Washington State Institute for Public Policy
110 Fifth Avenue SE, Suite 214 • PO Box 40999 • Olympia, WA 98501 • 360.664.9800 • www.wsipp.wa.gov

July 2025

Improving Evaluations of Programs Offered by the Department of Corrections

The Washington State Institute for Public Policy (WSIPP) often receives assignments from the legislature to conduct outcome evaluations of programs offered by the Department of Corrections (DOC). Ideally, outcome evaluations can draw cause-and-effect conclusions that policymakers can use to inform their decisions. However, researchers are sometimes unable to draw such conclusions.

The success of an outcome evaluation depends on a sequence that begins with the design and implementation of the program, continues with the creation of a research assignment, and concludes with the execution of that assignment. The objective of this report is to identify ways that WSIPP, in collaboration with DOC and the legislature, can improve the quality of its outcome evaluation research.

The report is organized in five sections. Section I explores the historical context of research in prison settings. Section II explores different outcome evaluation methods and their prevalence in research involving incarcerated individuals in Washington State. Section III presents an evaluability assessment of DOC programs. Section IV explores administrative practices that could improve the quality of future outcome evaluations. Section V concludes with implications for policymakers.

Summary

Causal outcome evaluations are conducted to quantify program impacts, allowing policymakers to know whether public investments are producing meaningful results. However, researchers are often unable to draw cause-and-effect conclusions.

We find that researchers evaluating DOC programs face challenges regarding incomplete data collection, the lack of quality assurance systems, and the use of subjective criteria for determining program eligibility. It is important to note that these factors do not imply that programs are ineffective, but that they are more challenging to evaluate.

We identify six DOC programs that show promise regarding the use of natural experiments to conduct outcome evaluations.

We also consulted with DOC to identify practices that could improve research quality.

- Additional data tracking and digitization would be less burdensome to implement.
- Data tracking for larger programs would require additional resources.
- Randomized controlled trials are the most rigorous tool for determining program effects. They face legal and ethical hurdles, but perspectives on their appropriateness in prison settings are evolving.

Suggested citation: Gibson, C., Liu, L., & Whichard, C. (2025). *Improving evaluations of programs offered by the Department of Corrections* (Document Number 25-07-1901). Olympia: Washington State Institute for Public Policy.

I. Background

Rehabilitation Programs in State Prison

Adults convicted of serious crimes are often sentenced to serve time in state prison. A noteworthy feature of the prison system is that it is not designed to serve a single goal or objective but is intended to serve multiple purposes. Although this report focuses on *rehabilitation* (i.e., providing assistance to individuals to encourage prosocial behavior), it is important to recognize that prisons can also be designed for retribution, incapacitation, and deterrence.[1]

It is also worth noting that the meaning of the term "rehabilitation" and the methods used to achieve this goal have changed throughout history.[2] For this report, we focus on the current approach to rehabilitation that is used in state prison systems across the country (including Washington State). This approach has been referred to as the *medical model*, which was "founded on the belief that trained experts could administer individualized assessment and treatment that would 'diagnose' and 'treat' the causes of criminality in the way that medical doctors were able to cure other forms of illness."[3]

The medical model of prison rehabilitation can be traced back to the early 1900s, when innovations in psychology and other social sciences popularized the idea that criminal behavior was the result of biological or psychological deficits that required treatment.[4] The defining feature of this perspective is the belief that individuals engage in crime because of internal characteristics (e.g., attitudes, temperament, habits, skills) that can be modified through targeted interventions administered within prison facilities.[5]

The specific practices associated with the medical model of prison rehabilitation have changed over time, but the basic framework remains a dominant force in U.S. corrections. In recent years, this is reflected in the widespread use of the "Risk-Needs-Responsivity" paradigm, risk assessment instruments, and Cognitive Behavioral Therapy in correctional programming.[6] This model also informs how DOC in Washington State approaches prison rehabilitation, which we describe below.

[1] Washington's Sentencing Reform Act of 1981 directly acknowledges retribution, incapacitation, deterrence, and rehabilitation as the rationale for sending people to prison. Sentencing Reform Act of 1981, Wash. Rev. Code § 9.94A.010 et seq. (1981). See also Fallen, D.L. (1993). The evolution of good intentions: A summary of Washington State's sentencing reform. *Federal Sentencing Reporter*, *6*(3), 147-151.

[2] Pifferi, M. (2024). The historical origins and evolution of rehabilitative punishment. *Crime and Justice*, *53*(1).

[3] Phelps, M.S. (2011). Rehabilitation in the punitive era: The gap between rhetoric and reality in US prison programs. *Law & Society Review*, *45*(1), 33-68.

[4] Cullen, F.T., & Gilbert, K.E. (2015). *Reaffirming rehabilitation* (2nd ed.). Routledge.

[5] Critics of the medical model argue that this perspective places too much emphasis on individual-level characteristics and overlooks the role of environmental factors that may contribute to criminal behavior. See Grasso, A. (2017). Broken beyond repair: Rehabilitative penology and American political development. *Political Research Quarterly*, *70*(2), 394-407.

[6] Bonta, J. (2023). The Risk-Need-Responsivity model: 1990 to the present. *HM Inspectorate of Probation* and Fazel, S., Hurton, C., Burghart, M., DeLisi, M., & Yu, R. (2024). An updated evidence synthesis on the Risk-Need-Responsivity (RNR) model: Umbrella review and commentary. *Journal of Criminal Justice*, *92*.

Soon after being admitted to the prison system, incarcerated individuals are assessed using the Washington Offender Needs Evaluation (Washington ONE). The Washington ONE is designed to collect information on "criminogenic needs" (i.e., characteristics that may increase the risk of criminal behavior) and assign individuals scores across eight domains.[7] DOC officials then use the results from the Washington ONE to develop an individualized treatment plan.[8]

For example, if an individual enters the prison system as a result of criminal behavior stemming from a lack of employment, then they might be assessed as "high needs" on the employment domain. They might then be directed to participate in vocational training programs. If these programs are effective, then the individual would be more likely to become employed after leaving prison. Insofar as the individual's criminal behavior was driven by their lack of employment, then they should also be less likely to engage in crime after leaving prison. In theory, this is how rehabilitation programs could (indirectly) reduce recidivism.

Outcome Evaluations of Prison Programs

Rehabilitation programs can be sorted into three broad categories: 1) programs that are effective at achieving their intended goal and have a beneficial impact on participants, such as lowering their risk for recidivism; 2) programs that are not effective and have no discernable impact on participants whatsoever; and 3) programs that are not effective and actually have negative impacts on participants, such as increasing their risk for recidivism. Programs that fall in this third category are known as "iatrogenic," where the intervention inadvertently causes harm.[9] Although individuals responsible for funding and implementing rehabilitation programs may have strong feelings about the value of a particular program, it is difficult to predict whether the effect of a given program will be helpful, neutral, or harmful. To resolve this dilemma, it is necessary to conduct outcome evaluation research.

When conducting outcome evaluations of prison rehabilitation programs, the standard approach is to gather data on individuals who participated in the program (i.e., the treatment group) and on individuals who did not participate but are otherwise similar (i.e., the comparison group). These data will include information on outcomes measured after prison release, such as whether the individual was convicted of a new crime (i.e., recidivism).[10]

[7] The eight domains are labeled as: residential, educational/vocational, employment, social influence, alcohol/drug use, mental health, aggression, and attitudes and beliefs.

[8] Bagdon-Cox, C. & Adams, G. (2023). *Overview of the Washington ONE Risk Assessment Tool*. Department of Corrections, Washington State.

[9] Although the term "iatrogenic" was originally developed by doctors to describe instances where medical treatment has a negative effect on patient health, it has also been used to describe criminal justice interventions (including prison rehabilitation programs) that unintentionally increase antisocial behavior. See Welsh, B.C., Yohros, A., & Zane, S.N. (2020). Understanding iatrogenic effects for evidence-based policy: A review of crime and violence prevention programs. *Aggression and Violent Behavior*, *55*, 101511.

[10] Most outcome evaluations of prison rehabilitation programs will examine recidivism. However, researchers may also examine other outcomes that are relevant for specific programs (e.g., employment/earnings for vocational programs, psychiatric hospitalization for mental health programs).

Next, the researcher will conduct analyses to investigate whether individuals in the treatment group are more or less likely to experience the outcome than individuals in the comparison group. Finally, the researcher will interpret the size and direction of these differences to assess program effectiveness.

Randomized Controlled Trials

Many research design options exist for conducting outcome evaluations. One of the most well-known is the randomized controlled trial (RCT). An overview of RCTs is provided in Section II.

Although RCTs have many advantages over other research designs and are widely regarded as the "gold standard" for conducting outcome evaluations, they are rarely used to evaluate prison rehabilitation programs. [11]

One reason RCTs are rarely used to evaluate prison programs is that research involving incarcerated people is heavily regulated and subject to restrictions that discourage experimental methods. These safeguards emerged from federal legislation in the 1970s that was introduced to protect incarcerated individuals from abusive research practices.[12]

Although it is now widely recognized that incarcerated individuals are vulnerable to coercion and should be afforded special protection as research participants, attitudes regarding the appropriate level of protection continue to evolve.

The authors of a 2007 federal report conclude that research protections for incarcerated individuals should not be overly restrictive, as this can lead to circumstances where incarcerated individuals are unfairly denied "access to the potential benefits that research has to offer."[13] In 2020, a report published by the National Institute of Justice called for more RCTs to be conducted on programs offered in state prisons.[14]

Another reason why RCTs are rarely used to evaluate prison programs is that corrections officials are reluctant to authorize research that involves "denying incarcerated individuals access to potentially beneficial treatment."[15] This occurs because individuals assigned to the comparison group do not participate in the treatment program that is being evaluated.[16] According to this perspective, it is unethical to conduct an RCT because it deprives incarcerated individuals of the opportunity to participate in a program that may help them. As a result, policymakers often express a desire for rigorous evidence but are reluctant to use the research designs most capable of producing it, citing ethical concerns.

---

[11] Bucklen, K.B. (2020). *Conducting randomized controlled trials in state prisons.* National Institute of Justice.

[12] Prior to the 1970s, incarcerated individuals were routinely recruited to participate in dangerous medical experiments. For example, researchers would expose incarcerated individuals to infectious disease in order to study the progression of illness and observe the effects of untested treatments. In many cases, research subjects were offered pardons (i.e., they would be released from prison) as an incentive to participate. See Hornblum, A.M. (1997). They were cheap and available: prisoners as research subjects in twentieth century America. *BMJ*, *315*(7120), 1437-1441.

[13] Pope, A., Vanchieri, C., & Gostin, L.O. (Eds.). (2007). *Ethical considerations for research involving prisoners.* National Academies Press. Page 116.

[14] Bucklen, (2020).

[15] Ibid.

[16] Certain RCT designs, including stepped wedge trials, eventually treat all study participants. See Appendix I for more information.

A concept from medicine helps illuminate this dilemma. *Clinical equipoise* refers to a state of genuine uncertainty within the expert community about whether one treatment is better than another.[17] Such cases have two ethical implications. First, experimental research is permissible because it is not yet clear whether participants would benefit from the intervention. Second, continuing to deliver untested programs may itself raise ethical issues due to the potential for harm[18] or waste.

In medicine, proven treatments often generalize well because they rely on stable biological mechanisms. By contrast, the effectiveness of social programs depends on various factors, including the details of local implementation, social settings, and demographic trends. As a result, equipoise may be even more common in public policy than in clinical medicine.

The principle of equipoise can offer policymakers an ethical framework for conducting research in sensitive settings such as prisons. If there is genuine uncertainty as to whether programs offered in prisons benefit participants, more rigorous evaluation methods may be the most appropriate tool.

Evaluability Assessments of Prison Programs

Due to these limitations on the use of experimental methods, researchers often turn to other methods when conducting outcome evaluations of prison programs. However, since programs are not typically designed with future research in mind, drawing causal conclusions can be difficult.

By the 1970s, there was a growing realization that there were costs associated with evaluating programs that were not yet ready for evaluation. These include the following:

- Waste of research resources,
- Inaccurate research findings,
- Failure to identify changes that could make programs more effective, and
- Funding or defunding programs based on inaccurate findings.[19]

One response to this issue was an increased emphasis on determining whether a program is well-suited to outcome evaluation research.

Evaluability assessment is a type of research that can be done prior to conducting an outcome evaluation. The goal is to determine what types of research are feasible given a program's design and implementation.[20]

[17] Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine, 317* (3), 141–145.

[18] "There have been examples of criminal justice programs that were based on a solid theoretical underpinning, and were intended and expected to produce positive outcomes, but were shown in RCTs to actually make participants worse off." Bucklen (2020).

[19] Van Voorhis, P., & Brown, K. (1997). *Evaluability assessment: A tool for program development in corrections*. University of Cincinnati.

[20] Ibid.

Evaluability assessments can examine a variety of program features, including:

- The program's objectives,
- Whether there is a theoretical link between a program and its goals,
- Whether the program is being delivered according to its design,
- The program's data collection practices, and
- Whether the program meets the requirements of different research designs.

Researchers typically engage with program administrators, collect information using interviews and administrative data, and identify feasible research designs.[21]

Once completed, evaluability assessments can inform a variety of decisions. As one example, policymakers could use the assessment to determine whether it is appropriate to conduct an outcome evaluation at all. If an outcome evaluation is conducted, the evaluability assessment would inform the study's scope and research questions. If not, policymakers could decide whether to implement changes to the program that would improve research options in the future.

[21] Craig, P., & Campbell, M. (2015). *Evaluability assessment: a systematic approach to deciding whether and how to evaluate programmes and policies.*

II. Outcome Evaluation Design

In this section, we explore different outcome evaluation research designs. First, it is important to distinguish non-causal research from causal research. While non-causal research can identify trends or associations, it cannot determine whether outcomes occurred *because of* an intervention. Causal research isolates the impact of a program, allowing legislators to know whether public investments are producing meaningful results.

The fundamental challenge for outcome evaluations is to establish whether a program is causing a change in an outcome. Often, simply comparing outcomes for participants and non-participants can be misleading. When individuals choose whether to participate in a program, there may be unobserved differences between these groups that explain differences in outcomes. This phenomenon is known as *selection bias*.[22] For example, a job training program for incarcerated individuals may show that employment outcomes are better for those who complete the program. However, selection bias may exist if the most motivated individuals signed up for the program. It may have been individuals' motivation that caused the improved outcomes, not the job training itself.

To remove selection bias, researchers have developed various methods, including RCTs and quasi-experimental designs, among others. The purpose of this section is to review these methods. For each method, we present a high-level overview.

At the end of this section, we also discuss some common practical impediments to these designs.

Exhibit 1 presents the basic setup of each research design. Technical details and extensions are provided in Appendix I.

Research Designs

The research designs in this section address the issue of selection bias in different ways. We classify research methods into four categories: experimental, quasi-experimental, selection on observables, and descriptive. The first three aim to establish causal effects, while the last is descriptive in nature. For each design, we describe the basic setup and provide a case study to illustrate how the design works in practice.

Experimental Design

Experimental design is a research process in which researchers determine who is assigned to receive the treatment. Researchers collect relevant information about treated and untreated individuals. Outcomes are then compared across treated and untreated groups to evaluate the effects of a program. The researcher retains relatively strong control over the design and data collection process.

Randomized Controlled Trials (RCTs). RCTs are widely recognized as the most credible design for establishing causal effects.[23] In an RCT, the researcher assigns participants to treatment and comparison groups using a random procedure that does not depend on the characteristics of the participants. This randomness eliminates the issue of selection bias.

[22] In statistical modeling, selection bias is one source of *endogeneity*, which arises when an explanatory variable is correlated with the error term. This violates a key assumption of many statistical methods and results in biased estimates.

[23] Bucklen (2020).

Exhibit 1

Overview of Research Designs

Category	Design	Description
Experimental design	Randomized controlled trials (RCTs)	RCT studies assign participants to treatment and comparison groups based on an experimental protocol and directly compare the average outcomes between the two groups to establish causal effects.
Quasi-experimental design	Instrumental variables (IV)	IV studies take advantage of extraneous factors, referred to as *instruments*, that affect participants' assignment to treatment and comparison groups. Analyzing how changes in the instrument affect outcomes enables the estimation of causal effects.
	Regression discontinuity design (RDD)	RDD studies take advantage of eligibility thresholds that affect participants' assignment to treatment and comparison groups. These studies account for the running variable and compare the average outcomes between the two groups to establish causal effects.
	Difference-in-differences (DID)	DID studies observe individuals over time. These studies compare the change in average outcomes for the treatment group before and after treatment with those in the comparison group in the same timeframe to establish causal effects.
	Synthetic controls (SC)	SC studies compare the average outcomes between a synthetic control group, created by combining several individual groups with similar characteristics to the treatment group, and the treatment group to establish causal effects.
Selection on observables method	Weighting methods	Weighting studies adjust the weights of individual participants to make the treatment and comparison groups similar regarding their average characteristics. These studies compare the average outcomes between the two groups to establish associations that may provide evidence in support of causality.
	Regression adjustment methods	Regression adjustment studies include participants' characteristics in the regression model to mitigate potential differences in outcomes due to differences in observed characteristics. These studies compare adjusted outcomes between the two groups to establish associations that may provide evidence in support of causality.
	Matching methods	Matching studies pair each individual in the treatment group with one or more similar untreated individuals. These studies compare the outcomes between the matched groups to establish associations that may provide evidence in support of causality.
Descriptive method	Comparison methods	Comparison studies compare outcomes for treated and untreated individuals without adjusting for differences in characteristics between groups. These studies establish associations that may not align with causality, but can be useful for understanding trends among different groups of participants.
	Summary statistics	Summary statistic studies summarize the statistical distributions (e.g., averages and variability in averages) of variables of interest to document patterns and trends that may inform outcome evaluation studies.

RCTs can establish causal effects for the groups and settings where they are conducted, but they are not without limitations. The impact of a program may vary across groups and time periods. To address this issue, researchers can conduct randomized experiments across multiple settings or conduct sub-analyses for different groups in the trial.

For example, Blattman et al. (2017)[24] studied the impact of cognitive behavioral therapy (CBT) on crime and violence among high-risk men in Liberia. The study used a lottery system to assign individuals to treatment, ensuring that selection bias was not an issue.

Quasi-Experimental Design

Unlike experimental design, program participation is not randomized in quasi-experimental design. Instead, researchers leverage *natural experiments* to establish causal relationships. Natural experiments are cases where randomness in who received treatment occurred by chance. For example, if different judges refer individuals to a program at different rates, this can create a situation where assignment to treatment is nearly random.

Because researchers typically have less control over the data collection process, studies utilizing quasi-experimental designs may face more data limitations than studies using experimental designs.

Instrumental Variables. Instrumental variable designs can be employed when program participation is affected by an extraneous factor that varies across individuals, referred to as an instrument. This allows researchers to estimate the causal effect, even though participants' treatment choices may be influenced by other factors.

For example, Drake and Aos (2012)[25] examined the impact of confinement on felony recidivism in Washington State. The authors established two key facts: some Community Corrections Officers (CCOs) applied confinement more frequently than others, and individuals were evenly distributed to CCOs based on risk level. These features enabled the authors to identify the causal effect of confinement on recidivism by using CCO assignment as an instrument.

Regression Discontinuity Design. The regression discontinuity design may be used when individuals receive a numeric score, and program eligibility is based on a threshold for that score. Individuals with scores above the threshold are eligible for the treatment, while those with scores below the threshold are not. When this situation arises, researchers compare outcomes for individuals who were just above and just below the threshold. It is important that the score, referred to as a *running variable*, and the threshold are well-defined, measured, and consistently applied.

[24] Blattman, C., Jamison, J.C., & Sheridan, M. (2017). Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia. *American Economic Review*, *107*(4), 1165-1206.

[25] Drake, E. & Aos, S. (2012). *Confinement for technical violations of community supervision: Is there an effect on felony recidivism?* (Doc. No. 12-07-1201). Olympia: Washington State Institute for Public Policy.

For example, Rose and Shem-Tov (2021)[26] studied the effect of incarceration on reoffending in North Carolina. The authors established that state sentencing guidelines changed discretely at specific criminal history score thresholds. This feature enabled the authors to identify the causal effect of incarceration on reoffending by comparing individuals just above and just below these thresholds.

Difference-in-Differences (DID). The DID design is typically used when two groups of individuals are tracked over an extended period of time. One group receives treatment at some point during the study timeframe, while the other group remains untreated. Researchers then compare how outcomes change for the two groups over time.

For a DID design to be valid, participants must not anticipate the treatment in advance and strategically change their behavior in ways that affect their likelihood of receiving the intervention. Additionally, while the DID design can accommodate some self-selection into treatment, any differences in outcomes due to this selection must remain stable over time between the treatment and comparison groups.

For example, Cannonier et al. (2021)[27] examined the impact of a reentry and aftercare program on recidivism in Tennessee. Program participation was voluntary, which introduced selection bias. The authors capitalized on the fact that the trends in recidivism rates for the treated and comparison groups were parallel in the years leading up to the program's implementation. This enabled the authors to use a difference-in-differences design. The authors compared the change in the recidivism rate for the treated group to that of the untreated comparison group, identifying the causal effect of the program on recidivism.

Synthetic Controls. Synthetic control methods are designed to estimate the effects of programs that are implemented at an aggregate level. For example, different states, counties, or facilities may implement various programs. Synthetic control designs compare aggregate outcomes for treated and untreated units.

Synthetic control designs use a weighted combination of untreated units to create a comparison group that resembles the treated units. The method uses a data-driven procedure to minimize differences in observed characteristics and outcomes prior to the intervention period. However, this approach may not be appropriate if the characteristics of the treated unit(s) cannot be adequately approximated by a weighted average of untreated units.[28]

[26] Rose, E.K., & Shem-Tov, Y. (2021). How does incarceration affect reoffending? Estimating the dose-response function. *Journal of Political Economy, 129*(12), 3302-3356.
[27] Cannonier, C., Galloway Burke, M., & Mitchell, E. (2021). The impact of a reentry and aftercare program on recidivism. *The Review of Black Political Economy*, *48*(1), 93-122.
[28] Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495-510.

For example, Lawrence et al. (2022)[29] studied the impact of correctional CCTV cameras on infractions and investigations in Minnesota. Because the treatment was implemented at the prison level rather than at the individual level, the authors created a synthetic control group using a weighted average of multiple prisons. They then compared the outcomes of the treated group to those of the synthetic control group. This method accounted for many, if not all, confounding factors across prisons, enabling causal analysis.

Selection on Observables Methods

Experimental and quasi-experimental methods use randomness in treatment assignment to overcome selection bias. The methods in this category seek to overcome it by accounting for all characteristics that affect treatment and outcomes. When they achieve this, the design can establish causal effects.

The assumption that *all* characteristics that affect treatment and outcomes are accounted for is a critical requirement for these methods, but it cannot be confirmed using data. As a result, relationships identified using these methods are often interpreted as associations rather than as causal relationships.

Regression Adjustment. This method establishes the relationship between two variables by running a regression. The coefficients in a regression model are often interpreted as the associations between two variables.

Weighting and Matching Methods. Weighting and matching methods both function by comparing similar treated and untreated individuals. Weighting methods achieve this by re-weighting individuals so that the treatment and comparison groups are similar to each other overall, while matching methods achieve this by pairing each treated individual with a similar untreated individual.

For example, Cramer and Gibson (2024)[30] investigated the impact of post-secondary education programs while incarcerated on post-release educational outcomes in Washington state. The authors accounted for confounding factors by reweighting the comparison individuals, ensuring that the treatment and comparison groups were similar in the available observable characteristics. This method addressed some, if not all, factors that could introduce selection bias.

Descriptive Methods

In contrast with the methods above, the goal of descriptive research is not to establish causal relationships. Instead, it seeks to provide a clear picture of current conditions, patterns, and practices. This type of research can be especially useful for programs that are large, have evolved piecemeal over time, or that vary significantly across facilities. It can also help identify demographic disparities, challenges faced by program administrators, and areas where more rigorous evaluation would be feasible.

[29] Lawrence, D.S., Peterson, B.E., Robin, L., & Shukla, R. (2022). The impact of correctional CCTV cameras on infractions and investigations: A synthetic control approach to evaluating surveillance system upgrades in a Minnesota prison. *Criminal Justice Policy Review*, *33*(8), 843-869.

[30] Cramer, J., & Gibson, C. (2024). *Postsecondary education programs in Washington prisons: An analysis of post-release outcomes* (Doc. No. 24-10-1902). Olympia: Washington State Institute for Public Policy.

For example, Knoth and Fumia (2021)[31] examined the institutional structure and funding mechanisms of postsecondary education programs for incarcerated individuals in Washington State. The objective of the report was not to establish causal relationships. Instead, the authors provided descriptive findings by analyzing trends in academic progress across racial and ethnic groups and annual enrollment cohorts.

Impediments to Different Research Designs

The research designs discussed above fall along a spectrum in terms of their ability to establish causal relationships. At one end are experimental designs, which represent the strongest opportunity to make causal claims. At the other end are descriptive designs. In general, designs that can make stronger causal claims also have stricter requirements regarding program features and available data. If these requirements are not met, the design may not be feasible.

Understanding impediments to different designs is an important first step in improving future outcome evaluation research. Exhibit 2 presents common issues that can inhibit correctional research.

Research on Incarcerated Individuals in Washington State

The types of research design that are feasible for a given program depend on the nature of the program and the available data. In this section, we assess the prevalence of different research designs in corrections research in Washington State. DOC's Research and Data Analytics (RDA) unit tracks all research that uses DOC data. This includes publications in peer-reviewed journals, government agency reports, and publications by DOC itself. Using RDA's research tracker, we identified 58 studies that analyzed outcomes for incarcerated individuals in Washington State. We then manually coded the research design used in each paper's primary analysis.

Exhibit 3 presents the distribution of research designs used in studies identified in our analysis. Our findings indicate that most of the research is descriptive (66%). Twenty-nine percent of the research uses selection on observables designs like regression and matching, while only 5% uses experimental and quasi-experimental designs. We identified only two studies using RCTs, both of which involved treatment after, rather than during, incarceration, and one study that used a quasi-experimental design. See Appendix I for more information.

[31] Knoth, L., & Fumia, D. (2021). *Postsecondary Program participation and completion patterns among individuals incarcerated in Washington State prisons* (Doc. No. 21-061901). Olympia: Washington State Institute for Public Policy.

Exhibit 2

Common Impediments to Research in Correctional Settings

Unclear definition of treatment: It is not always obvious what it means to be "treated" by a program. The number of sessions, duration, completion status, and type or version of the program can all vary between participants. If information on these factors is unavailable, researchers will need to rely on simplified measures of treatment, which can make it impossible to identify whether different types or amounts of treatment are more effective.

Lack of measurement regarding fidelity to program design: If programs do not track whether they are delivering treatment according to design, there is an increased risk that research findings will create "the impression that nothing worked when, in fact, nothing happened."†

Inconsistent application of eligibility criteria: If eligibility criteria are inconsistently applied, researchers will be unable to recreate the process that is used to assign individuals to treatment. This reduces the credibility of the comparison group and can make some quasi-experimental research designs infeasible.

Lack of data on eligibility: If information used to make eligibility determinations is not recorded, researchers will be unable to construct a comparison group. This issue can arise from the use of informal or subjective eligibility criteria or from inadequate data tracking practices.

Lack of data on control variables: Researchers use data on individual characteristics as control variables to ensure that individuals in the treatment and comparison groups are similar to each other. If relevant variables are unmeasured, researchers cannot ensure that comparisons are valid.

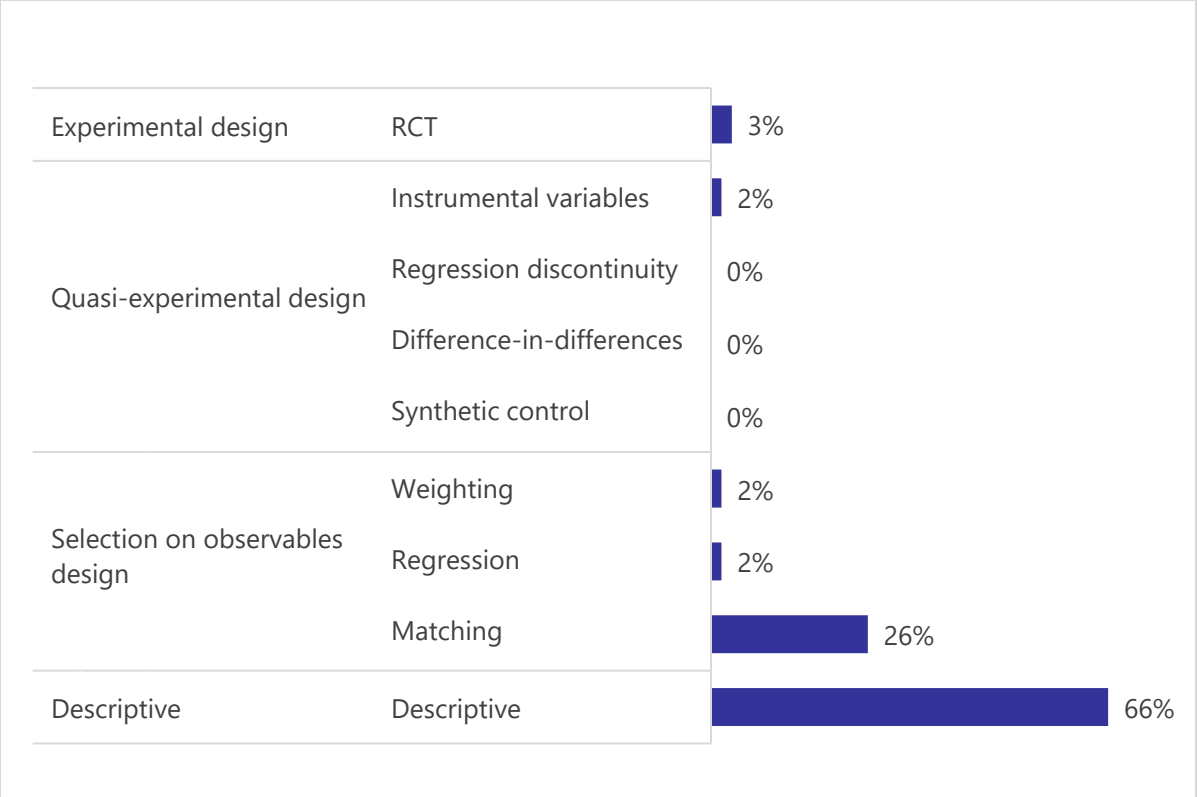
Lack of randomization: Experimental designs, including RCTs, were developed in part to eliminate the issue of selection bias. In the absence of intentional randomization, researchers can use quasi-experimental designs provided some randomization occurs by chance. If this is not the case, researchers are often limited to analyzing associations rather than causal relationships.

External validity: Program effectiveness depends on how the program is delivered, the characteristics of participants, and the social setting outside of prison, all of which change over time. This means that research findings may only apply to the specific settings that are analyzed. Changes between settings or over time may mean previous findings are no longer valid.

Note:
† Van Voorhis, P., & Brown, K. (1997).

Exhibit 3

Distribution of Research Designs Used in Studies of Washington Prisoners



Note:
Includes 58 papers from DOC Research and Data Analytics' article tracker published between 1995 and 2022.

III. Evaluability Assessment of DOC Programs

As discussed in Section I, evaluability assessment is a type of research that determines whether a program is suitable to be studied with an outcome evaluation. This involves determining whether program participation is well-defined and measured, and whether it is feasible to use causal research methods. The purpose of this section is to assess the evaluability of different DOC programs.

Defining Evaluability

To conduct a successful causal outcome evaluation, researchers need to compare outcomes between a *treatment group* and a *comparison group* in a way that overcomes *selection bias*, with a large enough *sample size* to detect program effects. These features represent four dimensions of evaluability that we use to assess current DOC programs.

Identifying the Treatment Group

Researchers need to identify individuals who participated in the program and received treatment. Researchers may want to account for factors such as whether individuals completed the program, how many sessions they attended, and the experience or skill of their program facilitator. If the program is delivered in an inconsistent fashion or if data regarding treatment is not maintained, researchers may be limited in their ability to identify a treatment group.

Identifying the Comparison Group

Researchers also need to identify non-participants who could have participated in the program and who closely resemble individuals in the treatment group. If the program has clear guidelines for determining eligibility, these can be used to identify a comparison group. If the program has no eligibility requirements or if eligibility is determined by subjective or undocumented assessments, researchers may be limited in their ability to identify a comparison group.

Overcoming Selection Bias

Selection bias exists when there are unobserved differences between individuals in the treatment and comparison groups that are also related to outcomes of interest. Researchers can overcome this issue when there is some degree of randomness in who receives treatment. When such natural experiments are present, researchers can leverage them to make causal claims. When they are not, researchers will be more limited in their research design options.

Sufficient Sample Size

In general, it will be easier for researchers to test whether a program is effective when it has a larger sample size.[32]

It is important to note that program evaluability is independent of program effectiveness. It may be the case that a feature that makes a program challenging to study also makes it more effective.

[32] Sample size also gives an indication of the impact and resource demands of a program, which may be relevant for policymakers when prioritizing which programs to evaluate.

For example, it may be difficult to identify a comparison group for a program that has no eligibility requirements. But this feature may also enable the program to serve more participants. Saying that a program is not well-suited to causal research does not imply that it is an ineffective program.

Next, we will identify DOC programs and assess their evaluability along these dimensions.

Programs Included in the Evaluability Assessment

DOC offers a range of programs to incarcerated individuals. These include educational and vocational training, substance abuse treatment, family-centered services, and many others. Because of the large number of programs and because available programming changes over time, we developed a set of criteria for inclusion in the evaluability assessment.

We began with the 60 DOC programs identified in WSIPP's 2024 Adult Corrections Inventory preliminary report.[33] From this set of programs, we included those that:

1) Served at least 30 participants in both 2023 and 2024,
2) Are facilitated by DOC staff or contracted workers, and
3) Are not physical fitness programs.

We identified 19 programs that met these criteria. Exhibit 4 presents programs included in the evaluability assessment. See Appendix I for more information on program inclusion and exclusion decisions.

Methodology

We collected information on program evaluability using survey interviews with DOC staff. Appendix II presents more information on survey design and analysis.

Questions

We developed a set of closed-ended and open-ended questions to solicit information relating to the four dimensions of evaluability described above. Questions addressed eligibility criteria, program waitlists, dosage and completion, quality assurance systems, and whether the program has changed over time.

Recruitment

DOC's RDA provided us with an initial list of program area contacts. Through these, we identified contacts for each program who expressed a willingness and ability to complete our survey. Prior to meeting for full interviews, we held preliminary meetings with each contact to discuss the survey's scope, answer questions, and ensure that contacts had time to gather necessary information.

Interview Format

To ensure that we collected information consistently for all programs, we administered the survey using a computer-assisted personal interview (CAPI) format. This involved WSIPP researchers sharing the interactive survey form with respondents via videoconferencing software. Researchers read aloud each survey question, recorded responses, and confirmed with respondents that the information being recorded matched their intended response. Interviews ranged in duration from 25 to 90 minutes, with a median of 59 minutes.

[33] Goodvin, R., & Wanner, P. (2024). *Inventory of evidence-based, research-based, and promising programs for adult corrections: Preliminary report* (Doc. No.24-03-1901). Olympia: Washington State Institute for Public Policy.

Exhibit 4

DOC Programs Included in Evaluability Assessment

Program	Total participant count, 2015-2024
Cognitive behavioral therapy, SOTAP, and evidence-based programs	
Beyond Violence	494
Moving On	1,628
Sex Offense Treatment and Assessment Programs (SOTAP)	9,529
Thinking 4 a Change (T4C)	12,978
Education	
Basic skills (ABE, GED, ESL, HS)	49,740
Postsecondary education	41,869
Family-centered services	
Strength in Families	4,217
Mental health & life skills	
Skill Building Unit (SBU)	6,193
TBI Pilot-to-Program	192
Substance treatment	
Intensive outpatient	8,741
Therapeutic communities	8,627
Vocation	
Construction Trades Apprenticeship Preparation (CTAP)	1,633
Correctional Industries	186,459
DNR correctional camps	14,187
Sustainability in Prisons Project (SPP)	20,726
Vocational education	43,304
Wellness	
Getting it Right	1,296
Intensive Transition Program	685
Stress Anger Management	1,226

Notes:
Total participant counts are based on DOC Offender Management Network Information (OMNI) records from 2015-2024. Participants are counted once per program per year but may be counted in more than one year.
Categories are based on program classifications in OMNI records.
Some program names differ from those on WSIPP's Adult Corrections Inventory report. See Appendix II for a full list of programs and inclusion and exclusion decisions.

Analysis

We used data collected in the survey to create ratings for each program on each of the four dimensions of evaluability. We rated each program as "limited," "moderate," or "strong" on each dimension.

Using these ratings, we identified what research designs would likely be feasible for each program. Because the requirements regarding program features and data collection are generally stricter for causal research designs than for non-causal designs, our findings are hierarchical. For example, a program identified as potentially compatible with quasi-experimental designs would also be compatible with selection on observables and descriptive designs.

Results

First, we discuss some common themes that emerged in the interviews. Second, we provide a summary of evaluability for the 19 DOC programs we evaluated.

Common Themes

Identifying the Treatment Group. We began by assessing whether treatment was standardized and whether programs had the ability to track variation in treatment dosage. We identified several common strengths in this area. Most programs feature a standard duration and have a clear completion process. Those that do not tend to be in program areas like educational, vocational, and job programs. Most programs also track the number of sessions and completion status for participants through DOC's Offender Management Network Information (OMNI) system, along with information on who facilitated each session or sequence of the program.

The most common limitation we identified in this area is that most programs do not have a quality assurance system in place. This means that they currently do not record information on the extent to which the program is being delivered as designed, nor do they evaluate their facilitators for their fidelity to program design. This presents a challenge to researchers, as differences in treatment and facilitator quality are not measured. Additionally, programs that do track information on facilitators often use separate data software or paper records that would be onerous for researchers to use.

Identifying the Comparison Group. Next, we assessed the extent to which researchers could use DOC administrative data to construct a valid comparison group for each program. The most common strength we identified is that a majority of programs have at least some eligibility criteria that researchers could use as a basis for constructing a comparison group. Some of these criteria, like court orders, case manager referrals, Washington ONE scores, custody levels, and time to release, are measured in readily available data systems like OMNI.

However, a majority of programs base their eligibility determinations partly on screenings that involve subjective judgment from DOC staff, are not recorded, or both. This presents a challenge to researchers because the information in these screenings is also likely relevant to participant outcomes. For example, some respondents indicated that screenings assessed motivation, curiosity, and desire for change. This suggests that researchers would not be able to recreate the process DOC uses to assign individuals to treatment.

Additionally, a majority of programs do not track information about individuals on waitlists. This limits researchers' ability to use waitlisted individuals as a credible comparison group.

Overcoming Selection Bias. We identified nine natural experiments that could arise in a correctional program setting. We assessed whether each program had a key feature required for each natural experiment. The most promising cases were

1) The use of a first-come, first-served waitlist, which could enable researchers to identify a highly credible comparison group,
2) The use of court-ordered program participation, which could enable researchers to use an instrumental variables design, and
3) The use of a numeric assessment score cutoff for program eligibility, which could enable researchers to use a regression discontinuity design.

We identified eight programs that could be compatible with at least one of these three cases. The remaining 11 programs had features that could make them compatible with one of the other six less promising cases.

It is important to note that the success of these opportunities is not guaranteed. We identified whether each program has a key feature that is required for each design. However, it may be the case that on closer inspection, the design is infeasible for other reasons. A cautious interpretation of this finding is that we identified eight programs for which we cannot *rule out* the use of promising quasi-experimental designs.

Sample Size. Finally, we assessed whether each program had a sample size large enough to detect effects on recidivism with a high probability.[34] We determined that evaluations of 14 programs could detect a 5-percentage-point reduction in recidivism, seven of which could detect a 2-percentage-point reduction.

Summarizing DOC Program Evaluability
Exhibit 5 presents our combined evaluability assessment findings. We identified six programs that have the potential to be compatible with quasi-experimental research designs, eight that would likely be limited to selection on observables designs, and five that would be most well-suited to descriptive designs.

We discuss our rankings for each category in more detail in Appendix II.

To show what evaluability looks like in practice, we have selected three programs that illustrate how the different dimensions of evaluability interact to determine the types of research that are feasible.

Program with Strong Evaluability. Moving On is a cognitive behavioral change program for incarcerated women. This program features clear, measurable, and consistently implemented eligibility criteria, which makes identifying treatment and comparison groups feasible. The eligibility criteria utilize cutoff values based on scores on the Washington ONE, which could enable the use of a regression discontinuity design.

[34] In the context of an outcome evaluation, "statistical power" refers to the probability that a statistical test will detect a program's effect, assuming that an effect exists. See Appendix II for more information on power analysis.

Additionally, the sample size is sufficient to detect a moderate effect on recidivism with reasonable statistical power.

Program with Moderate Evaluability.
Intensive Outpatient is a type of substance abuse treatment program. This program features clear, measurable, and consistently implemented eligibility criteria. It also has a sample size sufficient to detect a moderate effect on recidivism. However, we did not identify any sources of randomness in assignment to treatment that would enable the use of quasi-experimental research designs. An outcome evaluation of this program would likely be limited to analyzing associations rather than causal relationships. This illustrates that even if a program is strong on some dimensions of evaluability, a limitation in another dimension may affect the types of research that are feasible.

Program with Limited Evaluability
Basic skills (ABE, GED, ESL, HS) are a set of educational services offered in all DOC prison facilities. This program allows anyone to participate, and the type and number of classes taken vary widely among participants. As a result, selection bias would remain an impediment to drawing any causal conclusions, and researchers would face challenges in constructing treatment and comparison groups for this program.

However, this program has a high participant count. When combined with the complexity of defining treatment, this program would be well-suited to descriptive research. Researchers could explore trends in who participates, common course sequences, and reasons for completing or not completing programs.

This illustrates that in some cases, features that make a program less well-suited to causal research make it more well-suited to descriptive research.

Limitations

It is important to reiterate that evaluability is not synonymous with effectiveness. The ratings in this section should not be interpreted as measures of program quality. Programs may have features that help them reach their goals, but that also present challenges to researchers.

It is also important to reiterate that evaluability in this context refers specifically to whether a program is well-suited to a causal outcome evaluation. Programs rated as having limited evaluability may be well-suited to descriptive research. Descriptive research can illuminate differences in program operations across facilities, demographic trends among participants, and can serve as the foundation for future outcome evaluations.

The measures of evaluability in this section are based on interviews with contacts who are directly involved in administering and facilitating DOC programs, but they should be treated as preliminary. Understanding program design and data collection practices in sufficient detail to conduct an outcome evaluation requires time and resources beyond the scope of this report.

Our analysis also does not identify opportunities for conducting experimental research. Current program administration and data collection practices do not support this type of research. We explore future opportunities regarding experimental research in the next section.

Nevertheless, the analysis in this section could also serve as a framework for thinking about evaluability in other areas of public policy, including juvenile justice programs, non-prison criminal justice programs, and other social service programs.

Exhibit 5

Evaluability Assessment Findings

DOC Program	Identifying treatment group	Identifying comparison group	Overcoming selection bias	Sample size	Feasible design
Moving On	Strong	Strong	Strong	Moderate	Quasi-experimental
Therapeutic communities	Strong	Strong	Strong	Moderate	Quasi-experimental
Thinking 4 a Change (T4C)	Strong	Moderate	Strong	Strong	Quasi-experimental
Beyond Violence	Strong	Strong	Strong	Limited	Quasi-experimental
Strength in Families	Strong	Moderate	Strong	Moderate	Quasi-experimental
Sex Offense Treatment and Assessment Programs (SOTAP)	Strong	Limited	Strong	Moderate	Quasi-experimental
Vocational education	Moderate	Moderate	Strong	Strong	Selection on observables
Intensive outpatient	Strong	Strong	Limited	Moderate	Selection on observables
Postsecondary education	Moderate	Moderate	Moderate	Strong	Selection on observables
Construction Trades Apprenticeship Preparation (CTAP)	Moderate	Moderate	Moderate	Moderate	Selection on observables
Intensive Transition Program	Moderate	Moderate	Moderate	Limited	Selection on observables
TBI Pilot-to-Program	Moderate	Moderate	Moderate	Limited	Selection on observables
Getting it Right	Moderate	Moderate	Limited	Limited	Selection on observables
Stress Anger Management	Moderate	Moderate	Limited	Limited	Selection on observables
DNR correctional camps	Limited	Moderate	Strong	Strong	Descriptive
Basic skills (ABE, GED, ESL, HS)	Moderate	Limited	Limited	Strong	Descriptive
Skill Building Unit (SBU)	Limited	Strong	Limited	Moderate	Descriptive
Sustainability in Prisons Project (SPP)	Moderate	Limited	Limited	Strong	Descriptive
Correctional Industries	Limited	Limited	Limited	Strong	Descriptive

IV. Opportunities for Improving Evaluability

The previous section assessed the evaluability of DOC programs as they currently operate. This section explores opportunities that could improve evaluability in the future.

During our interviews with DOC program contacts, we asked about nine hypothetical practices that could improve program evaluability. We asked respondents to rate how burdensome each practice would be to implement for their program, taking into consideration additional resource needs and legal or ethical issues.

These practices were not posed as recommendations, but as hypotheticals. During interviews, we reiterated that whether a program implements each practice does not imply anything about the effectiveness of the program.

After completing interviews with program contacts, we analyzed responses for trends. We grouped practices into three categories: low-burden practices, practices that would require additional resources to implement, and practices that would raise legal or ethical issues. Exhibit 6 presents our overall findings.

Low-Burden Practices

We identified practices that respondents said would not be burdensome to implement. These practices represent opportunities to improve program evaluability with minimal disruption to program operations.

Several respondents rated data collection and retention practices as being low burden to implement. These include digitizing facilitator data that is currently maintained using paper records, tracking data on participants who are on the waitlist for the program, and tracking data that program administrators use when making eligibility determinations. These practices could help researchers better identify individuals in treatment and comparison groups. Several respondents indicated that the reason they do not already have these practices in place is that they do not serve a purpose related to DOC's core mission, but that they are not opposed to implementing them in the future.

Three respondents indicated that implementing a quality assurance system that measures and tracks facilitator performance over time would not be burdensome. However, two of these respondents represented comparatively small programs. Representatives of larger programs tended to rate this practice as more burdensome.

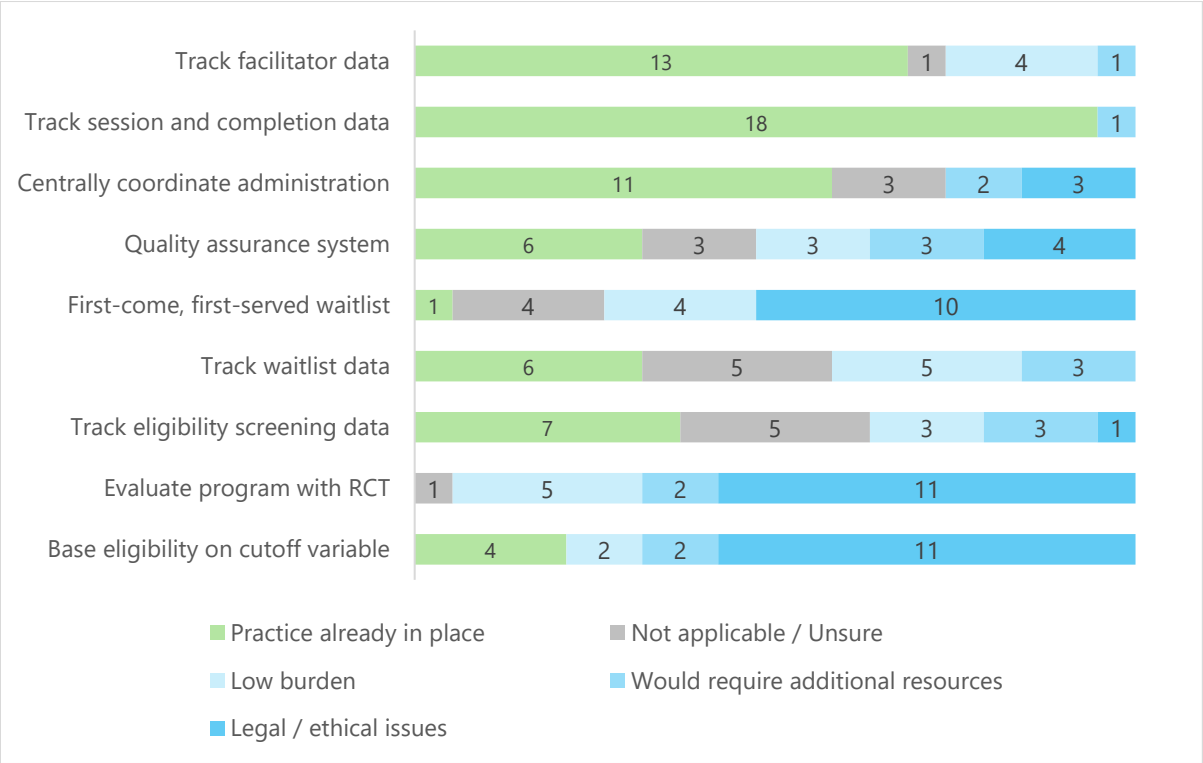
Practices Requiring Additional Resources

We also identified practices that respondents said were feasible, but that would require additional resources to implement.

Representatives of programs that are run in collaboration with organizations outside of DOC indicated that tracking participant waitlist data would raise challenges related to inter-agency data sharing. This practice reflects the additional resource requirements that arise for programs that involve collaboration with other organizations.

Exhibit 6

DOC Perspectives on Practices That Could Improve Evaluability



Practices Raising Legal or Ethical Issues

We also identified practices that respondents indicated could raise legal or ethical issues.

A representative for a mental health program noted that tracking medical data used by program administrators to make eligibility determinations could raise privacy issues.

Representatives of programs that are run in collaboration with other organizations indicated that some practices would be inappropriate given DOC's role in those arrangements. These included implementing a quality assurance system and centrally coordinating program operations.

Several respondents indicated that practices related to quasi-experimental research designs would be inappropriate for their programs. Respondents indicated that changing waitlists to be first-come, first-served would be illegal in some cases and would limit DOC's ability to prioritize participants for treatment based on individual needs. Respondents also indicated that modifying eligibility criteria to enable the use of a regression discontinuity design would be inappropriate given the goals of the program and the needs of participants.

Perspectives on Randomized Controlled Trials

We asked respondents for their perspectives on conducting randomized controlled trials of their program. We indicated that this would involve some eligible individuals receiving treatment, and others not receiving treatment. This practice elicited the broadest range of responses out of the practices we discussed.

A majority of respondents said that conducting an RCT would be burdensome for their program. This was largely due to legal and ethical considerations around withholding potentially beneficial treatment from some individuals. It was also partly due to concerns about the additional administrative demands of setting up randomization and data tracking processes.

However, several respondents indicated that conducting an RCT would not be burdensome. One respondent noted that if there is genuine uncertainty as to whether the program is effective or not, then an RCT would be the most effective way to investigate this.

These perspectives highlight the dilemma faced by policymakers and practitioners discussed in Section I.

Summarizing Opportunities

We identified program administration and data collection practices that could improve future program evaluability and collected perspectives on how burdensome they would be to implement for different DOC programs.

Practices that respondents rated as low burden tended to relate to data collection and tracking. In most cases, respondents stated that they did not view the practices as necessary for fulfilling their professional responsibilities, but they were not opposed to implementing them.

Practices that required additional resources to implement typically involved tracking data and conducting quality assurance for programs run in collaboration with organizations outside of DOC, as well as for program areas with a large number of sub-programs.

Some practices raised legal and ethical concerns, including data privacy regarding medical information, the need to honor inter-agency agreements regarding DOC's role in program operation and quality control, and not implementing eligibility criteria that are inconsistent with program design.

The use of randomized controlled trials, in particular, raised a range of perspectives. Several respondents expressed reservations about the legal and ethical implications of withholding treatment from eligible individuals. Others expressed openness to the idea because of the potential benefits of causal research for incarcerated individuals. This range of opinions reflects the recent evolution of thinking on the use of experimental methods in evaluating prison programs.

V. Conclusion

Prison programming is one tool that DOC uses to achieve the goal of rehabilitating incarcerated individuals. Outcome evaluations of DOC programs can help policymakers understand whether public investments are producing meaningful results. The purpose of this report has been to explore opportunities to improve the quality of WSIPP's research in this area.

To conduct a successful causal outcome evaluation, researchers must be able to identify program participants, similar non-participants, and overcome the issue of selection bias. Through interviews with DOC program contacts, we identified practices that could improve evaluability.

To help researchers better identify participants and non-participants, DOC could implement quality assurance systems, use objective criteria when making eligibility determinations, and maintain accessible data related to both of these practices.

Several options exist to overcome the issue of selection bias. Maintaining additional data on individual characteristics could help researchers control for more factors that affect outcomes, although this alone would likely not enable causal research.

Implementing practices related to natural experiments could enable the use of quasi-experimental research designs, but this is not feasible or appropriate for most programs. Randomized controlled trials are the "gold standard" for conducting causal outcome evaluations, but their use has been rare in corrections settings. There has been growing recognition that it is possible to conduct RCTs without violating ethical standards.

Research quality can also be improved by ensuring that research questions are appropriate to the program being studied. This involves collaboration between legislators, who know what questions they want answered, program administrators, who understand program operations, and researchers, who understand how to connect the two. Even when programs are not well-suited to causal research, descriptive research can answer meaningful questions.

Finally, researchers can improve the quality of their studies by understanding both the challenges and the opportunities presented by different programs. This requires an understanding of program structure, data systems, eligibility criteria, and awareness of natural experiments that could enable causal analysis.

Appendices

Improving Evaluations of Programs Offered by the Department of Corrections

Appendices

I. Technical Details of Research Designs

In this section, we provide technical information on the research designs discussed in the main text. First, we introduce a common requirement that each research design must satisfy to establish causal relationships. Next, we summarize the technical details of each research design and method, including the basic setup and extensions.

Stable Unit Treatment Value Assumption (SUTVA)

The stable unit treatment value assumption (SUTVA) is a fundamental requirement that every research design must satisfy. There are two pieces to this assumption.

1) No interference between units. That is, the outcomes of one unit do not depend on the treatment or outcomes of another unit.
2) No hidden variations in treatment. That is, the treatment should be consistent across all units.

Randomized Experiments and RCTs

Fundamentals

We begin by introducing completely randomized experiments, while other forms of randomized experiments will be covered in the discussion of extensions.

Setup. Experimental studies assign participants to treatment and comparison groups based on an experimental protocol and directly compare the average outcomes between the two groups to establish causal effects.
In completely randomized experiments, the treatment assignment for these subjects relies on a randomized procedure, in which some subjects are assigned to the treatment group to receive the intervention, while the remaining subjects are assigned to the comparison group to either receive no treatment or a placebo.

Estimand. The average treatment effect (ATE).

Identifying Assumptions/Requirements. There are two main requirements: the randomization assumption and the minimum sample size requirement.

- The randomization assumption: The treatment assignment is a randomized procedure (e.g., coin toss or computer-generated random number) that is unrelated to individual characteristics and, specifically, unrelated to potential outcomes. In other words, this method requires that the observed treatment is statistically independent of the potential outcomes.
    - In practice, it is common to conduct balance tests on the available covariates to test the randomization assumption.
- The minimum sample size requirement: The sample size must be large enough to detect the effect. The minimum sample size required to detect a pre-defined treatment effect size can be determined via power calculations. This can be calculated given a chosen statistical significance level (e.g., 0.05) and a predetermined level of power (e.g., 80%) against a given alternative hypothesis (e.g., ATE ≠ 0).

Estimation and Inference. Estimation can be conducted using differences in sample averages between treatment and comparison groups or through regression estimators.

Extensions

There are several extensions or variations to this design. Additional details on these extensions can be found in other studies.[35] Here, we briefly present some scenarios that may be encountered.

Randomized Experiments with Pre-Determined Covariates. There are two main reasons for considering the addition of covariates: 1) incorporating covariates may enhance the informativeness of analyses (e.g., precise inferences), and 2) adjusting for covariate differences may mitigate biases that arise when randomization is compromised (e.g., missing data).

Stratified and Paired Randomized Experiments. In a stratified randomized experiment, the population is first partitioned into multiple mutually exclusive strata based on relevant covariates. Then, subjects are randomly assigned to treatment and comparison groups within each stratum. This design aims to enhance the study's efficiency by preventing assignments that are likely to yield uninformative results.

Clustered Randomized Experiments. In a clustered randomized experiment, the population is first partitioned into multiple mutually exclusive clusters (e.g., schools or regions) based on the relevant covariates. The treatment is then randomly assigned to entire clusters, with all subjects within a given cluster receiving the same level of treatment. This design is motivated by the fact that interactions may exist between subjects or that the cost of randomizing at the cluster level is lower than the cost of randomizing at the individual level.

Stepped Wedge Trials. In stepped wedge trials, the population is first partitioned as in a clustered randomized experiment. All clusters are initially untreated and are then treated sequentially, so that by the end of the experiment, all clusters have been treated. This design is notable for not withholding treatment from any subjects, although this may present limitations in studies of prison programs where the outcomes of interest are often not observed until after release.[36]

---

[35] Athey, S., & Imbens, G.W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73-140). North-Holland.

[36] Hussey, M.A., & Hughes, J.P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, *28*(2), 182-191.

Instrumental Variables (IV)

Fundamentals.

IV studies take advantage of extraneous factors, referred to as instruments, that affect participants' assignment to treatment and comparison groups.

Setup. There is a source of exogenous variation (the "instrument") that results in a subset of the population being treated while leaving the remainder untreated. This variation does not affect the outcome, only the treatment. The treatment effect is assumed to be constant across participants.

Estimand. The local average treatment effect (LATE).

Identifying Assumptions/Requirements. There are two main requirements: the relevance assumption and the exclusion assumption.

- The relevance assumption: The instrument must be statistically significantly correlated with the endogenous treatment variable. In other words, the changes in the instrument should result in changes in the treatment variable.
- The exclusion assumption: The instrument affects the outcome of interest solely through its impact on the treatment variable. In other words, there should be no alternative channels through which the instrument affects the outcome of interest.

Estimation and Inference. Estimation is typically done using the two-stage least squares (2SLS) procedure.

Extensions

Heterogeneous Treatment Effects. The estimator introduced in Imbens and Angrist (1994)[37] allows treatment effects to vary across participants. The estimand is the local average treatment effect (LATE) rather than the ATE. This estimator has two additional requirements: the monotonicity assumption and the independence assumption.

- Monotonicity assumption: The instrumental variable should affect the probability of consistency receiving the treatment in a direction (e.g., weakly increasing) for all participants. In other words, the instrumental variable should have a uniform effect on the likelihood of treatment across different participants. For example, if an examiner is more likely to allow participants to join a program, this higher likelihood must apply to every potential individual assigned to her, compared to her counterparts who are stricter about accepting participants into the program.
- Independence assumption: The instrumental variable is independent of both potential outcomes and potential treatment assignments—that is, the instrument is as if randomly assigned.

Multiple Instrumental Variables. Refer to Mogstad et al. (2021)[38] to see a discussion about the causal interpretation of 2SLS in the case where there are multiple instruments.

Classical Selection Model. Refer to Gronau (1974)[39] and Heckman and Vytlacil (1998, 1999).[40]

---

[37] Imbens, G.W., & Angrist, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.
[38] Mogstad, M., Torgovitsky, A., & Walters, C.R. (2021). The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review*, *111*(11), 3663-3698.
[39] Gronau, R. (1974). Wage comparisons—A selectivity bias. *Journal of Political Economy*, *82*(6), 1119-1143.
[40] Heckman, J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources*, 974-987.

Regression Discontinuity Design (RDD)

Fundamentals

The regression discontinuity design applies when individuals receive a numeric score, and program eligibility is based on a threshold for that score. Individuals with scores above the threshold are eligible for the treatment, while those with scores below the threshold are not. When this situation arises, researchers can compare outcomes for individuals who were above and below the threshold.

Setup. There are three key components in the RD design: a score, a cutoff, and a discontinuous treatment assignment rule. Units (e.g., individuals, regions) in the study have a score value, with only those scoring above the cutoff being assigned to the treatment. A "sharp" design occurs when the treatment received perfectly matches the treatment assigned for all individuals, whereas a "fuzzy" design applies when there is an imperfect match between treatment received and treatment assigned for at least some individuals.

Estimand. The local average treatment effect (LATE).

Identifying Assumptions/Requirements. There are two main requirements: continuity and monotonicity.

- Continuity: The expected potential outcome is smooth around the cutoff point. In other words, individuals who have similar scores will have similar outcomes in the absence of treatment.
- Monotonicity: Fuzzy designs additionally require that participants do not "defy" their assigned treatment. An individual who defies their treatment is one who receives treatment if and only if they are ineligible to receive it.

External Validity. RD designs identify treatment effects only for individuals within a narrow window around the cutoff. Generalizing these effects to individuals outside of this window is challenging. Additional details on this issue can be found in other studies.[41]

Estimation and Inference. Estimation is typically performed using local polynomial methods, which involve selecting the polynomial order, kernel function, and bandwidth.

Extensions

The Local Randomization Framework. In the local randomization framework, the RD design is interpreted as a randomized experiment near the score cutoff value. Additional details of this approach can be found in other studies.[42]

Heckman, J.J., & Vytlacil, E.J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, *96*(8), 4730-4734.

[41] Wing, C., & Cook, T.D. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, *32*(4), 853-877; Cerulli, G., Dong, Y., Lewbel, A., & Poulsen, A. (2017). Testing stability of regression discontinuity models. In *Regression Discontinuity Designs* (Vol. 38, pp. 317-339); Angrist, J.D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331-1344; and Bertanha, M., & Imbens, G.W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, *38*(3), 593-612.

[42] Cattaneo, M.D., Frandsen, B.R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, *3*(1), 1-24.

Multidimensional RD Designs. In some cases, treatment may depend on multiple scores and cutoffs, and the treatment can take on multiple values. Additional details on multidimensional scores,[43] geographic cutoffs,[44] and multiple cutoffs on a single score[45] can be found in other studies.

Difference-in-Differences (DID)

Fundamentals

The DID design applies when two groups of individuals are tracked over an extended period of time. One group receives treatment at some point during the study timeframe, while the other group remains untreated. Researchers then compare how outcomes change for the two groups over time.

Setup. There are two time periods and two groups of individuals. The individuals in the treated group receive treatment in the second time period, while the individuals in the comparison group remain untreated in both time periods.

Estimand. The average treatment effect on the treated (ATT).

Identifying Assumptions/Requirements. There are two main requirements: the no-anticipation assumption and the parallel trends assumption.

- The no-anticipation assumption: There is no causal effect of participating in the treatment during the pre-treatment periods. In other words, individuals do not change their behavior in anticipation of upcoming treatment.
- The parallel trends assumption: The average outcome for the treated and untreated populations would have evolved in parallel in the absence of treatment. This assumption allows for the presence of some forms of selection bias due to non-random selection into treatment, provided that the selection bias remains constant over time.

Estimation and Inference. Estimation is most commonly done using difference-in-means estimators or two-way fixed effects (TWFE) regression estimators. Details on the use of clustering methods can be found in other studies.[46]

[43] Papay, J.P., Willett, J.B., & Murnane, R.J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, *161*(2), 203-207.
[44] Keele, LJ., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*(1), 127-155 and Keele, L., Titiunik, R., & Zubizarreta, J.R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(1), 223-239.
[45] Dong, Y., Lee, Y.Y., & Gou, M. (2023). Regression discontinuity designs with a continuous treatment. *Journal of the American Statistical Association*, *118*(541), 208-221 and Caetano, C., Caetano, G., & Carlos Escanciano, J. (2023). Regression discontinuity design with multivalued treatments. *Journal of Applied Econometrics*, *38*(6), 840-856.
[46] Liang, K.Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22 and Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249-275.

Extensions

Variation in Treatment Timing. Commonly, individuals do not all receive treatment at the same time. If treatment effects are assumed not to vary across individuals, estimation can proceed as above. Details on estimation under heterogeneous treatment effects can be found in other studies. [47]

Treatment is Not an Absorbing State. If treatment is an "absorbing state," this means that once individuals are treated, they remain treated throughout the study window. This contrasts with a situation where individuals may enter and exit treatment. Details on estimation in this situation can be found in other studies.[48]

Continuous Treatment Variable. In some cases, treatment is discrete (for example, when an individual either participates in a program or does not). In other cases, treatment may be continuous. Details on estimation in this situation can be found in other studies. [49]

Synthetic Controls (SC)

Fundamentals

Synthetic controls are designed to estimate the effects of programs that are implemented at an aggregate level. For example, when a state, county, or facility implements a program.

Setup. In synthetic control designs, there are two types of units: one treated unit and a set of untreated units. Pre-treatment and post-treatment outcomes are observed for both types of units, along with additional variables that could affect the outcome of interest. A comparison group is constructed using a weighted combination of untreated units that closely resembles the treated unit.

Estimand. The average treatment effect on the treated (ATT).

Identifying Assumptions/Requirements. The primary requirement is that it is possible to create a weighted combination of untreated units that resembles the treated unit. If the untreated units are too dissimilar to the treated unit, this may not be possible. The key constraint is that weights assigned to untreated units be nonnegative and sum to one. This ensures that estimates will not extrapolate beyond the available data.

Estimation and Inference. After constructing the comparison group, the observed outcome for the treated unit is compared to the weighted average of the outcomes of the units that comprise the comparison group. For a discussion of permutation methods for inference, see Abadie, Diamond, and Hainmueller (2010). [50]

---

[47] Callaway, B., & Sant'Anna, P.H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200-230; Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies*, *91*(6), 3253-3285; and Deb, P., Norton, E.C., Wooldridge, J.M., & Zabel, J.E. (2024). *A Flexible, Heterogeneous Treatment Effects Difference-in-Differences Estimator for Repeated Cross-Sections* (No. w33026). National Bureau of Economic Research.

[48] De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964-2996.

[49] Callaway, B., Goodman-Bacon, A., & Sant'Anna, P.H. (2024). *Difference-in-differences with a continuous treatment* (No. w32117). National Bureau of Economic Research.

[50] Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*(490), 493-505.

Extensions

Multiple Treated Units. Synthetic control methods were originally developed for cases with a single treated unit. Other studies have extended these methods to cases involving more than one treated unit. [51]

Selection on Observables Methods

Experimental and quasi-experimental methods use randomness in treatment assignment to overcome selection bias and other confounding issues. Selection on observables methods seek to overcome it by accounting for all observed characteristics that affect treatment and outcomes. When all relevant characteristics are controlled for, these methods can establish causal effects or provide evidence to support causal effects.

The methods in this category all feature two main requirements: unconfoundedness and overlap.

- Unconfoundedness: Treatment status is as good as random once individuals' relevant characteristics have been controlled for. The assumption that all characteristics affecting treatment and outcomes are accounted for is a critical requirement for these methods, but it cannot be confirmed entirely using the available data. As a result, relationships identified using these methods are often interpreted as associations rather than as causal relationships.
- Overlap: For every unit in the treatment group, there should be units in the comparison group with similar characteristics, and vice versa. This ensures comparability between treated and untreated individuals.

Regression Adjustment Methods

Setup. Information is collected on treatment status, outcome of interest, and relevant pre-treatment variables. The outcome of interest is modelled as a function of treatment status and other observed variables.

Estimation and Inference. Estimation is typically done using ordinary least squares regression or likelihood maximization methods.

Weighting & Matching Methods

Setup. Weighting and matching methods both function by comparing similar treated and untreated individuals. Weighting methods achieve this by re-weighting individuals so that the treatment and comparison groups are similar to each other overall, while matching methods achieve this by pairing each treated individual with a similar untreated individual.

[51] Abadie, A., & L'hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, *116*(536), 1817-1834.

Estimation and Inference. Weighting methods may assign weights to individuals based on propensity scores[52] or based on covariate balance restrictions.[53] Matching methods may pair observations based on their propensity scores[54] or based on other measures of similarity.[55] After weighting or matching, outcomes are compared for the treatment and comparison groups to estimate effects.

Research on Incarcerated Individuals in Washington State

As discussed in Section II, we analyzed 58 recent studies on incarcerated individuals in Washington State to identify the prevalence of different research designs. We present additional information here.

Treatment Conditions

For each paper, we identified the basic treatment type. We classified studies using the following categories:

- None: there was no treatment condition. This was typically true of descriptive studies that looked at outcomes for all subjects
- Pre-incarceration treatment: for example, one study used previous traumatic brain injury as the treatment condition
- Prison programming: treatment in structured prison programs
- Prison non-programming condition: for example, solitary confinement
- Post-incarceration treatment: for example, one study used post-release sex offender registration as the treatment condition

Outcomes

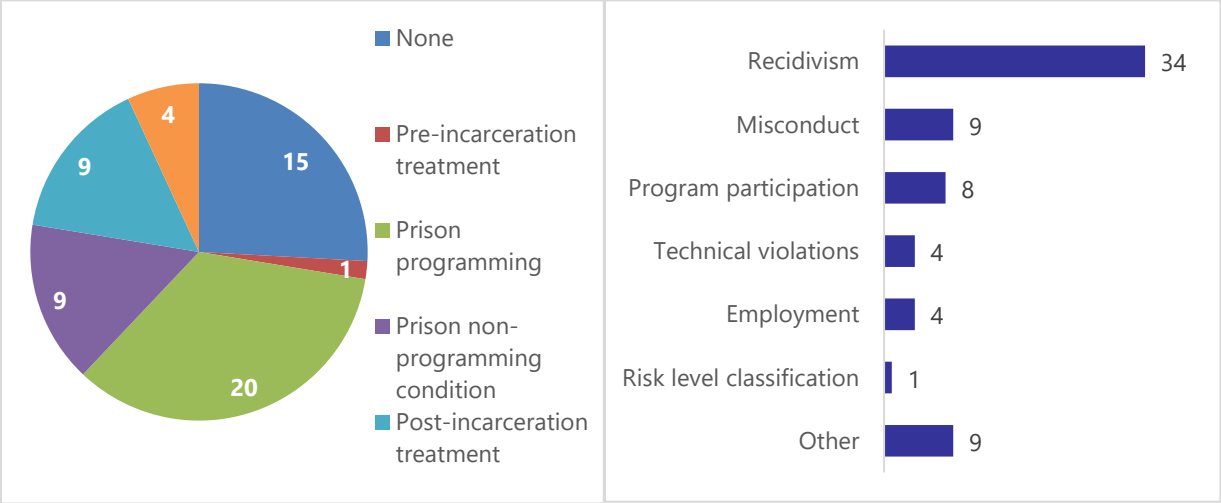
We also identified the outcome or outcomes analyzed by each paper.

Exhibit A1 presents information on treatment conditions and outcomes in studies included in our analysis.

[52] Robins, J.M., Rotnitzky, A., & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846-866.
[53] Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25-46.
[54] Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.
[55] Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 185-203.

Exhibit A1

Counts of Recent Studies of Incarcerated Individuals in Washington State Featuring Different Treatment Conditions (left) and Outcomes (right)



II. Evaluability Assessment

Programs Included in the Evaluability Assessment

We started with the programs identified in WSIPP's 2024 Adult Corrections Inventory (ACI) Preliminary Report.[56] The ACI delineated programs based on research areas, while we delineated programs based on DOC's definitions. Accordingly, some programs that were distinct in the ACI were combined in this report. We also excluded some programs that were included in the ACI. Exhibit A2 presents all DOC programs included in the ACI along with their status as of the evaluability assessment for this report.

Offender Management Network Information (OMNI) Data

DOC program participation is primarily tracked through the use of DOC's Offender Management Network Information (OMNI) system. OMNI "is the integrated software system used by DOC to record, track, and monitor information about offenders who are in the custody of the department."[57]

DOC provided WSIPP with annual program participation counts by facility and year, covering 2015-2024, as measured in OMNI records. These were used to estimate total program participation for each program included in the evaluability assessment in Section V.

Using this data comes with two limitations. First, each individual is counted once per program per year, but the same individual may be counted in more than one year. Second, some programs consist of multiple OMNI entries. For example, participants in the Strength in Families program may have OMNI entries for both "Parenting Inside Out" and "Walking the Line," which are both components of Strength in Families. As a result, our estimates are likely to overstate actual participation.

Survey Questions

We developed a survey to collect information on the evaluability of the program. A full copy of the survey, including introductory text, question formats, available response options, and skip logic, is available on WSIPP's website. Questions were divided into four sections.

1) Introduction: Program definition, objectives, and basic structure,
2) Who is the program for: Eligibility criteria, waitlist structure, and related data,
3) How is the program run: Program duration and completion, facilitators, quality assurance systems, and changes over time, and
4) Improving research readiness: Data tracking, waitlist, randomization, eligibility criteria, and quality assurance practices that could improve program evaluability.

Evaluability Assessment Methodology and Results

Here, we discuss in more detail how we used the survey to categorize programs in our four dimensions.

[56] Goodvin, R., & Wanner, P. (2024). *Inventory of evidence-based, research-based, and promising programs for adult corrections: Preliminary report* (Doc. No. 24-03-1901). Olympia: Washington State Institute for Public Policy.
[57] https://app.leg.wa.gov/committeeschedules/Home/Document/173205

Identifying the Treatment Group

We used survey data to measure the evaluability of treatment conditions by scoring DOC programs on two domains: 1) standardization in treatment delivery; and 2) ability to track treatment heterogeneity.

The first domain focused on whether the program was designed to deliver treatment in a way that was clearly defined and consistently implemented. The survey questions for this domain included whether the program had a defined endpoint, a standard duration or number of sessions, a quality assurance system, a measure of fidelity, and whether it was a discrete program versus a program area. Each time the respondent answered "yes" to one of these questions, the program received a point for this domain. This resulted in scores between 0 and 5, with higher scores corresponding to treatment conditions that are more clearly defined and consistently implemented.

Exhibit A2

Programs Included in Evaluability Assessment

Program name (Adult Corrections Inventory)	Updated program name
Anger management (other)	Stress Anger Management
Basic skills (ABE, GED, ESL, HS)	Basic skills (ABE, GED, ESL, HS)
Beyond Violence	Beyond Violence
CBT for individuals convicted of sex offenses	Sex Offense Treatment and Assessment Programs (SOTAP)
Construction trades apprenticeship preparation	Construction Trades Apprenticeship Preparation (CTAP)
Correctional industries/jobs	Correctional Industries
DNR jobs/fire camp	DNR correctional camps
Intensive outpatient	Intensive outpatient
Intensive Transition Program	Intensive Transition Program
Interactive journaling (e.g., Getting It Right)	Getting it Right
Life skills (general)	Skill Building Unit (SBU)
Moving On	Moving On
Parenting Inside Out Walking the Line	Strength in Families
Post-secondary	Postsecondary education
SPP - Dog training SPP - Horticulture SPP - Other SPP - Roots of Success	Sustainability in Prisons Project (SPP)
Therapeutic communities	Therapeutic communities
Thinking 4 a Change (T4C)	Thinking 4 a Change (T4C)
Traumatic Brain Injury (TBI) treatment/support	TBI Pilot-to-Program
Vocational education (general)	Vocational education
Excluded from the evaluability assessment	
Low participation or inactive: 99 Days & Get Up, Acceptance and Commitment Therapy, Aggression Replacement Training (ART), Beyond Trauma, Breaking Barriers, co-occurring disorder intensive outpatient, co-occurring disorder therapeutic communities, DBT skills programs, Decision Points, domestic violence support, emotion regulation/coaching, employment counseling/job training/search, IF Project, Inside Out Dads, Long Distance Dads, Men Facilitating Change, Moral Reconation Therapy (MRT), Moving Forward, New Freedom, parenting (other), Partners in Parenting, ReEntry And Community Health (REACH) program, reentry/release prep (other), residential parenting program, Seeking Safety, SMART Recovery, Tackling Anti-social Behavior, Transition to Life	
Not facilitated by DOC or contracted staff: Alcoholics/Narcotics Anonymous, Alternatives to Violence, Bridges to Life, Freedom Project, interpersonal skills training, Redemption Project, Toastmasters	
Fitness programs: Fitness/wellness, yoga/meditation	

The second domain focused on whether program administrators maintain records that could be used to measure differences in the quality and quantity of treatment that participants received. The survey questions for this domain included whether DOC tracked participants' completion status, the number of sessions participants attended, which facilitator led each session, information on facilitator credentials, and data on program fidelity. Each time the respondent answered "yes" to one of these questions, the program received a point for this domain. This resulted in scores between 0 and 5, with higher scores corresponding to more extensive data tracking practices.

We summed the scores for each domain to create our final measure of treatment evaluability, which ranges from 0 to 10.[58] In Exhibit A3, we show how the 19 DOC programs in our sample scored on this measure. In our main analysis, we categorized programs as limited if they scored between 0 and 4, moderate if they scored between 5 and 7, and strong if they scored 8 or higher.

One limitation of this scoring system is that it assigns equal weight to all factors. In practice, some factors may be more important than others, so two programs with the same score may not be directly comparable if they are missing different factors. It is also worth emphasizing that the treatment evaluability scores shown in Exhibit A3 do *not* reflect whether programs are effective. Instead, scores on this measure indicate the extent to which treatment conditions are conducive to rigorous outcome evaluation research.

Identifying the Comparison Group

We used survey data to assess the extent to which researchers could use administrative DOC data to construct valid comparison groups for an outcome evaluation. The survey asked whether the program had eligibility criteria, what factors influenced eligibility, and whether those factors were tracked in DOC data systems. After reviewing this information, we sorted each program into one of three categories:

1) Limited: We placed programs in this category if they had no eligibility criteria (i.e., anyone could join), if eligibility status was mainly determined by subjective and undocumented assessments (e.g., staff interviews with incarcerated individuals), or if participation was mandatory for all individuals convicted of particular types of crime (e.g., sex offenses). It may not be feasible to construct a valid comparison group for these programs.
2) Moderate: We placed programs in this category if there was partial or incomplete data on factors that impacted eligibility status. We also placed programs in this category if eligibility status was based on a wide variety of factors with limited guidelines. While it is feasible that researchers could use DOC data to construct comparison groups for these programs, it may be difficult to replicate fully the selection process.
3) Strong: We placed programs in this category if there were clear guidelines for determining eligibility status, and eligibility was based on a small number of factors that were measured in administrative data. It is highly feasible that researchers could use available DOC data to construct valid comparison groups for these programs.

Exhibit A3 shows how we used this measure to categorize the 19 DOC programs in our sample.

[58] The treatment evaluability measure has a mean value of 6.7 and a standard deviation of 2.9.

Again, we wish to emphasize that the information shown in Exhibit A3 does *not* reflect whether programs are effective. Instead, we have categorized programs based on how feasible it would be to construct comparison groups for outcome evaluation research.

Exhibit A3

Evaluability Assessment: Feasibility of Identifying Treatment and Comparison Groups

DOC Program	Identifying treatment group	Identifying comparison group
Beyond Violence	10 (Strong)	Strong
Intensive outpatient	10 (Strong)	Strong
Moving On	10 (Strong)	Strong
Strength in Families	10 (Strong)	Moderate
Therapeutic communities	10 (Strong)	Strong
Thinking 4 a Change (T4C)	10 (Strong)	Moderate
Sex Offense Treatment and Assessment Programs (SOTAP)	9 (Strong)	Limited
Construction Trades Apprenticeship Preparation (CTAP)	7 (Moderate)	Moderate
Getting it Right	7 (Moderate)	Moderate
Intensive Transition Program	6 (Moderate)	Moderate
Postsecondary education	6 (Moderate)	Moderate
Vocational education	6 (Moderate)	Moderate
Basic skills (ABE, GED, ESL, HS)	5 (Moderate)	Limited
Stress Anger Management	5 (Moderate)	Moderate
Sustainability in Prisons Project (SPP)	5 (Moderate)	Limited
TBI Pilot-to-Program	5 (Moderate)	Moderate
Correctional Industries	3 (Limited)	Limited
Skill Building Unit (SBU)	2 (Limited)	Strong
DNR correctional camps	1 (Limited)	Moderate

Overcoming Selection Bias

As discussed in Section III, selection bias is a common issue in outcome evaluations of programs where participation is not randomized. In these cases, researchers may be able to use quasi-experimental research designs, provided that there is some degree of randomness in who does and does not participate in the program. As part of our evaluability assessment, we identified nine potential sources of randomness that were present for one or more DOC programs. These opportunities were identified using the following criteria.

First-Come, First-Served Waitlist. Our survey asked whether each program used a waitlist that treated individuals in the order in which they joined the waitlist. We counted this opportunity if responses indicated that the program does not shift individuals' positions on the waitlist based on individual characteristics. We identified one program that uses a first-come, first-served waitlist. This enables researchers to reduce the impact of selection bias by constructing a comparison group of individuals who were not only eligible but also interested in participating in the program.

Court-Ordered Participation. Our survey asked whether each program uses court orders as an eligibility criterion. We identified four programs that are court-ordered for some participants. This enables researchers to use individuals' assignments to different judges as an instrumental variable.

Eligibility Based on Numeric Cutoff. Our survey asked how burdensome it would be to use eligibility criteria that depend on a cutoff based on a numeric variable. We counted this opportunity if responses indicated that this practice is currently in place. We identified four programs that determine eligibility using a cutoff value from a numeric variable on a screening assessment. For example, one program uses an assessment that results in a score ranging from -3 to 12, and individuals scoring three or higher are eligible.

Eligibility Based on Sentence length. Our survey asked whether each program uses sentence length as an eligibility criterion. We counted this opportunity if responses indicated that individuals must have enough time left until their earned release date (ERD) to participate in the program. We identified ten programs that require participants to have a minimum amount of time remaining until their earned release date. For example, one program requires individuals to have at least 18 months remaining so that they have time to complete the program before their release.

Eligibility Based on Custody Classification. Our survey asked whether each program uses security level as an eligibility criterion. We counted this opportunity if responses indicated that the program is not available to individuals at all security levels. We identified six programs that are only available to individuals at certain custody levels, where custody level is determined using cutoff points from a numeric variable called the Custody Review Score.

Change in Eligibility Criteria. Our survey asked whether eligibility criteria have changed over time. We counted this opportunity if responses indicated that this change could be operationalized by researchers using available data. We identified eight programs that substantially changed their eligibility criteria at some point in the last ten years. This enables researchers to analyze differences in outcomes for cohorts of individuals whose eligibility status changed over time.

Change in Program Content. Our survey asked whether each program has undergone significant changes in the last ten years. We counted this opportunity if responses indicated that the timing of the change was clear and researchers could identify individuals who received different versions of the program. We identified eight programs that substantially changed the content of the program at some point in the last ten years. This enables researchers to compare outcomes for cohorts of individuals who participated in different versions of the program.

Facility Closure. One DOC prison facility, Larch Corrections Center (LCC), closed in October 2023. We counted this opportunity if a program was offered at LCC prior to its closure. We identified seven programs that were offered at LCC prior to its closure in October 2023. This enables researchers to compare outcomes for individuals incarcerated at LCC who lost access to programs upon their transfer to other DOC facilities.

Timing of Program Rollout Across Facilities. Using OMNI participant counts, we identified which facilities offered each program between 2015 and 2024. We counted this opportunity if a program was phased in or phased out at different times at different facilities. We identified twelve programs that started or ended operations at some point between 2016 and 2023. This enables researchers to compare outcomes for individuals who gained or lost access to a program over time.

Exhibit A4 presents quasi-experimental design opportunities for programs included in the evaluability assessment. In our main analysis, we categorized programs as strong if they featured a first-come, first-served waitlist, court-ordered participation, or eligibility based on a numeric assessment score cutoff; moderate if they featured eligibility based on sentence length; and limited if they did not have any of the preceding features.

It is important to note that the opportunities discussed above are promising, but not definite. A program may have a key feature that is necessary to use a quasi-experimental design. However, it may be the case that on closer inspection, the design is infeasible for other reasons.

Sample Size

In the context of an outcome evaluation, "statistical power" refers to the probability that a statistical test will detect a program's effect, assuming that an effect exists. Power analysis is a type of preliminary analysis used to estimate the power of one or more proposed statistical tests. It depends on several factors, including sample size, the threshold of statistical significance, and the magnitude of the effect size. In general, larger sample sizes lead to higher statistical power.

One goal of this report was to estimate the statistical power that researchers can expect for different DOC programs, assuming that their goal is to determine whether the program reduces recidivism. To calculate statistical power, we set values for several inputs.

Type of Statistical Test. Statistical power depends on the type of statistical test being used. For simplicity, we assume the use of a two-proportion z-test.

Baseline Recidivism Rate. Statistical power also depends on the baseline value of the outcome of interest. Outcome evaluations of programs offered in prisons often analyze recidivism within three years of release. To set a hypothetical value for this parameter, we used results from WSIPP's 2019 recidivism report.[59] The three-year recidivism rate for the FY 2014 prison release cohort was 51.9%.

Magnitude of the Program's Effect. Statistical power also depends on the magnitude of the program's effect. To set hypothetical values for this parameter, we used WSIPP's benefit-cost results for adult crime programs.[60] We identified 26 programs for general adult prison populations with effect size estimates for recidivism. We divided these into the lowest nine, middle eight, and highest nine based on the effect size. The average effect size values in these groups were -0.272, -0.106, and -0.025, respectively. Next, we calculated the percentage point reductions in recidivism that would be associated with these effect sizes. These values were 11.1, 4.4, and 1.0 percentage points, respectively. A one-percentage-point reduction in recidivism would require a total sample size of nearly 80 thousand individuals, so we diverge from that value. In the interest of interpretability, we round these values to 10, 5, and 2 percentage point reductions as our hypothetical values for large, moderate, and small program effects.

Sample Size. Statistical power also depends on sample size. As discussed earlier, we used OMNI records to estimate program participation between 2015 and 2024. We assume that the treatment and comparison groups would be equal in size, so that the sample size is twice the participant count for each program.

---

[59] Knoth, L., Wanner, P., & He, L. (2019). *Washington State recidivism trends: FY 1995– FY 2014.* (Doc. No. 19-03-1901). Olympia: Washington State Institute for Public Policy.

[60] https://wsipp.wa.gov/BenefitCost?topicId=2

Threshold for Statistical Significance. Statistical power also depends on the threshold for statistical significance. We followed standard practice by using a significance level of 0.05.

Method for Calculating Statistical Power. We used the Stata –`power twoproportions`– command to calculate statistical power using the above inputs.[61]

Exhibit A5 presents estimated participant counts for each program along with whether we would expect to be able to detect reductions in recidivism of 2, 5, and 10 percentage points. In our main analysis, we classify programs as strong if they can detect a 2 percentage point effect, moderate if they can detect a 5 percentage point effect, and limited otherwise.

It is important to note that these calculations are dependent on several factors, and actual statistical power could vary depending on the assumptions and estimators used in any future analysis.

Overall Evaluability Classifications

We rated programs according to their potential to use different research designs. We classified programs as potentially compatible with quasi-experimental designs if they were rated "strong" on both treatment group identification and overcoming selection bias, selection on observables if they were not rated "limited" for either treatment group or comparison group identification, and descriptive otherwise.

[61] StataCorp. 2023. *Stata Statistical Software: Release 18*. College Station, TX: StataCorp LLC.

Exhibit A4

Quasi-Experimental Design Opportunities

	Comparison group identification	Instrumental variable (IV)	Regression discontinuity design (RDD)			Difference-in-differences (DID)			
DOC Program	First come, first served waitlist	Court ordered-participation	Eligibility based on cutoff variable	Eligibility based on sentence length	Eligibility based on custody classification	Change in eligibility criteria over time	Change in program content over time	Facility closure	Timing of program rollout across facilities
Basic skills (ABE, GED, ESL, HS)						✔	✔	✔	
Beyond Violence			✔	✔		✔			✔
Sex Offense Treatment and Assessment Programs (SOTAP)		✔	✔	✔	✔		✔		
Construction Trades Apprenticeship Preparation (CTAP)				✔	✔	✔	✔		✔
Correctional Industries								✔	
DNR correctional camps			✔		✔	✔		✔	
Intensive outpatient							✔		✔
Intensive Transition Program				✔	✔	✔	✔		
Getting it Right									✔
Skill Building Unit (SBU)							✔		✔
Moving On			✔	✔					
Postsecondary education				✔		✔			✔
Strength in Families	✔	✔		✔			✔	✔	✔
Stress Anger Management					✔				✔
Sustainability in Prisons Project (SPP)								✔	
Therapeutic communities		✔			✔		✔		✔
Thinking 4 a Change (T4C)			✔	✔		✔		✔	✔
TBI Pilot-to-Program				✔					✔
Vocational education		✔		✔		✔		✔	✔

Exhibit A5

Sample Size and Statistical Power

Program	Participant count, 2015-2024	Analysis could detect the program's effect with 80% statistical power if the program reduced recidivism by:		
		2 percentage points	5 percentage points	10 percentage points
Correctional Industries	186,459	Yes	Yes	Yes
Basic skills (ABE, GED, ESL, HS)	49,740	Yes	Yes	Yes
Vocational education	43,304	Yes	Yes	Yes
Postsecondary education	41,869	Yes	Yes	Yes
Sustainability in Prisons Project (SPP)	20,726	Yes	Yes	Yes
DNR correctional camps	14,187	Yes	Yes	Yes
Thinking 4 a Change (T4C)	12,978	Yes	Yes	Yes
Sex Offense Treatment and Assessment Programs (SOTAP)	9,529		Yes	Yes
Intensive outpatient	8,741		Yes	Yes
Therapeutic communities	8,627		Yes	Yes
Skill Building Unit (SBU)	6,193		Yes	Yes
Strength in Families	4,217		Yes	Yes
Construction Trades Apprenticeship Preparation (CTAP)	1,633		Yes	Yes
Moving On	1,628		Yes	Yes
Getting it Right	1,296			Yes
Stress Anger Management	1,226			Yes
Intensive Transition Program	685			Yes
Beyond Violence	494			Yes
TBI Pilot-to-Program	192			

Note:
In the context of an outcome evaluation, statistical power is the probability that a statistical test detects a program's effect, assuming that an effect exists.

Identifying Opportunities for Improvement

As part of our evaluability assessment, we asked program representatives about nine program administration and data collection practices that could improve evaluability. We asked them to rate each practice on a scale of 1 to 5 in terms of how burdensome it would be to implement:

Standardization in treatment delivery & ability to track treatment heterogeneity

- Track the number of sessions and the completion status of each participant
- Track which facilitator(s) delivered the program for each participant
- Implement a quality assurance system that includes 1) a clear definition of fidelity and 2) facilitator assessments that track fidelity over time
- Centrally coordinate program administration so that all facilities are implementing the program in a standardized manner

Feasibility of constructing a comparison group

- Implement a first-come, first-served waitlist
- Maintain data on past waitlist participation
- Track individual characteristics that screening committees use to refer individuals to the program

Randomization in treatment assignment

- Randomize referral to this program among eligible individuals (RCT)
- Change eligibility criteria to depend on a cutoff based on a numeric variable. For example, a domain needs level classification on the Washington ONE

We classified responses to each practice as follows:

- Low-burden practices: respondent rated the practice as 1 or 2
- Practices requiring additional resources: respondent rated the practice as 3 or higher and indicated that resource constraints were the primary reason for their rating
- Practices raising legal or ethical issues: respondent rated the practice as 3 or higher and indicated that legal or ethical issues were the primary reason for their rating

Acknowledgements

The authors would like to thank the following Washington State Department of Corrections staff for generously sharing their time and information: Courtney Bagdon-Cox, PhD, Senior Research Manager; Stefanie Bloomingdale, Associate Superintendent—SCCC; Jamie Dolan, Director—Correctional Industries; Cassandra Due, Program Specialist 5; Donald Feist, Administrator—Cognitive Behavioral Change; Paul French, Administrator—Substance Abuse; Bobby Greene, Reform Program Manager; Cathi Harris, Director—SOTAP; Donald Holbrook, Assistant Secretary—Prisons Division; Amber Medina, Sex Offender Treatment Supervisor; Kristen Morgan, Senior Administrator; Geraldine Newman, Correctional Program Manager; Jennifer Parsons, Corrections Mental Health Counselor 3; Scott Speer, Superintendent B— OCC; and Jeffrey Tatro, Classification Counselor 3.

For further information, contact:
Colin Gibson at 360.664.9076, colin.gibson@wsipp.wa.gov

Document No. 25-07-1901