clark County (WA) Status Offender Deinstitutionalization Experiment
(1976) Florida Project Crest Experiment
(1976) Memphis Drunk Driving Sanctioning Experiments--two experiments.
(1976) North Carolina Butner Correctional Facility Experiment
(1977) Memphis Juvenile Diversion Experiment
(1977) Wayne County (MI) Project Start Experiment
(1977) Vera Institute (NY) Pretrial Adult Felony Offender Diversion Experiment
(1977) Hamilton (Canada) Juvenile Services Project Experiment
(1977) San Quentin (CA) Squires Program Experiment
(1978) Illinois Juvenile Tours Experiment
(1978) Michigan Juvenile Offenders Learn Truth (JOLT) Experiment
(1978) New Jersey Juvenile Awareness Program (Scared Straight) Experiment
(1981) National Restitution Experiments--4 experiments
(1981) Minneapolis Domestic Violence Experiment
(1981) Ramsey County (MN) Community Assistance Program Experiment
(1983) Police Foundation Shoplifting Arrest Experiment
(1984) Ontario (Canada) Social Interaction Experiment