ASSESSMENT OF IN-PRISON DRUG TREATMENT

*Sheldon Zhang**

I. INTRODUCTION

This paper is written with correctional officials in developing countries in mind, specifically management staff and treatment program administrators who work in prisons or jails with inmates who have substance misuse problems. It is intended for non-academic and non-technical audiences. However, some prior exposure to research methods training will be helpful in understanding the materials. If anyone who wants to read more or explore specific topics within the literature of substance misuse research and treatment, additional resources can be provided.

A. Evaluation Research†

Evaluation research, in a nutshell, is any systematic assessment to determine if something is of value or worthwhile. Just like what the word—"evaluation"—means, it seeks to attach a value to something, be it an object, program, person, service, activity, need, policy, or piece of technology. Its goal is to produce knowledge that can be used immediately, or useful feedback. Evaluation research is not meant to discover generalizable knowledge, but to improve or make better use of something in existence. Also, the word "research" means systematic endeavors to examine or study, not judgment based on anecdotal stories or some haphazard personal observations or subjective assessment.

There is nothing mysterious about evaluation research because it uses all the standard research methods in social science. In the Western world, evaluation research is a booming business and employs many people who specialize in this field and make a living doing it for various government agencies and community organizations. This is because we live in a world where governments have to answer to tax payers about how their money is spent, or foundations or private donors want to account for their investment in social programs, and philanthropists want to know if their money achieved their intended goals. Such evaluative questions require researchers to frame their research questions that can put a relative value (not necessarily in monetary terms) and construct their data collection methods specifically to answer them.

The most fundamental question any evaluation researchers ask about an intervention program (or treatment services) is: does it work? Yes, this is a very simplistic way to describe evaluation research. And evaluation research differs from other types of research. It is intended to generate practical and specific knowledge that seeks to enhance decision making regarding a particular program or a particular type of services. The end goal is improvement of the program being evaluated over time. One unique feature of evaluation research is its tendency towards continuous monitoring and assessment, in which program administrators can provide feedback to guide another round of evaluation, thus making continuous program improvement possible.

B. Why Bother to Evaluate Your Treatment Program?

Why not? There are two ways to think about treatment evaluation. First, you firmly believe what you are doing (your services and treatment protocols) is effective and there is no need to waste any resources on evaluation. In this case, what you are doing is not much different from activities carried out by faith-based organizations, whose work does not require empirical verification. There are actually many social service providers and treatment programs that share such believes. They seek neither validation nor verification for

* Sheldon Zhang is Professor and Chair of the School of Criminology and Justice Studies at the University of Massachusetts Lowell.

† A free online introduction to evaluation research can be found at: http://www.socialresearchmethods.net/kb/evaluation.php; another general overview paper is at: https://www.ideals.illinois.edu/handle/2142/3666.

what they do. This is perfectly fine in today's pluralistic society. These organizations are organizing their social services around their belief system.

Second, if policy makers or government administrators appreciate the complexity of human behavior and maintain a healthy dose of skepticism towards claims made by others, evaluation can be a valuable tool for program implementation and improvement. These program managers typically want to strengthen the quality of their programs and improve outcomes, or simply have a curious mind that wants to see evidence of effectiveness in any particular treatment protocol before making a long-term commitment to resource allocation.

Evaluation by definition seeks to prove if something is working or how effective a particular treatment approach works; and if effective, for whom. Substance misuse is a complex social as well as personal problem. Most treatment approaches work for some but not others, in one context but not in others. Program evaluation answers basic questions about a program's effectiveness, and evaluation data can be used to improve program services.

Frankly speaking, evaluation research can be threatening, as it is often mistaken for auditing. Because evaluation research involves a detailed examination of administrative and financial records to document what has been done and how much money was spent at what stages. Most people do not like to have their work inspected. That's just human nature.

We can have a long conversation about the best ways to plan an evaluation study. It often depends on the organizational culture and the leadership of an agency. In general, it is important to bring in the stakeholders early, preferably at the planning stage of any new initiatives or treatment programs. Without the buy-in from the organizational leadership and stakeholders, any externally imposed evaluation efforts can be easily stalled and even sabotaged.

C. How Much Does Evaluation Usually Cost?

As a rule of thumb, the evaluation component should be about 10% of the total program budget. It is difficult to overemphasize the importance of evaluation, which enables continuous improvement of a program both in its efficacy and cost-effectiveness. It often costs more when a new program is implemented as opposed to an on-going evaluation of an existing treatment program. Oftentimes when a new treatment protocol has a potential for widespread implementation, dedicated funding for evaluation is usually required. In the US, typically government agencies or large private foundations fund these large-scale evaluation studies with wide implications. The 10% mentioned here is just for routine on-going program assessment and feedback for small and incremental improvement.

However, most, if not all, evaluation research literature and researchers themselves will tell you the money devoted to program evaluation is grossly inadequate, particularly in the criminal justice system. Let's just say most government agencies or program administrators do not think it is necessary to spend the money. As discussed earlier, treatment program administrators usually believe what they are doing is helpful and the limited resources should be best devoted to the delivery of services. However, studies time and again have shown that's not the case. Substance misuse is more than just an addiction disorder. It is supported and enabled by a host of complex social and personal factors, making treatment particularly difficult. For instance, naltrexone is an effective blocker to opioid receptors and has been around for three decades. How come such an effective opioid antagonist has not made much dent to the heroin addiction problem in the US or anywhere else in the world? This is because heroin users will do anything to avoid taking the pill or getting the injection. They want to feel the euphoria provided by heroin. Adherence to treatment protocol has remained a challenge. The same goes to methadone and buprenorphine. Without psychosocial interventions and other support (familial and peer), pharmacotherapies are not that effective.

D. Must an Agency Have Professionally Trained Researchers?

No. Evaluation research can be done by trained government staff or prison management. However, this short answer must be contextualized. As long as one has capable research-oriented and well-trained staff, one does not need to spend much money to hire outside consultants or research team to conduct treatment evaluation. Further, one can spend some money to set up an evaluation system with key performance indicators built into routine administration data collection. Routine statistical procedures can also be

established so cost for additional data processing can be kept to minimum.

The key is to establish an assessment-oriented data system to enable on-going feedback for the performance of a treatment program. For large organization with a dedicated research division, this task can be easily accomplished. For small justice agencies or administrative jurisdictions that operate small prisons and jails, the best approach is to team up with others of similar sizes and pool together resources to enable evaluation of treatment programs.

E. Prison Settings Have Unique Advantages for Evaluation Research

One of the main challenges for evaluation research on substance misusers is the attrition problem—study participants drop out at high rates from the program making the remaining sample biased towards motivated subjects. Because of the coercive nature of the prison environment, treatment services can be delivered to the substance misusers more effectively and efficiently, and the follow-up evaluation can be carried out easily also. Inside the prison, inmates' movements are monitored and any disciplinary problems are recorded. If any contraband drugs are suspected of being smuggled into the prison, urine samples can be collected with little resistance. In general, attrition is typically not a problem for evaluation research inside the prison. What is often cited is the tension between service provider/evaluator and prison management. Because of different occupational mandates, prison officials are most concerned about safety and order of the inmates. Lockdowns due to riots, inmate disruptive behaviors, searches for contraband or removal of inmates to different cells can all cause disruptions to treatment service delivery and evaluation activities.

Assessment of in-prison treatment programs inevitably needs to extend beyond the prison walls. Ultimately it is the behavior outside the prison that demonstrates the treatment effects inside the prison. It is quite different out in the community where treatment compliance is often the number one challenge for community-based services. High dropout rates often plague community-based treatment programs. Research has shown that for justice system-involved substance misusers (i.e., drug abusing prisoners), addiction assessment and treatment should begin during incarceration, and prison-based treatment is most effective when aftercare services are followed upon release.

There has been much discussion on how to improve retention both in treatment phase and follow-up phase for this population, such as the use of familial network and social media as venues to keep in touch with study participants. However, study participants may not want to be found if the research team represents the justice agency, or worse, prison officials. Suffice it to say, much planning is needed once inmates are released into the community, and the follow-up phase of the evaluation needs to develop multiple strategies in reaching the study participants.

II. EVALUATION DESIGNS

A. Basics in Evaluation Research

Evaluation research can be divided into two broad categories: *process evaluation* and *outcome evaluation*. Process evaluation focuses on the implementation and operation of a treatment program. Outcome evaluation focuses on the impact of a treatment program. There are ways to further divide evaluation research into different specialty areas. For instance, *cost-benefit* analysis is also commonly associated with outcome evaluation so that we not only look at the impact of a treatment protocol but also whether it is cost effective. A treatment can be effective in bringing about the anticipated reduction on substance misuse, but if it is too expensive, it will be difficult for wide implementation. Furthermore, before an evaluation begins, the researcher typically also examines whether a program is ready for evaluation. This is called *evaluability* assessment. Evaluability refers to the state of a program that has completed (or nearly completed) its intended design and is ready to produce the intended outcomes.

B. Key Performance Indicators

One of the key tasks in evaluation research is the establishment of key performance indicators. There are multiple ways to construct these indicators and most of the time the process is unique to each organization. Here are a few common indicators for one to consider.

Recidivism. This is probably the most important outcome indicator that most, if not all, justice agencies are concerned about. Although services may vary in their treatment goals, orientations or durations, justice

agencies are all concerned if, following the treatment services, these prison inmates will get rearrested and returned to prison. The central question is whether the treatment will result in fewer criminal activities, particularly those related to substance misuse. There are more nuanced ways to look at the recidivism indicators. One can look at the overall re-arrest rates, or how different programs or durations may result in differences in re-arrests.

Recidivism, a central theme in most correctional evaluation studies, can be defined in different, methodologically valid, ways. One can look at re-arrests, irrespective of convictions. One can look at the severity of new arrests. One problem with official arrest records is that most crimes are neither detected nor acted upon by authorities. Official arrest records often show little difference between treatment and comparison groups. This is because within the short follow-up period arrests do not occur enough times for the comparison to be valid. This is not to discredit the use of official data, but to point out the importance of including self-report measures to complement official statistics.

Self-report data can provide much richer information on the spread and frequency of criminal behavior among the offender population. Besides, self-report methods have been shown to be reliable with a remarkable degree of uniformity between self-reported answers and official data. A more recent study of drug dealers that traced self-reports of arrests from interviews through criminal records found about an 80% match between the two data sources. Still, an 80% match still leaves out 20% inaccuracy. Besides, this is only US experience. There is little research in non-English countries on the reliability of self-report criminal data. However, self-report data collection relies on the offender's memory, which fades over time and also fluctuates depending on the intention of the respondent.

For prison management purposes, one can also look at any unfavorable movement of an offender, once released, in and out of parole supervision or back to prison. Depending on the post-release conditions in each country's correctional system, inmates can be released outright without any further monitoring by the justice agencies, or one may be supervised for some length of time after release, such as parole supervision in the US. If inmates are released with supervision condition, then recidivism can be extended to include any condition violations, suspension from parole due to absconding, or newly convicted offenses.

Relapse in substance misuse. For substance abuse treatment programs, this is probably the most important outcome indicator. Relapse is an easy concept to understand but not always easy to measure. The best way and probably the most valid way is to obtain biological samples such urine at predetermined intervals or random schedules. To use biological samples to ascertain one's drug use, a program administrator needs to have access to qualified laboratory facilities and the money to pay for the analysis of bio-samples (urine, hair, saliva, etc.). Although urine analysis has become much cheaper these days, it can still add up if one operates a large prison and aftercare program. Over time, there need to be dedicated resources, staff and money, to capture accurate information on relapses among patients who have gone through the treatment program. These bio-samples also need to be collected frequently over the observation period as some illicit substance passes through the body quickly.

Aside from biological samples, researchers also use self-reports to ask study participants to report their drug use in different time periods, while incarcerated or post release. Self-reports are inexpensive to collect but not very accurate. There are many factors that can influence one's recall accuracy.

Other outcome indicators. Aside from recidivism and relapse, there are other outcome indicators that a researcher should consider. These include prosocial activities, such as job training, gainful employment, school attendance, stable residence, reunification with family and children, and participation in other prosocial activities. Although peripheral to the core mission of the prison management, these are also powerful indicators that can foretell the prognosis of an offender in his reintegration effort. Relapse and recidivism are indicators of particular events while these prosocial activities can reflect a more stable personal growth and improvement in recovery.

Program "effectiveness" means a lot of things to different people; at the minimum it means more than mere measurement in most evaluation research as arrests or parole/probation violations. Therefore, evaluation researchers need to work with stakeholders, treatment participants, and service providers to agree up and develop a set of outcome measures.

C. Data Collection

Data for an evaluation study needs to be planned out up front. For prison management or any other justice agencies, there are typically two major data sources. The first is the official management information system. Typically, a prison data system maintains criminal records and his/her stays in the prison. Upon the entry of the inmate into the prison, the data system also keeps tracks of his/her movements through different quarters, such as substance abuse treatment units or regular inmate units, medical histories, and any disciplinary records. The prison record keeping also contains some criminal history and background demographics about the inmate. Depending on the sophistication and longevity of the prison management information system, the data may be of great value for various evaluation considerations, which allow for both historical (trends over time) and biographical (patterns within individuals) analyses.

The second data source is typically provided by treatment service providers. In the US and many Western countries, substance misuse treatment services are contracted out to particular agencies specialized in treating substance abuse disorder among the criminal population. These treatment providers always maintain records of service utilization. These data can provide information about the numbers of inmates who have used the treatment services, the specific services used, and the outcomes of these service contacts. Once the data sources are identified, an inspection of what data are kept is necessary. Routinely collected data are the easiest place to start planning the evaluation. However, when routine administrative data or service data are not adequate, additional collection must be planned with all stakeholders involved. Prison officials and substance abuse treatment staff are generally reluctant to alter their routine activities and probably will find ways to resist inputting or gathering the data outside their job classification.

Closely associated with planning the data collection is the determination of *observation period*. Usually the longer the observation period the more we can find out about the effects of a treatment protocol on its participants. But there is also a pragmatic side to all evaluation efforts—the amount of money that can support observation over time. As a rule of thumb, 6 months following the exit of a treatment episode is the minimum required for outcome evaluation purposes; typically, one year is needed to examine recidivism rate for prison populations.

D. Common Designs

When someone says a treatment program is effective, one should always ask: compared to what? In other words, there needs to be a comparison of sorts. Without any comparison or contrast, there is no way to tell if something is working. There is quite a bit of science in constructing ways to do such comparisons. The following lists the three most common ones.

1. Pre-and-Post Test

This is probably the easiest evaluation design. Essentially one conducts a baseline assessment of a cohort of prison inmates at the entry of their treatment program. Then at the end of the treatment or a few months following the completion of the treatment, another assessment is conducted to detect any differences on the main outcome indicators.

A pre-and-post evaluation design using official and/or self-report methods is an easy way to enter the evaluation research business. Within the prison environment official records are easy to utilize for evaluation purposes, although self-report data may face challenges in validity and reliability. However, depending on the quality and independence of the research evaluator, there are no insurmountable barriers to collecting self-report data. Self-report data collection is routinely done in the US and other Western countries. However, this may not be the case in other countries.

While easy to understand and implement, pre-and-post test as an evaluation design has many limitations. First, research has shown that patients who participate in treatment tend to be more motivated than non-participants. This selection bias is very difficult to overcome because one cannot tell if the improvement in the end can be attributed to the motivation factor or the treatment effect itself or some other factors. Second, without a comparison, there is no way to tell if the treatment protocol has produced anything better than the status quo (or existing) treatment services or no treatment at all. In other words, the results one obtains from a pre-and-post design stand alone with no external reference.

2. Randomized Controlled Trial (RCT)

The most rigorous evaluation design is the use of randomized controlled trial (RCT). In this design, all eligible patients are randomly assigned to either the treatment group or the control group, hence randomized controlled trial. This is the strongest research evaluation. Its strength lies in the ability of researchers to infer cause – the program caused differences in outcomes rather than preexisting differences. One can infer that groups that are truly equivalent in all aspects going into the interventions have different outcomes only if the interventions have different impacts on the participants.

Randomized experiments are highly valued because they allow for such causal inferences, but they are not infallible. Random assignment must be thoughtfully implemented so that truly randomly equivalent groups are set up, and the equivalence of the groups must be protected over the course of the experiment so that the random equivalence has not eroded before the causal inference can be made.

There are several ways to establish and protect randomly equivalent comparison groups. First, one must *avoid differential consent*. Differential consent can lead to a subtle self-selection bias that corrupts random equivalence from the start. If consent to participate in the new intervention or remain in status quo program takes place *after* random assignment, there can be differences in the characteristics of inmates that consent to the different options. Inmates that refuse to consent to one condition or another may differ in attitudes, prior experiences with substance abuse treatment, or any number of ways not immediately apparent. If inmates who consent to the new treatment protocol tend to be slightly more functional, slightly more in control of their prison activities or more amenable to controlling this time, or different in any number of ways, then the random assignment is defeated. Inmates not consenting to be part of the new program may infuse an element of self-selection that could be related to outcome differences. The inferences that any differences between the new treatment and status quo are due to the program and not preexisting dispositions will be threatened.

Therefore, consent must be obtained *before* random assignment. Those refusing consent will be eliminated from both groups. The comparison groups will be formed from exactly the same pool of consenting inmates. There will be no possibility of differential consent. Consent letters (or instructions) will explain that two approaches are being compared. Inmates will be asked to consent to be randomly assigned to one, and to agree to cooperate with program and evaluation requirements. Consent will be requested at the time of the first assessment.

Second, on must *avoid resentful demoralization of controls*. Consent to be assigned to alternative treatment programs can be accomplished without provoking "resentful demoralization of controls" that occurs in some studies. In this situation, the alternatives can be described evenly and honestly as different approaches without values attached. In other words, it is important not to suggest in any way that one treatment approach is superior to another, or one treatment protocol is more current than the other. Such value-loaded descriptions of any treatment protocols will unwittingly influence inmates' preference to one treatment program over the other. Inmates are asked to consent to cooperate with either approach. In both programs, they will be required to participate fully in all program activities.

Third, one should *avoid instrumentation differences*. Probably the most common threat to validity in a true RCT design is from instrumentation differences that correspond with group assignment. Even when the same measurement tools are employed, if they are "calibrated" differently or applied differently in one group than in another, differences may stem from the measurement process rather than the treatment. For example, a prison guard's judgement of an inmate's substance misuse severity may be very different from that rendered by a professional treatment staff. If the intake assessments for inmates participating in a study are conducted by different sets of personnel using different instruments, the outcomes could easily be interpreted as tainted by instrumentation bias. Therefore, assessment instruments should be the same for both treatment and control groups and carried out by the same trained personnel.

Fourth one should *avoid contamination between treatment and control groups*. Another concern in protecting the random equivalence of the comparison groups has to do with preventing the re-assignment of inmates from one condition to another or interaction between the two groups. This is sometimes difficult to implement and will require commitment from the prison management to ensure both assignment conditions are not crossed. If the inmates can be kept separate physically, i.e., locked up in different facilities, then

contamination can be prevented.

Fifth, whenever possible, one should *strengthen random equivalence by assigning within key strata.* We know from logic, simulations, and mathematical laws that random assignment always results in equivalent groups *if the pool is large enough to overcome the heterogeneity within it.* Overall this will surely be true, but less certain for breakdowns within prisons or service areas, and on key characteristics. Within group distributions on a small number of key characteristics that may strongly affect outcomes need not be left entirely to chance. For example, if marital status has a strong influence on outcomes and we know that married inmates are less common than unmarried inmates, it would be desirable to have roughly equal numbers of married inmates receive each treatment without tainting random assignment. Since service areas (i.e., prison settings) have different populations that may react differently to the treatment options, it would be desirable to have roughly equal numbers of married inmates in treatment and control groups within each of the seven areas. Random assignment is preserved and the chances of differentially skewed distributions (on these key characteristics) within treatment groups reduced, when inmates are randomly assigned to treatments within prison and marital status groupings.

Therefore, once the eligible pool of participants is established, key characteristics known to be strongly related to outcomes need to be identified. If marital status or gender or age of offending onset are key predictors of outcomes, one will need to create sufficient strata, within which random assignment takes place. This approach strengthens statistical tests between randomly equivalent groups, provides opportunities for stronger subgroup analyses (i.e. by prison, by gender, by marital status, or by age of first offense), without in any way compromising the random equivalence of groups and the opportunity to make causal inferences.

Although RCT is an easy concept, there is a lot to be discussed in terms of proper execution of the design. The aforementioned represents a very rigorous form of RCT. Not all prison institutions are set up to do them this way. However, simple random assignment with sufficient participants can still produce results with much greater confidence than any other evaluation designs.

3. Comparison Group and Case Matching

When an RCT design is not possible, researchers often fall back on a quasi-experimental design of using other inmates for comparison purposes. The use of comparison groups is an old strategy in evaluation research. When RCT is not possible, researchers must find something to compare to in order to establish the efficacy of the treatment protocol. There are different ways to construct one's comparison group. The most common method is sometimes called quasi-experimental design, essentially constructing a comparison group using some recruitment criteria. For a long time, case matching (or blocking) is the method, in which researchers select a group of subjects that are the same as the treatment group on some key descriptive characteristics, such as gender, age, prior incarceration, marital status. Using this method, a group of similar inmates is then used to compare with the treatment group.

Prior to computer-assisted statistical analysis, case-matching was often done manually. As one can see, as the number of descriptive variables increase, it becomes very difficult, if not impossible, to do the matching by hand. In recent years, researchers employ a technique called propensity scoring to create a statistically equivalent comparison group in order to detect discernible patterns and treatment effects. Essentially a propensity scoring index is created through statistical analysis that simultaneously consider all known descriptive variables among the non-treatment subjects, thus using all the information available to compare the treatment subjects against this statistically equivalent group.

Propensity score matching offers the most robust alternative to a true randomized controlled trial because of the sophisticated statistical procedures. In a non-randomized, comparative study, the estimated treatment effect is likely to be biased due to confounding variables. This bias is called *the treatment assignment bias.* The best method for eliminating this bias is by random assignment of the treatment as commonly practiced in clinical trials. But such a rigorous design is often impossible or impractical to implement in criminal justice research, especially incarceration settings. The propensity score method is a statistical technique to approximate the randomized assignment design.

Researchers who developed this technique have shown that by matching on the propensity score calculated from multiple confounding variables, the distribution of these variables will be the same in the

treatment and the control group. Just like random assignment, matching on propensity score will balance the two groups on the confounding variables. The major strength of the propensity score method is its dimension reduction capability in the sense that it can achieve multivariate matching by using a single score (i.e., the propensity index). If done properly, researchers have shown the bias reduction of propensity score matching: sub-classification based on the quintiles of the estimated propensity scores can reduce 90% of the bias in the mean difference. In the 1990s, the propensity score method gained wide popularity among social scientists.

One major shortcoming of the case matching method (or propensity score indexing) is that it can only provide some control over descriptive variables (e.g., race, gender, age, and prior incarcerations), known to be related to recidivism or relapse in substance misuse. Some researchers argue that, given sufficient sample sizes, these case matching variables can be easily controlled through multivariate statistical procedures. More importantly, drawing samples (irrespective of the sampling techniques) means some loss of information about the population. Therefore, if one has access to the entire target population, say, the entire substance misuse population in the prison system, one can conduct parametric analysis by using the entire treatment participant population and the entire non-treatment population. By using the population data, assessment of the program, impact should be more precise and stable.

The case-matching method, or using the entire non-treatment population for comparison purposes, still faces the issue of selection bias—the possibility that those who are enrolled in the treatment services are somehow qualitatively different (e.g., either due to self-motivation or favorable prognosis by the prison management staff) from those who do not participate. There are, however, statistical methods to mitigate the problem of selection bias through different weighting schemes. There are also other factors that may mitigate the selection bias. For instance one can examine whether there are any special administrative incentives, perks, or advantages associated with providing the treatment protocol. Second, how the in-prison treatment services are distributed and whether all substance misuse inmates have access to them. If the services are not evenly distributed across all prisons in the jurisdiction, access to these services varies from prison to prison, so many other substance misusers still rely on other forms of treatment or no treatment at all. The comparison population then is made up of inmates who may be equally motivated but unable to access the new treatment protocol. In this case, the assessment of the new treatment protocol compares against the status quo treatment services.

III. CONCLUSION

Evaluation research is important for developing and improving substance misuse treatment in prison and community. Although many treatment strategies have been developed in the West, evaluation research continues to show their many limitations. It is important to point out that evaluation research, particularly those involving criminal justice agencies in the US and many Western countries, has long been plagued by weak designs and poor execution. There is still much room for improvement in terms of evaluation research on substance misuse treatment inside prison or out in the community. Unlike the medical world where research and evaluation are very much the backbone that supports and enhances our improvement in treatment technologies, procedures, and medications, substance misuse treatment inside prison or in the community does not catch much attention from policy makers.

There are two major problems in our assessment of substance misuse treatment inside prison or out in the community. First, design weaknesses in most evaluation research have hampered building knowledge on the effective treatment. Most of what we know about treatment comes from meta-analysis, that is, analyzing a body of literature to detect the treatment effects overall. An important movement that has been building momentum in the US and other Western countries is the use of true experimental trials of intervention programs in the criminal justice system to find out what works. These researchers argue that much of our past research has been plagued by weak research designs. Few question the rigors of randomized controlled trials, but such designs remain the exception rather than the norm in criminal justice research. Few of us in the trenches are ever afforded the luxury to implement a true experimental design to evaluate a correctional program.

Justice agency officials often oppose the "randomized administration of justice" or reject the idea on the basis that it is unethical to withhold services or certain types of treatment from eligible offenders. Evaluation researchers thus often adjust their designs to accommodate the demands or resistance of program

administrators. Oftentimes researchers become involved well after the program has already been in operation or well beyond the point of making any suggestions for programmatic changes to accommodate research activities. As a result, evaluation studies in correctional fields come in all shapes and forms; and findings are often so mixed and even confusing that they make little sense to either the public or policy makers.

Second, aside from the lack of rigorous designs, many evaluation studies are also plagued by small samples or highly localized populations to assess program effectiveness, making generalization to the larger population difficult. The sample size problem arises because correctional programs are often funded for small numbers of offenders in a catchment area either due to limited budgets or for "demonstration" purposes. Although sound in the logic that any large-scale operations should begin with rigorously designed demonstrations, few of these correctional efforts ever survive to see widespread replication in any jurisdiction irrespective of their outcomes. The lack of statistically significant findings (either positive or negative) in many evaluation studies is thus exacerbated by these small samples that are often low in recidivism anyway. Although statistical procedures can mitigate some of these design handicaps, the inherent problem of small samples and small participant populations is intransigent for making impact statements on a system-wide basis.

Finally, evaluation research is what we have been able to advance substance misuse treatment. Through decades of research, we have come to recognize that there are no silver bullets for substance misuse, and effective treatment often consists of multiple strategies, ranging from risk/needs assessment to pharmacotherapy to psychosocial therapies. Moreover, what is effective for one inmate population may not be so effective in a different setting. Appreciating the complexity of the substance misuse problem is the first step towards effective treatment.

BIBLIOGRAPHY

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation. Chicago, IL: Rand McNally.

Copas, Andrew J. and Farewell, Vern T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 161, 385-396.

Cronbach, L.J. (1982). Designing evaluations of educational and social programs. San Francisco, CA: Jossey-Bass.

Cronbach, L.J. (1982). Prudent aspirations for social inquiry. In W.H. Kruskal (Ed.), The social sciences: Their nature and uses (pp. 61-81). Chicago, IL: University of Chicago Press.

D'Agostino, Ralph B., Jr and Rubin, Donald B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95, 749-759

Dehejia, Rajeev H. and Wahba, Sadek (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs
Journal of the American Statistical Association, 94, 1053-1062.

Dennis, M.L. (1990). Assessing the validity of randomized fixed experiments: An example from drug abuse treatment research. *Evaluation Review*, 14(4): 347-373.

Duncan, Kristin Blenk and Stasny, Elizabeth A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27, 121-130.

Hahn, Jinyong (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects (STMA V40 1920). *Econometrica*, 66, 315-331.

Imbens, GW (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706-710

Lu, Bo, Zanutto, Elaine, Hornik, Robert and Rosenbaum, Paul R (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96, 1245-1253.

Powell, R.R. (2006). Evaluation Research: An Overview. *Library Trends* 55(1): 102-120. Available at: http://hdl.handle.net/2142/3666

Rosenbaum, P.R. and Rubin, D.B. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70, 41-55.

Rosenbaum, Paul R. and Rubin, Donald B. 1984. "Reducing bias in observational studies using subclassification on the propensity score." *Journal of the American Statistical Association*, 79, 516-524.

Rubin, Donal B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2, 169-188.

Rubin, Donald B. and Thomas, Neal (2000). Combining propensity score matching with additional adjustments for prognostic covariates *Journal of the American Statistical Association*, 95, 573-585.

Stern, E. (Ed) (2005). *Evaluation Research Methods*. Thousand Oaks, CA: Sage Publications.

Zaccaro, Daniel J., Wolfson, Mark and Preisser, John S. (2000). Use of propensity scores in a non-randomized community trial: Evaluating the enforcing underage drinking laws program. *ASA Proceedings of the Epidemiology Section*, 74-79

American Statistical Association (Alexandria, VA).

Zanutto, Elaine, Lu, Bo and Hornik, Robert (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30, 59-73.