CHAPTER 2: A REVIEW OF RECENT STUDIES OF HIGH INTENSITY ADULT CORRECTIONAL DRUG TREATMENT PROGRAMS — THE PROBLEM OF SELECTION BIAS AND POSSIBLE SOLUTIONS

Our previous report on 6-month outcomes provided a literature review of high intensity drug treatment programs in prisons. The focus of this review was to identify the common methodological problems faced in evaluating correctional treatment programs by examining the commonly cited studies. Before we examined these studies, we discussed the nature of evaluation research in an applied setting and what we viewed as the most significant methodological problem — selection bias. Our goal was to represent the difficulties of applied correctional research, to describe the organizational pressures that determine which inmates receive treatment, to depict these influences in an understandable model of selection pressures,[1] and to offer potential solutions — both analytical and methodological — to these problems. We then used our model of selection pressures to critique the research design, analyses, and interpretations of results contained in the most commonly cited studies.

For the benefit of readers who are not familiar with our previous report, we include our previous analysis of the methodological problems faced by evaluation researchers and our critique of the most commonly cited studies up until 1998. However, this chapter builds upon our previous discussion by updating our previous literature review. Thus, our critique of the literature includes the same research cited in our previous report as well as research that has been published since our previous review in 1998.

We complete our review of the literature with a discussion of methodological developments in the field of evaluation and critical commentaries of drug treatment evaluation research.[2] This literature suggests that the field of criminal justice drug treatment evaluation has been slow in addressing the methodological problems faced in quasi-experimental evaluation research.

Overall, our review suggests that methodological problems associated with evaluating residential drug treatment programs create important obstacles in interpreting the results of this research. We believe that, for the most part, the research we reviewed suffers from inferential problems associated with disentangling treatment effects from selection bias effects. We argue that it would be prudent to temper strong conclusions about successful treatment outcomes — which are often portrayed in the literature — with a bit of skepticism, born from a closer look at the methodological problems. We also describe different solutions for overcoming the problem of selection bias.

---

[1]Throughout this chapter we refer to the various selection pressures as a means of describing the various elements of selection bias.

[2] This was not included in our previous report.

Selection Bias and the Evaluation of Prison Drug Treatment Programs

There is no question that conducting evaluations in an applied setting is very difficult. Correctional systems are coercive by their very nature, and even when treatment is endorsed and carried out by well-trained, motivated providers, there is typically a tension between the necessities of custody practices and the goals of a therapeutic setting. Custody practices are necessarily rigid and uniform, while treatment delivery must be personalized and flexible.

The ideal model for any assessment is a clinical trial in which we can control the timing, dose (amount of exposure to), and administration of treatment. Using random assignment allows us to discount client characteristics when drawing inferences about the effects of treatment. Unfortunately, there are very few situations in which it is practical to carry out a well-controlled, random assignment design of drug treatment. In most correctional settings, control over who gets treatment and when they get it rests with the treatment providers or some administrative authority. Often there are policies that determine eligibility as well. Under these conditions, the best we can achieve is a quasi-experimental design but even these will vary in their rigor. Our emphasis is on the difficulty of doing either random assignment or quasi-experimental designs in a correctional setting.

We raise these cautions because the internal and external validity of evaluation studies in correctional settings can be compromised by the vagaries of correctional environments and possible differences in the characteristics of the clients involved in these studies. Rather than ignoring or avoiding these problems, we address them directly and offer some solutions that we used in the current study.

Fletcher and Tims (1992) have outlined the kinds of threats to internal and external validity that can occur in evaluation studies performed in a correctional setting. Their critique is thorough, but does not give any color or texture to the scope of problems. In this chapter, we try to characterize the nature of some of the problems that occur when a variety of administrative decisions and local practices can contaminate the research design.

Rather than repeat the Fletcher and Tims (1992) critique, we focus on what we believe is the most troublesome methodological problem in an applied setting, in general, and in the correctional drug treatment literature in particular: understanding and controlling for selection bias. In a simple two-group design, experimental versus control, we want to be able to assume that whatever effect we observe is attributable to the treatment and not to differences in the characteristics of the subjects in the two groups. Selection bias results from processes that change the composition of the two groups in such a way that we are unable to make a clear inference as to whether the effects we observe are due to the treatment or to the different group compositions.[3]

[3] For a technical discussion of sample selection bias, *see* Berk, 1983.

Adopting a skeptical perspective, we could conclude that the selection process prevents us from drawing any conclusions about treatment effectiveness regardless of whether the original design used randomization to assign offenders to treatment groups. From this perspective, program terminations, both voluntary and involuntary, cause the treatment group to "boil down" to only those participants who are ready and capable of succeeding when released to the community. Thus, the "effect" of treatment may be nothing more than the process of "weeding out" those more likely to fail from those more likely to succeed, and treatment has no additional value to those who remain in treatment.

A more sanguine view is that the selection process results in a motivated group of program participants whose treatment results in even greater success than would be the case had no treatment occurred. The problem becomes choosing a research design that can distinguish between outcomes that are due solely to the selection process and those that are due to both this selection process *and* treatment. Furthermore, the research design must be able to differentiate the effects attributable to the selection process from those attributable to treatment.

A simple conceptual device for understanding this problem is to treat it as an additive process. We assume a baseline group of untreated comparison clients similar in background to our treatment clients. For conceptual purposes, we can envision treatment subjects who "fall out" of treatment and those who remain. We assume that those clients who remain, on average, would be more successful than the comparison subjects even without treatment because they are a more select, motivated subgroup. But, we also assume treatment has benefits, naturally, and that it "pushes" the success of these motivated individuals higher than it would have been without treatment. The inferential problem comes in identifying the "push" from motivation from the "push" from treatment. In some cases, these causes may be so entangled that the separate influences are extremely difficult to reconstruct.

Although our discussion focuses on selection processes that bias results in favor of finding a treatment effect, it is possible that selection processes can affect group composition in a manner that biases results *against* finding a treatment effect. For example, there might be an incentive structure that would encourage higher risk offenders, rather than lower risk offenders, to enter treatment. Another possibility is that treatment selection is tightly controlled by providers who reserve treatment beds for the most difficult cases.

A Model of Sample Selection Process

To understand the complexity of the problem, we have attempted to represent in Figure A the most important selection processes that can occur in the research design when evaluating drug treatment in a correctional setting. In this context, we use the word "selection" to describe the processes that differentiate who enters treatment, as well as the processes that determine who exits treatment prematurely. This latter process is also called "attrition." Figure A indicates the kind of selection pressures (filters) that operate within an environment in which treatment is available and an additional selection process that occurs when researchers try to follow up on

inmates who have been released to the community. There are four prominent in-custody selection filters: self-selection, administrative — or clinical — selection, treatment selection, and

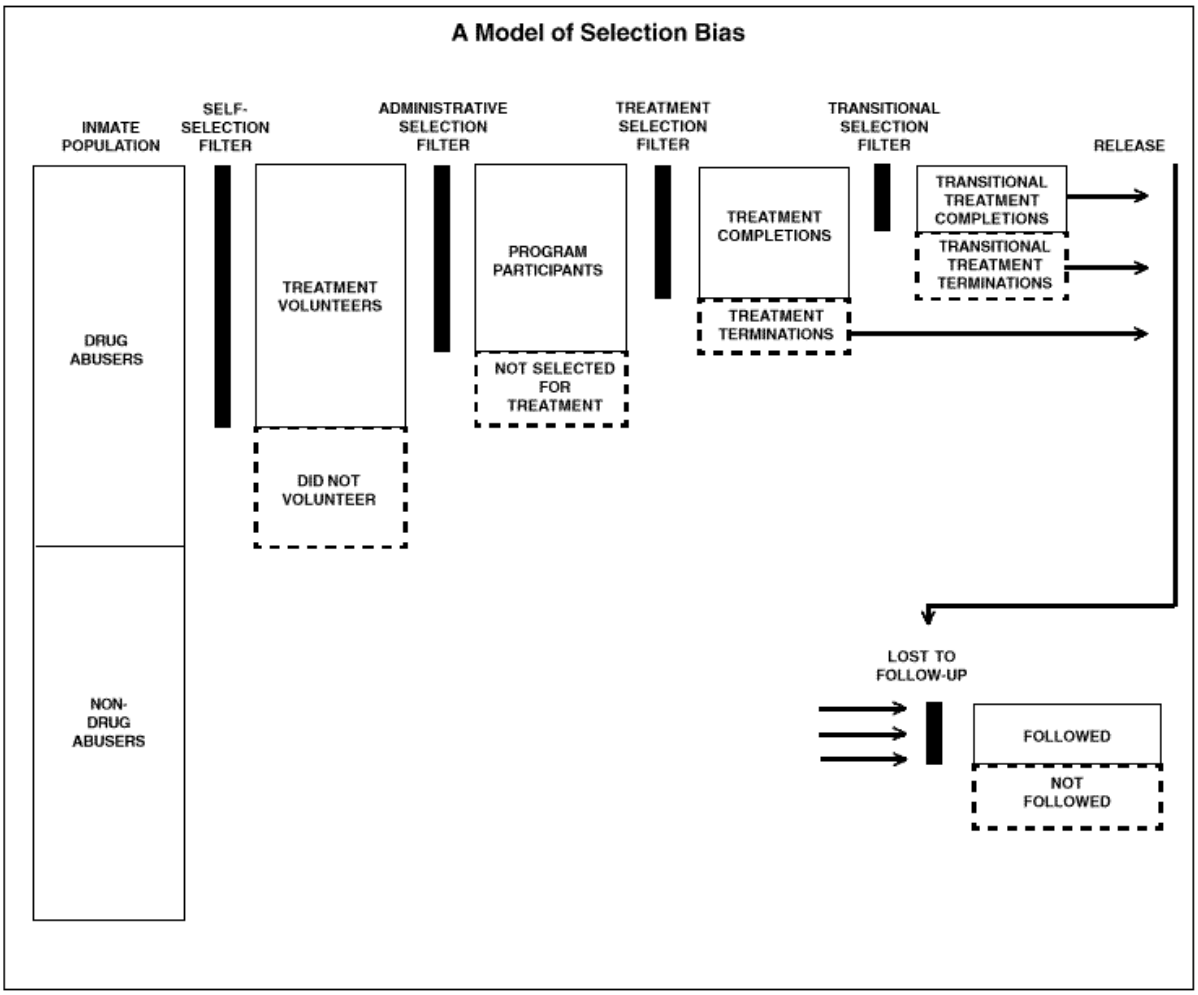

Figure A -- A Model of Selection Bias

transitional treatment selection. The last selection pressure occurs when there are biasing processes that determine which clients are lost to follow-up.

The first process, self-selection, is based on either internal motivational states or external incentives that predispose some people to volunteer for treatment.[4] The second process,

[4] One of the reviewers of our 6-month report asked us to address the issue of voluntary participation in these in-custody criminal justice programs. All of the programs reviewed in this section, as well as the drug treatment program within the Bureau of Prisons, were composed of voluntary participants. We are unaware of the extent to which criminal justice-based drug treatment programs are voluntary, mandatory, or "coerced." Although there is some literature on whether mandatory or coerced treatment can be successful, this is an aside for the present

administrative selection, reflects the clinical judgment of treatment providers and other administrators who determine whether someone is chosen for a program.[5] The third process, treatment selection, differentiates between clients who can and cannot meet the program demands. As illustrated in our review of the research on prison therapeutic communities, treatment selection can result in high numbers of inmates failing the treatment.

Another opportunity for selection occurs during the transitional care phase of a multi-phase treatment approach. Participants can be terminated by staff or they can withdraw themselves (in the case of voluntary treatment participation). It appears from our study and some of the others that selection in this phase is usually not very significant. However, for situations in which the conditions of supervision for treatment subjects were very different from those for control subjects (e.g. higher expectations regarding exemplary behavior among treatment participants), this additional level of selection might need to be considered. Finally, there may be conditions that affect which study participants are lost to follow-up. For example, if follow-up interviews are voluntary, there may be self-selection bias introduced by those characteristics that are correlated to the individual's willingness or unwillingness to be interviewed.

The problem of selection bias becomes readily apparent from Figure A when one focuses on the end of the selection process and sees who remains in treatment. If the study design assesses only those offenders who have made it through every selection filter, it is very difficult to construct a legitimate control group composed of only those non-treated offenders who also would have made it through the same selection process had they had the opportunity to do so.

What may not be readily apparent from Figure A is the way selection pressures can also affect the comparison group. Let us assume we begin a study with a pool of drug dependent clients. From this pool a sample of clients is selected into treatment. As we have already noted, the selection pressures may operate in one of two ways. Clients in treatment may have characteristics that predispose them to more successful treatment outcomes (case 1) or they may have characteristics that dispose them to more unsuccessful outcomes (case 2). In both cases, if we have to draw our comparison sample from the individuals remaining after others have been selected for treatment, we may bias our design in a subtle way.

In case 1, the residual group of untreated clients, on average, may be less disposed toward successful post-release outcomes. In that case, we have "creamed" the treatment clients and the

---

purposes. Even if drug treatment were mandatory, the attrition process would still affect the internal and external validity of the evaluation. Furthermore, the mandatory nature of drug treatment may mean that the selection process is removed from the client and handed to an administrator.

[5] These selection pressures can come from external sources, such as judges who strongly recommend candidates for treatment, or from internal sources, such as the pressures to fill treatment beds in a crowded prison system.

residual pool is composed of the "sour" remnants. A sample drawn from the residual pool will likely have less successful outcomes than will a randomly drawn sample of drug dependent clients composed of both motivated and unmotivated individuals. In case 2, the residual group of clients may be more disposed toward successful post-release outcomes. A sample drawn from this residual pool will be more likely to have successful outcomes than will a randomly drawn sample of drug dependent clients. Thus, in case 1, the residual comparison group will introduce bias in favor of finding a positive effect, while in case 2 it will introduce bias against finding such an effect.

Methodological Solutions to Selection Bias

There are several ways to attempt to handle the problem of selection bias in the absence of random assignment. None may be completely satisfactory. The first is to assess *all* client characteristics that can be used to adjust the treatment outcomes. Thus, if there are differences between two groups due to selection pressures, we can control for these differences in a multivariate analysis and adjust our outcomes accordingly. This approach will fail if there are important unmeasured variables that distinguish the treated from untreated samples and those differences affect recidivism. This approach also requires a thorough theoretical understanding of the selection processes.

We can speculate on some of the processes that may be affecting who is selected (or self-selected) to go into treatment. Client motivation or commitment may determine who volunteers for treatment. Perhaps a board or group of administrators choose clients based upon the seriousness of the subjects' drug dependencies and the extent to which the selecting personnel believe clients will benefit from treatment. Attrition may also have its unique determinants. Perhaps some clients are unreceptive to specific treatment approaches. Perhaps those inmates most entrenched in a criminal lifestyle or most embedded in an inmate subculture are the most likely to withdraw from treatment. All of these processes can affect outcomes. Many researchers in this domain have attempted to control for client characteristics by measuring such variables as age, race, sex, criminal history, and drug dependency and then use these variables in a multivariate analysis. However, unless these variables control for the processes that affect both selection into treatment and recidivism, this technique will fail to control for selection bias.

Our argument is that considerable thought should go into understanding and measuring selection pressures so that we can observe and control for these processes. One straightforward approach to addressing selection bias is an instrumental variable approach. This approach was considered appropriate because a federal prisoner s assignment to a specific prison had nothing to do with whether or not he needed substance abuse treatment (DAP) and some prisons did not offer treatment. So selection bias does not affect a comparison of the outcomes for a comparison group which did not have treatment available – DAP control – and the *combined* outcomes for both a treated and non-treated group of individuals who were offered treatment – DAP treatment

and DAP comparison.[6] To further illustrate this approach, suppose that every prison holds identical populations comprising: those who would enter treatment if offered to them and those who would not enter treatment if offered to them. When treatment is offered, these populations can be identified, and when treatment is not offered, they cannot be identified. Let:

P_{accept}	The percentage of a prison population that would accept treatment if given the opportunity. Call this group A.
$1-P_{accept}$	The percentage of a prison population that would decline treatment if given the opportunity. Call this group B.
F_{accept}	The fraction of group A who would recidivate if treatment were not provided.
$F_{decline}$	The fraction of group B who would recidivate.

Then if treatment were provided to no one, the rate of recidivism among group A and group B combined can be written:

$$F_{untreated\ population} = P_{accept}\ F_{accept} + (1-P_{accept})F_{decline}.$$

This is the expected value of the observed proportion of recidivists in the control group who did not have treatment available to them.

Suppose that, on average, treatment reduced the proportion of inmate who recidivate by an amount D. If treatment were provided to everyone who would accept it:

$$F_{treated\ population} = P_{accept}\ (F_{accept} - D) + (1-P_{accept})F_{decline}$$

Here D is the treatment effect. $F_{treated\ population}$ is the expected value of the observed proportion of recidivists in the combined group of individuals from treatment sites, some of whom received treatment (DAP treatment) and some who did not (DAP comparison groups). A test of treatment effectiveness can be based on the differences between two observables: $F_{treated\ population}$ and $F_{untreated\ population}$. Some algebra shows that the expected value of the effect from treatment is:

$$D = (F_{untreated\ population} - F_{treated\ population})/P_{accept}.$$

This is one illustration of an instrumental variable approach to quasi-experimental design. It affords an estimate of the average treatment effect D and a measure of its statistical significance despite the fact that the treated and untreated groups may have failure rates that differ from each other for reasons that have nothing to do with the receipt of treatment.

---

[6] *See* Chapter 5 for a more complete description of our treatment and comparison groups.

The instrumental variable approach to evaluating treatment effectiveness is not much complicated by introducing control variables and using regression models. The introduction of control variables has three benefits. By reducing unexplained variance, the regression can reduce the standard error of the estimate for the treatment effect. Second, the control variables can help adjust for any population difference between DAP (treatment) and non-DAP (non-treatment) facilities. And third, the parameters associated with control variables have policy relevance for the Bureau.

A second approach for addressing selection bias, called herein the *Heckman selection bias* approach (Heckman, 1979; Maddala, 1983) is somewhat more difficult to apply than is the instrumental variable approach. It requires the analyst to *jointly* model the selection into the sample and the post-release outcome. Here, note that the selection bias approach has much in common with the standard instrumental variable approach, and if the analyst is willing to limit his analysis to a linear-additive regression model, there may be little to recommend the selection bias approach over the instrumental variable approach. However, as explained by Maddala (1983, p.261), the Heckman selection bias model can be used to study more complicated models where treatment interacts with other variables.

We also adopted a second approach because of the well known problem with the instrumental variable approach. The estimated regression parameter associated with the dummy variable will be biased and inconsistent if the dummy variable and the error term are not independent. Independence seems unlikely if any unmeasured factor (such as motivation) affects both the receipt of treatment and the outcome variable.

We measured a number of background characteristics we believed were related to post-release recidivism and drug relapse. However, because we were still unsure of the exact nature of all the selection pressures operating in our study, we adopted the two procedures described above. The mathematics of the Heckman model are described in Appendix B. The selection bias adjustment was made to the survival function associated with the time to failure (e.g., time to recidivism, time to drug use). By modeling selection bias explicitly, we were able to test whether selection bias increased or decreased the survival time. If it increased the survival time, this was evidence that there were pressures that selected lower-risk defendants into treatment. If the selection bias parameter was negative, this suggested that there were pressures that selected higher-risk defendants into treatment, which would in turn lower their survival times. If the selection bias parameter was not significant, we could conclude there were no such selection bias pressures operating.[7]

It is not unusual to find, in previous research, that program completers are more successful than

[7] The derivation and computation of these parametric survival models with correction for selection bias are not available in typical statistical packages. We engaged an econometrician consultant, William Rhodes of Abt Associates, to help us derive the appropriate models and estimate them using GAUSS.

are controls, who in turn are more successful than are program terminators. If it was possible to classify correctly control group offenders' outcomes into hypothetical "completions" and hypothetical "terminations," and if we were to assume that treatment is effective, we would expect to observe that those who completed treatment had more successful outcomes than those in the control group who hypothetically completed the treatment. Further, we would expect to observe that those who did not complete treatment had equivalent outcomes to their counterparts in the control group who would hypothetically be expected not to complete the program. However, in the absence of our ability to classify hypothetical "completers" and "terminators" in the control group, the combined outcomes of program completers and terminators should be significantly better than those of the control group.

Some treatment proponents might argue that even if treatment is, in fact, nothing more than a weeding out process, this is still a useful result of the treatment process because it identifies the individuals who are more likely to succeed. The problem with this logic is that the same goal might be accomplished by simply improving our risk-classification devices in the absence of treatment. Furthermore, it is important to know whether it is treatment per se, risk selection, or both, that accounts for better outcomes in the treatment group. We cannot hope to improve treatment or understand how it works if all we accomplish is the risk selection of inmates.

Additional Selection Concerns

Figure A also can be used to conceptualize the selection bias issue by depicting the problem of choosing an appropriate control group and the level of generalizability inherent in the research design. It is clear from Figure A that treatment terminations, whether in-prison or community-based, cannot be ignored if we are to make any sense of a program's effects. Furthermore, it is clear that by choosing a comparison group of volunteers, our level of generalization is restricted to treatment for "motivated" treatment participants. Under these circumstances, it would be important to know how representative the pool of treatment volunteers is relative to the non-volunteers. The practical import is that if treatment volunteers represent a small fraction of the drug-dependent population, then treatment impact is minimized.

Another problem becomes apparent from Figure A. Consider a design in which researchers choose a control group (composed of drug abusers) that is a combination of volunteers and non-volunteers. Unless one models the selection process and incorporates it into the analysis, the outcome differences between a self-selected or administratively selected group and an "unselected" group may be attributable entirely to the level of motivation of the volunteers and have nothing to do with the treatment provided.

Our procedures ensured that we collected follow-up data on all inmates who began treatment and were selected into our "convenience" samples. Thus, regardless of whether an inmate completed or was terminated from the program, data were collected on the individual's post-release outcomes. We also collected data from comparison subjects, some of whom had treatment available and some of whom did not. These data collection procedures are described in great

detail in subsequent chapters.

To summarize, depending on the analysis, we treated comparison subjects from DAP sites in different ways. In the first analysis, to replicate previous research, we contrasted the DAP treatment group which completed treatment with the DAP treatment subjects who did not complete treatment, the DAP comparison subjects from the treatment sites and the non-DAP control subjects from sites at which no treatment was available. If highly motivated inmates were entering (and completing) treatment at DAP sites, the pool of drug dependent comparison subjects who participated in our study would have been composed of less motivated, perhaps more risky, clients. Under these circumstances, the DAP comparison subjects may have had lower success than the non-DAP control subjects. Clearly, quite the opposite would happen if the residual pool of DAP comparison subjects had been composed of more motivated inmates. Both of these hypothetical outcomes rest on the assumption that the DAP comparison group is composed of inmates representative of all the groups we have described and that they are represented in the same proportions. Thus, the non-DAP control group is hypothetically composed of completers, disciplinary terminations, program withdrawals, treatment noncompleters, and the residual comparison subjects.

In the second analysis, we combined data from the treatment groups with data from comparison subjects from DAP sites. We reasoned that the DAP comparison clients were inmates who had treatment available but chose not to participate. We would expect that these same types of inmates would be represented among our non-treatment site comparison subjects. Thus, research subjects in the non-treatment sites should consist of all levels of inmates who would have volunteered for treatment, as well as inmates who would have declined treatment. Thus, the appropriate test of treatment in our design is the combined test of (1) all inmates who were selected for treatment, as well as (2) the inmates who did not volunteer for treatment but who had treatment available (were housed at a DAP site), contrasted with the control subjects who did not have treatment available (were not housed at a DAP site).

Finally, in a third approach, we explicitly model the selection process, using non-treated subjects from both DAP and non-DAP sites as "controls." The difference between the two non-treated groups is that comparison subjects from DAP sites were subject to selection bias while control subjects from non-DAP sites were not. By explicitly modeling the selection process, we could statistically capitalize on all comparison subjects to increase the power of our treatment versus non-treatment contrast.

In the following sections we critically review the most commonly cited research on in-custody therapeutic communities using our conceptual model of selection bias as a heuristic device.

A Critical Review of Prison Drug Treatment Research

Our review of the literature focuses primarily on five programs that have received considerable attention in recent reviews of the literature on prison residential drug treatment studies. These programs include Stay 'N Out, Cornerstone, Key/Crest, New Vision at Kyle Unit, and the Amity Right Turn Project.

We examine the published and unpublished reports on these programs in some detail. Our general conclusion is that these studies suffer from the inferential problems associated with disentangling treatment effects from selection bias effects. Although we are somewhat critical of the research in this domain, we realize how difficult it is to conduct a program evaluation and how easily controls, intended to introduce rigor into the evaluation, are easily compromised. Our critique attempts to assist future program evaluators in this research area and help them avoid some of the same mistakes that both we and other evaluators have made.

Stay 'N Out Program

Wexler and colleagues have published a series of articles that report on the effectiveness of the Stay 'N Out drug abuse treatment program used by the Department of Corrections in New York State (Wexler and Chin, 1981; Wexler, Falkin, and Lipton, 1988; Wexler, Falkin, and Lipton, 1990; Wexler et al., 1992; Wexler and Williams, 1986). We focus primarily on the recidivism outcome results reported in Wexler, Falkin, and Lipton (1990) as adapted and slightly modified for a National Institute on Drug Abuse Research monograph (Wexler et al., 1992).

Their evaluation of the Stay 'N Out therapeutic community (TC) contrasted male inmates who participated in that program with inmates in two other drug treatment programs (milieu and counseling treatment) and a control group of inmates who had volunteered for the therapeutic community but were never admitted to the program because of time constraints. The volunteer control group was used to minimize selection bias issues.

Female TC participants were contrasted with those in a drug counseling treatment group and those in a control group composed of women who volunteered for the TC program but changed their minds prior to admission into the program. Unlike the male control group, the female control group could easily have been composed of unmotivated women who would be the least likely to succeed following release and thus bias any contrast between program and non-program participants.

Wexler, Falkin, and Lipton (1990) reported their results first by way of a series of group contrasts among the mean differences in the outcome variables without controlling for background differences, and then by using a multivariate analysis that controlled for background variables. Although Wexler and his colleagues argued that their study provided "convincing evidence that prison-based TC treatment can produce significant reductions in recidivism," (p. 89) we found several shortcomings in the study's analysis and methodology.

Female inmates were excluded from the multivariate analysis because, according to the authors, there were too few to analyze. The outcome variables reported by Wexler and his colleagues included the percent of inmates arrested after release to parole supervision, the number of months before such arrests occurred, and the percent having favorable parole outcomes. This last measure was based on whether an inmate completed parole without a revocation, arrest, or rule infraction.

Wexler, Falkin and Lipton (1990) reported that their multivariate analysis of percentage arrested and parole outcome produced no significant results. There was no effect tied to background characteristics or time-in-treatment. Wexler, Falkin and Lipton (1990) did find significant predictors in their multivariate analysis of time-to-arrest, and they reported those results. We are puzzled by the fact that age and criminal history, which were influential predictors of time-to-arrest, were not significant predictors of percentage arrested or parole outcome. These variables typically are the most influential predictors of any measure of post-release criminal recidivism (Harer, 1994). This is a minor point relative to their interpretation of the multivariate analysis involving time-to-arrest.

In addition to the background characteristics of age and criminal history, Wexler, Falkin and Lipton (1990) included the following variables in their regression analysis: the duration of parole supervision, a dummy code for each type of treatment, time-in-treatment for each of the treatments, the duration of parole supervision, the amount of time an inmate spent in prison after completing treatment but before release, and the square of the amount of time an inmate spent in the therapeutic community.

In their analysis of time-to-arrest, Wexler, Falkin and Lipton (1990) interpreted the linear and quadratic regression coefficients for time-in-treatment. However, they failed to interpret the program participation variables. Although only the dummy variable for the TC treatment was significant, all of the treatment dummy variables demonstrated that regardless of the type of drug treatment given to inmates, having any drug treatment shortened the period between release and arrest relative to the control group. Thus, the treatment dummy codes demonstrated that inmates in treatment were arrested sooner than were inmates in the control group. In addition to the fact that TC inmates failed sooner than did control group participants, the relationship between TC treatment and failure is quadratic. That is, time-to-arrest increased with the amount of treatment up to a point, then declined thereafter. Wexler et al. (1992) emphasized this finding while disregarding the dummy-coded treatment effects.

The other major finding emphasized in this study, as well as in secondary sources that refer to this study (*see*, for example, Lipton, 1995), was that when the treatment effect was examined without accounting for the other background variables, male inmates were less likely to be arrested if they participated in the TC drug treatment. For female participants, none of the group contrasts reached conventional statistical significance. The percentage of male inmates arrested after release from prison varied by treatment group. Among TC inmates, 26.9 percent were arrested after release. For milieu and counseling inmates, 34.6 and 39.8 percent, respectively,

were arrested. Among no-treatment controls, 40.9 percent were arrested after release. In light of the fact that the multivariate analysis of this outcome measure failed to reach statistical significance, we argue that these results probably were attributable to differences in background characteristics of the groups and not to a treatment effect. But there are other reasons why these group differences are possibly not meaningful.

The different groups had different risk periods, with the TC group having the shortest average risk period (34.7 months). The other groups each received, on average, about 41 months of parole supervision. Thus, each subject in the TC group, on average, had 6 fewer months of parole supervision and, therefore, much less time in which to be arrested. Another difficulty with this analysis of percent arrested by treatment group concerns the extent to which Wexler et al. (1992) should have adjusted their findings for people who were censored. If inmates were technically violated, rather than arrested, this would have removed them from the risk set. Thus, fewer arrests could mean greater parole violations. Without an explicit explanation of the censoring process, we cannot rule out this possibility.

In general, when one is analyzing the time to an event — whether relapse, arrest, or conviction — it is much more appropriate to use event history techniques that allow one to treat different risk periods by censoring observations and removing them from the risk set. This is a problem not only with this set of analyses, but with most other studies in this research domain as well.

Cornerstone Program

Field has published several studies evaluating the Cornerstone Program, a residential program for alcohol- and drug-dependent inmates within the Oregon correctional system. A key component of the Cornerstone Program as described by Field (1985) is that inmates who are admitted to the program must be willing to commit to at least 6 months of follow-up treatment in the community. Another program admission criterion requires that the inmates be granted minimum-security status by the prison superintendent. At first glance, this would seem to be a very narrow selection criterion that would exclude all but the lowest-risk candidates for drug treatment and would have profound implications for possible selection bias effects. However, Field (1985) described the treatment clients as having, on average, about 12 prior arrests, 6 prior convictions, and 6 years of adult incarceration. Also, these clients were described as having chronic substance abuse histories.

In addition to the follow-up treatment in the community, Cornerstone graduates "have a job, a place to live, and a drug-free support network before discharge" (Field, 1985, p. 52). Thus, the community aftercare component of this program went far beyond focusing on drug relapse.

To compare program graduates, Field retrospectively chose three comparison groups. Group I was composed of Cornerstone dropouts, Group II was composed of Oregon parolees with some history of drug abuse, and Group III came from a follow-up study in Michigan that Field chose because the study followed a "similar population over a similar time frame" (Field, 1985, p. 52).

There was a uniform 3-year follow-up period, and Field assessed recidivism in two ways. Recidivism was defined as a return to prison within 3 years and, separately, as a conviction within 3 years. Among Cornerstone graduates, 29.2 percent returned to prison within 3 years. Among the comparison groups, 74.1 percent of the dropouts were recommitted, 37.1 percent of the group composed of Oregon parolees with a history of substance abuse were recommitted, and about 43 percent of the Michigan release cohort were recommitted. Those reconvicted within 3 years consisted of 45.8 percent of the Cornerstone graduates, 85.2 percent of the Cornerstone dropouts, and 74.7 Oregon of the parolees with a substance abuse problem. No reconviction data were available for the Michigan cohort.

There are three major problems with interpreting the results of this study. The first is that we have no basis for comparing Cornerstone graduates with the three comparison groups on any variables related to recidivism, such as age, criminal history, degree of substance abuse, and family social support. Thus, we have no guarantee that the groups were equivalent with respect to their risk of recidivism. Secondly, there are two significant program components to Cornerstone — the first is institution based and the second is community based. Even if this program is influential in reducing relapse and criminal recidivism, we cannot disentangle which program component was the more important one. Finally, as the recidivism data showed for the program dropouts, and as Field noted, program participants simply may have been highly motivated inmates who would have succeeded with or without Cornerstone.

Apparently, the dropout rate at Cornerstone was extremely high. Field (1992) enumerated the dropout rate in a recidivism study of 220 inmates who had been admitted to Cornerstone over a 2-year period. Of those 220 admissions, 65 withdrew after spending one to two days in the program, 58 withdrew after spending between 2 and 6 months in the program, 43 withdrew after spending at least 6 months in the program, and 43 graduated. Thus, there was a far greater number of dropouts than program graduates.

Field used these differential dropout rates to make a point about the duration of treatment. Field (1992) reported on the criminal recidivism of these groups, showing that the longer an inmate was in the program, the less likely he or she would be arrested, convicted, or recommitted to prison following release from prison. Although Field acknowledged that the length of a subject's treatment may have acted merely as a proxy for his or her level of motivation, he argued that pre-treatment incarceration data demonstrated that all four groups were equivalent in their pre-program arrest, conviction, and commitment rates.[8] In other words, by controlling for pre-program levels of criminal history, Field was satisfied that the dropout pattern represented treatment effects and not motivation or other selection effects.[9]

---

[8] Please note that pre-program data were not available in the 1985 study comparing program graduates to the three comparison groups.

[9] As M. Douglas Anglin, one reviewer of our previous 6-month report pointed out, motivation is not constant over time. Rather, it is episodic. Anglin argued that treatment outcome is determined by a host of factors, including motivation, treatment retention, and type of services

Even though Field demonstrated equivalence among the treatment groups (categorized by duration of treatment) with respect to prior criminal history, we know there are a host of other variables that also could affect the group outcomes in the absence of a treatment effect, none of which Field incorporated into his analysis. Furthermore, self-selection probably represents, among other things, the level of motivation and commitment one has to maintaining a drug-free lifestyle. Commitment to change may be quite unrelated to one's criminal history; in fact, it may even be inversely related.

As we argue below, program dropouts contaminate the interpretation of treatment effects in more ways than one. Especially for programs in which the dropout rate is extremely high, there arises the possibility that a program is simply selecting out high-risk-of-failure candidates rather than changing or rehabilitating low risk candidates. Another way of viewing this potential selection process is to approach it as a risk-assessment analysis. We consider two possible hypotheses. The first is that dropouts are more likely to have background characteristics that predict criminal recidivism. The second is that they are equivalent to "stayers" on objective measures of risk; however, by observing treatment subjects closely or by testing their motivation in a controlled, closely monitored environment, staff can further "weed out" higher risk inmates.

Key-Crest Program

The Key-Crest Program is a drug treatment intervention occurring in three phases. The Key component is a prison TC for inmates in the Delaware corrections system. Crest, the second component, involves inmates released to a community work-release center where they maintain jobs in the community but live in a facility where they continue their drug treatment in a modified TC. In the final component, offenders are released to the community, either under parole or some other form of supervision. In this stage, drug treatment consists of outpatient counseling and group therapy.

Four groups were evaluated. The first was composed of 43 inmates — selected by correctional counselors — who volunteered to participate in the prison-based TC. Because the Crest program had not yet been implemented, these inmates were the only Key program participants who did not subsequently participate in the Crest stage. The second group consisted of Key-Crest inmates who participated in both stages. Virtually all Key graduates were allowed to participate in Crest after it was implemented. The third and fourth groups were composed of inmates who had drug abuse problems, had not participated in Key, and were given the opportunity to participate in the Crest work-release program. On a random basis, half of these volunteers (the Crest-only group) were provided the Crest program, while the other half (the comparison group) participated in work-release in the absence of residential drug treatment. Thus, the comparison group for these analyses included inmates who had drug abuse problems, had volunteered for Crest, and

---

offered. These combine in some complex way to influence outcomes. Nevertheless, it is still the case that the resultant effects cannot be easily disentangled.

had not received in-prison TC drug treatment but had received AIDS/HIV prevention education.

There were two selection bias processes operating in the Key-Crest design. The first selection process involved selection into the Key and Key-Crest groups. For one, it appears that the selection involved staff evaluation of candidates for the program. The second selection process occurred as a result of the way baseline data were gathered. These data were gathered just prior to inmate releases from prison. Baseline data were collected on Key graduates, but not on Key terminations. Thus, only Key graduates were followed in the longitudinal design. Data were gathered on Crest and comparison subjects at baseline, in the absence of any knowledge about potential future attrition in these two groups. Thus, both the Key and Key-Crest groups were composed of inmates who were motivated enough to graduate from the Key component of this program.

Even though Key-Crest participants had the opportunity to drop out of the program while they were in the Crest stage, this group already was composed of a very select group of motivated individuals. As noted in Deleon, Inciardi, and Martin (1995), the Crest-only group was composed of some clients who "displayed negative attitudes toward the treatment program, which generally led to their quitting or being discharged from the Crest program" (p.88). However, all inmates in the Crest-only groups were still followed even though some had dropped out of the program (Inciardi, 1997, personal communication).

The Key-Crest program is being evaluated by Inciardi and his colleagues (Martin, Inciardi, and Saum,1995; Martin, Butzin, and Inciardi, 1995; Inciardi et al., 1997). Martin, Butzin, and Inciardi (1995) reported data based on interviews conducted 6 months after the inmates were released from prison. Most inmates who had participated in the Crest stage were probably still under supervision at the time of this 6-month interview. Thus, the results at this stage should be interpreted with a great deal of caution. Based on inmate self-reports, the data showed that 97 percent of the Key-Crest group and 84 percent of the Crest-only group said they had not been arrested within 6 months of release from prison. Among the Key-only participants, 74 percent reported they had not been arrested, while 60 percent of the comparison group claimed no arrests. The proportions reporting drug use were similar. When these proportions were adjusted for background characteristics, including time-in-treatment, the same ordinal relationship was obtained. Key-Crest participants were the least likely to self-report arrest and drug use, followed by Crest-only, Key-only, and comparison subjects.

An 18-month follow-up of the program (Inciardi et al., 1997) showed that 77 percent of Key-Crest participants reported being arrest-free at 18 months, while 57 percent of Crest-only, 43 percent of Key-only, and 46 percent of the comparison group reported being arrest-free. Drug use was measured by combining results of self-reports and urinalysis tests. The drug-free pattern corresponded to the arrest-free pattern. However, there is no indication that there was any attempt to check the veracity of the self-reported arrests.

Although several papers written by Inciardi and his colleagues before 1998 have emphasized that offenders should be receiving aftercare while they are under supervision, at the time of

their study in 1997 there was no formal aftercare (Inciardi, 1997, personal communication). Apparently, this study has no selection bias and no attrition operating in the Crest-only and comparison groups, although the authors have never reported the extent to which inmates withdrew or were terminated from the Crest program. Therefore, the reductions in self-reported arrest and actual drug relapse may be entirely attributable to the effects of transitional treatment. However, the Key-only and Key-Crest groups are composed of offenders who were either selected into treatment or who selected themselves out of treatment. Reductions in self-reported arrest and actual drug relapse in those groups are still potentially contaminated.

The more recent three-year results focused upon the effects of aftercare (Martin et al., 1999). Although Martin et al. (1999) initially used the same subject groupings as in the 6-month and 18-month results, they controlled for various demographic factors, such as race, sex, drug use, criminal history and previous drug treatment. These 3-year results showed no treatment effects for in-prison treatment on arrests but showed positive effects on drug use (Martin et al., 1999). The researchers concluded that the subject groupings were 'conservative' because Crest completers and dropouts were lumped together. Thus, they created different subject groupings. The revised groupings did not identify those who received in-prison treatment as a separate group. Rather the groupings were based on work release assignment and were as follows: work release dropouts (Crest dropouts), work release completers (Crest completers) without aftercare and work release completers with aftercare. The researchers subsequently found positive effects for work release: both Crest completer groups had more favorable outcomes than the Crest dropouts and the comparison group. Of course by separating the Crest dropouts from the Crest program completers, Martin et al., (1999) may have capitalized on selection bias to find significant effects.

Amity Right Turn Project

The Amity Right Turn program combines prison- and community-based therapeutic communities for inmates who volunteer for treatment. This program is funded by the California Department of Corrections in the R. J. Donovan medium security Correctional Facility in San Diego. The program is being evaluated by Wexler and his colleagues. Wexler, DeLeon, et al.'s (1999) initial evaluation of the program used reincarceration of subjects in the California prison system as their primary outcome. Reincarceration included a commitment for either a new offense or a technical violation of parole.

The researchers divide the subjects into five groups, with inmates who either had volunteered to be treated, had a drug problem, or were within 9 and 14 months of their parole release composing a waiting list of eligible participants. From this pool, inmates were randomly selected to participate in the prison TC.

There were a total of 715 research subjects. Inmates who were eligible but could not be treated prior to their release composed the control group (n=290). The remaining four groups consisted of the inmates who had been randomly selected for treatment in the prison TC. The composition of the four study groups depended upon whether they volunteered for post-release community-

based treatment and whether they completed the prison- or community-based program. Thus, the first study group was composed of inmates who volunteered for the prison program but who were terminated (prison treatment dropouts, n=95). The second study group consisted of those inmates who completed the prison drug program but did not volunteer for the community-based program (prison treatment completions, n=193). The third study group included inmates who volunteered and completed prison drug treatment and who volunteered and were terminated from the community-based program (prison treatment completions/community-based dropouts, n= 45). The fourth study group was composed of inmates who volunteered and completed the prison and the community-based program (prison completions/community-based completions, n= 92).

Wexler, DeLeon, et al. (1999) reported that the no-treatment control group had significantly higher reincarceration proportions at both 12 and 24 months after release from prison than did all of the other study groups combined. The 12-month comparison showed that the control group had 49.7 percent recidivism and that the combined study groups had 33.9 percent recidivism. At 24 months, these percentages were 59 and 42.6, respectively. When the combined result is separated into the control and four study groups, the five groups had the following reincarceration percentages at 12 months: control group, 49.7; prison treatment dropouts, 45; prison treatment completions, 40; prison treatment completions/community-based dropouts, 40; and prison treatment completions/community-based completions 6.5.[10]

Wexler and his colleagues also reported the number of days until reincarceration; however, for some reason these data were only compiled on 256 releasees for the 12-month follow-up and on 166 releasees for the 24-month follow-up period. Generally, the time-to-recidivism data mirrored the 12- and 24-month reincarceration data. A logistic regression of background factors, in conjunction with the treatment effect, indicated that reincarceration was 42 percent less likely for the combined treatment groups than for the control group. The background factors included age, ethnicity, criminal history, IQ, childhood problems, anti-social DSM-III-R diagnosis, distress, and social achievement. Unfortunately, there was no multivariate analysis that combined all of the background factors with dummy-coded representations of the different study groups. This may have given some indication that the combined effect was primarily attributable to the inmates who completed both the prison- and community-based programs.

Wexler, DeLeon, et al.(1999) acknowledge that their results were confounded by the fact that, during the post-release period, inmates who were receiving treatment in the community-based TC were at much lower risk than were other releasees simply by their residence in the TC. This would also affect the 24-month outcomes. If the risk periods were defined as beginning the day after release from the community-based facility or the day after release from prison for

[10] In their report, Wexler , Deleon, et al. (1999) did not provide the actual percentages of inmates who were reincarcerated for the prison dropout, prison completion, and prison completion/community-based dropout groups. We had to estimate these percentages from a bar chart represented as Figure A.

clients who did not participate in the community-based facility, the "risk environment" would have been more comparable for the different groups involved in the evaluation. It is clear from the analysis of the individual study groups that the dramatic differences between the combined study group and the control group were attributable to, primarily, the prison treatment completion/community-based completion group. Although no analysis was presented, there were much more modest differences between the control group and the three study groups composed of inmates who spent little or no time in the community-based aftercare facility.

There is another limitation to this study as well. In order to control for selection bias, the researchers used treatment volunteers exclusively. This not only limits their generalizations to volunteers (as it does in most of these studies), but it also gives us no indication how treatment results compare to outcomes of drug dependent prisoners who are unwilling to volunteer for treatment. Secondly, as the authors acknowledged, while they were able to control for selection bias at the prison treatment phase, they were unable to control for selection bias at the community-based treatment phase. The clearest conclusion that can currently be drawn from this study is that the longer an inmate volunteers and stays in treatment, the less likely is his or her reincarceration. Whether prison drug treatment was effective was ambiguous in this study, and whether community-based drug treatment was effective after release was largely untested.

Wexler, Melnick, et al. (1999) recently reported 3-year outcomes for the Amity in-prison program and attempted to test whether community-based drug treatment is effective. Consistent with the previous results for 12- and 24-month outcomes (Wexler, DeLeon, et al., 1999), Wexler, Melnick, et al. (1999) found the lowest recidivism rate among in-prison treatment completers who also received aftercare. Individuals who were eligible for and volunteered for drug treatment were randomly assigned to the in-prison program or a control group. We note that aftercare was available only to the in-prison treatment completers. When comparing the control group to all individuals who participated in in-prison treatment, there were no significant differences in post-release recidivism. The differences occurred among the various types of in-prison treatment participants – in-prison treatment dropouts, in-prison treatment completers with no aftercare, and in-prison treatment completers with aftercare. Furthermore, Wexler, Melnick, et al. (1999) reported that the decreased rates of recidivism associated with longer treatment duration were not replicated with the 3-year follow-up period. They concluded that time in treatment could not be separated from aftercare and therefore the study could not assess the independent and combined effects of in-prison treatment from post-release treatment.

In addition to univariate analyses, Wexler, Melnick, et al. (1999) conducted multivariate analyses (logistic regression for the dichotomous outcome variables and ordinary least squares for the number of days until reincarceration[11]). Age, ethnicity, criminal history, injection drug use, drug use severity, and motivation for treatment were used to control for the characteristics that may

[11] We note that using ordinary least squares is inadequate because of censoring problems. Event history methods are suitable for data with varying time points to failure.

affect self selection into treatment. The univariate results comparing the treatment group to the control group are potentially contaminated by the fact that only in-prison treatment completers had access to aftercare services. Furthermore, the additional breakdown into the various in-prison treatment completer groups also potentially contaminates the results because individuals selected themselves into aftercare. The multivariate analyses, although recognizing the issue of self-selection, does not adequately address the methodological issue. Regression methods assume that the error term is uncorrelated with the predictor variables. If this assumption is violated, then the estimated effect of the variable of interest, namely treatment, will be biased. Motivation for treatment, one of the predictors, has been shown to be a predictor of treatment entry. Thus a researcher cannot assume that the error term is uncorrelated with the treatment effect. As discussed earlier in this chapter (and in Chapter 8 of this report), selection bias requires other analytic methods to provide unbiased estimates of the treatment effect.

New Vision In-Prison Therapeutic Community, Kyle Unit

The New Vision In-Prison Therapeutic Community in Kyle, Texas, is only one component of a comprehensive Texas criminal justice initiative to treat criminal drug abusers. The Kyle unit is being evaluated by a team of researchers affiliated with Texas Christian University. There have been several reports of the evaluation conducted by Simpson and his colleagues (Simpson et al., 1994; Knight et al., 1995; Knight et al., 1997). Outcomes are available for inmates who had been released for 6 months.

The program's evaluation compared a control group to a treatment group composed of inmates who participated in a 9-month prison-based TC, followed by 3 months of community-based residential treatment, followed by a year of outpatient treatment. Program graduates agreed to provide urine samples for drug testing on a monthly basis.

The selection process for participants in the drug treatment program began with a drug-use screening assessment given to all inmates who entered Texas Department of Corrections facilities. A treatment referral committee reviewed the inmates' records, which included self-reported drug use. Inmates who had less than 9 months remaining on their sentences or who had committed an aggravated offense were excluded from further referral. Inmates who qualified for treatment had their cases forwarded to the Texas Parole Board for the final decision on placement in a drug program. Both comparison and treatment subjects in this study completed the initial referral process. However, the Parole Board rejected a certain number of inmates for treatment while still granting parole to these inmates. The reasons for these decisions were not specified by the authors. Thus, we have an initial selection process that differentiates treatment and comparison subjects. As it turned out, based on a composite risk assessment, treatment subjects were at higher risk for recidivism than were comparison subjects. Nevertheless, Parole Board members used their "clinical judgment" to further refine the selection process based on some unknown set of "clinical" criteria.

Also, treatment subjects were sent to halfway houses. There was no indication that comparison subjects were assigned to halfway houses after release from prison; nor was there any

measurement of their level of release supervision (including whether they were tested for drug use). As the authors indicated, in addition to drug treatment, halfway houses fulfill other social service needs and provide assistance in locating employment. Thus, potential differences between the treatment and comparison groups could be attributable to in-prison treatment, halfway-house drug treatment, halfway-house transitional assistance, the drug testing and close supervision of parolees in the treatment plan, or any combination of these factors. Although there does appear to be a selection process operating in the Kyle Unit evaluation, Simpson and his colleagues have described that process more thoroughly than has any other study we reviewed.

A possible, but significant, measurement problem with this study is that the risk sets for the treatment and comparison groups were quite different. Outcome assessment occurred at 6 months and will occur at 12 months after release from prison. However, for treatment subjects, 6 months after release from prison was only 3 months after release from the halfway house. That is, the 6-month risk set for treatment subjects included 3 months of halfway house placement and 3 months of parole supervision, while the risk set for comparison subjects included 6 months of parole supervision. In their future analysis, the risk set for the treatment group will consist of 3 months of halfway house and 9 months of parole supervision, contrasted with 12 months of supervision for the comparison group. Treatment outcomes will be severely biased in the direction of a more positive treatment effect, because halfway house supervision decreases the probability of arrest relative to parole supervision. Thus, differences between treatment and comparison groups may merely reflect differences in the level of supervision and thus level of arrest risk for the two groups, rather than any effect of treatment.

The attrition process for this evaluation was described comprehensively and provides a good indication of how difficult it is to conduct follow-up interviews for this population. Of 482 treatment referrals, 386 (80 percent) graduated; 29 inmates (6 percent) were transferred for medical reasons, outstanding warrants, or inappropriate classification of drug problems); and 67 (14 percent) were terminated for program non-compliance. Unfortunately, no attempt was made to follow up on the program terminations. Also, there was attrition among those who completed the program and those who constituted the control group, because inmates were not available at the time the 6-month follow-up data were collected. By that time, only 222 of the original 386 treatment graduates could be interviewed, and 75 of 121 control group inmates released to parole could be interviewed. Attrition was due to offenders who moved out of the area accessible to interviewers, who were recommitted to prison, who could not be located, or who refused to be interviewed. It is not clear why inmates who were recommitted to prison were not interviewed and did not enter into the outcome results. However, there was an equal percentage of recommitment for the treatment and comparison groups — about 10 percent. Not only was the attrition rate extremely high, there was no attempt to collect follow-up data on the program failures; thus, the results could be severely biased.

It is interesting to note that — at least in a set of univariate comparisons — program terminations and graduates were similar in background characteristics. Program graduates were equivalent to program terminations in terms of age, education, marital status, type of commitment offense, and

recidivism risk score. Whites were more likely to be removed from the program than were African Americans. It would be useful to know whether graduates and dropouts were comparable in a multivariate analysis. One of the limitations of this kind of research is the failure to learn what distinguishes program graduates from program failures. The more we can understand about this process, the better we might be in selecting participants for the program in the first place and the more we will understand the selection process. Further, it will aid us in tailoring programs to meet the individual and group needs of the participants.

Knight et al. (1997) also reported 6-month post-release outcomes without controlling for the many background characteristics they measured. Official Texas arrest records indicated that 7 percent of the treatment group members had been arrested, compared to 16 percent of the comparison group members. Treatment clients self-reported that they had engaged in illegal activities during an average of 11 days in the 6 months since their release from prison, while comparison inmates reported an average of 28 days. In reporting these comparisons, Knight et al.(1996) acknowledged how dissimilar the risk sets were for these two groups. The drug relapse data they reported were problematic for this same reason.

The dissimilarity in risk sets was acknowledged by Knight et al. in 1997, although no adjustments were made to the data. Knight et al. (1997) reported on considerable background data, including information on sociodemographic characteristics, criminal background, drug use history, HIV/AIDS risk behaviors, ratings of social and psychological functioning, ratings of treatment experience, clinical assessments of attention-deficit disorders, hopelessness, depression, and symptom reports. These data should have been analyzed with multivariate techniques.

If we ignore the many methodological problems with this study and assume that at the end of the 12-month post-release arrest period the treatment group had a lower drug relapse and lower criminal recidivism rate, the strongest conclusion we can make is that while offenders are *in treatment*, they are less likely to recidivate and return to drugs. To assess what happens to these offenders *after* treatment, Simpson and his colleagues must follow the treatment and control groups for a period after the outpatient counseling has ended.

Similar to the Wexler, Melnick, et al.'s (1999) study, the most recent 3-year outcome study of the Kyle New Vision program in Texas (Knight et al., 1999), separated in-prison treatment completers into those who either dropped out of or completed aftercare.[12] They attempted to address selection bias by controlling for background severity. Preliminary analyses indicated that the treatment sample had a greater proportion of high severity offenders than did the comparison group. Knight et al. (1999) classified individuals into low and high severity using the Salient

[12] We note that Wexler et al.'s group of treatment completers who did not receive aftercare included those who dropped out of aftercare within 90 days.

Factor Score[13] and conducted separate analyses for these two groups. The findings showed that, controlling for background severity, the in-prison treatment completers who also completed aftercare, had the lowest reincarceration rates. Similar to the previous study (Knight et al. 1997), this study also faced the problem of selection into in-prison treatment. Many of those eligible for treatment were selected into treatment by parole officers. In addition, the Kyle New Vision in-prison treatment subjects were placed in a halfway house for 3 months and comparison subjects were not. The aftercare treatment, not examined in the previous study, was available only to the Kyle New Vision program graduates. Lastly, there was also the self-selection out of aftercare. Although Knight et al. (1999), as did Wexler, Melnick, et al. (1999), recognized the issue of selection into treatment they did not appropriately describe or address the issue.

Texas In-Prison Therapeutic Community (IPTC) Programs

Several reports on in-prison therapeutic community outcomes have been prepared by the Texas Criminal Justice Policy Council staff for the Texas Legislature (Eisenberg and Reed, 1997 &1999). These reports included some of the same subjects – those from the Kyle New Vision program – that were included in the research reported by Knight, Simpson and colleagues. The Kyle New Vision program is the first of the In-Prison Therapeutic Community (IPTC) programs planned by the Texas Legislature in the early 1990's. The reports of the Texas Criminal Justice Policy Council followed all IPTC participants and thus included a larger sample size. The first report provided information on return to prison within 2 years for 1992 program admissions and return to prison within 1 year for 1993 program admissions (Eisenberg and Reed, 1997). The recidivism rate of the IPTC participants was contrasted with that of a comparison group consisting of similar offenders with substance abuse problems (e.g., eligible for treatment) who did not participate in the program. Eisenberg and Reed (1997) found that while treatment graduates had lower recidivism rates than did treatment dropouts, the overall recidivism rate for all participants was very similar to that of the comparison group. For example, 37 percent of the 1992 IPTC admissions returned to prison within two years of release as compared with 38 percent of the comparison group. These results were attributed to the fact that only 42 percent of the 1992 admission cohort completed treatment, which was defined as completing *both* the in-prison treatment and at least 4 months of aftercare treatment. Eisenberg and Reed's (1999) report on 3-year outcomes showed similar results. They found that, after controlling for variables where the treatment and comparison groups differed, there was no difference between the treatment and comparison groups in reincarceration rates. As with their 1997 report, the 1999 report on 3-year outcomes found differences only when separating out treatment graduates from treatment dropouts. Forty-two percent of both the 1992 IPTC admission cohort and the comparison group returned to prison within 3 years.

---

[13] The Salient Factor Score was originally developed by the U.S. Parole Commission to assess the severity of an individual's crime and drug-related problems. It includes nine items, five of which focus upon criminal history. The Texas Department of Criminal Justice version of the Salient Factor Score increases the emphasis on prior criminal involvement.

We do not clearly know the extent to which the approach of Eisenberg and Reed (1997 and 1999) faces selection bias problems because the comparison group is not clearly defined. We note, however, that Eisenberg and Reed's (1999) findings differed from those of Knight et al. (1999). Unlike Knight et al. (1999) their conclusions are based on grouping all treatment participants together, although they do report on differences between treatment completers and dropouts. The difference in the findings of Eisenberg and Reed (1997 & 1999) and Knight et al. (1999) suggest that selection out of treatment can confound treatment effects.[14]

Ozarks Therapeutic Community Program

In addition to the above mentioned studies which are the most commonly cited studies, there are unpublished findings for Ozarks, a prison-based drug treatment program in Missouri (Hartmann et al. 1997). The evaluation of the Ozarks therapeutic community program contrasted treatment subjects who graduated from an in-prison treatment program with comparison subjects who did not volunteer for treatment. The comparison subjects were matched by custody score, education score, health and time remaining on sentence. The follow-up time was between 5 and 12 months. Hartmann et al. (1997) found that the treated group had lower arrest rates than the control group. However, they did not find significant differences in drug use. Although their study matched treatment and comparison subjects on a few factors related to recidivism, as previously mentioned, this does not adequately address the methodological problem of selection bias. They did not address either the issue of selection into treatment or that of selection out of treatment.

Summary of Research Literature

From our close reading of studies evaluating in-prison drug treatment programs and the aftercare programs provided to such participants, we have found fundamental problems in the designs, analyses, and interpretations of results. However, the researchers who have conducted these studies have referred to each other's work as mounting evidence that in-prison drug treatment, especially in combination with post-release community-based treatment, can produce dramatic results. Furthermore, secondary references to many of these studies (*see*, especially, Lipton, 1995) minimize or fail to mention the methodological problems inherent in these studies and, instead, continue to report what appears to be a consistent set of results across different settings.

The clearest finding comes from the earlier 18-month study of the program being evaluated by Inciardi and his colleagues in the state of Delaware (Inciardi et al., 1997). By virtue of random assignment and a comprehensive follow-up of those who dropped out of the transitional care

[14]We also note, as do Knight, Hiller, and Simpson (1999), that comparisons of recidivism and drug relapse rates across studies are complicated by the various types of outcome measures used. Eisenberg and Reed (1999) used incarceration whereas Knight et al. (1999) used arrest for any offense – felony or misdemeanor – as their measure of recidivism.

component of the program, we can have confidence in the finding that offenders receiving transitional care in the absence of in-prison treatment are less likely to recidivate and relapse to drug use. Replication of this finding in other settings by other researchers could be very compelling.

The most recently published studies – 1999 – reflect longer term follow-up periods of many of the same programs previously evaluated but with a focus on aftercare programming. Although the methodology used is more likely to be multivariate in nature, these studies continue to suffer from the methodological problem of selection bias. Aftercare programming is generally provided only to those who participated in in-prison drug treatment[15] and is oftentimes voluntary in nature. The researchers neglect to address issues of both selection into and out of aftercare treatment.

In a 1999 two volume special issue of *The Prison Journal*, more recent analyses of several of these studies were reported. A paper by Martin et al. (1999) updated the Key-Crest program evaluation. Martin et al. (1999) reported a logistic regression assessing the relationship between treatment, background variables used to control for differences among inmates in the treatment groups, and the dependent variable – whether or not offenders had been arrested within three years after release. None of the treatment group effects were significant. Using the same background variables, the researchers examined the impact of the following treatment subgroups: Crest Dropouts, Crest Completers, and Crest Completers who also had a community aftercare component. Only the Crest Completers and Crest Completers with aftercare were less likely to be arrested within three years of release than the comparison group. While Martin et al. (1999) interpreted these findings as implying that a continuum of treatment is necessary to insure drug treatment efficacy, we suggest that it is equally plausible that the results were attributable to selection bias.

Wexler et al. (1999), writing in the same volume, described their most recent analysis of the Amity Right Turn Program. These evaluators found that only those inmates who completed the prison and aftercare programs were less likely to be returned to prison than the control group composed of inmates who volunteered for treatment but who had too little time remaining on their sentence to complete treatment. Similarly, Knight, Simpson, and Hiller (1999) also subdivided treatment groups into Aftercare Dropouts and Aftercare Completers. Only the Aftercare Completers had lower recidivism rates than the control group. In anticipation of critics of this post hoc decomposition of their treatment groups, Knight et al.(1999) argued "Indeed, imposing strict experimental controls on treatment conditions for studying a complex process like treatment and recovery for drug abuse is illusory, and even if possible, the design would potentially restrict the study of the very dynamics that promote recovery." (Knight et al., 1999 P. 349)

---

[15] With the exception of the Key-Crest program evaluation where individuals were randomly assigned to aftercare from among those in a work release program.

While acknowledging that these designs which look at treatment "survivors" may not be as strong as "strict experimental" designs, these evaluators fail to point out exactly why their interpretations of treatment effects are compromised. The context of an experimental design may be the easiest way to demonstrate why the studies that capitalize on selection and attrition provide a potentially biased picture of the effect of substance abuse programming. Suppose we randomly assign drug dependent offenders to one of three treatments depicted in Table A. There is a comparison group that receives no treatment, a "placebo"-treatment group receiving treatment that based on previous research has no impact on post-release outcomes, and a residential drug treatment group. Assuming the attrition rates from the placebo and residential drug treatment groups are the same, we might be in a position to assess the different effects that treatment and merely completing treatment have on post-release outcomes. As depicted in Table A, suppose those who complete the placebo group have more favorable outcomes than the comparison group; that difference represents the effect of completing treatment. Suppose that the residential completers have a more favorable outcome than the placebo completers; that difference gives us a separate estimate of treatment effectiveness over and above merely completing treatment. The hypothetical failure rates we represent in Table A could just as easily have been designed to show most or all of the impact attributable to completing treatment rather than the treatment itself.

While this design does assume that there is no interaction between treatment type and completion (perhaps residential treatment is more intense and has a much higher attrition rate), it at least would provide an explicit attempt at distinguishing the effects of treatment from the effects of completing treatment. Until we use such designs, or we use quasi-experimental designs with adjustments for selection bias, we cannot deduce the effects of treatment separate from the effects of selection biases.

Table A. Hypothetical Failure Rates of Treatment Completers and Dropouts – Percentage Arrested After Three Years.

Treatment Groups	Post-Release Outcomes of Both Treatment Completers and Dropouts (Percentage Arrested)	Post-Release Outcomes of Treatment Completers (Percentage Arrested)
Residential Drug Treatment	30 %	20 %
Placebo Drug Treatment	50 %	40 %
Comparison Group	50 %	50 %

While it is important to study unbiased estimates of treatment effectiveness, it also important to understand the processes that affect treatment retention. Thus, treatment attrition and treatment engagement is also important to study. Hiller, Knight, and Simpson (1999) report on predictors

of dropping out of treatment among probationers assigned to a therapeutic community in lieu of imprisonment. Blankenship, Dansereau, and Simpson (1999) studied motivational readiness techniques among probationers in a mandated residential program.

In the same two volume set of *The Prison Journal*, Pearson and Lipton (1999) have a paper detailing a meta-analysis of drug treatment studies that had been conducted between 1968 and 1997. Recidivism was used to produce the effect size for each study. There were 6 boot camp studies, 7 therapeutic community (TC) evaluations, and 7 group counseling studies. Pearson and Lipton found that the average effect size for boot camps was .05, for TC's .13, and for group counseling .04. Only the TC effect size was considered statistically significant. Using a BESD conversion, this would mean that the TC effect size corresponded to a 56.7 percent success rate for the TC groups versus a 43.4 percent success rate for the comparison group. Mitigating these findings was the fact that most of these studies were rated as having poor or low quality. Among the seven TC studies, 1 was rated as good, 3 as fair, and 3 as poor. A poor quality study meant that the raters thought there was "very low confidence" (Pearson and Lipton, 1999 P. 391) in the research methods. A fair rating meant "low confidence" in the research methods. It is impossible to tell whether poor quality biases the results of these studies in favor of or against finding a significant effect size, or whether there was any potential bias at all. Pearson and Lipton (1999) reported that the lowest quality studies had the lower effect sizes; however, since 6 of the 7 studies had below average quality ratings, there were too few higher quality studies to conclude the direction of bias quality might have on effect size.

In summary, we note that when treatment completers are not separated from dropouts, there is a decreased likelihood of finding a positive effect for treatment. Thus, the consistency in results showing better outcomes for treatment completers, whether they are completers of in-prison and/or aftercare programs, when compared to dropouts, may be a reflection of selection factors confounding treatment effects. This is further complicated by the voluntary nature – e.g., self-selection into treatment – of many treatment programs, both in-prison and aftercare programs. Although the results, excluding those of Eisenberg and Reed (1997 & 1999), are suggestive of effective drug treatment, this may merely reflect the culmination of a selection process that demonstrates that drug treatment — whether in prison or in the community — is a winnowing process. By the end of that process, only those most likely to succeed remain in treatment.

Commentary on Drug Treatment Evaluation Research Methods

Apart from Singer (1986), whose focus is methodologically oriented, and Aiken, Stein and Bentler (1994), the drug treatment evaluation field not only ignores the problem of selection into treatment but also ignores the problem of selection out of treatment. This failure in the drug treatment evaluation field occurs despite the fact that selection bias is a pernicious problem in conducting evaluation research and has been recognized as a methodological problem in program evaluation literature and sociological literature (Berk, 1983; Cook and Campbell, 1979; Heckman, 1979; Kisker and Brown, 1997; Moffitt, 1991; Mohr, 1988; Reichardt and Mark 1998; Rindskopf, 1986; Rosenbaum and Rubin, 1984; Rossi, Freeman and Lipsey, 1999; Shadish,

Cook, and Houts, 1986; Stolzenberg and Relles, 1997; Wainer, 1986; Winship and Mare, 1992). Examples of quasi-experimental evaluations which have addressed selection bias include studies of the effectiveness of mental health programs, spouse abuse prevention programs, Alcoholics Anonymous, youth employment programs, and preschool interventions, as well as studies of clinical trials, studies of the effects of organizational differences of nursing care on mortality and studies of medicine use by the elderly (Aiken, Smith, and Lake, 1994; Berk and Newton, 1985; Fortney et al., 1998; Greenhouse and Meyer, 1991; Grossman and Tierney, 1993; Grotzinger, Stuart and Ahern, 1994; Humphreys, Phibbs, and Moos, 1996; Reynolds and Temple, 1995).

With the exception of a very recent analysis reviewing four large-scale follow-up studies of non-prison based programs (Johnson and Gerstein, 1999), the issue of selection bias is usually not recognized or it is not adequately considered and addressed in evaluations of drug treatment programs. The problem of selection bias is often ignored by dividing subject groups into treatment completers and non-completers to make conclusions about the effectiveness of drug treatment. If one follows the logic of an experimental design, all experimental subjects assigned to treatment are compared to all subjects not assigned to treatment. The treated subjects are not categorized into completers and non-completers because we cannot know whom among the comparison subjects would have completed treatment or whom would not have completed treatment. This logic is often ignored in drug treatment evaluation studies. Furthermore, as pointed out by Harris, even when one has not found any difference between dropouts and program graduates, this does not necessarily confirm the absence of attrition which is problematic for internal validity (Harris, 1998). The common practice of measuring outcomes only for program graduates or comparing graduates to drop-outs can result in misleading conclusions about a program's impact.

Despite the historical lack of attention to the methodological problem of selection bias in the drug treatment evaluation field, recognition of this methodological problem is beginning to surface in the more recent literature. Although inadequate as a method of addressing selection bias, Knight, Simpson, and Hiller (1999) attempted to address selection issues by controlling for severity level and Wexler, Melnick, et al. (1999) by conducting multivariate analyses. Martin et al. (1999) mentioned the need to look at selection into treatment in future studies of correctional treatment programs.

The recognition of methodological problems and the need to approach the positive results of these studies with caution are made apparent in a recent report on comparative costs and benefits of correctional treatment programs (Aos et al., 1999). Five of 17 evaluation studies of therapeutic community prison-based treatment programs were not included in Aos et al.'s (1999) calculation of effect sizes because these studies included program graduates only.[16] Of the studies included in their review the effect size for treatment was significant in 6 of 12 reports. We don't know the extent to which the studies used in Pearson and Lipton's meta-analysis overlap with the studies

[16] The number of treatment programs was less than 17 since some studies represented a longer follow-up time frame than a previous study on the same program.

included in Aos et al.'s (1999) review. The effect sizes reported by Pearson and Lipton (1999) may only refer to people who complete treatment since many of the studies we reviewed focus upon such a contrast.

Landry (1997), in his review of addiction treatment effectiveness, points to the problem of bias due to selective admission, selective participation, and selective detection. He concludes that studies have a potential self-selection bias since they utilize subjects who are more prepared for treatment than the average addicted individual. Ager (1992) discusses the problems of adjusting for selection bias in non-experimental evaluation designs of drug treatment and prevention programs and discusses the use of various alternatives including Heckman's methods. The recognition of selection bias issues in the drug treatment evaluation field has recently been more clearly delineated in the context of evaluation of non-prison based drug treatment. Johnson and Gerstein (1999) conclude that internal validity may be compromised by selection into treatment even when the interest in conclusions is limited to the treatment population. They suggest using Heckman methods to address the problem of selection bias, as we do in this evaluation. In conclusion, we note that there is an increasing recognition of selection bias issues in the drug treatment evaluation field but there are few suggestions on how to address the problem.

In summary, although we found the evidence on drug treatment effectiveness to be less than compelling, after reviewing the recent literature on in-prison therapeutic communities and conceptually examining the processes that lead to subject selection and attrition, we developed a research design that we believed would address and rectify the major methodological problems. We acknowledge that it is extremely difficult to conduct random-assignment research designs in an applied setting. In the absence of random assignment, statistical techniques — such as those we adopted — are technically difficult, depend upon a great many assumptions, and may not always solve the problem. A complete design is presented in Chapter 5.