Which prisoner reentry programs work?

Replicating and extending analyses of three RCTs

Jennifer L. Doleac, Chelsea Temple, David Pritchard, Adam Roberts

January 7, 2020*

Conducting a randomized controlled trial (RCT) can be an ideal way to avoid omitted variable and selection biases that complicate other research designs. However, the way that the data from an RCT are collected and analyzed can unintentionally reintroduce those biases. In this study we replicate and extend the analyses of data from three RCTs related to prisoner reentry, to more cleanly identify the causal effects of treatment. In two of the three experiments, our conclusions differ substantially from those of the original studies. We discuss best practices for running and analyzing RCTs, and consider our extension results in the context of the prisoner reentry literature.

*We are extremely grateful to the authors of the original studies – Daniel J. O'Connell, John J. Brent, Christy A. Visher, Leonard A. Jason, Bradley D. Olson, Ron Harvey, and Grant Duwe – who generously shared their data and helped us understand their original analyses to facilitate replication. Thanks also to Sara Heller, Sally Hudson, Jason Lindo, and Emily Owens, as well as participants at the 2019 PELS Workshop for helpful comments and suggestions. Doleac (corresponding author): Texas A&M University, jdoleac@tamu.edu. Temple: Texas A&M University, temce21@tamu.edu. Pritchard: Texas A&M University, davidpritchard@tamu.edu. Roberts: Texas A&M University, adam.roberts@tamu.edu.

1 Introduction

Half of individuals who are released from prison are re-incarcerated within three years (DuRose, Cooper and Snyder, 2014). Practitioners and policy-makers across the country are working to reduce recidivism rates for those coming out of jail and prison in order to break this vicious incarceration cycle. Unfortunately, there is relatively little evidence to guide their efforts (Doleac, 2019a). In this paper, we replicate and extend the analyses from three evaluations of prisoner reentry programs, with the goal of learning as much as we can from existing evidence. These studies consider data from well-implemented randomized controlled trials (RCTs), where the original analyses made it difficult to conclude whether the programs of interest had causal effects on participants' outcomes.

RCTs are typically considered the gold standard when it comes to program evaluation. They allow us to quantify the effect of treatment relative to a control group, and make it easier to avoid confounding factors that can complicate other research designs. Designing and implementing an RCT requires significant effort and resources, as well as buy-in from practitioners; this combination of challenges limits how frequently this type of research can be done, particularly in the criminal justice context (where safety and security concerns are paramount). Even in cases where RCTs are successfully implemented, many studies do not present or analyze the data in a way that cleanly measures the intent-to-treat (ITT) and/or treatment-on-the-treated (TOT) effects. Our main goal in this paper is to extend existing studies by using up-to-date econometric methods to identify the causal effects of the programs being evaluated. Ensuring that estimates of treatment effects are unbiased allows us to add valuable information to a thin empirical literature.

This exercise also provides case studies on the extent of bias due to such problems as non-random attrition (which can introduce selection bias) and including endogenous control variables that are affected by the treatment. Economists tend to prioritize eliminating such biases, but reasonable researchers can disagree a priori about the likely magnitude of any bias. If – in the prisoner reentry context – selection and omitted variable biases are small

in practice, and do not meaningfully change the estimated effects of the program being considered, then economists' insistence on clean identification may lead us to unnecessarily dismiss valuable research evidence. On the other hand, if such biases are large, then many research studies in this area may be pointing us in the wrong direction.[1]

We replicate and extend three studies: one on a swift, certain, fair (SCF) program of graduated sanctions for drug-involved probationers; one on aftercare programs for recently-released, drug-involved offenders; and one on a comprehensive reentry program for inmates in Minnesota. We find suggestive evidence that the SCF program reduced recidivism, but estimates are too imprecise to draw clear conclusions. Our reanalysis suggests that endogeneity bias in the original study affected the magnitude and sign of some coefficients, but not statistical significance (although this is because the study is substantially underpowered). In the aftercare program analyses, we find that (1) Therapeutic Communities reduced employment and earnings, with suggestive evidence that they also increased time incarcerated; and (2) Oxford Houses increased days incarcerated, with suggestive evidence of increases in employment. These conclusions differ substantially from those of the original study. Lastly, for the Minnesota reentry program, using matched comparison groups instead of simply controlling for baseline characteristics leads to conclusions that are qualitatively similar to those of the original study (that MCORP reduced recidivism). However, the data available did not allow us to conduct standard analyses based on original treatment assignment (to avoid selection and omitted variable biases). We thus interpret these results with caution.

These three studies were part of a larger set (identified in the course of review of the literature on prisoner reentry; Doleac, 2019a) where concerns about the analysis made it difficult to interpret the results. However, these three were the only studies where authors were willing and able to provide data for replication and extension.[2] This set of studies may

[1]For instance, a recent review of the literature on wrap-around services suggests that selection into treatment substantially biases estimates in existing studies using matched comparison groups to evaluate program effectiveness (Doleac, 2019c; Doleac, 2019b).

[2]We contacted authors of six additional studies. Those authors were unwilling or unable to share their data.

therefore be positively selected. This also points to a broader challenge in this research space: while it is now common for economics journals to require that authors provide replication files (including data) as a condition of publication, this is not yet the norm in other disciplines. This makes exercises like ours difficult if not impossible in most cases. Given a natural progression of quantitative methods over time, even methods that are cutting-edge at the time of publication may be viewed as falling short in the future. Being able to replicate and extend those analyses at a later date (as we do here) will facilitate a more rapid accumulation of knowledge.

For each study, we replicate the original analysis, then extend the results in two ways, one step at a time: First, we adjust the functional form of the empirical model used, as needed, to ease interpretation of the results and facilitate comparison with the broader literature. Second, we adjust covariates and other factors relevant to identifying causal effects. We generally expect that the functional form will not have a substantive effect on the results, and show this step for the sake of transparency. We do expect that addressing identification concerns will matter, reducing bias in the estimates.

This paper proceeds as follows: Section 2 lays out recommendations for analyzing data from an RCT. Section 3 discusses the "Decide Your Time" SCF program; Section 4 discusses aftercare programs for recently-released, drug-involved offenders; and Section 5 discusses a holistic reentry program called MCORP. In each section, we discuss the original study, replicate the original results, then extend the analyses to more cleanly identify the effect of the program. Section 6 discusses our findings in the context of the broader literature on prisoner reentry programs, and Section 7 concludes.

2 A short guide to analyzing data from an RCT

Randomizing treatment assignment is the hard part of a rigorous evaluation. The priority of subsequent data collection and analysis should be to avoid reintroducing the selection and omitted variable biases that randomization eliminated. Others have written extensively on best practices for running RCTs and analyzing data from experiments (in particular see

the resources compiled by J-PAL at https://www.povertyactionlab.org/research-resources/introduction-evaluations). We provide the following summary for readers who may not be familiar with best practices in this area, and to frame the issues we discuss and address in our replication and extension analyses below.

- Before beginning the experiment, conduct a power analysis to be sure that you have a sufficient sample size to detect meaningful effects. A statistically insignificant effect is only valuable if it is precisely estimated: Large point estimates with large standard errors do not imply that an intervention had no effect, only that the effect is statistically indistinguishable from the null due to lack of statistical power. An experiment that is underpowered to rule out large effects may not be worth running.

- Whenever possible, use administrative data – particularly for the outcome measures – to improve accuracy and avoid selective attrition from the sample. Using survey measures requires finding and interviewing all participants at various points in time, and inevitably some will not respond. It is unlikely that non-response will be random, and so this will lead to selection bias in the estimates.

- Try to include non-binary outcome measures in addition to binary measures. This provides more variation in the outcome, which can make it easier to detect program impacts. It is also useful for cost-benefit analyses. For instance, in addition to a binary measure of whether a participant was incarcerated during the follow-up period, consider the number of days incarcerated.

- It is often useful to show short-, medium-, and long-term program impacts. Whenever possible, use cumulative outcome measures so that the long-term impacts include behavior from the short- and medium-terms. This makes it easier to interpret the results than if results reflect consecutive snapshots of mutually-exclusive time periods.

- Select a small number of outcome measures to be the outcomes of primary interest. With enough outcomes it is statistically likely that at least some regressions will show

(spurious) significant results, so it is helpful to narrow your focus to the ones that are most important or relevant, before beginning the analysis. Consider pre-registering the RCT with those outcomes highlighted. Also consider formal adjustment for multiple hypothesis testing if you focus on more than two or three key outcomes.

- Check for balance on observable characteristics. Note any imbalances in the writeup and control for unbalanced characteristics in the analysis. This is a next-best approach, relative to the ideal scenario of balanced treatment and control groups. Controlling for unbalanced characteristics may raise concerns about data mining, so choose covariates in a way that limits researcher discretion. (To avoid imbalances in key covariates, consider a strategy such as block randomization.)

- Conduct a simple comparison of means for the outcome measure(s), without any controls. Differences may be imprecisely measured, but should be unbiased estimates of the treatment effect (if the randomization 'worked' and the treatment and control groups were similar before the experiment).

- While more complex, nonlinear models may be appropriate in some settings, conducting Ordinary Least Squares (OLS) regressions alongside those regressions is helpful. OLS estimates marginal effects of treatment that are easy to interpret and easy to compare with results from other studies.

- Cluster standard errors at the level of the treatment, to adjust for correlations of errors within groups.

- Regress outcome measures on treatment assignment plus covariates that were determined before treatment assignment (to improve precision). Never include covariates that themselves may have been determined in part by the treatment assignment (this is colloquially known as "controlling for an outcome"). For example, do not control for how much someone participated in the program, or their completion of program steps.

- Include fixed effects to match the way treatment was assigned, to avoid omitted variable bias. For instance, include fixed effects for the relevant blocks if block-randomization was used, or time period if the probability of treatment varied across time.

- Keep all individuals in the dataset with their initial treatment/control assignment, even if they did not follow that assignment. Compare individuals as assigned to measure the ITT effect. To account for noncompliance, use treatment assignment as an instrument for actual treatment. This will give you the TOT effect. Never drop non-compliers or restrict the analysis to program participants or completers. This reintroduces selection bias that randomization avoided.

3 Study 1: Decide Your Time

3.1 The Original Study

O'Connell, Brent and Visher (2016) investigate the effects of Delaware's "Decide Your Time" (DYT) program —an alternative to traditional probation for high-risk probationers. This program is based on the "swift, certain, and fair" (SCF) approach to sanctions in which modest and graduated punishments are made clear to the probationer, then implemented quickly and reliably. For instance, those in violation of court rules would be immediately punished with a short (1-2 day) jail spell. (This contrasts with standard community supervision, where sanctions can be unpredictable but severe when finally applied.) The program targets probationers required by the court to abstain from drug use; frequent drug tests are therefore a key component to measure compliance with program rules. Despite previous work finding evidence that increasing detection of drug use violations combined with SCF sanctions works to decrease noncompliance and recidivism (e.g. studies of the HOPE program in Hawaii; Hawken and Kleiman, 2009), it is important to test whether the model can be replicated and scaled. O'Connell, Brent and Visher (2016) is one of several recent efforts to replicate the SCF model in other contexts (see also Hawken and Kleiman, 2011; Hamilton et al., 2016; Lattimore et al., 2016; and Davidson et al., 2019).

With the goal of reducing recidivism and drug use, the DYT program included three components: increased monitoring (in the form of frequent random drug testing), SCF sanctions, and treatment referrals. Importantly, DYT was not overseen by a judge, which differentiated it from other prominent programs that utilized an SCF approach. This could make the program easier to scale if it is effective. Like other SCF programs, DYT informed probationers what was required of them, what would happen if they failed to meet program requirements, and how to reduce their level of monitoring after violating requirements and receiving increased sanctions.

The study sample included 400 high-risk probationers with a history of substance abuse. Specifically, the sample was comprised of individuals under intensive supervision for a drug-related offense and individuals under intensive supervision for a non-drug-related offense who had failed a drug test during probation. These probationers were randomly assigned to DYT treatment or standard probation (the control), and observed for 18 months. Of the 400 participants, complete baseline and follow-up data were available for 377; we focus our analysis on this sample.[3] (Using administrative data to track employment would have avoided sample attrition over time.)

Summary statistics for both the full and split (treatment and control) samples are shown in Panel A of Table 1. Columns 1-4 show the 'full sample' – data on all participants. Columns 5-8 show the 'analysis sample' – individuals for whom complete data are available. In both samples, eighty-five percent of participants were men, and forty-six percent were white. The average age at first arrest (a proxy for criminal history) was 21, and the average age at randomization into the current experiment was 30. Columns 4 and 8 show the differences in means between the treatment and control groups. We conduct a series of t-tests and do not find any statistically significant differences, including in the likelihood of being in the analysis sample (that is, of having complete data available). We thus conclude that

[3]Employment information is missing for 18 participants. Age at randomization and age at first adult arrest information are missing for an additional five participants, which brings the final analysis sample down from 400 to 377.

the randomization 'worked': the two groups are balanced on all observable characteristics available pre-randomization, which gives us confidence that the groups are also balanced on unobservable characteristics (though of course we cannot test that directly).

The original study considers the effect of DYT on a variety of outcomes: any arrest, arrest for a new crime, arrest for a violation of probation, arrest for a technical violation, incarceration, and drug use. Each outcome is coded as a binary measure, and collected at 6, 12, and 18 months post-treatment assignment (these measures are cumulative). Data on recidivism come from administrative records, and drug use was measured by drug tests.[4] In addition to examining whether probationers passed or failed a drug test, the authors also collected data on the total number of drug tests received and the number of days between drug tests. This allows us to confirm that probationers in the treatment and control group did indeed experience different levels of drug testing, as designed.

The authors regress each of these outcome variables on a treatment indicator (assignment to DYT versus standard probation), while controlling for demographics (e.g., race, gender, age at randomization), age of first adult arrest (a proxy for prior criminal conduct), employment during participation, missed meetings with the probation officer, referral/enrollment in a drug treatment program, and whether a formal warning was given by the probation officer. Drug test failure is also included as a control in some specifications.

The original results suggest that DYT increased the likelihood of failing a drug test (presumably because DYT probationers were subject to more drug tests to begin with). They also suggest that recidivism decreased for the DYT probationers over the 6, 12, and 18 months following treatment assignment. While these estimates suggest economically meaningful effects, they are not statistically significant. The authors concluded that DYT had no beneficial effects for participants.

[4]Employment data come from the Corrections data system; these are likely as reported by probationers or probation officers. An alternative that would be more complete (and perhaps more accurate) measures of formal labor market participation is data from Unemployment Insurance records.

3.2 Replication

We begin our replication by presenting a simple comparison of means in Panel B of Table 1. The differences in column 8 suggest beneficial effects of the program: a 10% reduction in being arrested for a new crime, an 8% reduction in being incarcerated, and a 22% increase in being employed. However, none the differences in these key outcome measures are statistically significant.

Next, we replicate the authors' original analysis on recidivism. When examining recidivism, we focus on arrest for a new crime and incarceration as the outcomes of primary interest. Additionally, while the original analysis focuses on outcome separately at 6, 12, and 18 months post-randomization, we focus on the final, cumulative effects (i.e., outcomes measured at 18 months post-randomization). Following the original study, we use a multilevel logistic (MLL) regression, which accounts for the fact that the 400 probationers are assigned to (or "nested within") 61 probation officers. The model takes the following form:

$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta DYT_{ij} + \theta X_{ij} + \epsilon_j, \tag{1}$$

where

$$\pi_{ij} = E(y_{ij}) = Pr(y_{ij} = 1), \tag{2}$$

and y_{ij} includes binary measures of arrest for a new crime and incarceration recorded at 18-months post-randomization for probationer i with probation officer j. DYT_{ij} is an indicator variable that takes a value of one when the probationer is assigned to the DYT group. X_{ij} is a vector of control variables, including age at randomization, gender, race (white/black), employment, age of first adult arrest, number of missed appointments with a probation officer, drug treatment, and number of failed drug tests.[5] β is the "cluster-specific" effect of being in the DYT program (the treatment group), i.e., the effect of DYT on the log-odds of

[5]The original study also includes a control for whether or not the probationer received a formal warning from the probation officer. This variable was not in the dataset we received, so we do not include it in our analysis.

recidivating for probationers assigned to the same probation officer.

Results from the original paper are shown in Table 2, Columns 1 and 2 of Panel A; our replicated results are shown in Columns 3 and 4 of Panel A. We report results as odds ratios for a direct comparison with the original study. We also report the implied marginal effects, which can be directly compared to the OLS coefficients in Panel B (discussed in more detail below).[6] We are able to almost exactly replicate the original findings for arrest for a new crime and incarceration: our results differ slightly in magnitude (with smaller standard errors), and we find a marginally significant decrease in incarceration. The dataset we received did not include the *Formal Warning* control variable that was included in the original study's analyses, and so we were unable to include this variable in our replication. This likely explains the minor inconsistencies between the original and replication results. That said, as in the original paper, our replication results suggest that DYT reduced recidivism: post-randomization, DYT probationers were 4.7 percentage points (9.7% of the control group mean, not significant) less likely to be re-arrested for a new crime, and 10.3 percentage points (15.9%, $p < 0.10$) less likely to be re-incarcerated. However, these effects are imprecisely estimated. Original and replicated results for all outcome measures are in Panels A and B of Table A1.

3.3 Extension

We extend the original analysis in two ways. First, we alter the functional form used in the empirical analysis. The MLL regression used in the original analysis – while common in other disciplines – is less common in economics. O'Connell, Brent and Visher (2016) use this model in order to account for the fact that the 400 probationers are assigned to 61 probation officers. More common in economics is to simply cluster standard errors to allow for within-

[6]To calculate the implied marginal effect, we do the following: first, following Sribney and Wiggins (n.d.), we calculate the logistic coefficients by taking the log of the odds ratio. Next, following Gelman and Hill, 2007, we divide the coefficient by four, which yields an approximate marginal effect. To calculate the standard errors for these implied marginal effects, we first calculate the standard errors associated with the logistic coefficients by dividing the standard error of the odds ratio by the odds ratio itself (see Sribney and Wiggins (n.d.)). Then, per Gelman and Hill (2007), we divide the standard errors associated with the logistic coefficients by four.

group correlations in the error term. We thus run OLS with standard errors clustered at the probation officer level, which produces easy-to-interpret estimates of marginal effects, while allowing standard errors to be correlated across individuals who have the same probation officer. The coefficients from logistic regressions are less intuitive, and a back-of-the-envelope calculation must be done in order to back out the marginal effects.

Second, we alter the covariates. The original study controls for several variables that may have been affected by treatment (employment, missed meetings, referral to/enrollment in drug treatment, receiving a formal warning, and drug test failure). While each of these variables is useful as a potential outcome measure, including them as controls can bias the estimates. We remove these endogenous control variables and consider one (employment) as an additional outcome measure of interest.

These alterations yield three 'extension' specifications: MLL with exogenous controls only, OLS with all original controls, and OLS with exogenous controls only. This final specification is our preferred model. More formally, we estimate the following OLS specification:

$$Y_i = \alpha + \beta DYT_i + \theta X_i + \epsilon_i, \tag{3}$$

where Y_i is an outcome variable, DYT_i is our treatment indicator (assignment to the DYT group), and X_i is a vector of baseline characteristics that are included to increase precision. Our primary outcomes of interest are binary indicators for any arrest for a new crime, any incarceration, and employment during participation - all recorded 18 months post-randomization. Control variables include demographics (age at randomization, gender, and race) and age of first adult arrest (a proxy for criminal history). As discussed above, standard errors are clustered at the probation officer level.

Results from our extension are shown in Table 2. In Columns 3 and 4 of Panel B, we alter functional form only. Here, we run an OLS regression with all controls used in the original analysis. OLS estimates a 4.6 percentage point (9.5%, n.s.) decrease in the likelihood of arrest for a new crime and an 8.6 percentage point (13.3%, $p < 0.05$) reduction

in the likelihood of incarceration. Overall this functional form change makes little qualitative difference, as expected. The primary reason for this change is to ease interpretation.

In Columns 5 through 7 of Panel A, we use the original functional form but adjust the covariates. More specifically, we run the MLL model used in the original analysis but only include exogenous controls: demographics and age at first arrest. Compared to the original findings, excluding endogenous controls yields a nearly-identical estimate of the likelihood of arrest for a new crime (a 4.8 percentage point decrease compared to a 4.7 percentage point decrease), but a smaller estimate of the likelihood of incarceration (a 4.4 percentage point decrease compared to a 10.3 percentage point decrease). The estimated effect on incarceration is no longer statistically significant. The MLL result for employment (an outcome not examined in the original study) implies that DYT probationers were 9.9 percentage points (27.1%, n.s.) more likely to be employed during probation.

Finally, in Columns 5 through 7 of Panel B, we alter both functional form and covariates, running an OLS regression with exogenous controls only. This is our preferred specification, which we interpret as estimating the causal ITT effects of the DYT program. Using this specification, we find suggestive evidence that DYT improved probationers' outcomes, but the analysis is too underpowered to draw strong conclusions. On average, probationers in the DYT (treatment) group were 4.7 percentage points (9.7%, n.s.) less likely to be arrested for a new crime than were those in standard probation (the control group). Additionally, DYT reduced the likelihood that probationers were incarcerated during the 18-month follow-up period by 4.0 percentage points (6.2%, n.s.); depending on how many days each incarceration entailed, this could imply a meaningful cost savings. We also find that DYT probationers were 8.9 percentage points (24.4%, n.s.) more likely to be employed during probation. However, none of these estimates are statistically significant.

In Table A1 we show DYT's effects on other outcomes: failing a drug test, any arrest, arrest for a violation of probation, arrest for a technical violation of probation, completed probation, referral to/enrollment in drug treatment, share of drug tests failed, missed appoint-

ment with a probation officer, and absconded. Each of these outcome variables – except the share of drug tests failed – is a binary measure recorded at 18 months post-randomization.[7] In general these results show that the SCF approach (including administering more drug tests) in the DYT program led to more violations of parole conditions. This is perhaps unsurprising, as having more requirements gives probationers more opportunity to fail to meet those requirements. The intent of these requirements is to help probationers build a stable life free of criminal activity. When evaluating the overall effectiveness of the program, we focus on effects on new criminal behavior, incarceration, and employment. Despite (or perhaps because of) increases in technical violations, the main estimates suggest beneficial effects of the DYT program for reducing crime and incarceration, and increasing employment.

3.4 Discussion

In the above replication and extension, we find that removing the endogenous control variables reduced the estimated effects of the program – by more than half in the case of incarceration. However, it is difficult to draw clear conclusions given the imprecision of the estimates. While effect sizes are typically meaningful and suggest beneficial effects, standard errors are too large to rule out null effects or effects of the opposite sign. However, we believe it would be misleading to conclude that DYT had no impact on probationers: we cannot rule out large beneficial effects.

The challenge here is that the original study was substantially underpowered. Column 1 of Table 3 displays power calculations for recidivism (as measured by arrest for a new crime). With the original sample size of 400 (200 in DYT and 200 in standard probation), the smallest effect detectable at the 5% level is a 28.6% change relative to the control group mean. In order to detect a 5% change in recidivism at the 5% level, a total sample size of over 13,000 participants would be required. Additional study of this and similar programs, with much larger samples, would be valuable.

It would also be helpful to have continuous measures of some of the outcomes (par-

[7] Drug use is measured as both whether a probationer failed a drug test, and as the number of drug tests failed as a share of the total number of drug tests taken.

ticularly incarceration and employment), instead of simple binary measures of whether a probationer was ever incarcerated or ever employed. Knowing how many days someone was incarcerated or employed would facilitate a cost-benefit analysis; since incarceration is expensive, even a small reduction in days incarcerated could make a program cost-effective. Continuous measures would likely provide more variation in observed outcomes, which could make it easier to detect treatment effects.

4 Study 2: Aftercare

4.1 The Original Study

Jason, Olson and Harvey (2014) evaluate the impact of two aftercare programs for recently-released offenders, following inpatient community-based drug treatment. Participants from the Chicago area were randomly assigned to one of three treatments: Oxford Houses, Therapeutic Communities, or status quo community services (control). Participation was restricted to adults recovering from alcohol and drug dependence that had been released from incarceration within the previous two years.

Oxford Houses (OHs) are recovery homes for individuals dealing with substance abuse problems. No professional staff are involved; instead, residents live together in moderately-sized, single sex, single-family homes, and provide each other with a supportive, sober social network. Residents must pay rent (approximately $100 a week), abstain from any alcohol or drug use, and comply with assigned weekly chores.

Participants assigned to a Therapeutic Community (TC) were taken to a licensed, private organization that provides a structured, professionally-staffed, residential, sober-living program. Residents live in two to three person units and must follow a regimented program of recovery. Treatment evolves over time, but initially requires that participants obtain full or part-time employment, attend five self-help meetings per week, have four "recovery-related" phone calls to a sponsor per week, and submit to random drug tests.

Participants assigned to the control condition did not receive any intervention above what was previously available in the community. After being discharged from their inpatient

programs, they were left to find their own living accommodations. Follow-up surveys indicated that they were living in a variety of settings, including their own house or apartment, with friends or family, or in homeless shelters.

The authors followed up with participants every six months for two years, with one baseline (pre-period) survey and four follow-up (post-period) surveys. The data therefore include pre- and post-period observations, with a maximum of five observations per person.

All data is collected through extensive in-person interviews.[8] All questions focused on behaviors and outcomes that had occurred since the last survey, with most questions focusing on the last 30 days.[9] Because surveys were six months apart, this means that each survey wave represents a snapshot of time for that individual; in other words, the outcomes aren't measured cumulatively. There was substantial attrition, from both the treatment programs and in terms of survey responses. Our extension analyses will address both issues.

Table 4 shows summary statistics for the analysis sample.[10] The sample is 17% female, 74% black, and an average of 41 years old. Participants had low levels of education: 30% had graduated high school and 11% had ever attended college. The treatment and control groups are unbalanced on multiple baseline and demographic characteristics, as shown in columns 5 and 6. That is, unfortunately the randomization did not 'work'.

The authors consider the effects of Oxford Houses and Therapeutic Communities on a number of self-reported outcome measures: drug and alcohol use, incarceration, days of paid work, employment income, illegal income, legal issues, and psychiatric hospitalization. They conclude that staying in OHs or TCs for longer increased employment and reduced substance abuse. They also conclude that assignment to an OH increased income, number days of work, and continuous sobriety rates. They did not find any significant effect on incarceration for either treatment.

[8] Phone interviews were conducted in rare cases if an in person interview was not possible.

[9] Participants were asked to mark dates on a calendar, or asked specifically about the last 30 days. For example; participants were asked to mark each day that they had worked in the last 30 days on a calendar, and were asked how much income they earned from employment over the last 30 days.

[10] As described below, we restrict our analysis to participants for whom all necessary data are available, to maintain a consistent sample across regressions. Summary statistics for the full sample are in Table A2.

4.2 Replication

We begin with a simple comparison of means, shown in Panel B of Table 4. To ease interpretation, all results we present will be based on a consistent sample for which the necessary data were available for all analyses.[11] First we show the effect of treatment assignment on actual program participation: 70% of those assigned to OH and 51% of those assigned to TC participate in their assigned program for at least 30 days. On average, assignment to the OH group is associated with significantly better employment outcomes (days worked and earnings) and a reduction in days incarcerated. Assignment to the TC group is associated with significantly worse employment outcomes, and no difference in incarceration. Because the groups were unbalanced on observable characteristics (despite randomization), these differences in mean outcomes should not be interpreted as the treatment effects of the programs.

Using the following OLS regression model, we are able to exactly replicate the authors' original results:[12]

$$Y_{it} = \beta_0 + \beta_1 OH_i + \beta_2 TC_i + \theta_1(TC_i * Time_t) + \theta_2(OH_i * Time_t) + \rho Time_t + \eta Dose_i + \gamma Age_i + \epsilon_{it} \quad (4)$$

This model includes an indicator for each treatment assignment, a linear time trend, and an interaction between each treatment assignment and time that allows the effects of each treatment to change linearly across time (from the baseline through the follow-up periods; note that there is no dummy variable for the post-period, so this is not a difference-in-differences model). This specification measures whether participants in the treatment groups are on different trajectories than those in the control group. The specification also

[11] Our replication results based on this sample are quite similar to the original study's results, but obviously they are not identical. We present summary statistics and results based on all available data (where the sample changes from one specification to the next) in Tables A2 and A3.

[12] The original study did not report individual coefficients. We exactly match the p-values reported in the original paper, and exactly match the coefficients and standard errors from one set of regression results provided by the authors. This gives us confidence that we have matched their specification.

controls for dose, which is the time each treated individual spent in their assigned treatment, as well as participant age.

The original paper reports which groups of explanatory variables were significant in their regressions, as well as the direction of these effects, but does not report the estimated coefficients or the significance of most individual variables. These coefficients are useful for comparing the magnitude of effects between treatment groups, and also allows direct comparisons with other research on this topic. For this reason, we report both coefficients and standard errors throughout the replication and extension of the paper. The original authors provided us with detailed results from part of their analysis, which was very helpful during the replication process.

Table 5, Columns 1-3 of Panel A, shows our replication of the original results. We use the authors' specification and all original covariates, but restrict the sample to individuals where the necessary data were available for all analyses (as described above). Column 1 shows the effects of the two treatments, relative to the control group, on the number of days worked. Those assigned to OHs work 2.1 fewer days per month on average ($p < 0.10$). However, the average days worked increases by 1.1 days per month ($p < 0.05$), relative to the time trend for the control group. Those in the TC group work 3.0 fewer days per month on average ($p < 0.05$), and the difference between the TC group and the control group does not change over time. Columns 2 and 3 show effects on income earned and days incarcerated, respectively. Because of differences in baseline characteristics across groups, the differential time trends are the outcomes of primary interest here, but recall that these are not difference-in-difference coefficients so it is difficult to tell if these measure the causal effects of treatment. Replication results for other outcome measures are in Table A4, Panel A, columns 1-5.

4.3 Extension

Our extension of the authors' analysis includes a number of changes, as follows:

Difference-in-Differences Ideally we would compare the cumulative outcomes across treatment and control groups at the end of the follow-up period. Because outcomes aren't measured cumulatively (due to a reliance on surveys rather than administrative data), we retain the panel nature of the data, but switch to a difference-in-differences framework.

We noted above that, due to a small sample size, the treatment and control groups are not balanced in terms of individual characteristics. Similarly, levels of the outcome variables also vary across groups in the baseline (pre-treatment) survey. For example, the OH group spent 75% less time incarcerated in the month prior to treatment than the control group did. Estimates of incarceration that fail to account for differences in pre-treatment levels will be confounded by these pre-period differences, resulting in biased estimates.

We adjust our model to control for these pre-treatment differences between groups. The resulting model follows a difference-in-differences framework and is specified as follows:

$$Y_{it} = \beta_0 + \lambda_t + \beta_1 TC_i + \beta_2 OH_i + \theta_1(TC_i * Post_t) + \theta_2(OH_i * Post_t) + \gamma X_{it} + \epsilon_{it} \qquad (5)$$

where Y_{it} is an outcome measure for individual i in survey wave t, TC_i and OH_i indicate treatment group assignment and λ_t are survey waves fixed effects. X_{it} are individual-level covariates. $TC_i * Post_t$ ($OH_i * Post_t$) is an interaction between those assigned to the TC (OH) treatment and an indicator for whether the survey was conducted after treatment assignment.

Endogenous Controls and Omitted Variable Bias One of the main results of the original paper was that individuals who stayed longer in either TCs or OHs had increased employment and reduced alcohol and drug use. However, this was tested by simply including length of stay directly in the regression as an explanatory variable. This approach is problematic because individuals choose how long to stay in the program, and their choice/eligibility to stay depends in part on their successful completion of program requirements (that is, the variable is an endogenous function of treatment). The current dose variable may be serving

as a proxy for motivation and success of the program, rather than simply an indicator of amount of treatment received. We drop the dose variable, and instead use program participation as a first-stage outcome in a two-stage least squares (2SLS) analysis (described below).

While endogenous variables should be removed to avoid potential biases, adding exogenous controls can increase the precision of the estimates. Adding controls can also adjust for any baseline imbalances in observable characteristics. As shown in Table 4, several baseline characteristics are statistically different across treatment groups. Including controls for age, gender, race, and education level would be appropriate, though in this case we opt to include individual fixed effects (described next) that will absorb these individual controls.

Individual Fixed Effects The original study uses survey data as outcome measures (instead of administrative data). This leads to the common problem of survey non-response: participants drop in and out of the dataset over time, thus changing the composition of people included in the analysis across survey waves. We add individual fixed effects to the analysis to account for this. These fixed effects absorb average differences across people, so the results can be interpreted as within-person effects of treatment assignment.

The final specification that we use to measure the ITT effects of the OH and TC programs is:

$$Y_{it} = \theta_0 + \alpha_i + \lambda_t + \theta_1(TC_i * Post_t) + \theta_2(OH_i * Post_t) + \epsilon_{it}, \tag{6}$$

where α_i are individual fixed effects, and everything else is as defined above. Note that the α_i absorb indicators of treatment group assignment (TC and OH) as well as baseline demographic characteristics (the vector X_i in Equation 5).

Clustering Standard Errors By collecting survey data every six months, the original authors constructed a panel data set with five observations per person. Each observation represents an individual's survey response in that specific time period. Thus, although the

study only included 270 participants, the analysis dataset has 899 observations. The original analysis treats each of these observations as independent. However, observations for the same person are not independent draws from the distribution of potential outcomes. We account for this by clustering standard errors at the person level.

Instrumental Variables We exclude dose (length of stay) from our analysis out of concern that it is endogenous and may be introducing omitted variable bias. However, many people who were assigned to a treatment group did not actually participate – or participated for very little time – and it would be helpful to understand what the effects of the treatments were on those who were actually treated (that is, the TOT effect).[13] The length of stay in a treatment program likely contains two sources of variation: 1) variation that is random (based on treatment assignment) and useful for identifying the effects of participation, and 2) variation that is driven by omitted variables. We use an instrumental variables strategy to isolate the random variation in participation, using a stay of at least 30 days as the threshold for 'participation'. We do this using a 2SLS regression with the following specification:

$$Y_{it} = \beta_0 + \alpha_i + \lambda_t + \beta_1(\widehat{OH30days}) + \beta_2(\widehat{TC30days}) + \epsilon_{it}, \tag{7}$$

where $\widehat{OH30days}$ and $\widehat{TC30days}$ are generated by the following first stage regressions:

$$OH30days = \gamma_0 + \alpha_i + \lambda_t + \gamma_1(OH * Post_t) + \gamma_2(TC * Post_t) + u_{it} \tag{8}$$

$$TC30days = \delta_0 + \alpha_i + \lambda_t + \delta_1(OH * Post_t) + \delta_2(TC * Post_t) + w_{it} \tag{9}$$

As above, outcome variables for individual i in survey t are represented by Y_{it}, while α_i and λ_t are individual and survey wave fixed effects, respectively. $TC_i * Post_t$ ($OH_i * Post_t$) is an interaction between those assigned to the TC (OH) treatment and an indicator for

[13]The majority of participants had left their treatment facilities by the first follow up survey (six months after treatment assignment).

whether the survey was conducted after treatment assignment. $OH30days$ and $TC30days$ are indicators of whether an individual participated in their assigned program for at least 30 days.

This method first identifies the effect of being assigned to a certain treatment group on participation (staying at least 30 days), and then, using only the variation in participation that was caused by treatment assignment, estimates the effect of participation on the outcome of interest. Since treatment was assigned randomly, isolating the variation in participation caused by treatment assignment allows us to circumvent any potential omitted variable bias. This allows us to estimate TOT effects.

The TOT effect represents the programs' effects on the compliers – that is, the type of people who participate in the programs when given the opportunity. These effects may not generalize to the full sample. However, they can be interpreted as suggestive evidence on what might happen to the full sample if program administrators can find a way to increase participation rates.

4.3.1 Extension Results: ITT effect

Table 5, Columns 1-3 of Panel B, shows results after changing the functional form from a comparison of intercepts and slopes to a difference-in-differences design. This design adds a *Post* variable that distinguishes the baseline/pre-treatment observations from post-treatment observations. This makes the results easier to interpret. Column 1 shows the results for days worked. Assignment to the OH group increases time worked by 2.6 days per month (56%, $p < 0.10$). Assignment to the TC group reduces time worked by 2.2 days per month (48%, $p < 0.10$), despite the program's requirement that participants be employed.

Consistent with these employment results, Column 2 shows that assignment to the OH group increases income, by \$150 per month on average (52%, n.s.). Assignment to the TC group reduces income by \$220 per month (76%, $p < 0.05$).

Column 3 considers effects on days incarcerated. Assignment to the OH group increases

time incarcerated by 1.8 days per month (87%, n.s.). Assignment to the TC group increases time incarcerated by 0.91 days per month (43%, n.s.).

Columns 4-6 of Panel A in Table 5 use the original specification but remove the endogenous controls, add individual fixed effects, and cluster the standard errors by individual. The estimates change, sometimes substantially. The treatment group fixed effects drop out of the analysis, since they do not vary within individual over time. The treatment*time coefficients remain, showing how outcomes change differentially over time across groups.

Columns 4-6 of Panel B combine these changes: they use a difference-in-difference specification with the new set of control variables and clustered standard errors. These are our preferred results, and can be interpreted as ITT effects of the programs. Assignment to the OH group increases days worked by 1.1 days per month (24%, n.s.), increases income by $40 per month (14%, n.s.), and increases incarceration by 2.3 days per month (108%, $p < 0.10$). Assignment to the TC group reduces days worked by 2.3 days per month (50%, $p < 0.10$), reduces income by $238 per month (82%, $p < 0.05$), and increases days incarcerated by 1.6 per month (75%, n.s.).

ITT effects for other outcome measures are in Table A4, Panel B, columns 6-10.

4.3.2 Extension Results: TOT effect

The first stage effects of treatment assignment on participation (staying at least 30 days) are shown in Table A5. Assignment to the OH group increases the likelihood of participating in OH for at least 30 days by 65%; assignment to the TC group increases the likelihood of participating in TC for at least 30 days by 52%.

TOT effects are shown in columns 7-9 of Panel B of Table 5. Participation in the OH treatment for at least 30 days increases days worked by 1.7 days per month (37%, n.s.) and income by $62 per month (21%, n.s.). It also increases days incarcerated by 3.5 days per month (167%, $p < 0.10$).

Participation in the TC program for at least 30 days reduces employment by 4.5 days

per month (96%, $p < 0.10$), reduces earnings by $458 per month (159%, $p < 0.05$), and increases days incarcerated by 3.0 per month (145%, n.s.).

TOT effects for other outcome measures are in columns 11-15 of Panel B in Table A4.

4.4 Discussion

Using an RCT, Jason, Olson and Harvey (2014) study the effects of two aftercare treatment models on a variety of outcomes for justice-involved individuals with histories of substance abuse. Using their data, we replicate and extend their statistical analyses. We focus on improving causal identification by eliminating potential sources of omitted variable bias. We implement a difference-in-differences design to utilize the panel nature of the data, while accounting for baseline imbalances across groups. We add individual fixed effects to increase precision of our estimates and account for the unbalanced nature of the panel, and we cluster standard errors at the individual level. Finally, we drop the endogenously-determined dose (length of treatment) variable as a control and instead instrument for program participation (at least 30 days) with random treatment assignment, to estimate a TOT effect.

These changes affect the significance and magnitude of the results, and change the interpretation of the original study's findings. We find suggestive evidence that assignment to Oxford Houses increased employment and income, but we also find that it increased days incarcerated. Assignment to Therapeutic Communities reduced employment and income, and also may have increased days incarcerated. For both treatment groups, the TOT estimates imply that program participation caused 3-3.5 additional incarceration days per month, relative to a control group mean of 2.1 days. Unfortunately the standard errors on these estimates are wide; the study does not have sufficient statistical power to measure these effects with precision. Column 2 of Table 3 shows that with the original sample (270 participants), and assuming no attrition due to survey non-response (which could be achieved if administrative data were used for all outcome measures), the minimum detectable effect (at the 5% level) is an 85% change in days incarcerated. To detect a 5% change in days incarcerated, the study would have needed over 77,000 participants.

5 Study 3: MCORP

5.1 The Original Study

Duwe (2014) evaluates the effectiveness of the Minnesota Comprehensive Offender Reentry Plan (MCORP), a prisoner reentry project aimed at reducing recidivism. Launched in 2008, MCORP focused on improving the delivery of services and programming by forging a more collaborative relationship between institutional caseworkers and supervision agents in the community. This collaboration aimed to provide planning, support, and direction for offenders to address their strengths and needs in both the institution and the community.

The MCORP evaluation was designed as an RCT. Offenders meeting certain eligibility criteria were randomly assigned to MCORP or a control group that received standard reentry services. This set of requirements included: (1) have committed their original offense in one of the five pilot counties (Hennepin, Ramsey, Dodge, Filmore, and Olmsted), (2) be incarcerated at one of 7 participating correctional institutions (Shakopee, Lino Lakes, Stillwater, Rush City, Red Wing, Moose Lake, and St. Cloud), (3) have a scheduled release date from prison that precedes the end of the pilot program, (4) have at least six months of community supervision remaining on their sentence, and (5) not have a requirement to register as a predatory offender (all sex offenders were excluded from the study). On top of these, participants also had to meet four additional requirements: (1) be released from prison into one of the five counties, (2) not participate in one of the MNDOC's early release program, (3) be released to regular supervised release rather than intensive supervised release, and (4) not have any detainers, warrants, or holds that would jeopardize participation. Information relevant to these final four criteria was typically not available until after randomization occurred. This complicated the analysis.

After eligible offenders were randomly assigned to either the MCORP or control group, caseworkers established a transition accountability plan. This plan involved caseworkers' reviewing offender file information, administering a risk and needs assessment, and interviewing offenders to determine their motivation related to interventions based on their risk

and needs. Caseworkers developed guides for what offenders would need to accomplish while in prison to prepare for release. To promote greater case planning and management continuity between the institution and the community, the caseworker included the assigned supervision agent in the case planning process as early as possible during an offender's confinement. Due to the additional case planning, caseload sizes for caseworkers involved with MCORP were expected to be half that of regular caseloads. Under status quo reentry planning, supervision agents seldom have any contact with offenders on their caseloads until the offenders are released from prison.[14]

As mentioned above, information relevant to some of the eligibility criteria was not available until after treatment assignment. This means that some participants (concentrated in the treatment group) were excluded from the study once those criteria were checked. As a result, the original sample suffered from non-random attrition after treatment assignment, which may have introduced selection bias.[15] About 63% of the treatment sample and 51% of the control sample was dropped from the study, which suggests that greater scrutiny was applied to treatment group members. Because the reasons for being dropped from the study appear correlated with risk level, it is likely that the treatment and control groups are no longer balanced in terms of their propensity to reoffend.

In Panel A of Table 6 we compare baseline characteristics for individuals ultimately included in the treatment and control groups. (All information is based on administrative data from the Department of Corrections, and so the study avoids sample attrition over time.) As expected, the remaining samples are unbalanced on several observable characteristics, including sex, age at release, and criminal history. To account for these imbalances, the original study controls for all observable characteristics. Of these, many are measured post-

[14]Those assigned to MCORP with only a few months remaining in their sentence were not exposed to the full program as designed. The original author codes those participants as in "Phase 1" (versus "Phase 2") of the program to account for this, and controls for Phase in the regressions. One could consider these to be different intensities of treatment and analyze the data accordingly; we follow the original study and simply control for Phase rather than considering an interaction of Phase with treatment assignment.

[15]The three most common reasons for a participant's being excluded were: 1) intensive supervised release (ISR) placement, 2) early release, or 3) released to supervision in a non-MCORP county.

randomization. Post-randomization variables could be affected by treatment assignment – that is, they might actually be outcomes. These variables are: release year, age at release, LSI-R score, the county an offender was released to, length of stay in prison, whether an offender received institutional discipline, whether an offender had a secondary degree at release, whether an offender entered a prison-based chemical dependency (CD) treatment program, and whether they had a release revocation.

Individuals are followed through the end of the experiment, regardless of their date of release. This means that the length of the post-release followup period (during which recidivism is possible) varied across participants. This would not necessarily be a problem if the followup periods were balanced across treatment and control groups, but Table 6 shows that the treatment group is released significantly earlier (0.14 years, $p < 0.05$) than the control group. This means that the treatment group had more time to recidivate than the control group did; this could bias results toward finding detrimental effects of the program. In addition, releases occurred shortly before or during the Great Recession; this difference in release dates means that those in the treatment group were more likely to be released before the recession began. A number of studies show that being released at a time when the local labor market is strong reduces recidivism (Raphael and Weiman, 2002; Yang, 2017; Schnepel, 2018). This difference in release dates could bias results toward finding more beneficial effects of the program. We will control for release year to reduce these biases, despite its being determined post-randomization. (We will also also control for age at release, because age is an important predictor of recidivism risk.) However, we note that analyzing the data based on original treatment assignment (including everyone randomized, regardless of subsequent eligibility) would likely have avoided this problem.

To analyze how MCORP affected recidivism, Duwe (2014) implements a Cox regression model, arguing that survival models are preferable because they consider not only whether offenders recidivated, but also how long it took them to reoffend (i.e., fail to "survive" in

the community).[16]

The original study considers effects of MCORP on five measures of recidivism: whether a prisoner was arrested for a new offense after release, whether a prisoner was reconvicted for a new offense after release, whether a prisoner was incarcerated for a new offense after release, whether a prisoner was reincarcerated due to revocation of parole for a technical violation after their release, and whether a prisoner was incarcerated for any reason (revocation or a new offense) after release. The original findings suggest that MCORP significantly reduced four of these five measures. We focus on the results for rearrest, reincarceration for a new offense, and any return to incarceration in our main replication and extension analyses, and provide results for the other measures in Table A6). We focus on these three outcomes because they effectively summarize the broader set of outcomes available.

5.2 Replication

We begin with a simple comparison of means, shown in Panel B of Table 6. Within the study sample, assignment to MCORP is associated with a reduction in the likelihood of rearrest and the likelihood of return to prison due to a technical violation of parole. However, due to baseline imbalances between the treatment and control groups it is unlikely that these associations represent the causal effects of treatment.

In columns 1-3 of Panel A of Table 7, we reproduce the estimates from the original study. In columns 4-6 we show our replication of those results. We are able to replicate the original study's point estimates exactly, though our standard errors are off by a small amount (perhaps due to our using a different statistical analysis software). Coefficients are hazard ratios, so an estimate of 1 implies no effect. These replicated results suggest that MCORP lowered the hazard ratio for all recidivism outcomes between 18 and 23 percent (though the effect on incarceration for a new offense is not statistically significant). In other words, at any time t following release, participants in MCORP were 18 to 23 percent less

[16] In essence, Cox regression models are a class of survival models that relate the time that passes (prior to some event occurring) to variables that could be associated with that quantity of time. Cox regressions yield hazard ratios, which can be interpreted as the chance of an event occurring in the treatment group divided by the chance of the event occurring in the control group.

likely to recidivate, conditional on not yet having reoffended.

5.3 Extension

Panel B of Table 7 switches to an OLS regression model instead of the Cox hazard model. We do this largely because survival model estimates can be difficult to interpret, and we want to be able to directly compare estimates from this study to related studies. OLS produces easy-to-interpret estimates of the marginal effects of treatment. The outcome of interest is now whether an event occurred at any time during the follow-up period, rather than the time-to-event. Estimated effects of treatment are qualitatively similar, but smaller in magnitude: the estimates in columns 4-6 of Panel B imply that MCORP reduced recidivism by 5 to 8 percentage points (11-16% of the respective control group means).

The ideal method for estimating the causal effect of the MCORP program would require obtaining information on the complete original sample, including individuals' treatment assignments and outcomes. We would then compare the means of the treatment and control groups to calculate the ITT effect of MCORP, and use assignment to MCORP as an IV for MCORP participation to measure the TOT effect. Unfortunately, information on all original participants is unavailable in this case. We use matching methods as a next-best alternative, to somewhat improve upon the use of OLS with controls. These methods construct observationally-equivalent treatment and comparison groups from within the set of post-attrition participants; instead of simply controlling for observable characteristics, this approach restricts the sample to those who look similar at baseline.

The goal of matching is to compare people across treatment and control groups who have similar propensities to reoffend. However, matching on observable characteristics alone may not eliminate selection bias; there may still be differences in unobservable (to the researcher) characteristics that are related to recidivism risk. In this context, offenders were more likely to be identified as ineligible and excluded from the study if they had been assigned to the treatment group than if they had been assigned to the control group. This means that the control group likely contains individuals who should have been excluded based on their risk

level (which is not perfectly observable). Our goal is to limit the overall sample to those who would not have been excluded even if they had been assigned to the treatment group (where eligibility received closer scrutiny). Because people were originally randomized across groups, it is plausible that observationally-equivalent people in the treatment and control groups are equivalent in terms of unobservable characteristics as well. The identifying assumption of this exercise – that matched offenders are equivalent on unobservable characteristics – is more plausible than it might be if, for instance, initial treatment assignment had been based on motivation or good behavior.

Panels C and D of Table 7 show results when matched comparison groups are used. We use two common matching methods: Propensity Score Matching (PSM) and Inverse Probability Weighting (IPW).[17] Results based on PSM and IPW matching are qualitatively similar to OLS. Estimates in columns 4-6 of Panels C and D suggest that assignment to MCORP reduced the likelihood of a rearrest by 9-10 percentage points (11-13% of the control group mean, $p < 0.01$), the likelihood of reincarceration for a new offense by 4-6 percentage points (13-20%, n.s.), and the likelihood of any return to incarceration by 8-11 percentage points (15-21%, $p < 0.05$).

The other change we make in our extension analysis is to drop covariates determined post-randomization. We do this because these variables may themselves have been affected by treatment assignment (recall that the program involved working with participants while they were still incarcerated). In this context it is not obvious whether this was the optimal choice. It is possible that these characteristics were determined pre-randomization and were then used to determine eligibility (that is, they become the basis for selection into the final sample). Related, these characteristics may proxy for unobservable characteristics – such as motivation – that may have affected eligibility. In such scenarios it would be correct to include these covariates as controls. We cannot tell when exactly these variables were determined, and so opt to exclude them (with two exceptions, described below); the original

[17] More information on the matching methods used – along with supporting tables and figures – is provided in Appendix B.

author made the opposite choice. It is likely that using data on the full sample as initially randomized (that is, not excluding those deemed ineligible) would have avoided this dilemma.

Columns 7-9 in Table 7 amend each specification to drop these post-randomization covariates, with two exceptions. We control for release year to account for opportunity to reoffend as well as changes in the local labor market, as described above. We also include age at release, because age is an important predictor of recidivism. (A more clearly exogenous covariate would be age at randomization, but that is unavailable.) Columns 7-9 in Panel A show results with these amended covariates using the Cox hazard model; Panel B uses OLS, and Panels C and D use PSM and IPW matching methods, respectively.

Dropping post-randomizaton covariates has minimal effect on the Cox, OLS, and IPW estimates, but shrinks the PSM estimate substantially due to the change in the underlying weights. The PSM results suggest that participants were 5.3 percentage points less likely to be rearrested (7%, n.s.), 2.6 percentage points less likely to be reincarcerated for a new offense (8%, n.s.), and 5.9 percentage points less likely to be reincarcerated for any reason (11%, n.s.) than individuals in the control group. These PSM coefficients still suggest economically meaningful effects on recidivism, but they are not precisely estimated. The PSM, IPW, and OLS estimates are not statistically distinguishable from each other.

To help guide future research in this area, we perform power calculations based on the data from the RCT. In Table 3 we show that the minimum detectable effect in the original study (with 689 participants) is a 12.5% change in the likelihood of a rearrest (at the 5% level). To detect a 5% change in this outcome measure, this study would have needed over 4,300 participants.

5.4 Discussion

Consistent with the original study, our analysis provides evidence that the MCORP program significantly reduced participants' likelihood of being rearrested, incarcerated for a new offense, or incarcerated for any reason.

Interpreting these results as causal requires that inclusion in the MCORP (treatment)

group is uncorrelated with individuals' baseline propensity to reoffend. Through PSM and IPW methods, we match and weight offenders conditional on observables. However, we cannot test whether the samples are balanced on unobservable characteristics that may have been used to determine eligibility after treatment was assigned. Future research should make sure that outcome data are available for all offenders who were randomly assigned to either treatment or control, to enable standard ITT and TOT analyses based on original treatment assignment. Following all participants for the same length of time after release would also ease analysis and interpretation of results.

6 How these studies fit into the literature on prisoner reentry

Doleac (2019a) reviews the literature on desistance from crime, including existing empirical evidence on the effects of various programs and policies on prisoner reentry outcomes. The above analyses contribute new evidence to relatively thin literatures in three areas: SCF programs, aftercare programs for those with substance-use disorders, and wrap-around services.

A number of recent RCTs have attempted to replicated the initial success of the HOPE program in Hawaii. DYT was part of this batch of RCTs, and the authors of that evaluation concluded that DYT had no impact on participants. Combined with null effects from other RCTs of similar programs, this contributed to a sense that HOPE (and SCF more broadly) did not replicate in other contexts. Our results above suggest that this punchline may be misleading. The DYT experiment cannot rule out large beneficial effects of the program on participants, and in fact the point estimates suggest meaningful benefits.

Therapeutic Communities (TCs) are a popular form of treatment for people struggling with addiction. Existing rigorous studies consider the effects of TCs for people during and after incarceration, and results are mixed. The study re-analyzed above provides evidence that TCs substantially reduce days worked and income earned. It finds no significant effect on recidivism (days incarcerated), but the point estimate suggests a meaningful increase. This study therefore contributes evidence against TC's effectiveness.

Oxford Houses are another form of treatment for people with addiction, and this is the first rigorous evaluation we know of of this type of program for formerly-incarcerated individuals. Across the full population assigned to the OH group (the ITT effect), the current study finds suggestive evidence of increases in employment but also finds a large, statistically significant increase in days incarcerated. The estimated TOT effect implies that participating in OH for at least 30 days increases days incarcerated by 3.5 days per month. Future research should aim to understand these mixed results.

Finally, MCORP is a holistic program that fits into a growing literature on wrap-around services for people coming out of prison. Our extension analysis largely supports the initial study's findings that the program improved participants' outcomes (reducing recidivism). However, without data on all participants as originally assigned to the treatment and control groups, we were not able to conduct ITT or TOT analyses. It is possible that the estimates are still biased due to selection on unobservables and omitted variables such as the strength of the labor market at the time of release. All other RCTs of similar programs find null or detrimental effects (see Doleac, 2019c, for a review, and Doleac, 2019b, for a discussion of how these RCT results differ from results based on matched comparison group designs). The MCORP results therefore contrast with the existing literature. If this program is achieving the large gains estimated above, then this is an important finding and the program should be replicated elsewhere. A follow-up RCT with all data retained for complete ITT and TOT analyses would allow us to confirm that the results above represent the true causal effects of the program. After that, replication RCTs in other places would reveal whether similar programs can achieve similar gains in other contexts.

7 Conclusion

Our extended analyses provide unbiased (or less biased, in the case of the MCORP reanalysis) causal estimates of these three prisoner reentry programs. We show that selection and endogeneity biases matter: in two of the three studies, correcting for biases leads to conclusions that differ at least somewhat from the original studies. However, all three studies

were underpowered to detect meaningful effects on recidivism. Researchers in a position to conduct future RCTs should consider statistical power before investing time and financial resources in an experiment. Once an experiment is complete, they should be careful to analyze the data in a way that avoids introducing selection bias. And in all cases they should make their data available to other researchers, to allow replications and extensions such as the ones we've conducted here, and facilitate more rapid accumulation of knowledge.

References

Cunningham, Scott. 2018. Causal Inference: The Mixtape (V. 1.7). Tufte-Latex.GoogleCode.com.

Davidson, Janet, George King, Jens Ludwig, and Steven Raphael. 2019. "Managing Pretrial Misconduct: An Experimental Evaluation of HOPE Pretrial."

Doleac, Jennifer L. 2019a. "Encouraging desistance from crime." Working paper.

Doleac, Jennifer L. 2019b. ""Evidence-based policy" should reflect a hierarchy of evidence." Journal of Policy Analysis and Management, 38: 517–519.

Doleac, Jennifer L. 2019c. "Wrap-around services don't improve prisoner reentry outcomes." Journal of Policy Analysis and Management, 38: 508–514.

DuRose, Matthew R., Alexia D. Cooper, and Howard N. Snyder. 2014. "Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010." Bureau of Justice Statistics Special Reprot, NJS 244205.

Duwe, Grant. 2014. "A randomized experiment of a prisoner reentry program: updated results from an evaluation of the Minnesota Comprehensive Offender Reentry Plan (MCORP)." Criminal Justice Studies, 27.

Gelman, Andrew, and Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Hamilton, Zachary, Christopher M. Campbell, Jacqueline van Wormer, Alex Kigerl, and Brianne Posey. 2016. "Impact of Swift and Certain Sanctions: Evaluation of Washington State's Policy for Offenders on Community Supervision." Criminology & Public Policy, 15: 1009–1072.

Hawken, Angela, and Mark A. R. Kleiman. 2009. "Managing Drug Involved Probationers with Swift and Certain Sanctions: Evaluating Hawaii's HOPE." DOJ report number 229023, available at https://www.ncjrs.gov/pdffiles1/nij/grants/229023.pdf.

Hawken, Angela, and Mark A. R. Kleiman. 2011. "Washington Intensive Supervision Program: Evaluation Report."

Jason, Leonard A., Bradley D. Olson, and Ronald Harvey. 2014. "Evaluating Alternative Aftercare Models for Ex-Offenders." Journal of Drug Issues, 45.

Lattimore, Pamela K., Doris Layton MacKenzie, Gary Zajac, Debbie Dawes, Elaine Arsenault, and Stephen Tueller. 2016. "Outcome Findings from the HOPE Demonstration Field Experiment: Is Swift, Certain, and Fair an Effective Supervision Strategy?" Criminology & Public Policy, 15: 1103–1141.

O'Connell, Daniel J., John J. Brent, and Christy A. Visher. 2016. "Decide Your Time: A Randomized Trial of a Drug Testing and Graduated Sanctions Program for Probationers." Criminology and Public Policy, 15: 1073–1102.

Raphael, Steven, and David Weiman. 2002. "The Impact of Local Labor Market Conditions on the Likelihood that Parolees are Returned to Custody." Working paper.

Schnepel, Kevin T. 2018. "Good Jobs and Recidivism." Economic Journal, 128: 447–469.

Sribney, William, and Vince Wiggins. n.d.. "Standard errors, confidence intervals, and significance tests for ORs, HRs, IRRs, and RRRs." StataCorp LLC resources and support, available at https://www.stata.com/support/faqs/statistics/delta-rule/.

Yang, Crystal S. 2017. "Local labor markets and criminal recidivism." Journal of Public Economics, 147: 16–29.

8 Figures and Tables

Table 1: DYT: Summary Statistics

	Full Sample				Analysis Sample			
	All	DYT (Treatment)	Standard Probation (Control)	Difference	All	DYT (Treatment)	Standard Probation (Control)	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Baseline Characteristics								
Age at Randomization†	29.77	29.77	29.75	0.028	29.76	29.63	29.89	-0.260
	(9.041)	(9.182)	(8.924)	(0.910)	(9.072)	(9.181)	(8.988)	(0.935)
Male	0.848	0.855	0.840	0.015	0.848	0.853	0.844	0.009
	(0.360)	(0.353)	(0.368)	(0.036)	(0.359)	(0.355)	(0.364)	(0.037)
White	0.463	0.455	0.470	-0.015	0.455	0.442	0.469	-0.027
	(0.499)	(0.499)	(0.500)	(0.050)	(0.500)	(0.498)	(0.500)	(0.051)
Age at First Adult Arrest†	20.88	20.71	21.05	-0.342	20.74	20.46	21.02	-0.559
	(4.609)	(4.339)	(4.866)	(0.464)	(4.316)	(3.822)	(4.741)	(0.444)
Panel B: Outcomes								
In Analysis Sample	0.955	0.950	0.960	-0.010				
	(0.208)	(0.218)	(0.196)	(0.021)				
Arrest for New Crime	0.470	0.450	0.490	-0.040	0.461	0.437	0.484	-0.048
	(0.500)	(0.499)	(0.501)	(0.050)	(0.499)	(0.497)	(0.501)	(0.051)
Incarceration	0.623	0.600	0.645	-0.05	0.623	0.600	0.646	-0.046
	(0.485)	(0.491)	(0.480)	(0.049)	(0.485)	(0.491)	(0.480)	(0.050)
Employment†	0.403	0.442	0.365	0.078	0.403	0.442	0.365	0.078
	(0.491)	(0.500)	(0.483)	(0.050)	(0.491)	(0.498)	(0.483)	(0.050)
Failed Drug Test	0.713	0.780	0.645	0.135***	0.723	0.784	0.661	0.123***
	(0.453)	(0.415)	(0.480)	(0.045)	(0.448)	(0.412)	(0.474)	(0.046)
Arrest for Any Crime	0.758	0.760	0.755	0.005	0.754	0.758	0.750	0.008
	(0.429)	(0.428)	(0.431)	(0.043)	(0.431)	(0.429)	(0.434)	(0.044)
Arrest for Violation of Probation	0.708	0.710	0.705	0.005	0.704	0.711	0.698	0.013
	(0.455)	(0.455)	(0.457)	(0.046)	(0.457)	(0.455)	(0.460)	(0.047)
Arrest for Technical Violation of Probation	0.288	0.310	0.265	0.045	0.293	0.321	0.266	0.055
	(0.453)	(0.464)	(0.442)	(0.045)	(0.456)	(0.468)	(0.443)	(0.047)
Completed Probation†	0.500	0.473	0.525	-0.052	0.503	0.466	0.536	-0.071
	(0.501)	(0.501)	(0.501)	(0.051)	(0.501)	(0.500)	(0.500)	(0.052)
Drug Treatment	0.473	0.485	0.460	0.025	0.474	0.474	0.474	0.000
	(0.500)	(0.501)	(0.500)	(0.050)	(0.500)	(0.501)	(0.501)	(0.051)
Percent Drug Tests Failed	0.637	0.441	0.832	-0.391***	0.639	0.448	0.828	-0.380***
	(0.353)	(0.340)	(0.239)	(0.029)	(0.351)	(0.343)	(0.240)	(0.030)
Missed Appointment with Probation Officer	0.355	0.430	0.280	0.150***	0.356	0.426	0.286	0.140***
	(0.479)	(0.496)	(0.450)	(0.047)	(0.479)	(0.496)	(0.453)	(0.049)
Absconded	0.043	0.070	0.015	0.055***	0.045	0.074	0.016	0.059***
	(0.202)	(0.256)	(0.122)	(0.020)	(0.206)	(0.262)	(0.124)	(0.021)
Observations	400	200	200	400	382	190	192	382

Note: Columns 1-4 include all participants where data are available. Columns 5-8 restrict attention to the participants included in our analysis, where data are available for all necessary variables. Columns 4 and 8 show the difference in average values between Columns 2 and 3 and Columns 6 and 7, respectively. The outcome measures in Panel B are binary indicators based on an 18-month followup period. Standard deviations/errors in parentheses. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. †Data on these variables are missing for some participants. Number of observations in columns 1-4 are as follows: Age at randomization – 396 total, 196 treated, 200 control. Age at first adult arrest – 395 total, 196 treated, 199 control. Employment – 382 total, 190 treated, 192 control. Completed probation – 384 total, 184 treated, 200 control.

Table 2: DYT: Main Results

	Original Results		**Our Results**				
			Original Covariates		Amended Covariates		
	Arrest for New Crime (1)	Incarceration (2)	Arrest for New Crime (3)	Incarceration (4)	Arrest for New Crime (5)	Incarceration (6)	Employment (7)
Panel A: MLL							
Odds Ratios							
DYT	0.88	0.66	0.828	0.662*	0.825	0.839	1.486
	(0.22)	(0.17)	(0.185)	(0.159)	(0.172)	(0.182)	(0.364)
Implied Marginal Effects							
DYT	-0.032	-0.104	-0.047	-0.103*	-0.048	-0.044	0.099
	(0.063)	(0.064)	(0.056)	(0.060)	(0.052)	(0.054)	(0.061)
Panel B: OLS							
Coefficients/Marginal Effects							
DYT			-0.046	-0.086**	-0.047	-0.040	0.089
			(0.043)	(0.041)	(0.047)	(0.040)	(0.055)
Control Group Mean	0.484	0.646	0.484	0.646	0.484	0.646	0.365
Observations	377	377	377	377	377	377	377
Controls:							
Sex	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X
Age at randomization	X	X	X	X	X	X	X
Age at first adult arrest	X	X	X	X	X	X	X
Employed	X	X	X	X			
Missed appointments	X	X	X	X			
Drug treatment	X	X	X	X			
Failed drug tests	X	X	X	X			

Note: Coefficients show the effect of assignment to the DYT group on various outcomes (listed at the top of each column). Panel A uses an MLL model as in the original study. Coefficients are odds ratios, so 1 implies no effect. Implied marginal effects are included to ease comparison with Panel B, which uses an OLS model. All outcomes are binary measures based on an 18-month followup period. Standard errors are in parentheses; in the OLS regressions they are clustered by probation officer. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Power Calculations
Recidivism

	DYT (1)	Aftercare (2)	MCORP (3)
Total Sample Size in Original Study	400	270	689
Smallest Percentage Effect Detectable w/Original Sample	28.6%	85.0%	12.5%
Sample Needed Per Group to Detect 5% Effect	6,531	25,758	2,152
Total Sample Needed to Detect 5% Effect	13,062	77,274	4,303

Note: Each column displays power calculations for the DYT, Aftercare, and MCORP studies, respectively. They are based on the following recidivism outcomes: for the DYT study, we use arrest for a new crime; for the Aftercare study, we use days detained or incarcerated; and for the MCORP study, we use re-arrest. Calculations assume 80% power and a level of significance of 5%.

Table 4: Aftercare: Summary Statistics

	All (1)	Oxford House (2)	Theraputic Community (3)	Control (4)	OH: Difference from Control (5)	TC: Difference from Control (6)
Panel A: Baseline Characteristics						
Age	40.59	39.04	43.16	39.48	-0.436	3.685**
	(0.615)	(1.030)	(0.982)	(1.131)	(1.533)	(1.497)
Female	0.172	0.234	0.173	0.113	0.121**	0.060
	(0.025)	(0.059)	(0.042)	(0.036)	(0.060)	(0.055)
White	0.218	0.260	0.160	0.238	0.022	-0.077
	(0.027)	(0.050)	(0.041)	(0.048)	(0.069)	(0.063)
Black	0.739	0.675	0.778	0.763	-0.087	0.015
	(0.029)	(0.054)	(0.046)	(0.048)	(0.072)	(0.067)
Graduated High School	0.298	0.429	0.198	0.275	0.154**	-0.077
	(0.030)	(0.057)	(0.045)	(0.050)	(0.076)	(0.067)
Attended College	0.105	0.078	0.099	0.138	-0.060	-0.039
	(0.020)	(0.031)	(0.033)	(0.039)	(0.050)	(0.051)
Days of Alcohol Use	21.89	17.25	22.54	25.71	-8.466	-3.169
	(2.778)	(4.151)	(4.892)	(5.285)	(6.753)	(7.198)
Days of Drug Use	44.98	46.68	44.07	44.27	2.400	-0.201
	(3.793)	(6.305)	(6.717)	(6.729)	(9.236)	(9.508)
Earnings from Employment	80.73	85.44	46.98	110.38	-24.93	-63.40
	(18.45)	(32.89)	(26.12)	(36.22)	(49.03)	(44.57)
Illegal Earnings	62.60	110.5	37.65	41.75	68.76	-4.096
	(21.05)	(60.17)	(15.70)	(17.55)	(61.68)	(23.53)
Days of Paid Work	1.605	1.442	1.198	2.175	-0.733	-0.977
	(0.345)	(0.548)	(0.493)	(0.727)	(0.916)	(0.877)
Legal Problems	0.173	0.166	0.157	0.197	-0.031	-0.039
	(0.012)	(0.021)	(0.018)	(0.022)	(.031)	(0.029)
Days Detained or Incarcerated	2.765	1.325	3.049	3.863	-2.538**	-0.813
	(0.468)	(0.521)	(0.805)	(0.999)	(1.139)	(1.282)
Psychiatric Hospitalizations	1.122	1.273	0.667	1.438	-0.165	-0.771
	(0.263)	(0.441)	(0.161)	(0.638)	(0.781)	(0.654)
Participants	238	77	81	80	157	161
Panel B: Main Outcomes						
Participate for 30+ Days		0.699	0.507			
		(0.054)	(0.061)			
Days of Paid Work	7.762	10.50	4.966	8.138	2.365**	-3.172***
	(0.390)	(0.726)	(0.560)	(0.694)	(1.004)	(0.886)
Earnings from Employment	468.6	677.1	238.4	515.9	161.2*	-277.5***
	(32.01)	(72.61)	(31.66)	(54.62)	(90.37)	(62.09)
Days Detained or Incarcerated	1.093	0.545	1.397	1.291	-0.745*	0.107
	(0.177)	(0.203)	(0.337)	(0.345)	(0.405)	(0.483)
Observations	661	209	234	218	427	452

Note: Columns 1-4 display average values by treatment assignment. Columns 5 and 6 display the difference in means from the Control for Oxford House and Therapeutic Community, respectively. Baseline Characteristics were measured prior to treatment assignment; Main Outcomes represent the average value of those variables across all post treatment surveys. In Panel B, the unit of observation is a participant-survey-wave. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Aftercare: Main Results

	Our Results								
	Original Covariates			Ammended Covariates			TOT effects (partic. 30+ days)		
	Days Worked	Income	Days Incarcerated	Days Worked	Income	Days Incarcerated	Days Worked	Income	Days Incarcerated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: OLS									
Oxford House	-2.054*	-111.6	-1.719**						
	(1.196)	(95.07)	(0.739)						
Oxford House*Time	1.145**	63.01*	0.288	0.864*	44.35	0.386			
	(0.472)	(37.53)	(0.292)	(0.441)	(33.11)	(0.299)			
Therapeutic Community	-2.985**	-173.6*	-0.225						
	(1.168)	(92.80)	(0.721)						
Therapeutic Community*Time	-0.001	-36.59	0.07	-0.095	-48.53	0.197			
	(0.469)	(37.29)	(0.290)	(0.439)	(32.89)	(0.297)			
Panel B: Difference-in-Difference									
Oxford House*Post	2.614*	149.6	1.833	1.125	40.11	2.276*	1.739	62.06	3.513*
	(1.455)	(121.9)	(1.200)	(1.487)	(129.7)	(1.268)	(2.259)	(198.1)	(1.960)
Therapeutic Community*Post	-2.235*	-220.1**	0.912	-2.335*	-238.0**	1.579	-4.490*	-457.8**	3.039
	(1.296)	(95.04)	(1.403)	(1.272)	(100.2)	(1.503)	(2.489)	(199.0)	(2.887)
Control Group Mean	4.728	288.8	2.060	4.728	288.8	2.060	4.728	288.8	2.060
Observations	899	899	899	899	899	899	899	899	899
Controls:									
Age	X	X	X						
Time spent in program	X	X	X						
Individual FEs				X	X	X	X	X	X

Note: Panel A shows results using the authors' original OLS specification. Panel B shows our extended analysis results using a difference-in-differences model. Outcomes are indicated by the column titles. Columns 1-6 represent ITT effects; columns 7-9 show TOT effects, using treatment assignment as an IV for whether individuals spent at least 30 days in their assigned program. Standard errors are shown in parentheses; in Panel B they are clustered at the individual level. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: MCORP: Summary Statistics

	All	MCORP	Control	Difference
	(1)	(2)	(3)	(4)
Panel A: Baseline Characteristics				
Male	0.930	0.949	0.901	0.048**
	(0.009)	(0.010)	(0.018)	(0.019)
Minority	0.722	0.696	0.762	-0.066*
	(0.017)	(0.018)	(0.016)	(0.034)
Age at Release (years)	35.05	36.12	33.43	2.694***
	(0.385)	(0.509)	(0.574)	(0.781)
Prior Supervision Failures	1.751	1.951	1.448	0.502***
	(0.079)	(0.109)	(0.108)	(0.160)
Prior Convictions	6.544	7.031	5.806	1.224***
	(0.197)	(0.277)	(0.260)	(0.401)
LSI-R Risk Assessment Score	27.05	26.85	27.35	-0.503
	(0.272)	(0.346)	(0.439)	(0.556)
Admission Type: New Commitment	0.595	0.614	0.565	0.048
	(0.018)	(0.023)	(0.029)	(0.038)
Admission Type: Probation Violation	0.269	0.267	0.273	-0.006
	(0.016)	(0.021)	(0.026)	(0.034)
Admission Type: Release Violation	0.134	0.118	0.160	-0.042
	(0.013)	(0.015)	(0.022)	(0.026)
Offense Type: Violent	0.227	0.228	0.226	0.002
	(0.015)	(0.020)	(0.025)	(0.032)
Offense Type: Property	0.275	0.296	0.244	0.051
	(0.017)	(0.022)	(0.026)	(0.034)
Offense Type: Drug	0.198	0.171	0.240	-0.069**
	(0.015)	(0.018)	(0.025)	(0.030)
Offense Type: DWI	0.123	0.122	0.124	-0.001
	(0.012)	(0.016)	(0.019)	(0.025)
Offense Type: Other	0.171	0.178	0.160	0.017
	(0.014)	(0.018)	(0.022)	(0.029)
County of release: Hennepin	0.586	0.616	0.540	0.076**
	(0.018)	(0.023)	(0.030)	(0.038)
County of release: Ramsey	0.345	0.322	0.379	-0.056
	(0.018)	(0.022)	(0.029)	(0.037)
County of release: DFO	0.068	0.060	0.080	-0.020
	(0.009)	(0.011)	(0.016)	(0.19)
Length of Stay (months)	18.38	18.40	18.35	0.051
	(0.496)	(0.614)	(0.833)	(1.014)
Disciplinary Infractions	2.632	2.559	2.744	-0.185
	(0.117)	(0.140)	(0.206)	(0.240)
Secondary Degree at Release	0.783	0.824	0.722	0.101
	(0.015)	(0.018)	(0.027)	(0.031)
Entered Prison-Based Drug Treatment	0.261	0.274	0.240	0.033
	(0.016)	(0.021)	(0.025)	(0.034)
Release Year	2008	2008	2009	-0.140**
	(0.033)	(0.043)	(0.051)	(0.068)
Panel B: Outcomes				
Rearrest	0.725	0.701	0.762	-0.061*
	(0.017)	(0.022)	(0.025)	(0.034)
Reconviction	0.606	0.583	0.642	-0.059
	(0.018)	(0.024)	(0.029)	(0.038)
Reincarceration: New Offense	0.298	0.293	0.306	-0.012
	(0.017)	(0.022)	(0.027)	(0.035)
Reincarceration: Parole Revocation	0.335	0.306	0.379	-0.073**
	(0.017)	(0.022)	(0.029)	(0.036)
Reincarceration: Any	0.487	0.465	0.521	-0.056
	(0.019)	(0.024)	(0.030)	(0.038)
Observations	689	415	274	689

Note: Columns 1-3 are average values. Column 4 shows the difference in average value for MCORP and control. Standard errors are shown in parentheses. Significance levels in column 4 are indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: MCORP: Main Results

	Original Results			**Our Results**					
				Original Covariates			Amended Covariates		
	Rearrest	New Offense Reincarceration	Any Return	Rearrest	New Offense Reincarceration	Any Return	Rearrest	New Offense Reincarceration	Any Return
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Replication - Cox Model									
MCORP	0.801*	0.819	0.765*	0.801**	0.819	0.765**	0.816**	0.819	0.746**
	(0.095)	(0.150)	(0.116)	(0.076)	(0.123)	(0.089)	(0.072)	(0.112)	(0.081)
Panel B: Extension - OLS									
MCORP				-0.083**	-0.049	-0.074*	-0.071**	-0.046	-0.082**
				(0.034)	(0.035)	(0.038)	(0.033)	(0.034)	(0.038)
Panel C: Extension - PSM									
MCORP				-0.100***	-0.062	-0.109**	-0.053	-0.026	-0.059
				(0.037)	(0.046)	(0.043)	(0.038)	(0.041)	(0.040)
Panel D: Extension - IPW									
MCORP				-0.086***	-0.039	-0.077**	-0.077**	-0.046	-0.092**
				(0.033)	(0.034)	(0.038)	(0.032)	(0.034)	(0.037)
Control Group Mean	0.762	0.306	0.521	0.762	0.306	0.521	0.762	0.306	0.521
Observations	689	689	689	689	689	689	689	689	689
Controls:									
Phase	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X	X	X
Criminal/supervision history	X	X	X	X	X	X	X	X	X
Age at release	X	X	X	X	X	X	X	X	X
Release year	X	X	X	X	X	X	X	X	X
LSI-R score	X	X	X	X	X	X			
County of release	X	X	X	X	X	X			
Disciplinary infractions	X	X	X	X	X	X			
Drug treatment	X	X	X	X	X	X			
Secondary degree	X	X	X	X	X	X			
Length of stay	X	X	X	X	X	X			
Release revocation	X	X	X	X	X	X			

Note: Coefficients show the effect of assignement to MCORP on recidivism (specific outcome listed at the top of each column). Panel A shows hazard ratios, so a coefficient of 1 implies no effect. Panel B uses Ordinary Least Squares (OLS), Panel C uses Propensity Score Matching (PSM), and Panel D uses Inverse Probability Weighting (IPW); the coefficients in all three represent marginal effects. Standard errors are in parentheses. Significance levels are indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A Additional tables and figures – for online publication only

Table A1: DYT: Additional Outcomes

VARIABLES	Failed Drug Test (1)	Any Arrest (2)	Arrest for Violation of Probation (3)	Arrest for Technical Violation of Probation (4)	Completed Probation (5)	Referral to/ Enrollment in Drug Treatment (6)	% Drug Tests Failed (7)	Missed Appointment with Probation Officer (8)	Absconded (9)
Panel A: Original Results (MLL)									
Odds Ratios									
DYT	1.58	0.77	0.78	0.90					
	(0.58)	(0.22)	(0.24)	(0.25)					
Implied Marginal Effects									
DYT	0.114	-0.065	-0.062	-0.026					
	(0.092)	(0.071)	(0.071)	(0.069)					
Panel B: Our Replicated Results (MLL, original covariates)									
Odds Ratios									
DYT	0.783	0.892	0.932	1.162					
	(0.256)	(0.245)	(0.237)	(0.295)					
Implied Marginal Effects									
DYT	-0.061	-0.029	-0.018	0.037					
	(0.082)	(0.069)	(0.064)	(0.063)					
Controls for Panels A & B:									
Sex	X	X	X	X					
Race	X	X	X	X					
Age at randomization	X	X	X	X					
Age at first adult arrest	X	X	X	X					
Employed	X	X	X	X					
Missed appointments	X	X	X	X					
Drug treatment	X	X	X	X					
Failed drug tests	X	X	X	X					
Panel C: Our Extended Results (OLS, amended covariates)									
Coefficients/Marginal Effects									
DYT	0.131***	0.020	0.025	0.067	-0.065	0.007	-0.388***	0.146***	0.055***
	(0.042)	(0.054)	(0.049)	(0.053)	(0.055)	(0.102)	(0.029)	(0.049)	(0.020)
Controls for Panel C:									
Sex	X	X	X	X	X	X	X	X	X
Race	X	X	X	X	X	X	X	X	X
Age at randomization	X	X	X	X	X	X	X	X	X
Age at first adult arrest	X	X	X	X	X	X	X	X	X
Control Group Mean	0.661	0.750	0.698	0.266	0.536	0.474	0.828	0.286	0.016
Observations	377	377	377	377	362	377	377	377	377

Note: Panels A, B, and C, show original, replicated, and extended results, respectively, for six drug use and recidivism outcomes. Panels A and B use an MLL model as in the original analysis. Coefficients are odds ratios, so 1 implies no effect. Implied marginal effects are included to ease comparison with Panel C, which uses an OLS model. All outcomes are based on an 18-month follow-up period; except for column 7, all outcomes are binary measures. Standard errors are in parentheses; in Panel C they are clustered by probation officer. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Aftercare: Summary Statistics (full sample)

	All (1)	Oxford House (2)	Theraputic Community (3)	Control (4)	OH: Difference from Control (5)	TC: Difference from Control (6)
Panel A: Baseline Characteristics						
Age	40.43	39.19	43.28	38.83	0.356	4.444***
	(0.579)	(0.946)	(0.911)	(1.087)	(1.441)	(1.418)
Female	0.170	0.244	0.156	0.111	0.133**	0.044
	(0.023)	(0.046)	(0.038)	(0.033)	(0.056)	(0.051)
White	0.211	0.244	0.144	0.244	0.000	-0.100 *
	(0.025)	(0.046)	(0.037)	(0.046)	(0.064)	(0.059)
Black	0.741	0.689	0.789	0.744	-0.056	0.044
	(0.027)	(0.049)	(0.043)	(0.046)	(0.067)	(0.063)
Graduated High School	0.296	0.422	0.189	0.278	0.144**	-0.089
	(0.028)	(0.052)	(0.041)	(0.047)	(0.071)	(0.063)
Attended College	0.100	0.078	0.1000	0.122	-0.044	-0.022
	(0.018)	(0.028)	(0.032)	(0.035)	(0.045)	(0.047)
Days of Alcohol Use	20.07	16.00	20.71	23.53	0.885	0.929
	(2.510)	(3.667)	(4.451)	(4.859)	(8.514)	(8.906)
Days of Drug Use	44.80	45.08	45.12	44.19	-7.534	-2.823
	(3.521)	(5.755)	(6.313)	(6.279)	(6.080)	(6.584)
Legal Issues - Composite Score	0.173	0.168	0.153	0.198	-0.030	-0.045
	(0.012)	(0.021)	(0.018)	(0.021)	(0.030)	(0.028)
Psychiatric Hospitalizations	1.120	1.218	0.764	1.378	-0.159	-0.614
	(0.241)	(0.395)	(0.195)	(0.571)	(-0.698)	(0.606)
Illegal Earnings	60.52	100.1	33.89	48.88	51.23	-14.99
	(19.13)	(53.35)	(14.17)	(18.94)	(55.88)	(23.66)
Days of Paid Work	1.458	1.314	1.102	1.944	-0.630	-0.842
	(0.312)	(0.492)	(0.455)	(0.650)	(0.821)	(0.796)
Earnings from Employment	77.99	84.93	45.61	103.67	-18.74	-58.06
	(16.80)	(30.22)	(23.71)	(32.67)	(44.57)	(40.37)
Days Detained or Incarcerated	2.692	1.186	3.182	3.663	-2.477**	-0.481
	(0.434)	(0.468)	(0.778)	(0.907)	(1.031)	(1.196)
Observations	266	87	89	90	177	179
Panel B: Main Outcomes						
In Analysis Sample	0.594	0.560	0.634	0.589	-0.025	0.044
	(0.015)	(0.026)	(0.025)	(0.026)	(0.037)	(0.036)
Days of Paid Work	7.750	10.45	4.950	8.167	2.281**	-3.217**
	(0.383)	(0.717)	(0.550)	(0.680)	(0.988)	(0.871)
Earnings from Employment	464.6	681.3	236.6	502.6	178.7**	-266.0***
	(31.52)	(72.60)	(30.90)	(52.79)	(89.02)	(60.32)
Days Detained or Incarcerated	1.121	0.659	1.436	1.218	-0.559	0.218
	(0.177)	(0.234)	(0.334)	(0.327)	(0.407)	(0.468)
Observations	688	214	243	231	445	474

Note: Columns 1-3 display average values by treatment assignment. Columns 4 and 5 display the difference in means from the Control for Oxford House and Therapeutic Community, respectively. Baseline Characteristics were measured prior to treatment assignment while Main Outcomes represent the average value of those variables across all post treatment surveys. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Aftercare: Main Outcomes (full sample)

	Our Results								
	Original Covariates			Amended Covariates			TOT effects (participated 30 days)		
	Days Worked (1)	Income (2)	Days Incarcerated (3)	Days Worked (4)	Income (5)	Days Incarcerated (6)	Days Worked (7)	Income (8)	Days Incarcerated (9)
Panel A: OLS									
	(1.135)	(89.97)	(0.708)						
Oxford House	-2.056*	-94.88	-1.432**						
	(1.135)	(89.97)	(0.708)						
Oxford House*Time	1.157**	64.27*	0.245	1.008**	49.18	0.411			
	(0.455)	(35.96)	(0.283)	(0.424)	(30.88)	(0.286)			
Therapeutic Community	-2.907***	-158.4*	0.054						
	(1.108)	(87.73)	(0.691)						
Therapeutic Community*Time	-0.020	-37.206	0.031	-0.076	-42.79	0.192			
	(0.452)	(35.74)	(0.282)	(0.423)	(30.73)	(0.284)			
Panel B: Difference-in-Difference									
Oxford House*Post	2.506*	168.8	1.941*	1.598	62.84	2.534**	2.352	234.0	-1.997**
	(1.387)	(124.2)	(1.104)	(1.384)	(114.6)	(1.169)	(1.868)	(161.4)	(0.852)
Therapeutic Community*Post	-2.534**	-221.2**	0.676	-2.478**	-216.7**	1.461	-5.383**	-430.3***	0.136
	(1.220)	(91.07)	(1.321)	(1.158)	(89.57)	(1.400)	(95.16)	(0.049)	(0.605)
Control Group Mean	4.560	273.1	2.003	4.560	273.1	2.003	4.560	273.1	2.003
Observations	945	949	951	916	919	924	945	949	951
Controls:									
Age	X	X	X						
Time spent in program	X	X	X						
Individual FEs				X	X	X	X	X	X

Note: Panel A shows results using the authors' original OLS specification. Panel B shows our extended analysis results using a difference-in-difference model. Outcomes are indicated by the column titles. Columns 1-6 represent ITT effects; columns 7-9 show TOT effects, using treatment assignment as an IV for whether individuals spent at least 30 days in their assigned program. Standard errors are shown in parentheses; in Panel B they are clustered at the individual level. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Aftercare: Additional Outcomes

	Our Results														
	Original Covariates					Ammended Covariates					TOT effects (partic. 30+ days)				
	Days of Alcohol Use (1)	Days of Drug Use (2)	Illegal Income (3)	Legal Issues (4)	Psych. Hosp. (5)	Days of Alcohol Use (6)	Days of Drug Use (7)	Illegal Income (8)	Legal Issues (9)	Psych. Hosp. (10)	Days of Alcohol Use (11)	Days of Drug Use (12)	Illegal Income (13)	Legal Issues (14)	Psych. Hosp. (15)
Panel A: OLS															
OH	-2.555	1.439	71.03	-0.021	-0.179										
	(5.550)	(6.981)	(44.30)	(0.024)	(0.316)										
OH*Time	-0.563	-3.419	-27.21	-0.006	-0.004	-1.151	-2.181	-28.34	-0.007	-0.104					
	(2.191)	(2.756)	(17.49)	(0.009)	(0.125)	(1.747)	(2.411)	(17.67)	(0.009)	(0.110)					
TC	1.703	-1.743	1.595	-0.032	-0.604*										
	(5.417)	(6.815)	(43.24)	(0.023)	(0.309)										
TC*Time	1.939	-1.093	1.087	0.015	0.153	0.009	-2.150	-7.309	0.014*	0.075					
	(2.177)	(2.738)	(17.38)	(0.009)	(0.124)	(2.178)	(2.745)	(17.36)	(0.009)	(0.110)					
Panel B: Difference-in-Difference															
OH*Post	2.935	-14.50	-76.85	-0.001	-0.093	-0.602	-13.71	-63.63	-0.010	-0.017	-0.930	-21.15	-98.22	-0.016	-0.027
	(6.954)	(10.71)	(76.84)	(0.035)	(0.733)	(7.442)	(11.36)	(89.15)	(0.037)	(0.746)	(11.44)	(17.46)	(135.6)	(0.057)	(1.145)
TC*Post	10.42	-9.668	-5.333	0.044	0.603	2.349	-16.28	-14.48	0.052	0.564	4.519	-31.33	-27.87	0.101	1.085
	(7.454)	(11.02)	(43.91)	(0.033)	(0.604)	(7.431)	(11.05)	(43.55)	(0.034)	(0.635)	(14.00)	(22.71)	(86.43)	(0.067)	(1.216)
Control Mean	20.58	38.18	55.36	0.150	0.759	20.58	38.18	55.36	0.150	0.759	20.58	38.18	55.36	0.150	0.759
Observations	899	899	899	899	899	899	899	899	899	899	899	899	899	899	899
Controls:															
Age	X	X	X	X	X										
Time in program	X	X	X	X	X										
Individual FEs						X	X	X	X	X	X	X	X	X	X

Note: Panel A shows results using the authors' original OLS specification. Panel B shows our extended analysis results using a difference-in-differences model. Outcomes are indicated by the column titles. Columns 1-10 represent ITT effects; columns 11-15 show TOT effects, using treatment assignment as an IV for whether individuals spent at least 30 days in their assigned program. Standard errors are shown in parentheses; in Panel B they are clustered at the individual level. Significance levels indicated by: * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A5: Aftercare: First Stage

	Stayed 30 or More Days in OH (1)	Stayed 30 or More Days in TC (2)
Random Treatment Assignment		
Oxford House*Post	0.648***	0.000
	(0.058)	(0.000)
Therapeutic Community*Post	-0.000	0.520***
	(0.000)	(0.060)
Observations	899	899

Note: Each column is a separate regression using treatment assignment as an IV for whether individuals spent at least 30 days in their assigned program. Individual fixed effects are included and standard errors (shown in parentheses) are clustered at the individual level. Significance levels indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: MCORP: Additional Outcomes

	Original Results		**Our Results**			
			Original Covariates		Amended Covariates	
	Reconviction (1)	Tech. Violation Revocation (2)	Reconviction (3)	Tech. Violation Revocation (4)	Reconviction (5)	Tech. Violation Revocation (6)
Panel A: Replication - Cox Model						
MCORP	0.790*	0.748*	0.790**	0.748**	0.809***	0.713**
	(0.103)	(0.139)	(0.082)	(0.104)	(0.082)	(0.098)
Panel B: Extension - OLS						
MCORP			-0.095***	-0.077**	-0.082**	-0.086**
			(0.036)	(0.037)	(0.035)	(0.036)
Panel C: Extension - PSM						
MCORP			-0.103**	-0.100**	-0.054	-0.080**
			(0.044)	(0.045)	(0.043)	(0.040)
Panel D: Extension - IPW						
MCORP			-0.088**	-0.082**	-0.084**	-0.098***
			(0.035)	(0.037)	(0.034)	(0.036)
Control Group Mean	0.642	0.379	0.642	0.379	0.642	0.379
Observations	689	689	689	689	689	689
Controls:						
Phase	X	X	X	X	X	X
Sex	X	X	X	X	X	X
Race	X	X	X	X	X	X
Criminal/supervision history	X	X	X	X	X	X
Age at release	X	X	X	X	X	X
Release year	X	X	X	X	X	X
LSI-R score	X	X	X	X		
County of release	X	X	X	X		
Disciplinary infractions	X	X	X	X		
Drug treatment	X	X	X	X		
Secondary degree	X	X	X	X		
Length of stay	X	X	X	X		
Release revocation	X	X	X	X		

Note: Coefficients how the effect of assignement to MCORP on recidivism (specific outcome listed at the top of each column). Panel A shows hazard ratios, so a coefficient of 1 implies no effect. Panel B uses Ordinary Least Squares (OLS), Panel C uses Propensity Score Matching (PSM), and Panel D uses Inverse Probability Weighting (IPW); the coefficients in all three represent marginal effects. Standard errors are in parentheses. Significance levels are indicated by: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Matching analysis for MCORP study – for online publication only

Implementing matching requires estimating propensity scores (i.e., the probability of being assigned to treatment conditional on observable baseline characteristics) followed by selecting an algorithmic method to estimate the average treatment effect (ATE). Common matching algorithms are propensity score matching (PSM) and inverse probability weighting (IPW).[18] We show results based on both methods.

In Figure A1, we show the distribution of estimated propensity scores separately for the MCORP and control samples. In this particular figure, the propensities are calculated using only the covariates included in the regression for rearrest (note we dropped post-randomization variables from the model to produce these figures with the exception of release age and release year); distributions using other baseline covariates look similar. While the distribution for MCORP is shifted to the right (likely because a larger share of low-risk individuals were dropped from treatment after randomization occurred), the two groups have almost identical distributions. This is evidence that, based on observables – and despite *actual* treatment assignment – individuals in both the control group and MCORP were equally likely to be *assigned* to treatment and that characteristics of individuals in the treatment group are similar to those in the control.[19]

Figure A2 displays the densities of propensity scores for the MCORP and control group prior to matching (left panel) and after matching (right panel). As seen in the right panel, after matching the treatment and control groups have perfectly balanced propensity scores.[20]

Tables A7 and A8 display the covariate balance summary of raw data next to those using the PSM and IPW methods. Specifically, Table A7 displays the number of observations pre-

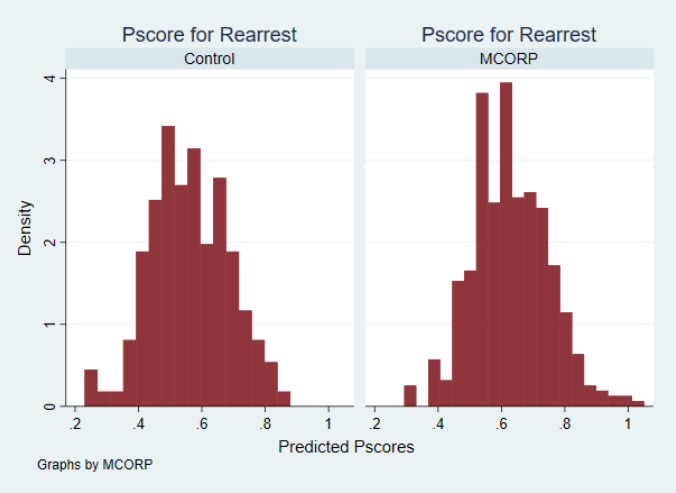
[18] PSM estimators impute the missing potential outcomes for each subject by using an average of the outcomes of similar subjects that received the other treatment level. The treatment effect is computed by taking the average of the difference between the observed and potential outcomes for each subject. IPW weights each treatment and control subjects by the predicted propensity score.

[19] See Cunningham (2018) for more information regarding propensity scores and propensity score matching.

[20] Again, we show densities for propensity scores calculated using covariates from the regression on rearrest, but note that the densities using other baseline covariates follow a similar pattern.

and post-matching, and Table A8 displays the standardized differences in means between the two groups pre- and post-matching (we exclude post-randomization variables here). Comparing columns (1) and (2) of Table A8 reveals that the matching procedure achieved balance on observables; after matching, the standardized differences between MCORP and the control group are nearly zero for all covariates.[21] Given that 1) both PSM and IPW yield observationally-equivalent comparison groups and 2) individuals were initially randomly assigned to treatment and control, the identification assumption that these comparison groups are also balanced on unobservable characteristics is plausible, albeit not directly testable.

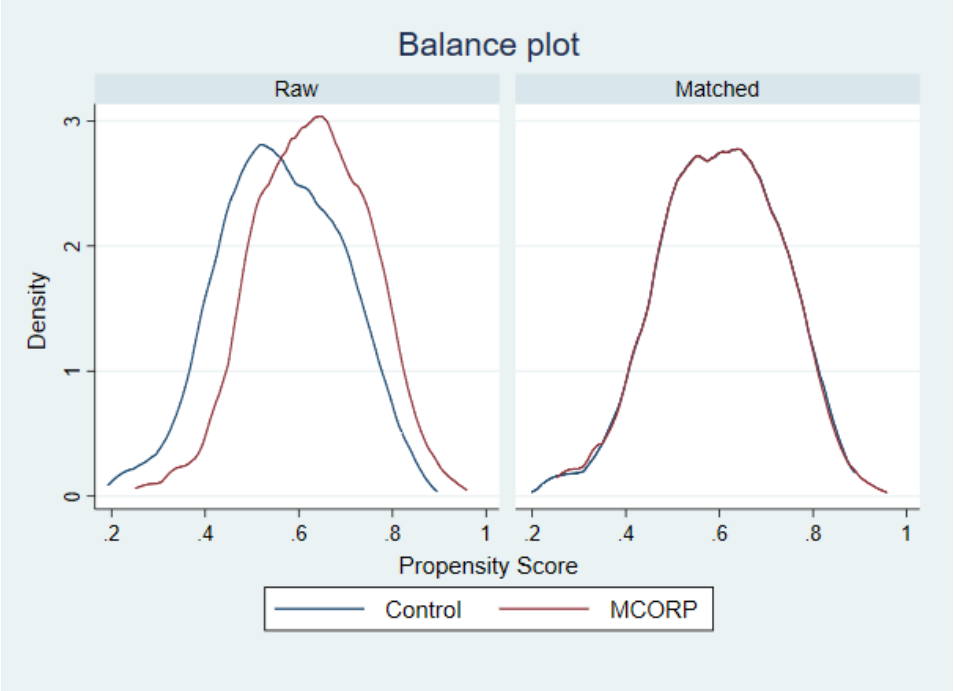
Figure A1: MCORP: Histogram of pscores for Rearrest as outcome



Note: Here, we show the density of estimated propensity scores using the covariates from the regression for rearrest. In the left panel, we show the densities of propensity scores using the raw data. In the right panel, we show densities using matched observations with the PSM method. These results come from models that exclude post-randomization covariates with the exception for release age and release year.

[21]Note that these tables examine the covariates used in the regression of rearrest. Results for other covariates look similar.

Figure A2: MCORP - Density of Propensity Scores Pre- and Post-Matching



Note: Here, we calculate propensity scores using the covariates from the regression for rearrest. In the left panel, we show the densities of propensity scores for both the MCORP and control groups prior to the matching exercise. In the right panel, we show the densities of propensity scores for both groups after implementing matching with the PSM method. These results come from models that exclude post-randomization covariates with the exception of release age and release year.

Table A7: MCORP: Number of Observations for PSM and IPW

	Raw (1)	PSM (matched) (2)	IPW (Weighted) (3)
Number of obs =	689	1,378	689
Treated obs =	415	689	343.3
Control obs =	274	689	345.7

Note: This table uses rearrest as an outcome; results for other outcomes look similar. These observations come from models that exclude post-randomization covariates with the exception of release age and release year.

Table A8: MCORP: Covariate Balance Summary: Standardized Differences

	Raw (1)	PSM (2)	IPW (3)
Phase	-0.162	-0.144	-0.022
Male	0.182	-0.017	-0.002
Minority	-0.149	0.038	0.015
Prior Supervision Failures	0.248	0.019	-0.003
Prior Convictions	0.243	0.014	0.006
Probation Violator	-0.014	-0.069	-0.015
Release Violator	-0.122	0.026	0.001
Property	0.116	0.003	0.018
Drug	-0.172	0.114	0.001
DWI	-0.003	-0.029	0.007
Other	0.047	-0.038	0.001
Release Age	0.270	-0.033	0.013
Release Year	-0.160	-0.113	-0.004

Note: This table uses rearrest as an outcome; results for other outcomes look similar. These results exclude post-randomization covariates with the exception of release age and release year.