U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

NATIONAL INSTITUTE OF JUSTICE

Conducting Randomized Controlled Trials in State Prisons

Kristofer Bret Bucklen, Ph.D.

June 2020

NIJ.OJP.GOV | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

U.S. Department of Justice
Office of Justice Programs
810 Seventh St. N.W.
Washington, DC 20531

David B. Muhlhausen, Ph.D.
Director, National Institute of Justice

This and other publications and products of the National Institute of Justice can be found at:

National Institute of Justice
Strengthen Science • Advance Justice
NIJ.ojp.gov

Office of Justice Programs
Building Solutions • Supporting Communities • Advancing Justice
OJP.gov

The National Institute of Justice is the research, development, and evaluation agency of the U.S. Department of Justice. NIJ's mission is to advance scientific research, development, and evaluation to enhance the administration of justice and public safety.

The National Institute of Justice is a component of the Office of Justice Programs, which also includes the Bureau of Justice Assistance; the Bureau of Justice Statistics; the Office for Victims of Crime; the Office of Juvenile Justice and Delinquency Prevention; and the Office of Sex Offender Sentencing, Monitoring, Apprehending, Registering, and Tracking.

This paper was prepared with support from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, under contract number GS-00F-219CA. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily represent those of the Department of Justice.

Conducting Randomized Controlled Trials in State Prisons

Introduction

State prisons nationwide house approximately 1,306,000 inmates, which is more than half (57%) of the total population of incarcerated individuals on any given day in the United States.[1] The incarceration rate for those in state prisons increased from 87 sentenced inmates per 100,000 residents in 1970 to a peak of 447 sentenced inmates per 100,000 residents in 2008, more than a 400% increase. Since 2008, the state prison incarceration rate has decreased slightly, to 390 sentenced inmates per 100,000 residents.[2] Approximately 562,000 inmates are admitted to state prisons each year, of which about two-thirds are sentenced by the court and one-third are community supervision violators. Approximately 573,000 inmates are released from state prisons each year, of which nearly 80% are conditionally released to community supervision and the remaining 20% are unconditionally released.[3] The average sentence length for those sentenced to state prisons is 6.4 years, and the average actual length of stay in prison is 2.6 years.[4] Just over half of state prison inmates (55%) were originally committed for a violent offense, 18% for a property offense, 15% for a drug offense, 11% for a public order offense (e.g., weapons), and 1% for some other type of crime.[5] According to a recent national estimate, about 83% of inmates released from state prisons are rearrested within nine years of release.[6]

Program evaluation is essential to ensuring that state prison systems adopt effective programs and policies. The "gold standard" methodology for evaluating outcomes of programs and policies is the randomized controlled trial (RCT). This paper presents an overview of the RCT design as a program evaluation method, describes examples of RCT evaluations in both criminal justice generally and a state prison context specifically, and also discusses considerations and challenges to be addressed when seeking to conduct an RCT evaluation in a state prison. The takeaway from this paper is that, while conducting RCT evaluations can be met with skepticism and challenges in a state prison environment, it is possible to overcome these challenges and to conduct RCT evaluations in state prisons in order to best determine the impact of state prison policies and programs.

The State Prison Context

State prisons are generally designed as a postconviction incarceration sentence for more serious criminal offenders. Virtually all inmates in state prisons are felony offenders and usually serve sentences of one year or longer. State prisons are generally funded by state governments; for the most part, they represent one of the top line items of spending in

a state government's budget. Because inmates generally serve longer periods of time in state prison compared to local jails, inmates in state prisons may have more opportunities to be involved in treatment programming and rehabilitative efforts. State prison systems include an assessment and classification process that inmates must undergo upon admission in order to assess and understand an inmate's specific risks and needs that may need to be addressed during his or her stay in prison. A state prison typically operates like a small town, with common functions such as health care and pharmacy services, educational/vocational services, religious services, housing, food services, security, leisure activities, and treatment programming.

As the number of state prison releases continues to increase each year, prisoner reentry services are also a major emerging component of state prison operations. Since these inmates typically spend longer periods of time incarcerated compared to local jail inmates, there are likely more barriers to their return home, which makes reentry services all the more important. Some state prison systems manage community corrections, including parole supervision and halfway houses. Prison systems also usually include centralized functions such as human resource management, inmate transportation management, records management, staff training, information technology services, internal investigations and intelligence, inspections/audits, legal counsel, and budget/operations management. Some state correctional departments may include a press/communications office, a policy office, a legislative affairs office, and a research and statistics office.

The state prison context provides many opportunities for researchers to become involved in program and policy evaluation. State prison systems are complex organizations. As with any complex organization, research and development should play a crucial role in helping to move state prison systems forward in their mission, goals, and operations. Emerging areas of interest for state prison systems include the use of solitary confinement, drug/contraband interdiction efforts, responding to the opioid crisis, managing staff overtime and improving staff wellness, reentry planning, the use of risk assessment instruments, and implementing treatment programming and educational services that reduce recidivism. All of these areas and others provide opportunities to evaluate, learn about, and improve the operations, management, and outcomes of state prisons.

Causal Inference and the Randomized Controlled Trial

When evaluating the impact of a particular program or policy, an evaluator is typically attempting to draw a causal link between the program or policy (X) and a specific outcome (Y), independent of any other external influences (Z). This is referred to as causal inference. To develop a strong causal link between program or policy X and outcome Y, at least three criteria must be satisfied: (1) X must precede Y temporally, (2) X and Y must covary together, and (3) there can be no other factor Z that explains the relationship between X and Y.

As an example, if the causal impact of a drug treatment program on the rate of drug relapse is being explored, the drug treatment must be provided before the period of time that drug relapse is measured (i.e., temporal order), there must be a relationship between receiving the drug treatment or not and whether or not drug relapse occurs (i.e., covariance), and there cannot be variation in any other factors that affect drug relapse — such as a person's internal motivation to change

before receiving drug treatment (i.e., confounding variables).

In practice, in order to develop a strong causal link between a program or policy and an outcome, an evaluator must identify what is referred to as a "counterfactual." In other words, the evaluator must contrast what actually happened to a subject under treatment with what would have happened to that subject without treatment. Again using the example of a drug treatment program, the evaluator wants to know the rate of drug relapse for a subject who participates in the drug treatment program compared to the rate of drug relapse if that same subject did not participate in the program (i.e., the counterfactual). The immediate problem with this counterfactual logic is that causal effects cannot be calculated for individuals because each individual is observed at any point in time as either receiving the treatment (i.e., the program/policy being evaluated) or not receiving the treatment, but not both. Instead, an evaluator has to estimate the average effect across a population or sample of individuals by comparing a treatment group (those receiving the program or policy) with a comparison group (those not receiving the program or policy).

Developing a credible comparison group (i.e., a counterfactual) is at the heart of establishing the causal impact of a program or policy. While various statistical options exist for attempting to model the counterfactual when evaluating a program or policy, perhaps the best place to start is by describing the strongest design for developing a credible counterfactual. An evaluator's best method for doing this is the experimental design, specifically the RCT. As a method of program/policy evaluation, the RCT has the highest internal validity for establishing a causal link between a specific program/policy and an outcome. For this reason, the RCT is often referred to as the "gold standard." It is not only the strongest design, but it is often the simplest and most intuitive design as well.

In an RCT, the evaluator controls the assignment of subjects to a treatment (i.e., the program or policy being evaluated), and assignment is made by the evaluator at random. Subjects are randomly assigned to a treatment group (those who receive the program or policy) or a control group (those who do not receive the program or policy). Random assignment uses pure chance to form comparable groups. The treatment and control groups are comparable because they are said to be "balanced" in terms of both observable and unobservable/unmeasured characteristics. In the example of the drug treatment program, if eligible participants are first randomly assigned to a group that either receives the treatment program or does not receive the program, with a large enough sample of individuals studied one can be confident that any difference in drug relapse rates between the two groups is completely due to whether or not they received the program or to chance alone. Traditional statistical tests (such as T-tests) can effectively rule out chance alone as an explanation, leaving the evaluator with the strongest possible ability to isolate the impact that the drug treatment program itself has on drug relapse rates. Thus, on average, evaluators can get groups that are essentially identical (in both observable and unobservable factors), by chance, using the RCT design. This is why the RCT design is so appealing.

Why not use a quasi-experimental alternative to the RCT?

For evaluating the causal impact of a program or policy on an outcome, the alternative to an RCT is an observational study, in which the evaluator has no control over the treatment assignment.

The evaluator must rely on a retrospective look at the individuals who have already participated in the program or policy, and try to identify a suitable comparison group of those who did not participate, in order to develop a credible counterfactual. This is difficult because program participants often self-select into treatment, or treatment is assigned based on predetermined criteria. Differing outcomes after the treatment may be due to preexisting differences between the groups as opposed to a causal effect of the treatment itself. This is often referred to as selection bias. There are two categories of selection bias – those biases that the evaluator can observe and those that the evaluator cannot observe.

To use the example of a drug treatment program again, we can assume that decision-makers are not willing or able to conduct an RCT evaluation of the drug treatment program. Instead, they identify those who went through the drug treatment program and were released from prison last year, and attempt to compare them to a group of others who did not participate in the program and were also released from prison last year. But what if those who went into the drug treatment program had a more severe history of drug abuse than those who did not go through the program? It might be expected that the drug treatment group would have a higher rate of drug relapse, not necessarily because the drug treatment program did not work, but because the program was already selecting those with a more severe history and thus a higher likelihood of relapse apart from the program. This would be an example of a potential selection bias on an observable factor (i.e., drug abuse history). However, what if participation in the drug treatment program was voluntary, and individuals already more internally motivated to change and to take opportunities to change were more likely to sign up for the program? In that case, if the evaluator found that relapse rates were lower for drug treatment participants, one would have to ask whether that was because the program caused them to be lower, or because those who volunteered for the program were already more motivated and likely to change even without the program. This would be an example of a potential selection bias on an unobservable factor (i.e., internal motivation to change).

Evaluators have a number of alternative statistical methods available for creating comparable groups with observational data in an attempt to isolate the causal impact of a program or policy from other external factors. These methods are referred to as quasi-experimental designs and include regression modeling, propensity score matching, and instrumental variable modeling. However, none of these methods can produce estimates of the causal impact of a program or policy that are as unbiased and consistent as the estimates produced by an RCT design. In short, they deliver less credible counterfactuals, thus the label "quasi-experimental."

Several characteristics of quasi-experiments make them less than ideal for evaluating the causal impact of a program or policy. First, even with very careful matching, quasi-experimental designs only allow the evaluator to address observed biases. Unobserved biases, by definition, are either not measured or not measurable. Even when sophisticated statistical techniques are used to reduce or eliminate observed bias (e.g., prior drug abuse history), unobserved biases (e.g., prior motivation to change) will not be unaddressed. Only the RCT design can rule out both observed and unobserved biases. In many criminal justice and correctional environments, there are strong selection effects for who receives certain programs or policies, and much of this selection is theoretically unobserved. Quasi-experimental designs are thus less optimal for drawing causal inference when compared to the RCT design.

Second, even if very few or no unobserved selection biases existed, the statistical methods required to address observed biases in good quasi-experimental designs are quite sophisticated and complex, requiring a fair amount of statistical knowledge and experience to perform correctly with a reasonable level of credibility. In general, the results from these quasi-experimental statistical models (e.g., odds ratios) are also more difficult to interpret or communicate to policymakers and other lay consumers of evaluation research. RCTs tend to provide much simpler and more intuitive outputs, such as the difference in the rate of drug relapse between drug treatment participants and nonparticipants. The logic of an RCT design itself tends to be more intuitive to the layperson than the complexities of quasi-experimental methods.

Third, it is fairly well documented that quasi-experimental methods tend to exaggerate the size of the effects found in criminal justice studies, thus producing misleading results of which programs or policies work and by how much. For instance, criminologist David Weisburd and colleagues examined the relationship between research design and study outcomes in 308 studies that were included in a broad review of research evidence on crime and justice commissioned by the National Institute of Justice.[7] They found a systematic effect, where weaker designs (such as quasi-experiments) were more likely than stronger designs (such as RCTs) to report a result in favor of treatment and less likely to report a harmful treatment effect.

Similarly, Brandon Welsh and colleagues reviewed evidence across 136 criminal justice studies and found that weaker evaluation designs were more likely to report desirable effects.[8] This is likely due to the strong selection effect in many criminal justice interventions, which is unobservable and thus missed in many quasi-experimental evaluations but is accounted for in an RCT design. Thus, we should suspect that to the degree that evaluation research in a state prison context relies on quasi-experimental designs, it is likely exaggerating the true impact of programs and policies on outcomes. This may lead policymakers to commit too many resources to programs and policies that do not work or that have very minimal impacts.

A Word of Caution on Evidence-Based Practices

A popular term in criminal justice and correctional settings is evidence-based practices (EBPs). Policymakers have learned that there must be some evidence of the effectiveness of the programs and policies being delivered in state prisons if resources are to be allocated to those programs or policies. It has become popular for policymakers to claim that they are implementing only EBPs.

In theory, this is a good thing. Evidence of effectiveness should be used in part to drive the management of state prison resources, programming, and policies. In practice, however, EBPs are only as good as the quality of the evidence behind them. In some cases, the evidence base for some EBPs is thin, exaggerated, or weak. For example, some so-called EBPs are indeed based on dozens of studies documenting their effectiveness; however, a closer look reveals that most of those studies used weaker quasi-experimental designs with all of the limitations previously mentioned, including exaggerating the true impact. A program or policy cannot be declared an EBP only by virtue of the quantity of evaluations documenting its positive impact. Rather, programs and policies should be judged based on both the quantity and quality of the evaluations.

Adopting programs or policies that have been labeled as EBPs based on research conducted elsewhere can also

be a way to avoid evaluating the impact of a jurisdiction's programs or policies as actually implemented in that jurisdiction. For instance, prison staff may decide to adopt a specific drug treatment program that has received the EBP label due to its positive findings from evaluations in other prisons or jurisdictions, and then claim that there is no need to evaluate it locally because it is already known to be an EBP. The problem with this logic arises when a program or policy labeled as an EBP does not actually work when it is transplanted to a context different from the one in which it was originally evaluated (e.g., in a different prison or among a different population), or when it is implemented differently. Conversely, programs or policies that are not labeled as EBPs may actually be effective in specific environments or prisons, or when implemented in a certain way.

For example, national research has mostly found that correctional boot camps are ineffective in reducing recidivism, which would indicate that this program is not an EBP. However, several evaluations of one correctional boot camp at the Pennsylvania Department of Corrections have consistently found that this program reduces recidivism rates there.[9] This is why it is important not to adopt programs or policies based solely on an EBP label, but to evaluate them at the prison or location where they are actually being delivered — and moreover, to use a rigorous program evaluation design like the RCT. RCTs are the gold standard for internal validity when evaluating the causal effect of a specific program or policy, but they offer no added advantage to addressing external validity. External validity describes the extent to which the results of a particular evaluation can be generalized to other settings or locations. This is again why it is important to evaluate programs locally rather than relying on evaluation results from other jurisdictions.

Common Objections to Conducting RCTs in a State Prison Setting

Several objections are often raised for why an RCT evaluation design should not be used in a state prison setting. This section discusses the most common objections and provides a response to them. The first objection is that it is unethical to assign participants to a program or policy on a random basis. Practitioners will often say they are concerned that if an RCT evaluation is conducted, someone in need of a program will be withheld from that program based on being randomly selected not to receive it. For instance, an inmate soon to be released from prison may be withheld from a new reentry services program because he or she was randomly assigned to a nontreatment control group as part of an RCT evaluation. However, it is important to understand that this type of assertion assumes that solid evidence already exists regarding the program's effectiveness in producing its intended impacts. That is rarely the case in corrections due to some of the limitations mentioned previously, such as a lack of research, weak evaluation designs, and a lack of replications. The real impact of a program or policy is often not truly known. Even if a program was evaluated previously and found to be effective, program implementation may change over time, or the target population and their responses may change over time. Thus, evaluating a program periodically to monitor emerging changes that may impact its outcomes is important. The reason that an RCT evaluation is proposed in the first place is because there is uncertainty as to the impact of the program or policy in question. If it was in fact already known with a reasonable degree of certainty that a program or policy worked, then it would be easier to argue that it is unethical to randomly hold back a control group from receiving the program or policy.

It is important to keep in mind that when faced with uncertainty and a lack of solid evidence as to whether or not a program or policy is effective, there is a possibility that it may actually harm rather than help. There have been examples of criminal justice programs that were based on a solid theoretical underpinning, were intended and expected to produce positive outcomes, but were shown in RCTs to actually make participants worse off. One example is the longitudinal Cambridge-Somerville Youth Study, in which approximately 500 boys in Massachusetts were randomly assigned either to participation in a prevention/mentoring program or to a control group that received no services. The treatment group received family counseling, tutoring, and mentoring services, which were all based on what was considered a solid theory of delinquency/crime prevention. In a 30-year follow-up of this program, Joan McCord found that the treatment group did significantly worse than the control group across at least seven measures, including lifespan, criminal behavior, mental health, physical health, alcoholism, reported job satisfaction, and reported marital satisfaction.[10] This is an example of a well-intentioned intervention that actually ended up harming the participants. If this finding had been known at the time of the study, it would have actually been unethical to assign the boys to receive this program.

Another example is the evaluation of seven Second Chance Act (SCA) adult demonstration programs.[11] In this study, 966 individuals eligible for SCA reentry services were randomly assigned to either a treatment group that received individualized SCA services or a control group whose members could receive already available reentry services but not individualized services. At 30 months after randomization, those in the SCA program group were no less likely to be rearrested, reconvicted, or reincarcerated, and they did not have fewer total days incarcerated (including time in both prisons and jails). They actually had a slightly higher number of rearrests and reconvictions than the control group, although the authors speculate that enhanced case management for the treatment group might have increased the likelihood of detection, rather than an increase in actual crime. The point is that when there is uncertainty about the impact of a program or policy, it cannot be assumed that participants would benefit; in fact, participation may actually be detrimental.

A second objection that is often raised when considering an RCT evaluation in a state prison context is that conducting RCTs is expensive and slow. The RCT design involves the prospective tracking of participants to observe outcomes over time, rather than a retrospective analysis of previous program participants using observational data in a quasi-experimental design. Policymakers sometimes do not want to wait for results from an RCT study to make the policy and programmatic decisions they are currently facing. Not only do they have to wait for results, but they have to provide funding for researchers to design and conduct the evaluation and track study participants over time. In a typical RCT program evaluation model, state prison staff might identify an academic partner to work with in order to conduct an evaluation of a particular program or policy, collaborate on an application for a grant to fund the evaluation, run the experiment if the grant is received, and finally track program outcomes during and after program participation. By the time the evaluation report is written and the findings are circulated to policymakers, several years may have passed. A typical program evaluation takes three to five years. This is far from ideal for policymakers, who work in a rapidly changing environment and must make resource allocation decisions today. By the time results are in, they may no longer be relevant, or interest in continuing the program or policy may be gone. This is a legitimate concern.

Fortunately, RCT evaluations do not necessarily need to follow this traditional model; they do not need to be expensive and slow. Alternative RCT evaluation models are emerging that involve rapid cycle testing and the use of existing staff to minimize cost and time. One example of this new model is an organization at New York University named BetaGov, whose mission is to help policymakers and government agencies identify problems, develop innovative solutions, and test them rapidly using rigorous research methods.[12] BetaGov is a fully funded organization that provides evaluation assistance to government agencies for free, eliminating the cost to the government agency of hiring an evaluator to conduct an RCT. Further, its staff train practitioners and government employees to become directly involved in the experimental process. These trained practitioners are referred to as "pracademics." BetaGov has helped government organizations in criminal justice and other public sectors conduct dozens of rapid RCT evaluations. The typical RCT evaluation at BetaGov is concluded in three to six months. Its evaluation model is drawn from the private sector, which has long relied on simple, pragmatic RCTs to improve efficiency and performance. In Pennsylvania, for example, BetaGov has helped the state department of corrections conduct several rapid trials targeted at reducing inmate violence; some of them have been successful in reducing the department's inmate assault rates in recent years.

The BetaGov model is only one alternative, among others, that can help practitioners who are attempting to conduct relatively low-cost and more operationally driven RCT evaluations. Some important research questions will still require a longer-term evaluation design with a larger investment in time and money, much like the traditional model. But policymakers do not need to wait until the end of a long-term RCT evaluation to make decisions. Evaluation results should be examined incrementally. For example, if a prison is conducting an RCT evaluation of a drug treatment program in which it is proposing to follow program participants for three years after release from prison to examine recidivism rates, prison officials need not wait until the end of the third year to make decisions. If a six-month follow-up is conducted and recidivism rates for the drug treatment group are double that for the control group, decision-makers may decide to discontinue the program — or to revise it, if implementation fidelity is a concern or if the program could be better matched to a theoretical model of success. This is not to say that longer-term follow-ups are unimportant though. Examples exist of programs in which positive outcomes did not materialize until after longer follow-up periods of several years post-intervention.[13]

Considerations When Conducting RCTs in State Prisons

This section contains some practical considerations that may help those who are seeking to conduct RCT evaluations within a state prison setting. First, because prison inmates are protected human subjects, it is important to consider all legal and ethical issues related to human subjects research. If an external evaluator (such as an academic partner) has primary responsibility for conducting an RCT in a prison, human subjects protection coverage is usually provided through oversight by the evaluator's institutional review board (IRB). IRB reviews for RCT evaluations conducted among prison inmates can be intense, burdensome, and time consuming. An alternative option may be to make the case to the IRB that the experiment is being conducted by prison staff to improve internal operations, and the external evaluator is simply being provided data access in order to compile evaluation results and is not actually leading the design and

implementation of the experiment itself. If an RCT is being conducted internally by correctional agency staff, it might not need full IRB review if it does not meet the definition of research.

For IRB review purposes, research is defined as a systematic investigation designed to develop or contribute to generalizable knowledge. Generalizable knowledge means that the evaluation is designed to be applicable and/or shared beyond the population or situation being studied, including publication of results in peer-reviewed journals. Often, however, agency staff are not interested in generalizable knowledge outside of their agency and have no plans to publish the results of their evaluation in a journal or other outlet. They simply want to know if their specific program or policy works, or to understand what they are getting in return for the resources they are committing to a particular program or policy. Internal resource management is the purpose, not research.

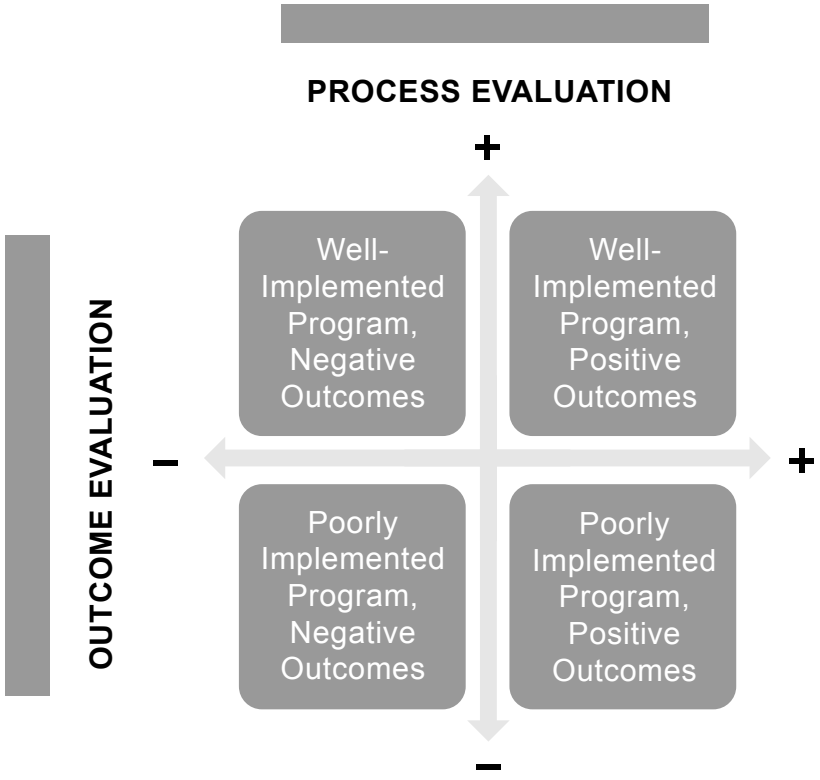
An agency's legal staff should be involved in the decision on whether IRB review is needed and where the IRB review should come from (i.e., from an outside entity, a research partner's university, another government agency, etc.). Legal staff should also be educated in IRB and human subjects protection requirements; often, prison attorneys have received little to no training or education in these areas. The agency should look for training opportunities to educate legal staff as the agency begins to become involved in RCT evaluation work. If agency attorneys are not adequately familiar with these requirements, they will typically provide risk-averse counsel, which may unnecessarily tie up or hamper RCT evaluation work.

Second, the agency needs to foster a culture that promotes learning and where failure is not punished but is seen as an opportunity to improve. BetaGov staff refer to such an organization as a "learning organization." If a learning culture is not already in place in the agency, this can present a challenge, since changing agency culture is difficult and takes time. A prerequisite for establishing a learning organization is having agency leadership who support such learning. Learning organizations greatly facilitate conducting evaluations using an RCT design.

One aspect of a learning organization as it relates to conducting RCT program evaluations is how to respond to negative findings. This is where process/implementation evaluation becomes important. It is generally not advisable to abandon a program or policy based on initial negative results. A more important question to ask is "why are negative results being found here?" It may be that the program was not implemented as intended or with a high degree of fidelity to the program model. In that case, it is easy to understand why negative outcomes occurred. A process/implementation outcome evaluation can help to shed light on whether this is a possible explanation for the negative results. On the other hand, it may be the case that the program itself is simply an ineffective model (or at least ineffective among the target population and/or in the target setting). Exhibit 1 provides a simple matrix for combining the results of a process/implementation evaluation with the results of an outcome evaluation to help make sense of the findings.

If a program is based on a solid theoretical framework, is implemented well and with a high degree of fidelity to the model, and produces positive outcomes, it makes sense intuitively (see top right quadrant in the matrix). It also makes sense intuitively if a program is poorly designed, is poorly implemented, and produces negative outcomes (see bottom left quadrant in the matrix). The matrix oversimplifies the

Exhibit 1. Possible Evaluation Outcomes



results of process and outcome evaluations given that results are more on a continuum rather than discrete, but it highlights how process and outcome evaluations can work together in interpreting results.

The situations that need further investigation are those in which the program is designed and implemented well but produces negative outcomes (top left quadrant), and those in which the program is designed and implemented poorly but produces positive outcomes (bottom right quadrant). If a program is designed and implemented well but produces negative outcomes, the program itself may simply be ineffective (or at least ineffective in its current context). The situation where a program is designed or implemented poorly but produces positive outcomes is unusual. It may be that the program is based on a new model with little evidence or theory behind it, and thus is an innovation without previous evidence that turns out to be effective. It could also be that the program somehow performed well despite poor implementation, and that better implementation might further improve program outcomes. It might also be the case that an unidentified outside factor not considered to be part of the "treatment" was the actual driver of the positive results.

Proper messaging can also help in selling and implementing an RCT evaluation design within a state prison setting. For example, sometimes there are not enough program slots to accommodate the number of inmates who are interested in, or assessed as in need of, a particular program. This is an opportunity to present an RCT evaluation of that program as analogous to a lottery system. When there are more inmates who are interested in or who staff think need a program than can be accommodated, then the fairest way

to assign inmates to those limited slots is through a lottery system with random selection.

A final consideration that may help facilitate the implementation of RCT evaluations in a state prison context is for agency staff to partner with an external expert who understands the technical requirements of an RCT and can assist with the design, implementation, data collection, and reporting. This expert might be an academic partner or an organization like BetaGov. Assistance from an external expert does not need to be expensive, and it can often be obtained for free. As mentioned previously, organizations such as BetaGov are already funded and will provide technical assistance for free. Academics are often looking for opportunities to gain access to state prison data in order to conduct research, and thus will often trade their technical assistance services simply for access to the agency or prison and its data.

RCT Evaluations in Policing

The field of corrections lags behind the field of policing in embracing the use of RCTs to evaluate programs, practices, and policies. RCTs have generally caught on earlier and faster in policing than in corrections. This section provides some early examples of bold policing experiments that should serve to stimulate correctional agencies to further embrace RCT evaluation designs.

One early experiment in policing was the Minneapolis Domestic Violence Experiment (MDVE). Before this evaluation in the early 1980s, domestic violence was viewed as a private family matter that need not involve criminal justice intervention. The MDVE study sought to examine various policies that police were beginning to use in order to reduce domestic violence. The study involved 51 police officers in the Minneapolis Police Department. Based on random selection of which approach to use, each police officer in the study was asked to use one of three different approaches when handling domestic violence calls where probable cause existed that an assault had occurred: (1) send the abuser away for eight hours, (2) advise and mediate the dispute, or (3) arrest the abuser. Again, the specific approach to be used for a given call was selected randomly, making this an RCT evaluation. In this study, arrest was found to be the most effective police response. Abusers who were assigned to be arrested reoffended at a lower rate than abusers who were assigned to mediation or who were temporarily sent away (19% for arrest, 37% for mediation, and 34% for being sent away).[14] This was a landmark study that subsequently led many states and law enforcement agencies to enact policies of mandatory arrest for domestic violence cases.[15]

A second policing experiment was the Kansas City Preventive Patrol Experiment.[16] This RCT was designed to test the assumption that the presence of police officers in marked cars reduced the likelihood of a crime being committed. The experiment took different police beats in Kansas City and randomized them to varying patrol routines: (1) no patrol routine but only reactive calls from residents, (2) a normal level of patrol, or (3) two to three times the normal level of patrol. The study found that the rate at which crime was reported did not vary across the different patrolling routines, nor did citizen perceptions of crime vary across the routines. This groundbreaking study in part moved modern American policing away from random preventive patrolling to more proactive and targeted patrolling.

These examples of policing experiments should motivate the field of corrections to catch up with the policing field in

embracing and conducting RCTs that can move the field forward significantly.

An Example From One State Prison Jurisdiction: Pennsylvania

This section provides an example of one state correctional jurisdiction that has invested significantly in conducting RCT evaluations. Over the past 15 years, the Pennsylvania Department of Corrections (PA DOC) has conducted several program evaluations using an RCT design, including evaluations of a reentry program, a life skills program, a therapeutic community program, a medication-assisted-treatment program for inmates with an opioid use problem, and a post-release community relocation program. Until 2015, PA DOC followed a conventional model for conducting RCT studies. PA DOC research and evaluation staff would identify a specific internal need for evaluating a program or policy, identify an external research partner (typically based in an academic institution) who has interest and expertise in evaluating the particular program or policy, and then work together with the evaluator to identify a source of third-party funding (e.g., a federal grant) for supporting the evaluation. These evaluation projects typically took three years or longer to conclude.

While this model worked well for the department in certain cases, it also suffered limitations noted previously, including being expensive and taking a long time to complete. An example that highlights these limitations to the traditional RCT evaluation model was PA DOC's evaluation of a therapeutic community program at State Correctional Institution — Chester. This was a federally funded evaluation in which prison inmates with a diagnosed substance abuse problem were randomly assigned to participate in either a 12-month intensive therapeutic community program or a much less intensive outpatient program.[17] The department could not afford to wait for the evaluation results before deciding whether to reduce the duration of the program, due to waiting lists and the number of inmates assessed as needing the program. In the interim, the department reduced the duration from 12 months to six months, and then reduced it again to four months. At the end of the evaluation, the finding was that recidivism and relapse rates were no different between the 12-month therapeutic community program and the outpatient program. However, because the program had already changed in the interim, the evaluation was somewhat irrelevant; the department had no evidence as to whether the new four-month therapeutic community program produced any better or worse outcomes than the outpatient program.

In 2015, PA DOC partnered with BetaGov to start using a rapid-cycle model for conducting RCT evaluation and experimentation around three agency goals: (1) reducing in-prison violence, (2) reducing the use of restrictive housing, and (3) improving staff wellness. All staff at every level in the agency were invited to submit ideas for experimenting with new programs, practices, and policies around these three goals. Since 2015, more than 200 trial ideas have been submitted and at least three dozen RCT evaluations have been completed. Trials include ideas such as varying rates of inmate pat searches, visitor notification of the consequences for bringing in contraband, colored bed sheets for inmate bed linens as an alternative to the traditional white bed sheets, aromatherapy, a swift and certain inmate discipline system in response to minor misconduct, an anxiety reduction "chill plan" program for female inmates, use of virtual reality as an incentive for good behavior, the introduction of an intelligence officer staff position at the prison, unit dogs, suicide prevention training, and crisis intervention team training for working with mentally ill inmates.

Most of these trials took three to six months. In many cases, the intervention was tested on one side of a prison housing unit, with a nonintervention control group on the other side of the same prison unit, where inmates are randomly assigned to a bed on one side or the other. Trials are conducted by prison staff, with free support from BetaGov staff. The model has worked so well that in 2018 it was extended to evaluating community corrections and reentry interventions, with the primary goal of finding the best ways to reduce recidivism. This model has allowed PA DOC to rapidly develop broad and strong evidence around what works and what does not work for furthering agency goals. Through this process, PA DOC has developed into a "learning organization," where experimentation is encouraged, strong RCT evaluation designs are promoted, and failure is viewed as an opportunity to learn and improve. To this end, a staff innovations award was created and presented for the first time at the department's annual employee recognition ceremony in May 2019. The department plans to present this award annually to a staff member who shows leadership as a pracademic in developing a new innovation or trial. Standout innovations are also recognized by the department through a podcast they developed. During each podcast episode, a pracademic within the department is interviewed about his or her innovation.

Of course, not all programs will work. In the trials that PA DOC has completed so far, approximately 40% of the interventions were found to work in achieving the desired outcomes of reduced in-prison violence, reduced use of solitary confinement, reduced recidivism, or improved staff wellness. Another 40% were found to be ineffective, producing either no better outcomes or worse outcomes. The remaining 20% could be classified as promising, with some mixed evidence of effectiveness.

PA DOC still uses the traditional RCT evaluation design for larger interventions that take more time to evaluate. Currently, the department is conducting several large-scale RCT evaluations in addition to its BetaGov experiments, such as an evaluation of providing Pell grants for funding in-prison college courses for inmates, an evaluation of a program for teaching inmates financial management skills, and an intervention where released inmates are provided with overdose-reversing naloxone kits before release.

Conclusion

If state corrections professionals are interested in understanding the true causal impact of various policies and programs, the RCT evaluation design provides the strongest model for doing so. State prison programs and policies should be evaluated locally rather than relying on evidence in other jurisdictions. Such evaluations do not need to be expensive or drawn out over long periods of time. Despite common objections, RCT evaluations are also ethical and can be conducted in a state prison setting. The field of corrections has lagged behind other criminal justice fields (such as policing) in embracing RCT designs for evaluating programs and policies, but this can change. State prison departments should commit to fostering a learning organization where the strongest possible evidence is generated for making decisions about what programs, policies, and practices to use or not use. Just as experimentation has progressed in private-sector organizations, experimental evaluations can also help state prisons better achieve their mission and goals.

About the Author

Kristofer Bret Bucklen, Ph.D., is the director of the Office of Planning, Research and Statistics at the Pennsylvania Department of Corrections.

Notes

1. Wendy Sawyer and Peter Wagner, *Mass Incarceration: The Whole Pie 2019* (Northampton, MA: Prison Policy Initiative, 2019), https://www.prisonpolicy.org/reports/pie2019.html.

2. Margaret Werner Cahalan, *Historical Corrections Statistics in the United States, 1850-1984* (Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics, 1986), https://www.bjs.gov/content/pub/pdf/hcsus5084.pdf; and Jennifer Bronson and E. Ann Carson, *Prisoners in 2017* (Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics, 2019), https://www.bjs.gov/content/pub/pdf/p17.pdf.

3. Bronson and Carson, *Prisoners in 2017.*

4. Danielle Kaeble, *Time Served in State Prison, 2016* (Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics, 2018), https://www.bjs.gov/content/pub/pdf/tssp16.pdf.

5. Sawyer and Wagner, *Mass Incarceration.*

6. Mariel Alper, Matthew R. Durose, and Joshua Markman, *2018 Update on Prisoner Recidivism: A 9-Year Follow-Up Period (2005-2014)* (Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics, 2018), https://www.bjs.gov/content/pub/pdf/18upr9yfup0514.pdf.

7. David Weisburd, Cynthia M. Lum, and Anthony Petrosino, "Does Research Design Affect Study Outcomes in Criminal Justice?," *The Annals of the American Academy of Political and Social Sciences* 578 (2001): 50-70, doi:10.1177/000271620157800104.

8. Brandon C. Welsh, Meghan E. Peel, David P. Farrington, Henk Elffers, and Anthony A. Braga, "Research Design Influence on Study Outcomes in Crime and Justice: A Partial Replication with Public Area Surveillance," *Journal of Experimental Criminology* 7 (2011): 183-198, doi:10.1007/s11292-010-9117-1.

9. It should be noted that the multiple evaluations of the PA DOC boot camp program were quasi-experimental evaluations, and thus the results might potentially be influenced by the types of biases in nonexperimental research that are noted in this paper. Kristofer Bret Bucklen, Nicolette Bell, and Joseph Hafer, *Quehanna Motivational Boot Camp 2016 Performance Report* (Commonwealth of Pennsylvania: Pennsylvania Department of Corrections, 2016), https://www.cor.pa.gov/About%20Us/Statistics/Documents/Reports/2016%20Boot%20Camp%20Report.pdf.

10. Joan McCord, "A Thirty-Year Follow-Up of Treatment Effects," *Journal of American Psychology* 3 no. 33 (1978): 284-291, doi:10.1037/0003-066X.33.3.284.

11. Ronald D'Amico and Hui Kim, "Evaluation of Seven Second Chance Act Adult Demonstration Programs: Impact Findings at 30 Months," Final report to the National Institute of Justice, grant number 2010-RY-BX-0003, May 2018, NCJ 251702, https://www.ncjrs.gov/pdffiles1/nij/grants/251702.pdf.

12. BetaGov, http://www.betagov.org/index.html.

13. See, for example, the SVORI long-term outcome evaluation in Pamela K. Lattimore, Kelle Barrick, Alexander Cowell, Debbie Dawes, Danielle Steffey,

Stephen Tueller, and Christy A. Visher, *Prisoner Reentry Services: What Worked for SVORI Evaluation Participants?: Final Report* (Washington, DC: U.S. Department of Justice, National Institute of Justice, February 2012), https://www.rti.org/sites/default/files/resources/svori-final-report.pdf.

14. Lawrence W. Sherman and Richard A. Berk, *The Minneapolis Domestic Violence Experiment* (Washington, DC: Police Foundation, 1984), https://www.policefoundation.org/publication/the-minneapolis-domestic-violence-experiment/.

15. It should be noted that in later replications of this study the evidence showed somewhat mixed results, including possible iatrogenic effects over time. Because of this finding, the lead author has advocated for repealing mandatory arrest laws.

16. George L. Kelling, Tony Pate, Duane Dieckman, and Charles E. Brown, *The Kansas City Preventive Patrol Experiment: A Summary Report* (Arlington, VA: Police Foundation, 1974), https://www.policefoundation.org/publication/the-kansas-city-preventive-patrol-experiment/.

17. Wayne N. Welsh, Gary Zajac, and Kristofer Bret Bucklen, "For Whom Does Prison-Based Drug Treatment Work? Results from a Randomized Experiment," *Journal of Experimental Criminology* 10 no. 2 (2013): 151-177, doi:10.1007/s11292-013-9194-z.

NCJ 254767