ESSAYS ON THE ECONOMICS OF CRIME

A Dissertation

by

CHELSEA EVAN TEMPLE

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jennifer Doleac
Committee Members,	Mark Hoekstra
	Jonathan Meer
	Kalena Cortes
Head of Department,	Steve Puller

May 2022

Major Subject: Economics

ABSTRACT

In this work, I present three essays on the economics of crime. The first paper (joint with Jennifer Doleac, David Pritchard, and Adam Roberts) replicates and extends the analyses of data from three randomized controlled trials (RCTs) related to prisoner reentry in order to more cleanly identify the causal effects of treatment. This is important given that the way data from an RCT are collected and analyzed can unintentionally reintroduce omitted variable and selection biases even though RCTs are designed to avoid such biases. 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. The second paper evaluates prosecutorial reform. While there is a breadth of evidence showing prosecutors' abilities to affect case outcomes, little is known about whether prosecutors affect criminal justice contact in the first place. I answer this question in the context of decriminalization in Seattle, Washington. My results do not indicate any significant effects of prosecutorial reform on recidivism. The third paper (joint with Maya Mikdash) studies crisis intervention team (CIT) units, which aim to reduce police use of force against and unnecessary incarceration of individuals with mental illnesses - particularly those in crisis. Using data from El Paso, Texas, we find suggestive evidence that dispatching a CIT unit to mental health crisis calls reduces the likelihood of arrest, but increases the likelihood of low-level force. Additionally, our empirical approach provides a model for evaluating CIT units and similar interventions in other police departments.

DEDICATION

To God be the glory. The pages shown here are but a tiny piece of all that has happened in the past six years. Thank you, Lord, for carrying me throughout this program. I am thankful for all that You have done in and through me.

ACKNOWLEDGMENTS

In addition to my committee members and coauthors, I would like to thank the following people for supporting me throughout this program: my family (Mom and Dad, Drew, and Grandma); Grace Mears; Dennis Jansen and the Private Enterprise Research Center; Andie Kelly; my Department of Economics cohort (Benjamin Priday, David Pritchard, Adam Roberts, Mayra Pineda Torres, Hedieh Tajali, Adam Bestenbostel, Xiaoxiao Bai, Shuo Tian, and Mingze Huang); Chao Man, CarlyWill Sloan; Scott Cunningham; Courtney Collins; Valerie and David Webb; Marshall and Teresa Gramm; Tod and Erin Berkey; Kaye and Steve Tseng; Kaylah Kennard; Andrew and Rebecca Simon; Lauren Meyer; and my Rhodes track girls (Lindsay, Sadie, Mary Kate, Morgan, and Liz). I am immensely grateful for each of you. Last but not least, I would like to thank my wonderful husband-to-be, Ben Strickland. I love you and could not have finished this without you.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Jennifer Doleac (advisor), Mark Hoekstra, and Jonathan Meer of the Department of Economics and Professor Kalena Cortes of the Bush School of Government and Public Service.

The analyses in Chapter I were conducted in part by Jennifer Doleac, David Pritchard, and Adam Roberts of the Department of Economics. The analyses in Chapter III were conducted in part by Maya Mikdash of the Department of Economics.

All other work for the dissertation was conducted independently in collaboration with Jennifer Doleac, Mark Hoekstra, and Jonathan Meer.

Funding Sources

Graduate study was supported by fellowships from the Department of Economics and the Private Enterprise Research Center and dissertation research fellowships from the College of Liberal Arts and the Institute for Humane Studies.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. WHICH PRISONER REENTRY PROGRAMS WORK? REPLICATING AND EXTENDING ANALYSES OF THREE RCTS*

1.1 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-

*Reprinted with permission from: Doleac, Jennifer, Chelsea Temple, David Pritchard, and Adam Roberts. "Which prisoner reentry programs work? Replicating and extending analyses of three RCTs." 2020. International Review of Law and Economics, 62, Article 105902.

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 inter-

[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).

pret 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 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 1.2 lays out recommendations for analyzing data from an RCT. Section 1.3 discusses the "Decide Your Time" SCF program; Section 1.4 discusses aftercare programs for recently-released, drug-involved offenders; and Section 1.5 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 1.6 discusses our findings in the context of the broader literature on prisoner reentry programs, and Section 1.7 concludes.

1.2 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

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

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.

1.3 Study 1: Decide Your Time

1.3.1 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 1.1. 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 the randomization ‘worked’: the two groups are balanced

[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.

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.

1.3.2 Replication

We begin our replication by presenting a simple comparison of means in Panel B of Table 1.1. 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:

$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha + \beta DYT_{ij} + \theta X_{ij} + \epsilon_j, \tag{1.1}$$

where

$$\pi_{ij} = E(y_{ij}) = Pr(y_{ij} = 1), \tag{1.2}$$

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 recidivating for probationers assigned to the same

[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.

probation officer.

Results from the original paper are shown in Table 1.2, 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 A.1.

1.3.3 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-group correlations in the error term. We thus run OLS with standard errors clustered at the probation officer level,

[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.

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:

$$Y_i = \alpha + \beta DYT_i + \theta X_i + \epsilon_i, \tag{1.3}$$

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 1.2. 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 A.1 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 appointment 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.

1.3.4 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 1.3 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 (particularly incarceration and employment), instead of simple binary measures of whether a probationer was ever

[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.

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.

1.4 Study 2: Aftercare

1.4.1 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 1.4 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 A.2.

1.4.2 Replication

We begin with a simple comparison of means, shown in Panel B of Table 1.4. 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]

$$Y_{it} = \beta_0 + \beta_1 OH_i + \beta_2 TC_i + \theta_1 (TC_i * Time_t) + \theta_2 (OH_i * Time_t) + \\ \rho Time_t + \eta Dose_i + \gamma Age_i + \epsilon_{it} \quad (1.4)$$

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 controls for dose, which is the time each

[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 A.2 and A.3.

[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.

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 1.5, 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 A.4, Panel A, columns 1-5.

1.4.3 Extension

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

1.4.3.0.1 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:

$$Y_{it} = \beta_0 + \lambda_t + \beta_1 TC_i + \beta_2 OH_i + \theta_1(TC_i * Post_t) + \theta_2(OH_i * Post_t) + \gamma X_{it} + \epsilon_{it} \qquad (1.5)$$

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.

1.4.3.0.2 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 1.4, 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.

1.4.3.0.3 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:

$$Y_{it} = \theta_0 + \alpha_i + \lambda_t + \theta_1(TC_i * Post_t) + \theta_2(OH_i * Post_t) + \epsilon_{it}, \qquad (1.6)$$

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 1.5).

1.4.3.0.4 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.

1.4.3.0.5 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:

$$Y_{it} = \beta_0 + \alpha_i + \lambda_t + \beta_1(\widehat{OH30days}) + \beta_2(\widehat{TC30days}) + \epsilon_{it}, \tag{1.7}$$

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

$$OH30days = \gamma_0 + \alpha_i + \lambda_t + \gamma_1(OH * Post_t) + \gamma_2(TC * Post_t) + u_{it} \tag{1.8}$$

$$TC30days = \delta_0 + \alpha_i + \lambda_t + \delta_1(OH * Post_t) + \delta_2(TC * Post_t) + w_{it} \tag{1.9}$$

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 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

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

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.

1.4.3.1 Extension Results: ITT effect

Table 1.5, 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 1.5 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 A.4, Panel B, columns 6-10.

1.4.3.2 Extension Results: TOT effect

The first stage effects of treatment assignment on participation (staying at least 30 days) are shown in Table A.5. 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 1.5. 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 A.4.

1.4.4 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 1.3 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.

1.5 Study 3: MCORP

1.5.1 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

[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.

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 1.6 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-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 1.6 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

[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.

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 A.6). We focus on these three outcomes because they effectively summarize the broader set of outcomes available.

1.5.2 Replication

We begin with a simple comparison of means, shown in Panel B of Table 1.6. Within the study sample, assignment to MCORP is associated with a reduction in the likelihood of rearrest and

[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.

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 1.7, 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 likely to recidivate, conditional on not yet having reoffended.

1.5.3 Extension

Panel B of Table 1.7 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 1.7 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

[17] More information on the matching methods used – along with supporting tables and figures – is provided in Appendix A.1.

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 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 1.7 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 1.3 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.

1.5.4 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.

1.6 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.

1.7 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.

1.8 Figures and Tables

Table 1.1: DYT: Summary Statistics

	Full Sample				Analysis Sample			
		DYT	Standard Probation			DYT	Standard Probation	
	All	(Treatment)	(Control)	Difference	All	(Treatment)	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

†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. Reprinted with permission from Doleac et al. (2020).

Table 1.2: DYT: Main Results

	Original Results		**Our Results**				
			Original Covariates		Amended Covariates		
	Arrest for New Crime	Incarceration	Arrest for New Crime	Incarceration	Arrest for New Crime	Incarceration	Employment
	(1)	(2)	(3)	(4)	(5)	(6)	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table 1.3: 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%. Reprinted with permission from Doleac et al. (2020).

Table 1.4: Aftercare: Summary Statistics

	All	Oxford House	Theraputic Community	Control	OH: Difference from Control	TC: Difference from Control
	(1)	(2)	(3)	(4)	(5)	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table 1.5: Aftercare: Main Results

	Our Results								
	Original Covariates			Ammended Covariates			TOT effects (partic. 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									
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. Reprinted with permission from Doleac et al. (2020).

Table 1.6: 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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table 1.7: 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 assignment 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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

2. PROSECUTORIAL REFORM

2.1 Introduction

Prosecutors play a tremendous role in the criminal justice system. Between an officer's arrest and a judge's conviction, prosecutors are the primary decision-maker for every aspect of the case. Prosecutors determine whether charges will be filed, whether to recommend pretrial detention or bail, and the type and timing of evidence that will be revealed to the defense attorney. Prosecutors are also the main drivers in plea bargain negotiations, through which approximately 95% of cases are resolved (Devers, 2011). Additionally, prosecutors influence the makeup of the jury, which witnesses will take the stand, and the defendant's sentence. Undeniably, prosecutors hold a significant amount of power, and economists have begun to document the ways in which it is used. Regarding sentencing, studies indicate that prosecutors can either hinder or facilitate mandatory minimums (Bjerk (2005), Tuttle (2021)). Other papers have shown that prosecutorial discretion contributes to racial disparities in conviction and sentencing (Rehavi and Starr (2014), Tuttle (2021), Sloan (2022)). Additionally, Yang (2016) provides evidence that prosecutors respond to scarce resources by dismissing more cases, and Krumholz (2019) and Arora (2019) show that case outcomes are affected by the district attorney's party affiliation and race.

In recent years, prosecutors' offices have increasingly used their power to institute criminal justice reform once an individual has been charged - altering plea guidelines, pretrial detention protocol, cash bail, and which sentences to pursue – largely in efforts to decrease mass incarceration and create a more equitable system (Brennan Center for Justice, 2018). Prosecutors have also used their discretion earlier in the process by modifying charging standards. While current work documents the extent to which prosecutors matter once a case goes to court, little is known about whether prosecutors can effect change on the front end. The only evidence to date shows that when prosecutors choose not to prosecute nonviolent misdemeanor offenses, the likelihood of reoffending within two years falls - with particularly large reductions for first-time offenders (Agan, Doleac and Harvey, 2021). In this paper, I provide more evidence on the efficacy of prosecutorial reform before a case goes to court. Specifically, I ask: when prosecutors decriminalize minor drug possession, do affected individuals experience reduced criminal justice contact? To answer this question, I take advantage of a 1 gram threshold for filing charges in a difference-in-differences design. In comparing individuals holding just under 1 gram to individuals holding just over 1 gram, I find no significant effects on recidivism. This is in direct contrast to the existing work of Agan, Doleac and Harvey (2021) and points to the need for more research in this area.

My paper proceeds as follows. I provide background information on the reform in section 2.2. Section 2.3 outlines the data and section 2.4 explains the identification strategy. I discuss results in section 2.5 and section 2.6 concludes.

2.2 Background

On September 6, 2018, the King County, Washington prosecuting attorney announced that prosecutors would no longer file charges for the possession of *any* drug that is less than 1 gram, fewer than 5 pills, or a single syringe or less. There were, however, two caveats to this. First, charges would be filed if possession was committed with another felony offense or a DUI (King County Prosecuting Attorney's Office, 2019). Second, marijuana was formerly decriminalized in 2012: adults 21 and older can legally possess up to 1 ounce of useable marijuana.[1] It is also important to note that during my study period, the possession of any controlled substance was still illegal in the state of Washington (with some exceptions for marijuana). Thus, while drug possession remains illegal, prosecutors are choosing not to prosecute these cases.

The process resembles the following scenario. Suppose a Seattle Police Department (SPD) officer stops a suspect on the street holding some amount of a drug in her hands. The officer will arrest the individual on probable cause, drive her back to the precinct, and put her in a holding cell. Next, the officer conducts a field test to identify the drug, and then weighs the drug.[2] Prior to 2017,

[1] Marijuana was decriminalized in the state of Washington on December 9, 2012. Adults over 21 are allowed to purchase up to one ounce of usable marijuana, up to 16 ounces of marijuana-infused edibles in solid form and up to 72 ounces in liquid form, and up to 7 grams of marijuana concentrates (Washington State Liquor and Cannabis Board, 2021).

[2] The field test uses a negligible amount of the drug, so it should not have an effect on the drug weight.

two officers witnessed the weighing of the drug. Now that SPD officers wear body cameras, one officer records the weighing process. This can be done by either the arresting officer or another officer assisting the case. The scale displays the drug weight to the nearest tenth, and the officer records the amount as shown on the scale.[3] The officer then packages the drug, documents the arrest, and speaks to his/her sergeant, as a sergeant screens every arrest for SPD. The officer then books the suspect, and arranges for her to stand before a judge at a first appearance hearing. At this hearing, the judge informs the suspect whether the prosecution will be filing or charges or not. If the prosecution declines to file, the suspect is released. With the new filing standard, the prosecution should decline to file charges if the suspect is found in possession of drugs weighing less than 1 gram *total* and has not also been arrested for a DUI or felony offense.

Given that the majority of drugs in the data are weighed in grams, I focus on the 1 gram cutoff. Why decline to file, and why the 1 gram threshold? In a radio interview, King County Prosecuting Attorney Dan Satterberg cited two reasons: expense on behalf of the county and disruption of defendants' lives; in a given year, prosecuting the 800 cases of minor drug offenses involved warrants, jail, and multiple appearances in court, and cost the county $3 million. No explanation was given for the chosen threshold. For context, 1 gram is approximately the size of a Sweet'N Low packet. Additionally, drugs are typically sold on the street in single dose amounts, which are often less than 1 gram (depending on the specific drug). One concern could be that prosecutors were already doing this in practice prior to the actual filing change being made. When I inquired about this, the Prosecuting Attorney's Office stated that deputy prosecuting attorneys file cases according to their filing and disposition standards.

2.3 Data

I use two datasets from the Seattle Police Department (SPD): incident reports for all offenses (offense data) and incident reports for offenses involving drugs (drug data), where an incident is defined as an arrest made by an officer. The original offense dataset is at the person-incident-offense level. For each observation, I have information on the timing of the incident, the offense,

[3] In some cases, the drug is weighed in a plastic bag. This should be recorded in the suspect's case file.

and the suspect. Suspect information includes demographics (age, race/ethnicity, and gender) and a person number that uniquely identifies individuals over time. I hand-code offense categories and Revised Code of Washington (RCW) citations by comparing offense information in my dataset to the current RCW. From the RCW citations, I create RCW citation categories, which each encompass multiple offense categories. I then use these to create indicators for each offense type. I also create indicators for eleven major offense type categories: property, person, motor vehicle, white collar, drugs/alcohol, disturbance, weapon, mental health, land/wildlife, animals, and other. My final offense dataset is collapsed to the person-incident level.[4] The drug dataset is at the person-incident-drug item level. For each observation, I have information on the timing of the incident and a unique person number for the suspect. I also have information on the drugs attached to the incident report, including the type, item quantity, and measurement unit (e.g., grams, dosage unit/item, fluid ounce). I merge these datasets based on the incident number and collapse it to the person-incident level.[5] For each observation, I have indicators for offense type, drug type (if applicable), and total drug quantity (if applicable). It is important to note that, in some cases, observations without a drug offense have drugs attached to them. SPD informed me that not all individuals with drugs attached to their offense are arrested for a drug crime. Rather, if officers find drugs in the midst of an arrest, they should record the drug type and amount even if the arrest is for a non-drug offense.

This dataset covers incidents between January 2015 and February 2020. However, in May 2019 the police department switched to a new record management system that changed the final documentation of an incident. With the old system, the full list of offenses for a single incident was not kept in the final documentation. Thus, multiple offenses may be grouped under one final arrest used to clear the National Incident-Based Reporting System (NIBRS) case. With the new system, the full spectrum of seizure is retained in the final documentation. Because of inconsistencies in data recording, I restrict my sample to arrests through April 2019. I only use the data through

[4] Some observations are missing a master person ID. Given my interest in studying recidivism, I drop these observations for which I cannot track the same person over time.

[5] Not all observations in the drug data have a match with the offense data. I drop these observations.

February 2020 to construct a measure of recidivism. This allows me to calculate the ten-month recidivism rate for each arrest in my sample.[6]

My final dataset has 114,394 observations at the person-incident level between January 2015 and April 2019. Of these 114,394 observations, 7,091 are in my “drug sample", i.e. the sample of individuals for whom this filing standard would affect. With the new filing standard for drug possession, the prosecuting attorney’s office only files charges for individuals in possession of less than 1 gram, 5 pills, or a single syringe. Given that 95% of drugs in my dataset are non-pills weighed in grams, I focus on the first cutoff, excluding pills and liquids from the analysis. I also exclude marijuana, as the 1 gram limit does not apply. Thus, my drug sample indicator takes on a value of 1 for those carrying non-pill, non-liquid, non-marijuana drugs measured in grams. I restrict the data further to examine individuals with drug amounts just above and just below the 1 gram threshold. Specifically, I limit my analysis to observations with a total drug weight between 0.5 and 1.5 grams.[7] This final “analysis sample" has 1,814 person-incidents.

Tables 2.1-2.4 include summary statistics for five groups over the entire sample period: 1) all observations, 2) observations with drugs recorded, 3) all observations in my analysis sample (i.e., those holding small drug amounts between 0.5 and 1.5 grams) 4) observations in my analysis sample that are less than 1 gram (those affected by the reform), and 5) observations in my analysis sample that are 1 or more grams (those unaffected by the reform). Tables 2.1-2.4 show information on suspect characteristics, offense types, drug crimes, and drug types. All variables except for age are measured as indicators. Column 3 of each table displays information for my analysis sample. The majority of individuals are males in their mid-30s. Over 50% are white, while 30% are black. Possession is the most common drug offense, and common drug types include

[6] I first count the number of months until an individual’s subsequent arrest. If an individual is re-arrested (i.e., shows up in my dataset another time) within 10 months, then my recidivism indicator variable takes on a value of 1. I do this for every person-incident in my sample.

[7] The goal of this limitation is to identify a good control group. It is plausible to believe that individuals with *just* over 1 gram are similar to those with *just* under 1 gram. One natural cut in the data was to limit the analysis to total drug weights between 0 and 2 grams. This is because the vast majority of total drug weights are 2 grams or less. However, 2 grams is very different from 0.1 grams, for example. Thus, I limit the sample to drug weights as close to the 1 gram threshold as possible without trading off too many observations.

amphetamine/methamphetamine (58%), heroin (38%), and cocaine (22%).[8] Aside from drug offenses, the most common offense is property crime (34%). On average, individuals commit 2.5 offenses per incident and 9% reoffend within 10 months. The average number of new arrests within 10 months is 0.1. Column 4 of each table displays information separately for individuals holding just below 1 gram (the treatment group) and just above 1 gram (the control group). Importantly, these groups look similar on observables, providing confidence that the control group is a good counterfactual for the treatment group.

2.4 Identification Strategy

With the 1 gram threshold for filing charges, the natural first move would be to exploit this cutoff in a regression discontinuity design. I do not employ this method, however, due to data limitations arising from a record management system change in May 2019. When I limit my analysis to data through April 2019, the regression discontinuity design is not only underpowered due to a too-small sample size.

I instead utilize a difference-in-differences design. I define my treatment group as individuals with total drug weight between 0.5 and 0.9 grams, and my control group as individuals with total drug weight between 1 and 1.5 grams. With this model, I assume that individuals carrying just over 1 gram are good counterfactuals for those carrying just under 1 gram. Put differently, in the absence of the filing standard change, changes in recidivism would have been the same for those carrying just under versus just over 1 gram. My model takes the following form:

$$Y_{it} = \beta_1(LessThan1g_i) + \beta_2(Post_t) + \beta_3(LessThan1g * Post_{it}) + \beta_4(X_{it}) + \gamma_t + \epsilon_{it} \quad (2.1)$$

where Y_{it} is a measure of recidivism for individual i in time t; $LessThan1g_i$ is an indicator for treatment (holding drugs weighing less than 1 gram); $Post_t$ takes on a value of 1 after the filing standard change in September 2018; X_{it} is a vector individual-level controls, such as age, race, and

[8] An observation is placed in the analysis sample if it has a non-pill, non-liquid, non-marijuana drug measuring between 0.5 and 1.5 grams. Some observations, however, have multiple drugs attached to them. This is why there are nonzero percentages of observations in the analysis sample with marijuana, pills, and liquids.

gender; and γ_t is a set of time fixed effects. Specifically, I include day of month, month, and year fixed effects, which control for within-month, across-month, and across-year shocks to arrests. The coefficient of interest is β_3, which represents the difference in recidivism for those carrying small drug amounts (i.e., less than 1 gram) after the filing standard change. Standard errors are clustered at the individual level.

The validity of this design relies on the common trends assumption: the changes in recidivism between those carrying slightly less and slightly more than 1 gram must be similar in the period prior to the reform. I check for this graphically using event study plots. In figures 2.2 and 2.3 I examine two measures of recidivism. In figure 2.2, I use the ten-month recidivism rate. In figure 2.3, I use the number of subsequent arrests within 10 months as it is possible for the recidivism rate to remain unchanged even if the number of arrests changes. In addition to examining recidivism for the entire analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug who are arrested for any offense), I focus on the subsample of these individuals who are at least arrested for drug possession. While it is plausible that any individual carrying small drug amounts could be affected by the reform, I would expect a larger effect for individuals arrested for drug possession given that the filing standard change specifically applies to possession offenses.

For each event study, I regress the respective outcome on coefficients for twelve month leads and seven month lags, where time $t = 0$ represents September 2018. The $t - 12$ lead includes all data from twelve+ months prior to the reform (i.e., January 2015 through September 2017), while the $t + 7$ lag only includes data from seven months after the reform, given that the data ends in April 2019. All regressions include controls for age, race, and gender, as well as day-of-month, month, and year fixed effects. I then plot the coefficients and 95% confidence intervals from these regressions. Each coefficient in figures 2.2 and 2.3 represents the difference in recidivism between individuals holding just under 1 gram and individuals holding just over 1 gram, relative to what is expected based on pre-period trends. To be confident that the treatment and control groups were not diverging prior to the reform, the lead coefficients (those to the left of the vertical line) should be close to 0. In figures 2.2 and 2.3, this is the case for the most part. It is only in the third month prior to the reform that the coefficient jumps significantly above 0 for both samples. This indicates that the treatment group (those carrying less than 1 gram) diverged slightly from the control group (those carrying more than 1 gram) in June 2018. While any pre-period divergence is not ideal for identification, the June 2018 result is an isolated incident, lessening any worries that individuals carrying just over 1 gram are not good counterfactuals for individuals carrying just under 1 gram.

Another potential threat to identification is if the composition of the treatment and control groups is changing over time. If different individuals choose to hold different drug amounts after the reform, the effect of the reform on recidivism could be biased. Put differently, I would be less confident that I am isolating the causal effect of the reform on recidivism. I provide four pieces of evidence against this. First, I examine the distribution of total drug weight before and after the filing standard change in figure 2.1. Importantly, the distribution does not shift left following the reform, providing suggestive evidence that individuals are not strategically holding smaller drug amounts to avoid charges. The distribution also does not shift right, suggesting that officers are not manipulating the total drug weight to ensure charges.[9] Because the distribution looks similar over time, I am less concerned that the treatment and control groups are changing.

Second, I examine exogenous characteristics of these individuals before and after the reform in table 2.5. From the table, there are no notable differences in suspect characteristics within groups over time. Third, I use exogenous covariates to predict recidivism. Predicted estimates come from regressions on suspect characteristics (age, race/ethnicity, and gender) and day-of-month, month, and year fixed effects. Robust standard errors are used. In figures 2.4 and 2.5 I regress predicted outcomes on twelve month leads and seven month lags to create event study plots. Standard errors are clustered at the individual level as I would expect unobserved factors to be similar for the same individual over time. Lead and lag coefficients far from 0 would indicate differences in predicted recidivism between the treatment and control groups relative to pre-period trends. This would

[9] This supports anecdotal evidence from my contact at the department; my contact said that while theoretically possible, there are many checks in place to ensure officers are recording things correctly. Thus, the risk of being caught is very high.

be concerning, as it could indicate the baseline characteristics of individuals in these groups are changing after the reform. Because all coefficients are close to 0, I do not worry about this. Lastly, I regress predicted recidivism on equation 2.1 (excluding any controls or fixed effects). Estimates from these regressions are shown in table 2.6. None of the estimates are statistically different from zero, further increasing confidence that the composition of these groups is changing.

2.5 Results

I analyze the reform's effects on two measures of recidivism: the likelihood of reoffending within 10 months and the number of subsequent arrests within 10 months. Figures 2.2 and 2.3 show event study plots for these measures for two groups: all observations in my analysis sample (i.e., arrested for any offense) and observations in my analysis sample that have a drug possession arrest. As mentioned above, I include separate results for this subsample since the reform specifically targeted drug possession. Each of the lag coefficients (after the vertical line) represents the difference in recidivism between those holding slightly less than 1 gram and those holding just over 1 gram in the months after the reform. All coefficients in figures 2.2a-2.3b are near 0, indicating no effect. Average treatment effects are presented in table 2.7. Point estimates suggest decreases in ten-month recidivism rate. In column 1, I estimate that individuals arrested for any offense and holding just under 1 gram are 0.3 percentage points less likely to reoffend within 10 months (a 1.11% decrease off the pre-period treatment mean of 0.25). In column 3, I estimate that individuals arrested for drug possession and holding just under 1 gram are 2.5 percentage points less likely to reoffend within 10 months (a 9.36% reduction off the pre-treatment mean of 0.27). In columns 2 and 4, I estimate increases in the number of subsequent arrests by 0.05 (13.72%) and 0.02 (6.16%) for all observations and those arrested for drug possession, respectively. Coefficients, however, are not statistically different from zero and imprecise.

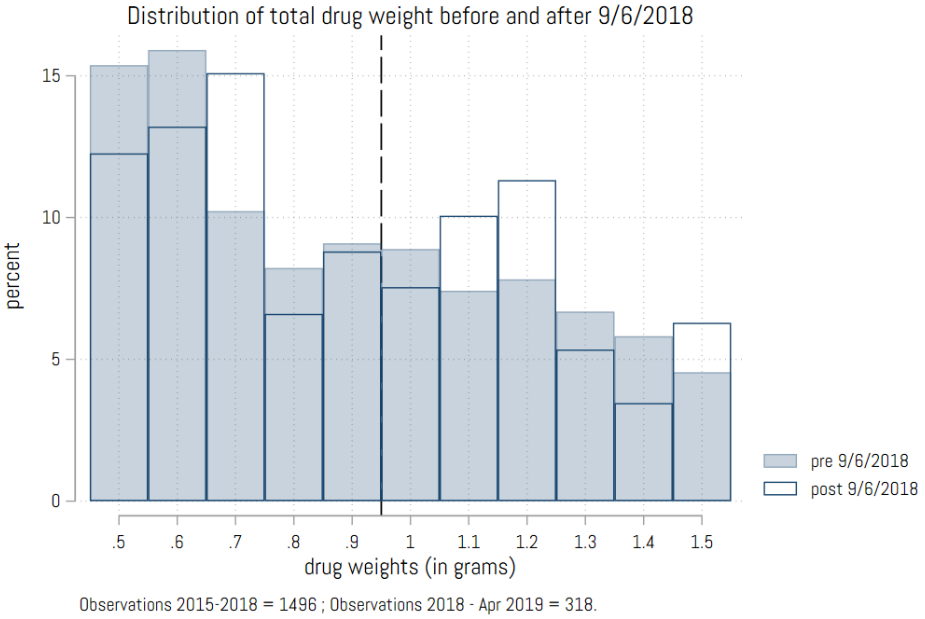
2.6 Conclusion

In this paper, I examine whether prosecutorial reform affects individuals before a case goes to court. I use data from Seattle, Washington to evaluate a filing standard change in September

2018 that effectively decriminalized minor drug possession: the King County prosecuting attorney decided not to file charges against anyone arrested for drug possession who was carrying less than 1 gram, fewer than 5 pills, or a single syringe or less. I focus on the 1 gram cutoff in a difference-in-differences strategy. Specifically, I test for differences in recidivism between individuals holding slightly less than 1 gram and individuals holding slightly more than 1 gram. Point estimates suggest a reduction in the likelihood of reoffending within ten months, but an increase in the number of subsequent offenses within ten months. If taken at face value, my results could indicate that this reform did not reduce criminal justice contact. However, because estimates are statistically insignificant, my analysis could also indicate that the reform did not matter for recidivism. Either conclusion contrasts with the only other research on this subject (**?**), pointing to the importance of additional work to better understand the effects of prosecutorial reform.

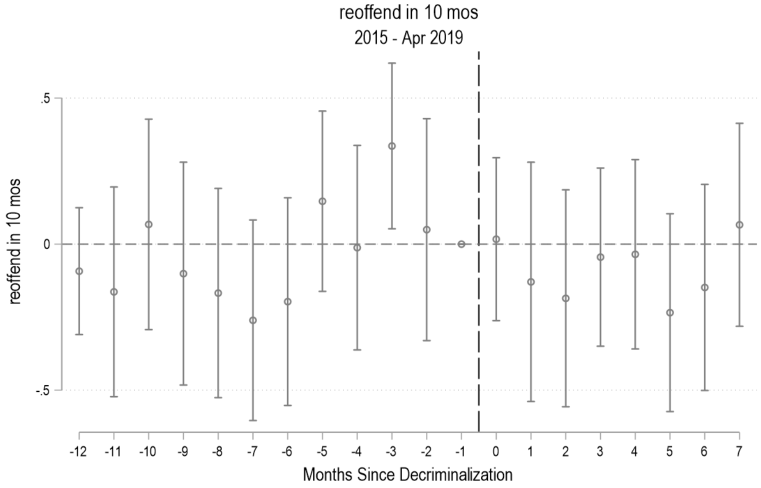
2.7 Figures

Figure 2.1: Distribution of Drug Weights

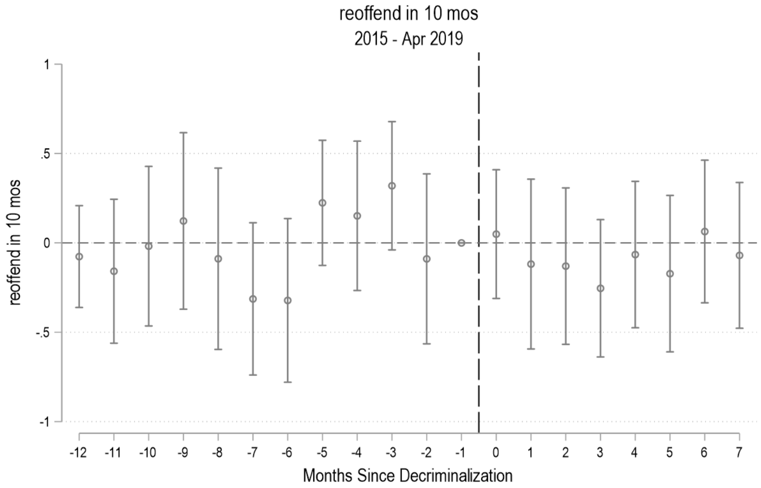


Notes: This figure displays the distribution of total drug weight before and after the filing standard change on September 6, 2018 using data between January 2015 to April 2019. The light blue shaded bars indicate the distribution of drug weights prior to the filing standard change; the white bars indicate the distribution of drug weights after the filing standard change. Total drug weights between 0.5 grams and 1.5 grams (my analysis sample) are shown. Each bar is centered over the drug weight value. The vertical dotted line indicates the 1 gram threshold.

Figure 2.2: Event Study Plot Examining the Likelihood of Reoffending Within 10 Months



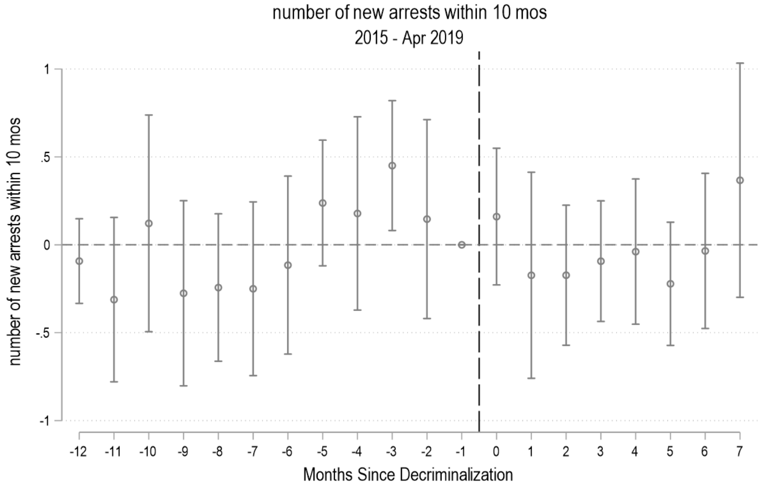
(a) Arrested for Any Offense



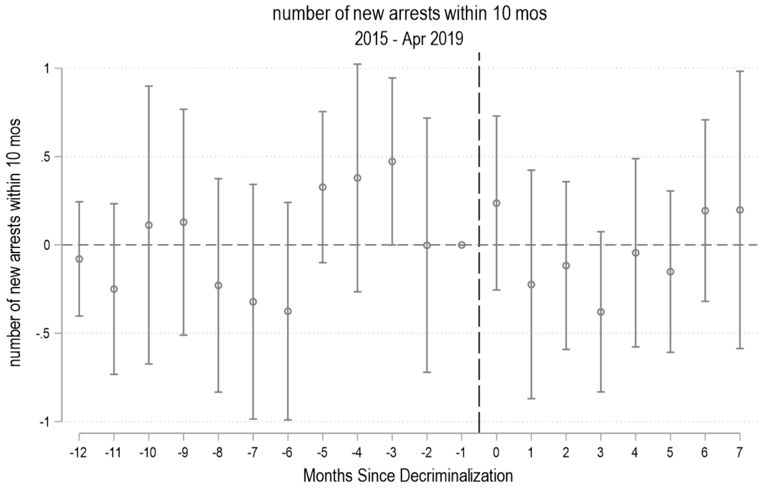
(b) Arrested for Drug Possession

Notes: Each subfigure shows the event study plot examining the likelihood of reoffending within 10 months. Subfigure (a) displays this for the full analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug who are arrested for any offense). Subfigure (b) displays this for the subsample of these individuals who are at least arrested for drug possession, as the filing standard change specifically applies to possession offenses. In each event study, I regress an indicator for reoffending within ten months on coefficients for twelve month leads and seven month lags, where time $t = 0$ represents September 2018. The $t - 12$ lead includes all data from twelve+ months prior to the reform (i.e., January 2015 through September 2017), while the $t + 7$ lag only includes data from seven months after the reform, given that the data ends in April 2019. All regressions include controls for age, race, and gender, as well as day-of-month, month, and year fixed effects. I then plot the coefficients and 95% confidence intervals from these regressions. Each coefficient represents the difference in recidivism between individuals holding just under 1 gram and individuals holding just over 1 gram, relative to what is expected based on pre-period trends.

Figure 2.3: Event Study Plot Examining the Number of Subsequent Arrests Within 10 Months



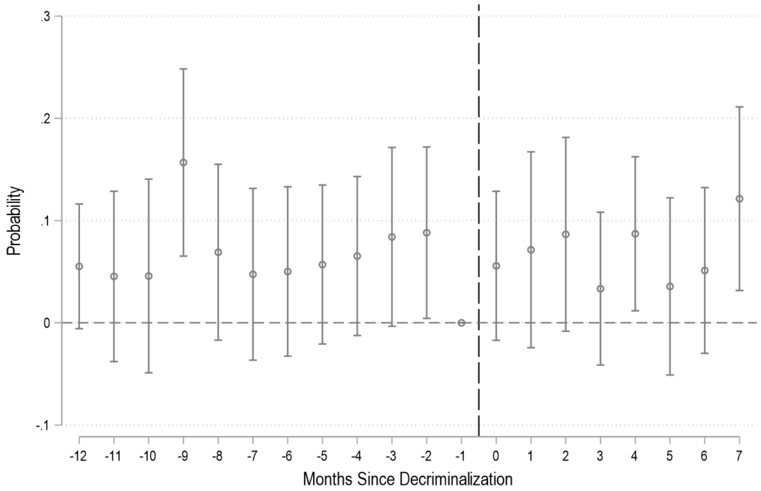
(a) Arrested for Any Offense



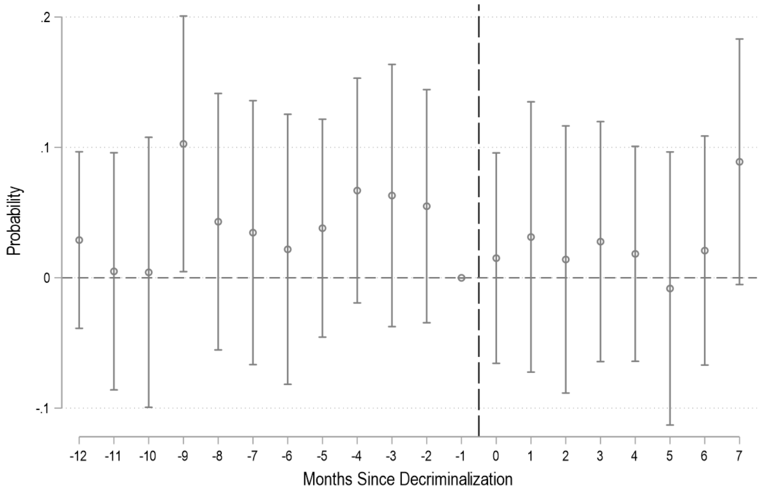
(b) Arrested for Drug Possession

Notes: Each subfigure shows the event study plot examining the number of subsequent arrests within 10 months. Subfigure (a) displays this for the full analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug who are arrested for any offense). Subfigure (b) displays this for the subsample of these individuals who are at least arrested for drug possession, as the filing standard change specifically applies to possession offenses. In each event study, I regress a variable for the number of new arrests within 10 months on coefficients for twelve month leads and seven month lags, where time $t = 0$ represents September 2018. The $t - 12$ lead includes all data from twelve+ months prior to the reform (i.e., January 2015 through September 2017), while the $t + 7$ lag only includes data from seven months after the reform, given that the data ends in April 2019. All regressions include controls for age, race, and gender, as well as day-of-month, month, and year fixed effects. I then plot the coefficients and 95% confidence intervals from these regressions. Each coefficient represents the difference in recidivism between individuals holding just under 1 gram and individuals holding just over 1 gram, relative to what is expected based on pre-period trends.

Figure 2.4: Event Study Plot Examining the Predicted Likelihood of Reoffending Within 10 Months



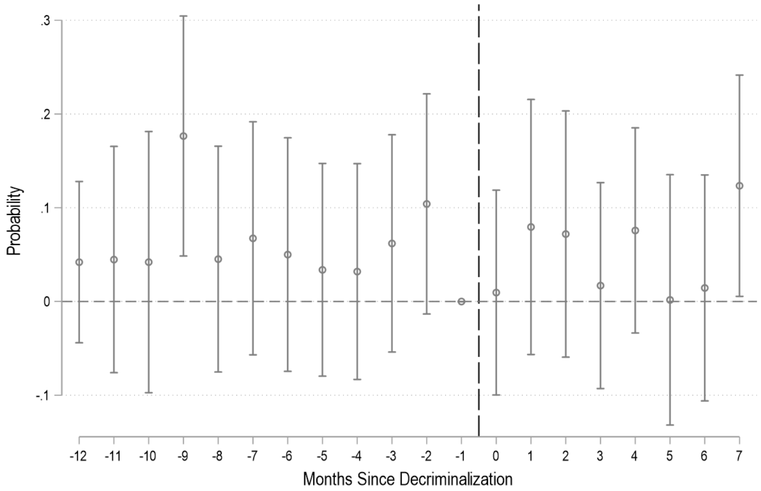
(a) Arrested for Any Offense



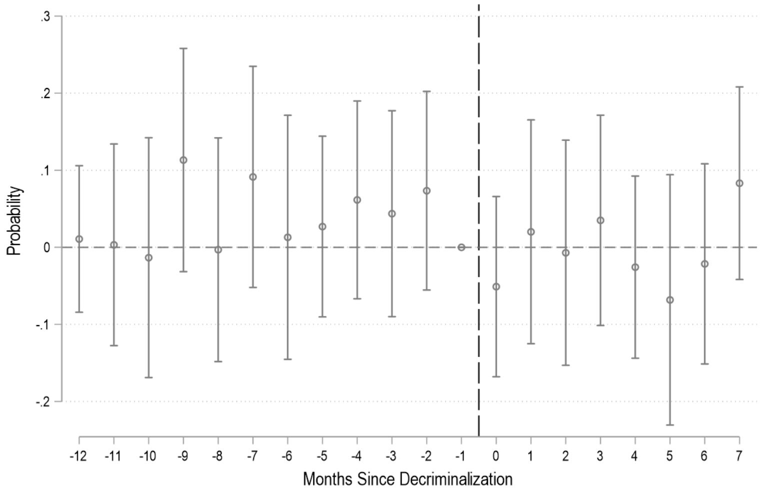
(b) Arrested for Drug Possession

Notes: Each subfigure shows the event study plot examining the predicted likelihood of reoffending within 10 months. Predictions come from regressions on exogenous characteristics (suspect age, race/ethnicity, and gender) and day-of-month, month, and year fixed effects. Robust standard errors are used. Subfigure (a) displays this for the full analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug who are arrested for any offense). Subfigure (b) displays this for the subsample of these individuals who are at least arrested for drug possession, as the filing standard change specifically applies to possession offenses. In each event study, I regress an indicator for reoffending within ten months on coefficients for twelve month leads and seven month lags, where time $t = 0$ represents September 2018. The $t - 12$ lead includes all data from twelve+ months prior to the reform (i.e., January 2015 through September 2017), while the $t + 7$ lag only includes data from seven months after the reform, given that the data ends in April 2019. I then plot the coefficients and 95% confidence intervals from these regressions. Each coefficient represents the difference in predicted recidivism between individuals holding just under 1 gram and individuals holding just over 1 gram, relative to what is expected based on pre-period trends.

Figure 2.5: Event Study Plot Examining the Predicted Number of Subsequent Arrests Within 10 Months



(a) Arrested for Any Offense



(b) Arrested for Drug Possession

Notes: Each subfigure shows the event study plot examining the predicted number of subsequent arrests within 10 months. Predictions come from regressions on exogenous characteristics (suspect age, race/ethnicity, and gender) and day-of-month, month, and year fixed effects. Robust standard errors are used. Subfigure (a) displays this for the full analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug who are arrested for any offense). Subfigure (b) displays this for the subsample of these individuals who are at least arrested for drug possession, as the filing standard change specifically applies to possession offenses. In each event study, I regress a variable for the number of new arrests within 10 months on coefficients for twelve month leads and seven month lags, where time $t = 0$ represents September 2018. The $t - 12$ lead includes all data from twelve+ months prior to the reform (i.e., January 2015 through September 2017), while the $t + 7$ lag only includes data from seven months after the reform, given that the data ends in April 2019. I then plot the coefficients and 95% confidence intervals from these regressions. Each coefficient represents the difference in predicted recidivism between individuals holding just under 1 gram and individuals holding just over 1 gram, relative to what is expected based on pre-period trends.

2.8 Tables

Table 2.1: Summary Statistics - Suspect Information

	(1) All	(2) Drugs Recorded	(3) Analysis Sample	(4) 0.5-0.9g	(5) 1-1.5g
Age	36.91	35.24	35.30	34.89	35.87
	(12.41)	(11.50)	(11.04)	(11.07)	(10.98)
Unk Age	0.02	0.00	0.00	0.00	0.00
	(0.14)	(0.04)	(0.03)	(0.04)	(0.00)
Male	0.75	0.81	0.81	0.80	0.82
	(0.43)	(0.39)	(0.39)	(0.40)	(0.39)
Female	0.25	0.19	0.19	0.20	0.18
	(0.43)	(0.39)	(0.39)	(0.40)	(0.39)
Unk Sex	0.00	0.00	0.00	0.00	0.00
	(0.05)	(0.04)	(0.02)	(0.03)	(0.00)
White	0.53	0.51	0.56	0.56	0.55
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Black	0.30	0.35	0.31	0.31	0.31
	(0.46)	(0.48)	(0.46)	(0.46)	(0.46)
Asian/Indian	0.08	0.09	0.09	0.08	0.10
	(0.27)	(0.28)	(0.28)	(0.27)	(0.30)
Unk Race	0.09	0.06	0.05	0.05	0.04
	(0.28)	(0.23)	(0.21)	(0.21)	(0.20)
Hispanic/Latino	0.04	0.05	0.06	0.06	0.07
	(0.20)	(0.23)	(0.24)	(0.24)	(0.25)
Unk Ethnicity	0.71	0.64	0.63	0.64	0.61
	(0.45)	(0.48)	(0.48)	(0.48)	(0.49)
Observations	114,394	8,983	1,814	1,058	756

Notes: This table displays average characteristics for suspects for five sets of observations: (1) all observations; (2) observations with drugs recorded; (3) the analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug); (4) treatment group (those in my analysis sample carrying just under 1 gram); and (5) control group (those in my analysis sample carrying 1 gram or just over 1 gram). Standard deviations are in parentheses. Data covers January 2015 through April 2019. Each variable (except for age) is measured as an indicator.

Table 2.2: Summary Statistics - Offenses

	(1) All	(2) Drugs Recorded	(3) Analysis Sample	(4) 0.5-0.9g	(5) 1-1.5g
Property	0.44	0.32	0.34	0.36	0.30
	(0.50)	(0.47)	(0.47)	(0.48)	(0.46)
Person	0.37	0.12	0.11	0.10	0.11
	(0.48)	(0.32)	(0.31)	(0.30)	(0.32)
Motor Vehicle	0.10	0.16	0.12	0.11	0.14
	(0.29)	(0.37)	(0.33)	(0.31)	(0.35)
White Collar	0.02	0.02	0.02	0.02	0.02
	(0.15)	(0.12)	(0.13)	(0.13)	(0.13)
Drugs/Alcohol	0.07	0.76	0.87	0.85	0.89
	(0.25)	(0.43)	(0.34)	(0.36)	(0.31)
Disturbance	0.06	0.05	0.05	0.05	0.05
	(0.24)	(0.23)	(0.22)	(0.22)	(0.22)
Weapons	0.02	0.07	0.06	0.06	0.06
	(0.15)	(0.26)	(0.24)	(0.24)	(0.23)
Mental Health	0.02	0.01	0.00	0.01	0.00
	(0.15)	(0.07)	(0.07)	(0.08)	(0.04)
Land/Wildlife	0.00	0.00	0.00	0.00	0.00
	(0.04)	(0.02)	(0.00)	(0.00)	(0.00)
Animals	0.00	0.00	0.00	0.00	0.00
	(0.04)	(0.02)	(0.04)	(0.03)	(0.05)
Other	0.04	0.10	0.12	0.12	0.12
	(0.20)	(0.30)	(0.32)	(0.32)	(0.32)
No. Offenses	1.61	2.42	2.52	2.45	2.63
	(0.88)	(1.42)	(1.38)	(1.32)	(1.46)
No. Suspects	1.28	1.57	1.44	1.38	1.52
	(1.09)	(1.41)	(1.02)	(0.70)	(1.35)
Reoffend within 10 mos	0.48	0.25	0.09	0.06	0.04
	(0.50)	(0.43)	(0.28)	(0.23)	(0.19)
No. New Arrests within 10 mos	1.64	0.35	0.10	0.07	0.04
	(3.11)	(0.72)	(0.34)	(0.29)	(0.20)
Observations	114,394	8,983	1,814	1,058	756

Notes: This table displays average offense characteristics for five sets of observations: (1) all observations; (2) observations with drugs recorded; (3) the analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug); (4) treatment group (those in my analysis sample carrying just under 1 gram); and (5) control group (those in my analysis sample carrying 1 gram or just over 1 gram). Standard deviations are in parentheses. Data covers January 2015 through April 2019. Each variable (except for age) is measured as an indicator. The offense type variables (Property, Person, ..., Other) and the Reoffend within 10 mos variable are measured as indicators. The remaining variables (No. Offenses, No. Suspects, and No. New Arrests within 10 mos) are counts.

Table 2.3: Summary Statistics - Drug Crimes

	(1) All	(2) Drugs Recorded	(3) Analysis Sample	(4) 0.5-0.9g	(5) 1-1.5g
VUCSA	0.07	0.77	0.87	0.85	0.89
	(0.25)	(0.42)	(0.34)	(0.36)	(0.31)
Possession	0.05	0.57	0.71	0.71	0.73
	(0.22)	(0.50)	(0.45)	(0.46)	(0.45)
Sale	0.02	0.19	0.15	0.13	0.17
	(0.12)	(0.39)	(0.35)	(0.33)	(0.38)
Manufacturing	0.00	0.01	0.01	0.00	0.01
	(0.02)	(0.08)	(0.08)	(0.03)	(0.12)
Smuggling	0.00	0.00	0.00	0.00	0.00
	(0.01)	(0.03)	(0.03)	(0.04)	(0.00)
Loitering	0.00	0.01	0.00	0.00	0.01
	(0.04)	(0.08)	(0.07)	(0.06)	(0.07)
Found Drugs	0.00	0.03	0.03	0.03	0.02
	(0.05)	(0.17)	(0.17)	(0.18)	(0.14)
Forgery	0.00	0.00	0.00	0.00	0.00
	(0.02)	(0.03)	(0.00)	(0.00)	(0.00)
Fraud	0.00	0.00	0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)
Precursor Drugs	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Observations	114,394	8,983	1,814	1,058	756

Notes: This table displays average drug crime characteristics for five sets of observations: (1) all observations; (2) observations with drugs recorded; (3) the analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug); (4) treatment group (those in my analysis sample carrying just under 1 gram); and (5) control group (those in my analysis sample carrying 1 gram or just over 1 gram). Standard deviations are in parentheses. Data covers January 2015 through April 2019. Each variable is measured as an indicator.

Table 2.4: Summary Statistics - Drug Types

	(1) All	(2) Drugs Recorded	(3) Analysis Sample	(4) 0.5-0.9g	(5) 1-1.5g
Offense with Drugs	0.08	1.00	1.00	1.00	1.00
	(0.27)	(0.00)	(0.00)	(0.00)	(0.00)
In Drug Sample	0.06	0.79	1.00	1.00	1.00
	(0.24)	(0.41)	(0.00)	(0.00)	(0.00)
Measured in Grams	0.07	0.89	1.00	1.00	1.00
	(0.25)	(0.31)	(0.00)	(0.00)	(0.00)
Marijuana	0.01	0.14	0.02	0.01	0.03
	(0.11)	(0.35)	(0.13)	(0.10)	(0.16)
Pills	0.02	0.20	0.09	0.09	0.10
	(0.12)	(0.40)	(0.29)	(0.28)	(0.30)
Liquids	0.00	0.04	0.02	0.01	0.02
	(0.06)	(0.20)	(0.13)	(0.11)	(0.14)
Prescription Drugs	0.00	0.01	0.00	0.00	0.01
	(0.03)	(0.12)	(0.06)	(0.04)	(0.07)
Cocaine	0.02	0.24	0.22	0.22	0.22
	(0.14)	(0.43)	(0.42)	(0.41)	(0.42)
Heroin	0.03	0.34	0.38	0.36	0.39
	(0.16)	(0.47)	(0.48)	(0.48)	(0.49)
Amphetamines/Meth	0.03	0.44	0.58	0.57	0.60
	(0.18)	(0.50)	(0.49)	(0.50)	(0.49)
Hallucinogen	0.00	0.01	0.00	0.00	0.00
	(0.03)	(0.10)	(0.03)	(0.03)	(0.04)
Other Narcotic	0.00	0.06	0.03	0.03	0.04
	(0.07)	(0.23)	(0.18)	(0.16)	(0.20)
Barbiturate	0.00	0.03	0.01	0.01	0.01
	(0.05)	(0.16)	(0.09)	(0.09)	(0.10)
Opium/Morphine	0.00	0.01	0.00	0.01	0.00
	(0.02)	(0.08)	(0.06)	(0.08)	(0.04)
Other Drugs	0.01	0.15	0.08	0.08	0.08
	(0.11)	(0.36)	(0.27)	(0.26)	(0.27)
Multiple Drug Items	0.00	0.01	0.00	0.00	0.00
	(0.03)	(0.09)	(0.03)	(0.00)	(0.05)
Observations	114,394	8,983	1,814	1,058	756

Notes: This table displays average drug type characteristics for five sets of observations: (1) all observations; (2) observations with drugs recorded; (3) the analysis sample (those carrying between 0.5 and 1.5 grams of any non-pill, non-liquid, non-marijuana drug); (4) treatment group (those in my analysis sample carrying just under 1 gram); and (5) control group (those in my analysis sample carrying 1 gram or just over 1 gram). Standard deviations are in parentheses. Data covers January 2015 through April 2019. Each variable is measured as an indicator.

Table 2.5: Composition of Treatment and Control Groups Over Time

	Treatment (0.5-0.9g)		Control (1-1.5g)	
	(1)	(2)	(3)	(4)
	Pre	Post	Pre	Post
Age	34.55	36.58	35.28	38.51
	(11.07)	(10.96)	(10.77)	(11.54)
Unk Age	0.00	0.00	0.00	0.00
	(0.05)	(0.00)	(0.00)	(0.00)
Male	0.80	0.79	0.83	0.76
	(0.40)	(0.41)	(0.37)	(0.43)
Female	0.20	0.21	0.17	0.24
	(0.40)	(0.41)	(0.37)	(0.43)
Unk Sex	0.00	0.00	0.00	0.00
	(0.03)	(0.00)	(0.00)	(0.00)
White	0.57	0.54	0.55	0.55
	(0.50)	(0.50)	(0.50)	(0.50)
Black	0.31	0.31	0.30	0.34
	(0.46)	(0.47)	(0.46)	(0.47)
Asian/Indian	0.07	0.12	0.11	0.08
	(0.26)	(0.32)	(0.31)	(0.27)
Unk Race	0.05	0.02	0.04	0.04
	(0.22)	(0.15)	(0.20)	(0.19)
Hispanic/Latino	0.06	0.04	0.07	0.05
	(0.25)	(0.19)	(0.26)	(0.22)
Unk Ethnicity	0.63	0.71	0.60	0.65
	(0.48)	(0.45)	(0.49)	(0.48)
Observations	880	178	616	140

Notes: This table displays average suspect characteristics for the treatment and control groups before and after the filing standard change in September 2018. Standard deviations are in parentheses. Data covers January 2015 through April 2019. Each variable (except age) is measured as an indicator.

Table 2.6: Predicted Recidivism Within 10 Months

	All Observations		Arrested for Drug Possession	
	(1) Predicted Reoffend (Y/N)	(2) Predicted # New Arrests	(3) Predicted Reoffend (Y/N)	(4) Predicted # New Arrests
Below 1g*Post	0.00956	0.00241	-0.00863	-0.0269
	(0.0112)	(0.0167)	(0.0130)	(0.0194)
Observations	1,812	1,812	1,294	1,294
Pre-Period Treatment Mean	0.25	0.35	0.25	0.34
Treatment Effect (%)	3.79	0.70	-3.44	-7.91

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Data covers my analysis sample between January 2015-April 2019. This sample includes all individuals carrying non-pill, non-liquid, non-marijuana drugs weighing between 0.5 and 1.5 grams. In columns 1 and 3 predicted recidivism is measured as an indicator equal to 1 if an individual ever reoffended within 10 months. In columns 2 and 4, predicted recidivism is measured as the number of subsequent arrests within 10 months. Columns 1 and 2 include all observations in the analysis sample. Columns 3 and 4 only include observations in the analysis sample that have an arrest for drug possession. Predictions come from regressions on exogenous characteristics (suspect age, race/ethnicity, and gender) and day-of-month, month, and year fixed effects. Robust standard errors are used.

Table 2.7: Recidivism Within 10 Months

	All Observations		Arrested for Drug Possession	
	(1) Reoffend (Y/N)	(2) # New Arrests	(3) Reoffend (Y/N)	(4) # New Arrests
Below 1g*Post	-0.00283	0.0480	-0.0249	0.0230
	(0.0526)	(0.0771)	(0.0629)	(0.0934)
Observations	1,812	1,812	1,294	1,294
Pre-Period Treatment Mean	0.25	0.35	0.27	0.37
Treatment Effect (%)	-1.11	13.72	-9.36	6.16

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Data covers my analysis sample between January 2015-April 2019. This sample includes all individuals carrying non-pill, non-liquid, non-marijuana drugs weighing between 0.5 and 1.5 grams. In columns 1 and 3 recidivism is measured as an indicator equal to 1 if an individual ever reoffended within 10 months. In columns 2 and 4, recidivism is measured as the number of subsequent arrests within 10 months. Columns 1 and 2 include all observations in the analysis sample. Columns 3 and 4 only include observations in the analysis sample that have an arrest for drug possession. Each specification includes individual-level controls (age, race, and gender) as well as day-of-month, month, and year fixed effects. Standard errors (in parentheses) are clustered at the individual level.

3. CRISIS AVERTED? THE EFFECT OF CRISIS INTERVENTION TEAMS ON ARRESTS AND USE OF FORCE

3.1 Introduction

There is broad policy interest in deescalating police interactions with civilians. One such group of civilians is those with mental illnesses: in the United States, approximately 23% of all fatal police shootings since 2015 have involved individuals suffering from a mental illness.[1] One possible solution is crisis intervention team (CIT) programs. There are two facets to these programs: crisis intervention training (CIT training) and crisis intervention team units (CIT units). With CIT training, all patrol officers receive more comprehensive mental health training. CIT units - the focus of this paper - pair a specially trained police officer with a mental health professional to specifically respond to calls involving mentally ill individuals - especially mental health crisis calls. While such interventions have become increasingly popular, they are difficult to evaluate because the assignment of CIT officers or units to calls is highly endogenous based on the characteristics of the calls. That is, CIT units are typically only dispatched to calls involving a mental health crisis. Simply comparing calls to which a CIT unit responded with other calls would not tell us whether CIT units were helpful because these sets of calls are so different.

In this paper we provide the first causal evidence on the effects of CIT units. The El Paso Police Department (EPPD) in El Paso, Texas established CIT units in December 2018. However, these teams are only available during certain hours of the day. This means that otherwise-similar calls will have very different probabilities of getting a CIT unit response based on the precise time of the call. Using EPPD data between December 2018 and February 2020, we use this quasi-random variation in the availability of CIT units to measure the causal effects of CIT units on call outcomes. Specifically, we exploit changes in the likelihood that CIT units respond to calls due to CIT shift schedules in a fuzzy regression discontinuity design.

[1] Washington Post (2022)

We focus on a subset of calls likely to involve a mental health crisis.[2] First, we show that such "MH crisis" calls that occur soon after CIT units come on duty at 8am are 3.4 percentage points more likely to ever receive a CIT unit response (a 368% increase relative to pre-8am calls). We then consider the effects of this CIT unit response on two outcomes of interest: the likelihood that the call resulted in an arrest and the likelihood that police used force during their response. Our 2SLS estimates indicate a 45 percentage point reduction in the probability of arrest (a 124% reduction relative to the complier mean of 0.37). With regard to use of force, we find a 4.3 percentage point increase in the probability force is used (a 158% increase relative to the complier mean of 0.03). While both of these estimates suggest large changes in police behavior, they are imprecisely estimated not statistically different from zero.

We demonstrate that our results are robust to varying bandwidths in our fuzzy RD design. As we would expect, standard errors increase as we decrease the bandwidth, but our estimated coefficients remain the same or increase in magnitude as we focus on times closer to the 8am cutoff. While our reduced form estimates remain imprecise, our first stage estimates are statistically significant across the various bandwidths, supporting the validity of our research design. We also show how adjusting our definition of MH crisis calls affects our results.

Our results are a bit puzzling, suggesting conflicting effects. CIT units appear to be increasing police use of force in incidents involving civilians in mental health crisis, but simultaneously appear to reduce the likelihood of arrest. The decreased likelihood of arrest could be evidence of CIT units diverting individuals in crisis from incarceration and instead connecting them to community services for treatment, which could reduce criminal justice contact. To understand the use of force results, we examine the type of force used in our MH crisis sample. All incidents of force include low-level "type 1" force. Higher-level "type 2" force is only used alongside type 1 force, which suggests that higher-level force is only used if low-level force was not sufficient. No deadly "type

[2] To determine the subsample of calls that are ex-ante most likely to receive a CIT response, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. Second, we exclude suicides since we do not expect these calls to result in use of force or end in an arrest.

3" force is used. Furthermore, the reasons for force are combative and/or non-compliant subjects. Overall, this could indicate that CIT units are successfully deescalating the situation.

With CIT becoming increasingly popular, criminologists and psychologists have studied its effectiveness. Numerous studies have examined CIT training on officer knowledge and attitudes through surveys and interviews (Peterson and Densley, 2018). For example, Ellis (2014) surveys 25 officers before and after receiving CIT training and finds that CIT increased officers' knowledge and perception of mental illness, and improved officers' attitudes towards individual with mental illness. Bonfine, Ritter and Munetz (2014) finds similar improvements in a survey of 57 officers who participated in the training. Compton et al. (2014a) compare a total of 586 officers with and without CIT training to study whether the training is associated with officers' abilities to respond to individuals with mental illness. Using an assessment, they examine attitudes towards mental health and treatment, de-escalation sills, referral decisions, and self-perceived readiness. They find that CIT-trained officers perform better on all measures of responsiveness – especially de-escalation and referral decisions.

Researchers have also studied the effects of CIT training on call outcomes. Compton et al. (2014b) examine 1,063 incidents involving 180 officers (91 with CIT training and 89 without CIT training), and find that the officers with CIT training were more likely to engage in verbal negotiation as the highest level of force, more likely to refer or transport the individual to mental health services, and less likely to arrest than officers without the training. It is unclear whether this difference in behavior is due to the CIT training or if different preferences over how to handle crisis calls led officers pre-disposed to de-escalation techniques to volunteer for CIT training in the first place. Other studies find similar results when comparing CIT officers to non-CIT officers and when comparing the same officers before and after receiving the training (Peterson and Densley, 2018). While studies that compare officers before and after they receive CIT training do control for pre-existing differences across officers, they do not account for the fact that officers with CIT training are often sent to different types of calls than officers without the training.

Overall, the current research suggests CIT training is associated with beneficial outcomes, but it is unclear if these studies are isolating the causal effects of CIT training. Moreover, while numerous researchers have examined CIT training, we are unaware of any studies that analyze CIT units – a more intensive intervention that sends a trained counselor alongside a police officer. There is good reason to expect that such CIT units could be more helpful than the standard, CIT-training for trained law enforcement. The ideal experiment to test the efficacy of these interventions would be to randomize whether a call involving a MH crisis is handled by a CIT-trained officer or a CIT unit. So far we do not know of any research that approximates this ideal experiment to measure the causal effect on call outcomes. Our paper fills this gap, and provides the first causal evidence of CIT units' ability to deescalate incidents between police and civilians.

Our paper contributes to several literatures. First and foremost, we consider the causal effects of CIT units on the escalation of MH crisis calls to arrest or use of force. There is little to no evidence on this relatively new, but increasingly popular, intervention. Second, we contribute to a broader literature on how to change police behavior.[3] There is remarkably little causally-identified evidence on the efficacy of police trainings and other interventions that aim to change what police do on the job. CIT units are one of many such interventions. Finally, we contribute to a growing literature on how individuals with mental illness interact with the criminal justice system.[4] Most of this evidence is descriptive (e.g. Frank and McGuire (2010)), though there is some recent evidence on how increasing access to mental health care affects criminal behavior (e.g. Jácome (2020)). We add evidence on the other side of this interaction: how CIT units might change the way police respond to civilians with mental illness.

Our paper proceeds as follows. Section 3.2 provides background on the CIT program and on the rollout of CIT units in El Paso. Sections 3.3 and 3.4 describe the data and empirical strategy, respectively. We present results in section 3.5 and discuss our findings in section 3.6. Section 3.7 concludes.

[3] Banerjee et al. (2021); Cheng and Long (2018); Ater, Givati and Rigbi (2014); Shi (2009); Doleac (2017); Anker, Doleac and Landerso (2021); Owens (forthcoming)

[4] Frank and McGuire (2010); Bondurant, Lindo and Swensen (2018); Deza et al. (2020); Hjalmarsson and Lindquist (2013); Aslim et al. (2022); Jácome (2020); Bencsik (2021)

3.2 Institutional Background

Following multiple lawsuits for using deadly force against individuals with mental illness, the El Paso Police Department (EPPD) established a Crisis Intervention Team (CIT) program on October 2, 2018.[5] After spending eight weeks in training, CIT personnel began responding to calls on December 2, 2018.[6] CIT serves two main purposes: 1) to better serve individuals suffering from a mental health crisis and 2) to improve law enforcement responses to *any* call involving individuals with mental illness and/or intellectual disability.[7]

CIT consists of four supervisors (1 lieutenant and 3 sergeants) and fourteen officers. Within CIT there are seven units (CIT units). Each CIT unit includes one officer and one licensed professional clinician from the Emergence Health Network (EHN).[8] CIT officers are specially trained to recognize mental illness, identify whether someone is in crisis, effectively communicate with mentally ill individuals, and deescalate the situation. CIT clinicians assist officers in identifying mental illness - not only upon arrival to a call, but also en route to a call by accessing electronic medical records.[9] CIT clinicians also perform on-site evaluations, provide solutions for these individuals, and act as a liaison between officers and community resources.[10]

CIT units operate seven days a week in two shifts: 8am to 6pm and 2pm to 12am, with two to four units operating during a given shift. No CIT units are scheduled between 12am and 8am. The availability of CIT units is critical to our research design. In the absence of CIT units between 12am and 8am, calls that warrant a CIT response do not receive a CIT unit solely because CIT units are off-duty. When on duty, there are three avenues through which CIT units can respond to calls. First, call takers can notify dispatch to send a CIT unit. Call takers are the first to communicate with the individual calling 911 (either the subject of the call or a second party caller). After asking a set of questions, call takers notify dispatch if the call warrants a CIT unit. If a CIT unit is available,

[5] According to El Paso Matters (2020), four lawsuits were filed between 2014 and 2018. Three incidents resulted in fatalities. The individual involved in the fourth incident survived multiple gunshot wounds.

[6] EPPD (2018)

[7] City of El Paso (2018)

[8] EPPD (2018); City of El Paso (2018)

[9] El Paso Times (2020)

[10] El Paso Times (2020); City of El Paso (2018)

dispatch will send them as the primary responding unit. If not, any available patrol unit on the shift is dispatched to the call. Second, patrol officers may request CIT units to assist them in handling individuals with mental illness and/or intellectual disability. Third, CIT units may follow-up with non-active calls.[11]

The goal of CIT units is to provide “safer and more effective responses" to the mentally ill and/or intellectually disabled and to “increase the number of persons diverted from incarceration when allowed by statutes."[12] A successful response could look like approaching individuals cautiously with a calm and friendly demeanor and practicing deescalation: “Officers should encourage communication, demonstrate empathy, and be aware of their body language so as not to be perceived as threatening or hostile."[13] Diversion includes connecting the individual to community-based support services or transporting the individual to a hospital via an emergency detention order.[14]

It is important to note that during the period covered by our data, all EPPD patrol officers have received training in crisis intervention and deescalation.[15] As a result, when we estimate the effects of a CIT unit response the counterfactual is a patrol officer that has received crisis intervention training, not a patrol officer without any sort of deescalation training. We are thus estimating the value-added of extra training (CIT officers receive eight weeks of training), specialization (CIT officers only work as a part of the CIT unit), and additional expertise (the partnership with a mental health professional).

[11] City of El Paso (2018)

[12] City of El Paso (2018)

[13] El Paso Police Department (2022)

[14] According to the Texas Health and Safety Code, emergency detention orders allow officers to take a person into custody if that officer believes the person is mentally ill and a danger to his/herself or others. Custody may look like an inpatient or an outpatient mental health facility. An emergency detention order is not involuntary commitment. Upon admission to a mental health facility, a mental health professional will decide whether treatment and/or commitment is necessary.

[15] Per the Sandra Bland Act (Texas S.B. 1849), as of April 1, 2018 all officers are required to take a 40-hour crisis intervention training course (either during police academy or as part of the continuing education curriculum). This is a more intensive course than the previously required 16-hour class.

3.3 Data

Our data comes from the El Paso Police Department and includes all 911 calls for service from December 2018 through February 2020. The full sample of calls totals 137,384. In addition to call characteristics such as the type, time and date, and priority, we are able to observe all units that respond to a call. While we can observe whether a CIT unit responded to a call, the data do not tell us *how* the CIT unit responded to the call (i.e., as the primary responding unit, as an assistant to the primary responding unit, or as a follow-up to a non-active call). Each call has a unique event number, and an incident number is available if a report is written by the on-scene officers, such as when an arrest is made or there is use of force. We link the arrest records and the use of force incidents to the 911 data using these unique incident numbers.

For our analysis, we cut the full sample in two important ways. First, because we are interested in the effect of receiving a CIT response, we are only interested in studying calls that would warrant one of these units - mental health calls ("MH calls"). Ideally, we would have a variable that flags these calls. Because we do not have such a variable, we determine the subsample of calls that are ex-ante most likely to receive a CIT response by predicting the likelihood of receiving a CIT response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We apply this prediction to calls during all hours (including when CIT units are off-duty), and then define MH calls as the subsample of calls with a predicted CIT response in the top 25%. Second, we exclude suicides (9% of MH calls) since we do not expect these calls to result in use of force or end in an arrest.[16] This reduces our sample to 31,289 MH calls (approximately 23% of the full sample of calls). Figure 3.1 shows the likelihood that each call type receives a CIT unit response. Put differently, the figure answers the following question: given that a call is of type *t*, what is the probability (on average) it will receive a CIT unit? Figures 3.1a and 3.1b show the actual likelihoods, while figures 3.1c and 3.1d show the predicted likelihoods from the above exercise. Comparing both sets of figures indicates the

[16] Our results are robust to including suicides.

accuracy of our predictions: we are able to correctly estimate which calls are ex-ante most likely to receive a CIT unit based on exogenous call characteristics.

Given that our research design exploits discontinuities in outcomes due to a CIT shift change at 8am, we focus in on the subset of calls that occur right around the threshold. Specifically, we limit our sample to calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020). Our final analysis sample consists of 4,571 calls.

Summary statistics our analysis sample are shown in table 3.1. This table includes statistics for exogenous covariates (indicators for weekend and call priority), treatment (CIT ever responds), and outcomes (arrest made and force used). On average, 30 percent of calls occur on the weekend (defined as 12:00am Saturday through 11:59pm Sunday). 29% of calls are priority 0/1 (most serious) and 57% of calls are priority 3 (less serious).[17] With regard to calls that occur during a non-CIT unit shift, 0.8% receive a CIT unit response, 4.6% end in an arrest, and 0.6% involve force. With regard to calls that occur during a CIT unit shift, 6.4% receive a CIT unit response, 2.8% end in an arrest, and 0.6% involve force.

Figure 3.2 shows the types of calls that receive a CIT unit response. Put differently, the figure displays the share of CIT unit responses that go to call type *t*. Figure 3.2d shows the call types represented in our final analysis sample. Almost 60% of CIT unit responses are to welfare calls. Domestic, assault, and assistance calls each account for between 10 and 20% of CIT unit responses. For comparison purposes, we also include this information for all calls in the full dataset (figures 3.2a and 3.2b) and all calls that occur during between 5:30am and 10:30am (figure 3.2c).

[17] Whenever a call comes in, the call taker is responsible for coding the priority and call description after assessing the situation. They might also add comments such as “weapon involved”. Call priority takes on the following values: 0-5, 7, and 9, with 0 referring to the most serious calls. Calls that are assigned the lowest priority (most serious priority) are the ones that are in progress, such as an assault in progress or robbery in progress, etc. It is important to note that in our analysis sample only 12 calls are given a priority of 0 and no calls are given priorities 5, 7, or 9. Thus, we create indicators for priority 0/1, priority 2, priority 3, and priority 4 calls.

3.4 Identification Strategy

The goal of this paper is to measure the effect of CIT unit responses on call outcomes. Simply comparing calls that receive CIT units to calls that do not is problematic because CIT unit responses are non-random. We thus take advantage of CIT unit shift changes to identify the causal effects of a CIT response. Because CIT only operates between the hours of 8am and 12am, we can exploit changes in CIT availability due to the timing of the call in a fuzzy regression discontinuity design. The identification assumption of this design is that the timing of the all acts as an as-good-as-random shock to getting a CIT unit. While this allows us to examine two cutoffs (CIT coming on duty at 8am and CIT going off duty at 12am), the raw data indicates a significant lag in CIT going off duty; there is no indication of a sharp change in the probability of receiving a CIT response right after midnight (see appendix figure B.1). Thus, we focus on the shift change at 8am at which CIT comes on duty.

Our reduced form equation is:

$$\begin{aligned} Y = \alpha + \beta \widetilde{CallTime} + \theta CITShift \\ + \gamma(\widetilde{CallTime} * CITShift) + \omega X + \lambda_{year} + \tau_{month}, \end{aligned} \tag{3.1}$$

where Y is a vector of outcome variables (arrest, use of force); $\widetilde{CallTime}$, a continuous measure of the time of the call, is our normalized running variable (0=8am, 1=9am, etc.);[18]; $CITShift$ is an indicator for calls that occur when CIT is on duty (after 8am); X is a vector of control variables; and λ_{year} and τ_{month} are year and month fixed effects. Standard errors are clustered by call time.

Given that the probability of receiving a CIT unit response changes at 8am, we formally estimate a two-stage least squares (2SLS) model to uncover the causal effect of a CIT unit response on outcomes:

[18] In our data, we have the hour, minutes, and seconds of the call. We take advantage of this in constructing our running variable. For example, calls that occur at 08:00:00am have a value of 0, while calls that occur at 08:00:01am have a value of 0.00028.

First Stage

$$\begin{aligned} CITUnitResponse =& \alpha + \beta \widetilde{CallTime} + \theta CITShift \\ & + \gamma(\widetilde{CallTime} * CITShift) + \omega X + \lambda_{year} + \tau_{month} \end{aligned} \tag{3.2}$$

Second Stage

$$\begin{aligned} Y =& \alpha + \beta \widetilde{CallTime} + \theta \widehat{CITUnitResponse} \\ & + \gamma(\widetilde{CallTime} * \widehat{CITUnitResponse}) + \omega X + \lambda_{year} + \tau_{month} \end{aligned} \tag{3.3}$$

From equation 3.3 our variable of interest is θ, which represents the discontinuity in outcomes for calls that receive a CIT response. The validity of this design relies upon the assumption that calls coming in near a CIT shift change are similar. For example, we would not expect to see large differences between the characteristics of calls coming in at 7:55am versus 8:05am. Put differently, any discontinuity in outcomes should only be coming from differences in the probability of a CIT response. This assumption is plausible for multiple reasons. First, the likelihood of callers precisely manipulating the call time is low. It is unlikely that civilians know the shift times of CIT, as only a few internet articles state the exact shift times, and most individuals most likely have not read them. Even if callers know when CIT operates, we would not expect callers to choose to time their call accordingly; individuals call 911 when they need police assistance, so it is unlikely they would withhold a call to try to get a specific patrol unit.

In addition to anecdotal evidence against manipulation, we can test for this empirically. First, we examine the raw distribution of calls around the threshold. If callers were waiting to call 911 until CIT came on duty, we would expect to see a mass of calls on the right side of the threshold (i.e., bunching in the distribution). In contrast, a smooth distribution would indicate that callers are not altering their behavior in order to guarantee a CIT unit response. This is exactly what we see in figure 3.3a. Furthermore, the McCrary (2008) density test does not estimate a statistically significant discontinuity at the threshold, which suggests callers are not timing their calls to receive a CIT unit response. We also examine observable call characteristics across the threshold: the priority of the call, whether the call occurred on a weekend, and the call type. If our identification assumption is valid, we should see a smooth distribution of these baseline characteristics across the threshold. Figures 3.3b - 3.3d show the smoothness of these exogenous covariates as CIT comes on duty.[19]

Furthermore, in figure 3.4 we use exogenous covariates to predict arrest and use of force, and plot these average predicted outcomes across call time. Our predicted estimates come from regressions on indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. The purpose of this exercise is to show the smoothness of predicted outcomes across the threshold. If we were to see a discontinuity, we would worry that something else besides a CIT shift change is happening at 8am and affecting call outcomes.

In addition to graphing raw data, we regress our reduced form equation (equation 3.1 above, but without controls or fixed effects) on exogenous call characteristics and predicted outcomes. In these regressions we cluster standard errors by call time, as we would expect unobserved factors to be similar across calls that happen at the same moment in time. Results from these regressions are shown in table 3.2. We do not estimate statistically significant discontinuities for any call characteristics.

3.5 Results

3.5.1 First Stage

Figure 3.5 shows the discontinuity in the probability of a CIT response across call time. Specifically, we plot the raw average likelihood of receiving a CIT response during thirty minute intervals between 5:30am and 10:30am. Figure 3.5 indicates that the likelihood calls receive a CIT response just after 8am rises between 3 and 4 percentage points. After 8:30am, the likelihood that calls receive a CIT response is approximately 6 percentage points higher, indicating some transition time in the first thirty minutes of the CIT shift. In panel A of tables 3.3 and 3.4, we formally estimate

[19] All four of these figures use raw data to plot the average likelihood across call time using 30 minute bins.

our first stage equation (equation 3.2 above) to examine the discontinuity in the probability of a CIT response across the threshold, and find a statistically significant increase of 3.4 percentage points off the control group (pre-8am) mean of 0.01. This estimate is robust to the inclusion of controls for call characteristics and year and month fixed effects.

3.5.2 Reduced Form/Second Stage

Figure 3.6 graphs the reduced form discontinuities in arrests and use of force across CIT shift changes. Each figure plots raw data, and markers represent the average likelihood of a given outcome during a thirty-minute interval between 5:30am and 10:30am. In figure 3.6a the raw data suggests a slight decrease in the likelihood of arrest after CIT comes on duty. Figure 3.6b suggests no discontinuity in the probability force is used; throughout the sample period, the likelihood of use of force jumps between 0 and 0.01. We discuss regression estimates below.

3.5.2.1 Arrests

In table 3.3 we estimate the effect of a CIT response on arrests. Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage), as discussed above in section 3.5.1. Panel B displays the estimated discontinuity in arrests from a CIT shift change (reduced form) and panel C displays the estimated discontinuity in arrests from a CIT response using call time as an instrument for the likelihood of a CIT response (local average treatment effect or “LATE"). Column (1) displays estimates from our baseline specification, while columns (2) and (3) add controls for call characteristics and year and month fixed effects, respectively. We focus on estimates from our preferred specification using controls and year and month fixed effects (column (3)), but note that coefficients are similar across all specifications. Regression results indicate a 45 percentage point decrease in the probability of arrest when a CIT unit responds to the call. To better understand our LATE, we calculate average arrest rates for compliers to estimate their average treatment effect.[20] We estimate that the average probability a complier is arrested when CIT is off duty is 37%. Our estimates thus suggest that a CIT response reduces the probability of arrest by

[20] See appendix B.3 for a discussion of compliers.

124% for compliers, though this is imprecise and not statistically different from zero.

3.5.2.2 Use of Force

We estimate the effect of a CIT response on use of force in table 3.4. Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage), as discussed above in section 3.5.1. Panel B displays the estimated discontinuity in use of force from a CIT shift change (reduced form) and panel C displays the estimated discontinuity in use of force from a CIT response using call time as an instrument for the likelihood of a CIT response (local average treatment effect or “LATE"). Column (1) displays estimates from our baseline specification, while columns (2) and (3) add controls for call characteristics and year and month fixed effects, respectively. We focus on estimates from our preferred specification using controls and year and month fixed effects (column (3)), but note that coefficients are similar across all specifications. We estimate a 4.2 percentage point increase in the likelihood that force is used when CIT responds to the call. Again, we calculate average use of force rates rates for compliers, and find that the likelihood a complier has force used when CIT is off duty is 3%.[21] In comparing this to our LATE, our estimates suggest that a CIT unit response increases the probability force is used by 158% for compliers. Estimates, however, are statistically insignificant and imprecise.

3.5.3 Robustness

One concern with regression discontinuity designs is bandwidth selection. Larger bandwidths allow for more precision, but can introduce bias as you incorporate data points further and further from the cutoff. There are also functional form concerns. With smaller bandwidths, you can be more confident that you are picking up an unbiased effect at the threshold and a local linear regression is often appropriate. However, the smaller sample size decreases precision. For this paper, the selection methods outlined by Calonico, Cattaneo and Farrell (2020) yield an optimal bandwidth of 2.5 hours (5:30am to 10:30am). In appendix B.2 we show that our results are consistent when we vary the bandwidth. Appendix figure B.2 plots the estimated coefficients and 95% confidence

[21] See appendix B.3 for a discussion of compliers.

intervals from various regressions of arrest (appendix figure B.2a) and use of force (appendix figure B.2b) on the probability that CIT ever responds to a call using call time as an instrument for a CIT response. These local average treatment effect estimates come from our preferred specification, which controls for exogenous call characteristics and includes year and month fixed effects. Each figure plots seven coefficients: starting with 5 hours, we decrease the bandwidth by 30 minutes until we get to 2 hours on either side of the threshold.[22] Coefficients range between -0.21 and -0.52 for arrest and between -0.03 and 0.08 for use of force, though none are statistically different from zero. As we would expect, the coefficients tend to increase as we decrease the bandwidth. In appendix tables B.1 and B.2 we show first stage, reduced form, and second stage estimates for arrest and use of force, respectively. Importantly, our first stage estimates are consistent and statistically significant across various bandwidths, providing further validity for our research design.

We also show that our results are robust to alternate MH crisis samples. In the main analysis, we define a set of MH crisis calls as those with predicted likelihoods of receiving a CIT unit response in the top 25%. We pick this sample for two reasons. First, we want to have a small enough sample so that we account for calls that would truly warrant a CIT unit response. Second, we want to avoid cutting our sample too small such that we only account for the highest risk calls. In appendix figure B.3 we show the coefficients and 95% confidence intervals from our main specification using alternate MH crisis samples. Specifically, we display estimates for calls with predicted likelihoods between the top 40% and the top 5% (excluding suicides). How do these alternate MH crisis samples compare to the MH crisis sample used in the main analysis? The types of calls in the top 30-40% MH crisis samples are almost identical to the call types in the MH crisis sample used in the main analysis (see figure 3.2d). The most notable difference is an increase in the number of domestic calls. Likewise, the only difference in call type between the top 15-20% and the main MH crisis sample is the loss of disturbance calls. The only call types in the top 5-10% are assistance and welfare calls. Thus, we can think of lower-risk calls as involving

[22] We do not exceed a bandwidth of 5 hours in order to avoid any confounding effects from CIT going off duty at 12am. The data shows that there was a significant lag in CIT ending their shift; because we see CIT responding to calls past 2am, we avoid estimating regressions until 3am. Additionally, while we could estimate regressions using bandwidths smaller than 2 hours, we do not show that here since the standard errors are so large.

more domestic incidents, while the highest-risk calls include assistance and welfare incidents.

As shown in appendix figure B.3a, estimates for MH crisis samples in the top 20-40% show consistent reductions in the likelihood of arrest (between 30 and 45 percentage points), indicating that CIT unit responses have the same effect on arrests when we include lower-risk calls and some (slightly) higher-risk calls. As we restrict to highest-risk calls (top 5-15%), however, estimates vary significantly. Estimates for the top 15% MH crisis sample suggest a 94 percentage point decrease. Moving to the top 10% and top 5%, estimates suggest a 63 percentage point increase and 4 percentage point decrease, respectively. Figure B.3b tells a similar story: estimates for MH crisis samples in the top 25-40% are very similar, only varying between a 3 percentage point decrease and a 4 percentage point increase in the probability force is used. The estimate for the top 20% MH crisis sample is not far off either; it suggests a 15 percentage point increase. Estimates for the top 5-15% have a much larger range.

In appendix tables B.3 and B.4 we show first stage, reduced form, and second stage estimates for arrest and use of force, respectively. It is important to note that first stage estimates are consistent and statistically significant for all MH crisis samples between the top 20-40% (columns 1-5). We do not, however, estimate a statistically significant first stage for MH crisis samples between the top 5-15% (columns 6-8) - i.e., for these samples, there is no discontinuity in the probability that a CIT unit responds to a call when CIT comes on duty. This could indicate that the highest-risk calls always get a CIT unit response: these calls that come in when CIT is on duty get an active response, while the calls that come in when CIT is off duty get a follow-up. Without a significant first stage, we cannot be confident that our LATE estimates the causal effect of a CIT unit response on outcomes. Thus, we do not worry that our estimates differ for these samples.

3.6 Discussion

Do CIT units meet their goals? One goal was to avoid incarceration when allowed by statutes.[23] Per EPPD policy,[24] nonviolent mentally ill individuals with no offense may be transported by

[23] CIT-EHN Agreement

[24] EPPD CIT Policy and Procedures

officers to a treatment facility if desired. Nonviolent mentally ill individuals who have committed an offense will should be brought to jail. For violent mentally ill individuals, officers may request a mental health evaluation and/or emergency detention order, regardless of whether an offense was committed.[25]

While we see whether an arrest was made, the data does not tell us why (i.e., we do not see the offense charged, only the initial call type). Furthermore, the data does not tell us why an arrest was avoided. Examining the types of MH crisis calls in our analysis sample, however, is informative. The vast majority of MH crisis calls involve welfare, domestic, assault, and assistance incidents. Ex-ante, welfare and assistance calls would be less likely to initially involve an arrestable offense. If, instead, arrests arise due to a combative or resistant subject, then a reduction in these arrests could indicate CIT units are deescalating the incident. On the other hand, domestic and assault calls would be ex-ante more likely to initially involve an arrestable offense. A reduction in arrests for these calls could indicate diversion from incarceration due to a mental health issue. Both of these scenarios would suggest that CIT units are successfully avoiding unnecessary incarceration.

A second goal of CIT units was to provide “safer and more effective responses."[26]. At first glance, our use of force results seem to contradict that: we find that the probability of force increases when CIT units respond to the call. All instances of force, however, exhibit type 1, low-level force (e.g., use of officer hands, striking the subject, restraints). Too, type 2, higher-level force (e.g., impact weapons, tasers, chemical agents, canine deployments) is only used in conjunction with type 1, which could suggest that type 2 force is only employed when type 1 force is insufficient. Type 3 (deadly force) is never used. Additionally, in all cases, the reasons for force are a combative and/or non-compliant subject. This could indicate that officers are successfully deescalating; while the goal is verbal diffusion, it is plausible that force may still need to be used.

[25] According to the Texas Health and Safety Code, EDOs allow officers to take a person into custody if that officer believe the person is mentally ill and a danger to his/herself or others. Custody may look like an inpatient or an outpatient mental health facility. An EDO is not involuntary commitment. Upon admission to a mental health facility, a mental health professional will decide whether treatment and/or commitment is necessary.

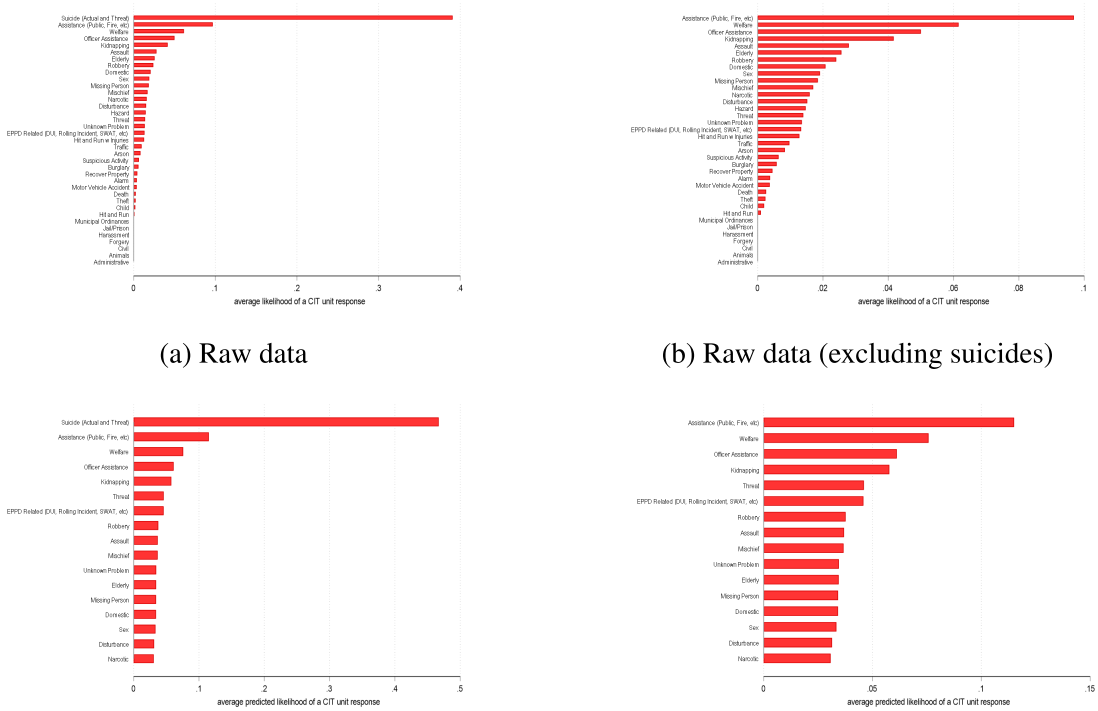
[26] CIT-EHN Agreement

3.7 Conclusion

In this paper, we study CIT units, which pair a specially-trained officer with a mental health professional to respond to calls involving mentally ill individuals. Using data from El Paso, Texas, we examine whether a CIT unit response affects the probabilities of arrest and use of force for MH crisis calls. To do this, we take advantage of variation in CIT unit availability due to CIT shift changes in a fuzzy regression discontinuity design using call time as an instrument for receiving a CIT unit response. First, we find that MH crisis calls that occur just after CIT comes on duty are 3.4 percentage points more likely to receive a CIT unit response - a statistically significant 368% increase. Next, we find a 124% reduction in the probability of arrest and a 158% increase in the probability low-level force is used when a CIT unit responds to a call. Both estimates, however, are imprecise and statistically insignificant. Though suggestive, we present the first causal evidence of CIT units' ability to deescalate police-civilian interactions. Taking point estimates at face value, our results indicate that CIT units may be meeting their goals of deescalating crisis calls and avoiding unnecessary incarceration.

3.8 Figures

Figure 3.1: Likelihood that a Call Type Receives a CIT Unit Response



(a) Raw data

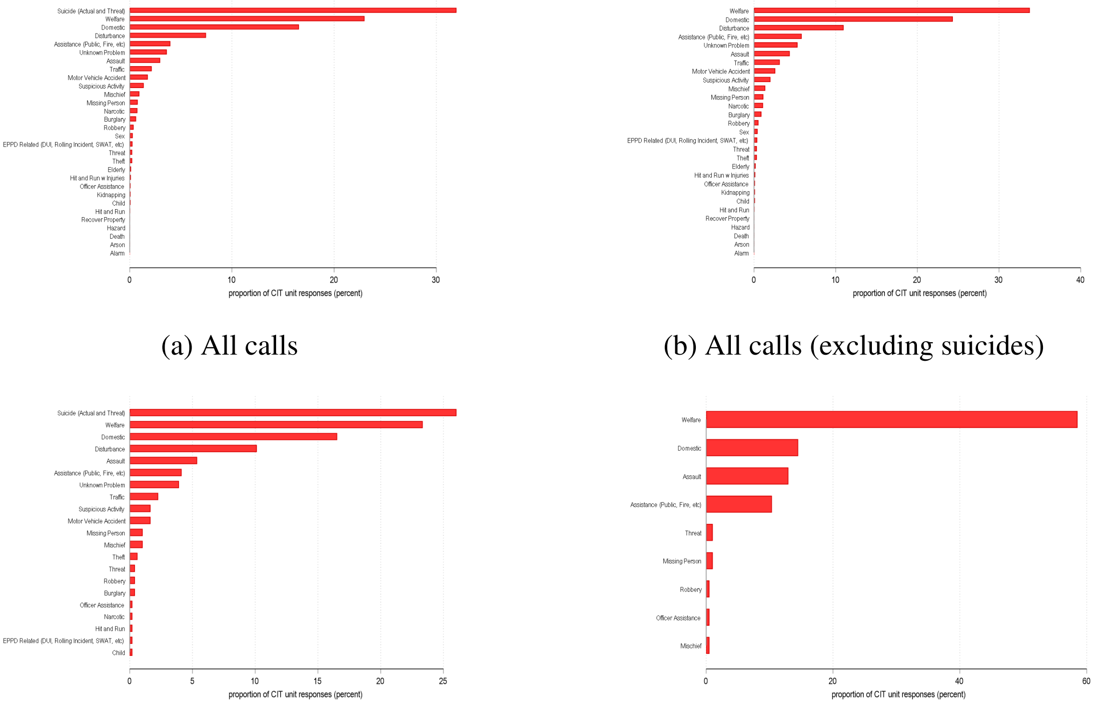
(b) Raw data (excluding suicides)

(c) Prediction in the top 25%

(d) Prediction in the top 25% (excluding suicides)

Notes: Each subfigure displays the likelihood that a call type receives a CIT unit response. Subfigures (a) and (b) display likelihoods for all calls in the full dataset. Subfigures (c) and (d) display the types of calls that are most likely to receive a CIT unit response (i.e., mental health crisis calls). To find this subgroup of calls most likely to receive a CIT response, we predict the likelihood of receiving a CIT response when CIT units are in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%.

Figure 3.2: Call Types that Receive a CIT Unit Response



(a) All calls

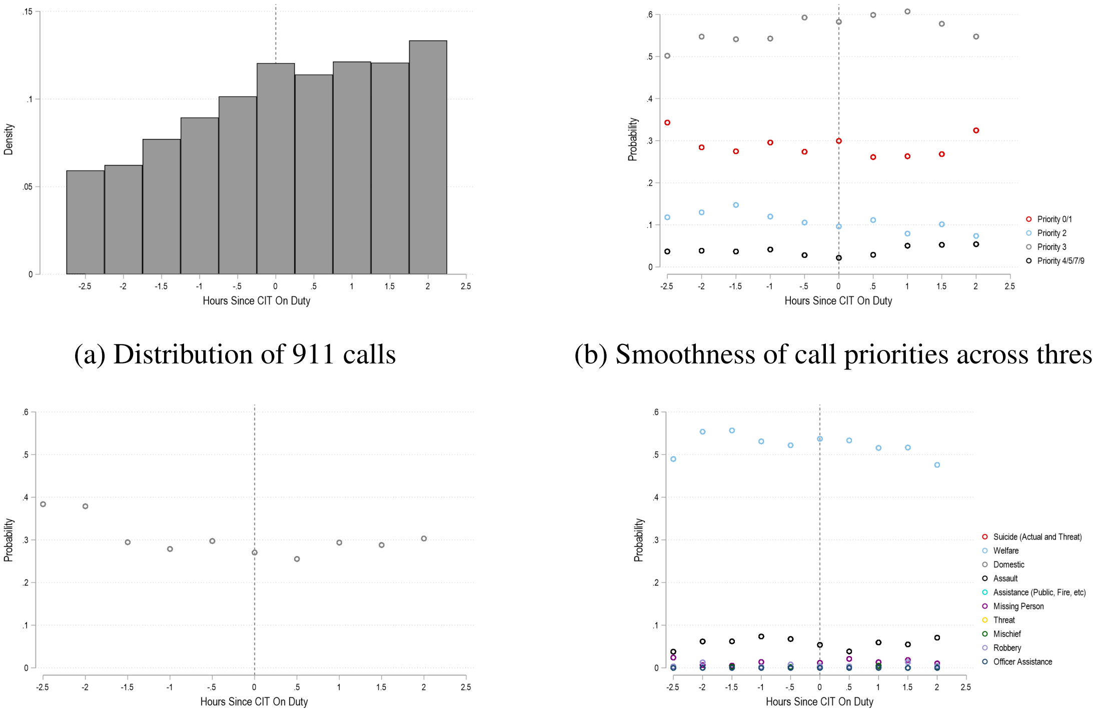
(b) All calls (excluding suicides)

(c) All calls that occur between 5:30am and 10:30am

(d) Calls in analysis sample

Notes: Each subfigure displays call types that receive a CIT unit response (based on raw data). More specifically, each subfigure shows the proportion of CIT unit responses that are of call type *t*. Subfigure (a) displays this for all calls in the full dataset; Subfigure (b) displays this for all calls in the full dataset except suicides. Subfigure (c) displays this for all calls in the full dataset that occur between 5:30am and 10:30am (the optimal bandwidth for our analysis). Subfigure (d) displays this for calls in our final analysis sample.

Figure 3.3: Testing the Validity of the RD Design I



(a) Distribution of 911 calls

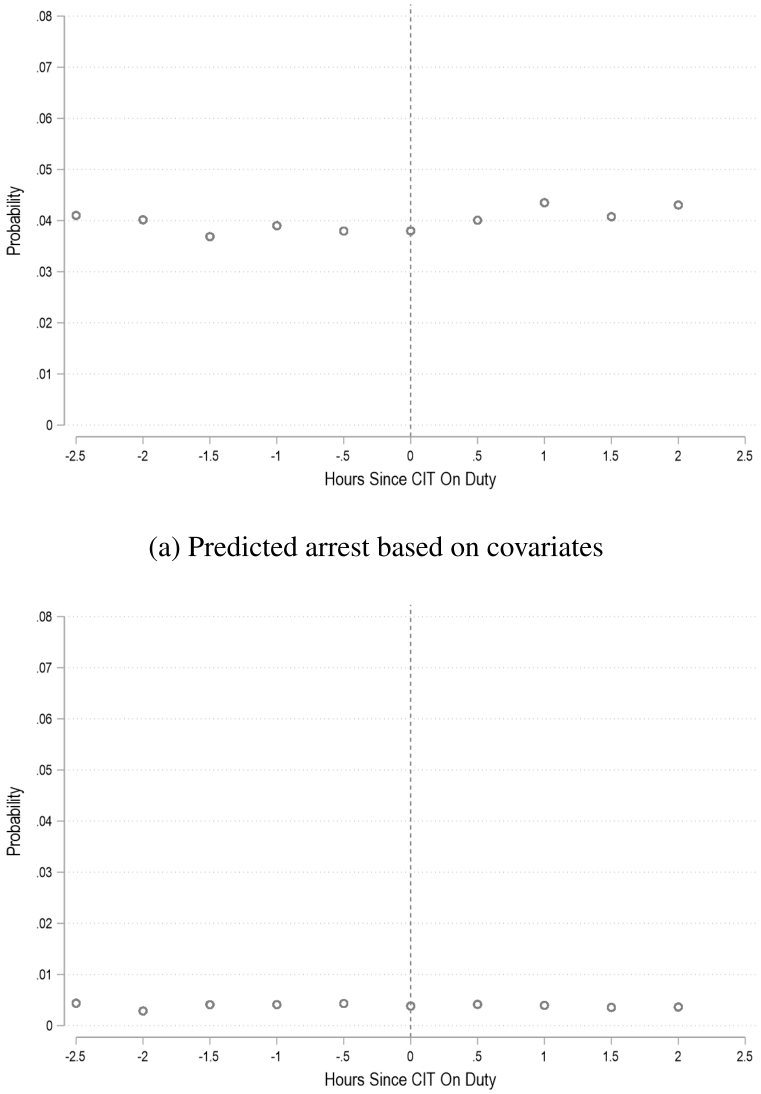
(b) Smoothness of call priorities across threshold

(c) Smoothness of weekend indicator across threshold

(d) Smoothness of call event types across threshold

Notes: Subfigure (a) displays the distribution of calls across call time. From the McCrary (2008) density test, the estimated discontinuity at the threshold is 0.215 with a standard error of 0.112. Subfigures (b)-(d) display the smoothness of baseline covariates across the threshold. Subfigure (b) plots the average likelihood of each call priority across call time, subfigure (c) plots the average likelihood the call occurs on a weekend across call time, and subfigure (d) plots the average likelihood of each call event type across call time. Time $t = 0$ corresponds to 8am, and each bin corresponds to thirty minutes. Each subfigure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

Figure 3.4: Testing the Validity of the RD Design II

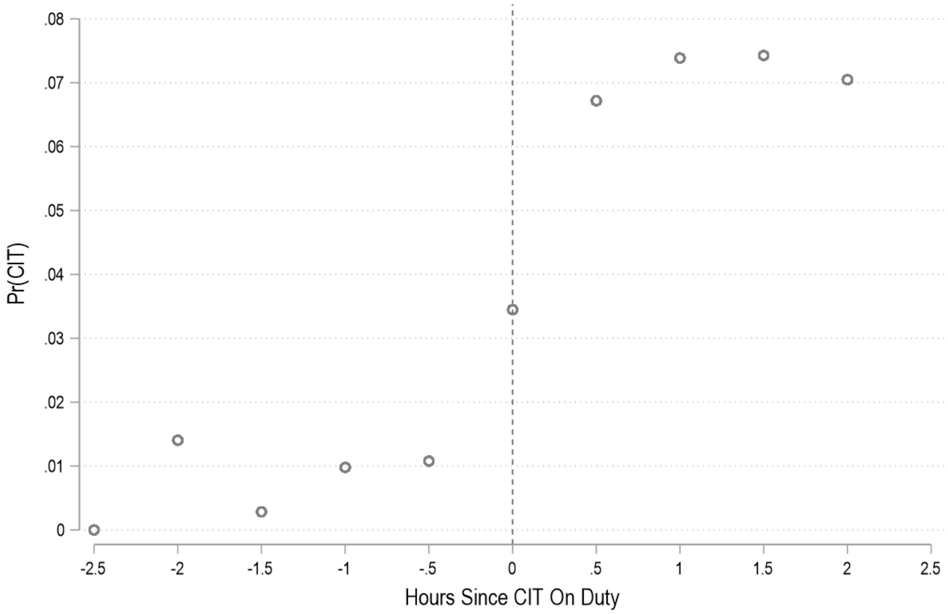


(a) Predicted arrest based on covariates

(b) Predicted use of force based on covariates

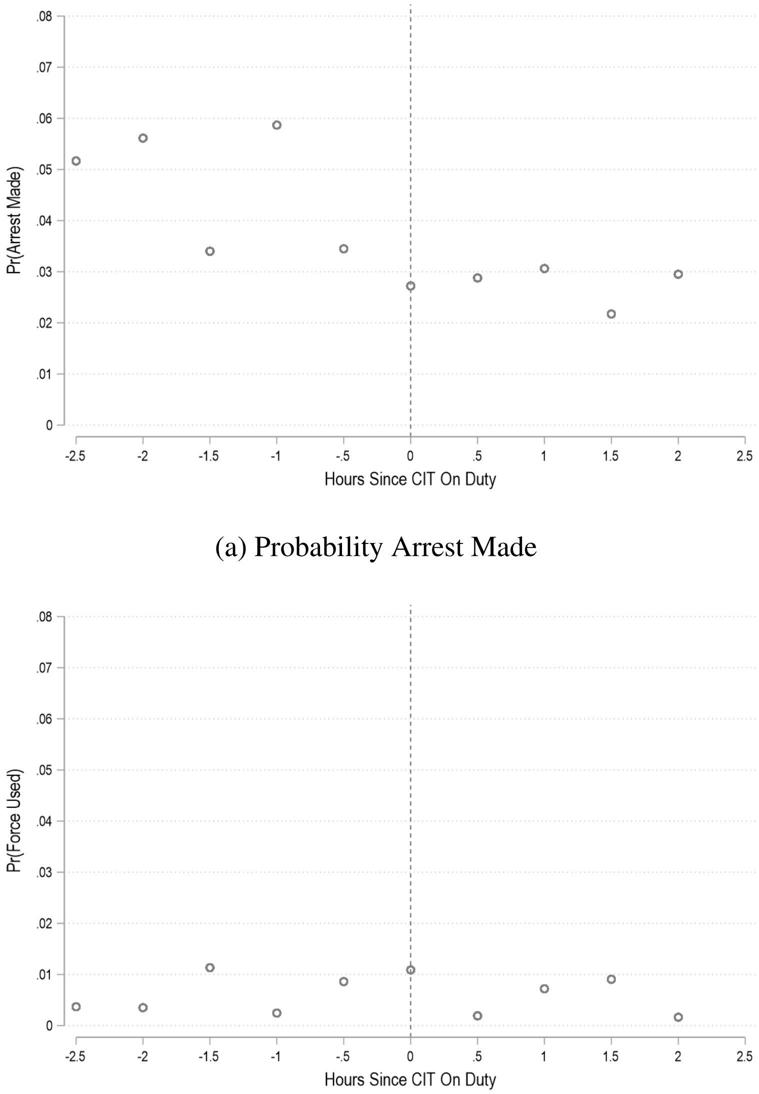
Notes: Subfigure (a) plots predicted arrest and subfigure (c) plots predicted use of force. Both predicted outcomes are based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. Time $t = 0$ corresponds to 8am, and each bin corresponds to thirty minutes. Each figure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

Figure 3.5: Discontinuity in the Probability of a CIT Unit Response Across Call Time (First Stage)



Notes: This figure plots the raw data for all calls against call time. Each marker represents the average likelihood of receiving a CIT unit response (y-axis) during a thirty-minute interval (x-axis). Time $t = 0$ corresponds to 8am. This figure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

Figure 3.6: Discontinuities Across CIT Unit Shift Changes (Reduced Form)



(a) Probability Arrest Made

(b) Probability Force Used

Notes: All figures plot the raw data for all calls against call time. Each marker represents the average likelihood of receiving a CIT unit response (y-axis) during a thirty-minute interval (x-axis). Time $t = 0$ corresponds to 8am. This figure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020). For subfigure (a), each marker represents the average likelihood of arrest during a thirty minute interval. For subfigure (b), each marker represents the average likelihood of use of force during a thirty minute interval. Time $t = 0$ corresponds to 8am. Each subfigure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

3.9 Tables

Table 3.1: Summary Statistics for Call Characteristics and Outcomes

	(1) All Shifts	(2) Non-CIT Unit Shift	(3) CIT Unit Shift
Weekend	0.297 (0.457)	0.319 (0.466)	0.283 (0.450)
Priority 0/1	0.287 (0.452)	0.291 (0.454)	0.284 (0.451)
Priority 2	0.104 (0.305)	0.123 (0.328)	0.092 (0.289)
Priority 3	0.569 (0.495)	0.550 (0.498)	0.582 (0.493)
Priority 4	0.040 (0.195)	0.036 (0.186)	0.042 (0.201)
CIT Ever Responds	0.042 (0.201)	0.008 (0.088)	0.064 (0.245)
Arrest Made	0.035 (0.183)	0.046 (0.210)	0.028 (0.164)
Use of Force	0.006 (0.078)	0.006 (0.078)	0.006 (0.078)
Observations	4,571	1,782	2,789

Notes: This table displays average call characteristics for three sets of calls: (1) all calls, (2) calls that occur during a non-CIT shift, and (3) calls that occur during a CIT shift. Standard deviations are in parentheses. We limit our analysis sample to calls who are most likely to get CIT. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. We further limit our analysis sample to calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

Table 3.2: Testing the Validity of the RD Design

	(1) Predicted Arrest	(2) Predicted Use of Force	(3) Weekend	(4) High Priority
CIT On Duty	0.000	0.000	-0.001	-0.004
	(0.004)	(0.001)	(0.026)	(0.029)
Observations	4,571	4,571	4,571	4,571
Outcome Mean	0.05	0.00	0.30	0.54
	Priority 0/1	Priority 2	Priority 3	Priority 4/5/7/9
CIT On Duty	-0.003	-0.001	0.019	-0.015
	(0.027)	(0.018)	(0.029)	(0.010)
Observations	4,571	4,571	4,571	4,571
Outcome Mean	0.19	0.35	0.43	0.04

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table provides estimates from our baseline regression on predicted arrest, predicted use of force, and exogenous call characteristics (including weekend and call priority). Standard errors are clustered by call time. Both predicted outcomes are based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We limit our analysis sample to calls who are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. We further limit our analysis sample to calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020).

Table 3.3: Estimating the Effect of a CIT Response on Arrests

	(1)	(2)	(3)
A: CIT Ever Responds (First Stage)			
CIT on Duty	0.0324***	0.0337***	0.0341***
	(0.00921)	(0.00915)	(0.00914)
Control Mean	0.01	0.01	0.01
Treatment Effect (%)	349.71	363.43	368.07
B: Arrest Made when CIT on Duty (Reduced Form)			
CIT on Duty	-0.0146	-0.0153	-0.0154
	(0.0110)	(0.0108)	(0.0108)
Control Mean	0.05	0.05	0.05
Treatment Effect (%)	-32.51	-34.02	-34.19
C: Arrest Made when CIT Responds (LATE)			
CIT Ever Responds	-0.452	-0.455	-0.451
	(0.362)	(0.343)	(0.338)
Control Mean	0.05	0.05	0.05
Treatment Effect (%)	-994.01	-1000.81	-993.38
Complier Mean	0.37	0.37	0.37
Complier Treatment Effect (%)	-123.70	-124.55	-123.62
Observations	4,571	4,571	4,571
Controls		X	X
Year & Month FE			X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT unit response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in arrests from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in arrests from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). Estimates in columns (1) come from our baseline regression equation with no controls or fixed effects. Column (2) includes controls, and column (3) adds year and month fixed effects. Controls include indicators for weekend, call priority, and call event type. In all specifications we cluster standard errors by call time. We limit our analysis sample to calls who are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. We further limit our analysis sample to calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020). Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

Table 3.4: Estimating the Effect of a CIT Response on Use of Force

	(1)	(2)	(3)
A: CIT Ever Responds (First Stage)			
CIT on Duty	0.0324***	0.0337***	0.0341***
	(0.00921)	(0.00915)	(0.00914)
Control Mean	0.01	0.01	0.01
Treatment Effect (%)	349.71	363.43	368.07
B: Force Used when CIT on Duty (Reduced Form)			
CIT on Duty	0.00109	0.00143	0.00145
	(0.00498)	(0.00509)	(0.00512)
Control Mean	0.01	0.01	0.01
Treatment Effect (%)	16.54	21.61	21.90
C: Force Used when CIT Responds (LATE)			
CIT Ever Responds	0.0338	0.0425	0.0425
	(0.153)	(0.150)	(0.149)
Control Mean	0.01	0.01	0.01
Treatment Effect (%)	505.82	635.75	636.16
Complier Mean	0.03	0.03	0.03
Complier Treatment Effect (%)	125.90	158.24	158.34
Observations	4,571	4,571	4,571
Controls		X	X
Year & Month FE			X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in use of force from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in use of force from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). Estimates in columns (1) come from our baseline regression equation with no controls or fixed effects. Column (2) includes controls, and column (3) adds year and month fixed effects. Controls include indicators for weekend, call priority, and call event type. In all specifications we cluster standard errors by call time. We limit our analysis sample to calls who are most likely to get a CIT unit response. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. We further limit our analysis sample to calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico, Cattaneo and Farrell (2020). Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

REFERENCES

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." Journal of Econometrics, 113.

Agan, Amanda, Jennifer L. Doleac, and Anna Harvey. 2021. "Misdemeanor Prosecution." Working paper.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association, 91.

Anker, Anne Sofie Tegner, Jennifer L. Doleac, and Rasmus Landerso. 2021. "The Effects of DNA Databases on the Deterrence and Detection of Offenders." American Economic Journal: Applied Economics, 13.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial Bias in Bail Decisions." The Quarterly Journal of Economics, 133.

Arora, Ashna. 2019. "Too tough on crime? The impact of prosecutor politics on incarceration." Working paper.

Aslim, Erkmen G., Murat C. Mungan, Carlos I. Navarro, and Han Yu. 2022. "The Effect of Public Health Insurance on Criminal Recidivism." Journal of Policy Analysis and Management, 41.

Ater, Itai, Yehonatan Givati, and Oren Rigbi. 2014. "Organizational Structure, Police Activity and Crime." Journal of Public Economics, 115.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Daniel Keniston Esther Duflo, and Nina Singh. 2021. "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incenives, Managerial Autonomy, and Training." American Economic Journal: Economic Policy, 13.

Bencsik, Panka. 2021. "Stress on the sidewalk: The mental health costs of close proximity to crime." Working paper.

Bjerk, David. 2005. "Making the Crime Fit the Penalty: The Role of Prosecutorial Discretion

Under Mandatory Minimum Sentencing." Journal of Law and Economics, 48.

Bondurant, Samuel R., Jason M. Lindo, and Isaac D. Swensen. 2018. "Susbstance abuse treatment centers and local crime." Journal of Urban Economics, 104.

Bonfine, Natalie, Christian Ritter, and Mark R. Munetz. 2014. "Police officer perceptions of the impact of Crisis Intervention Team (CIT) programs." International Journal of Law and Psychiatry, 37.

Brennan Center for Justice. 2018. "21 Principles for the 21st Century Prosecutor." Available at brennancenter.org.

Calonico, Sebatian, Matias D. Cattaneo, and Max H. Farrell. 2020. "Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs." Econometrics Journal, 23.

Cheng, Cheng, and Wei Long. 2018. "Improving police services: Evidence from the French Quarter Task Force." Journal of Public Economics, 164.

City of El Paso. 2018. "Interlocal agreement by and among the city of El Paso and El Paso MHMR D/B/A Emergence Health Network for operation of the Crisis Intervention Team." Agreement courtesy of El Paso Open Records.

Compton, Michael T., Roger Bakeman, Beth Broussard, Dana Hankerson-Dyson, Letheshia Husbands, Shaily Krishan, Tarianna Stewart-Hutto, Barbara M. D'Orio, Janet R. Oliva, Nancy J. Thompson, and Amy C. Watson. 2014a. "The police-based crisis intervention team (CIT) model: I. Effects on officers' knowledge, attitudes, and skills." Psychiatric Services, 65.

Compton, Michael T., Roger Bakeman, Beth Broussard, Dana Hankerson-Dyson, Letheshia Husbands, Shaily Krishan, Tarianna Stewart-Hutto, Barbara M. D'Orio, Janet R. Oliva, Nancy J. Thompson, and Amy C. Watson. 2014b. "The police-based crisis intervention team (CIT) model: II. Effects on level of force and resolution, referral, and arrest." Psychiatric Services, 65.

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

Dahl, Gordon B., Andreas Ravndal Kostol, and Magne Mogstad. 2014. "Family Welfare Cultures." The Quarterly Journal of Economics, 129.

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

Devers, Lindsey. 2011. "Plea and Charge Bargaining."

Deza, Monica, Johanna Catherine Maclean, , and Keisha T. Solomon. 2020. "Local Access to Mental Healthcare and Crime." NBER Working Paper 27619.

Doleac, Jennifer L. 2017. "The Effects of DNA Databases on Crime."

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.

Doleac, Jennifer L., Chelsea Temple, David Pritchard, and Adam Roberts. 2020. "Which prisoner reentry programs work? Replicating and extending analyses of three RCTs." International Review of Law and Economics, 62. Article 105902.

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.

Ellis, Horace A. 2014. "Effects of a Crisis Intervention Team (CIT) Training Program Upon Police Officers Before and After Crisis Intervention Team Training." Archives of Psychiatric Nursing, 28.

El Paso Matters. 2020. "El Paso taxpayers spend more than $1.7 million to defend police in four deadly force lawsuits." Available at `https://elpasomatters.org`.

El Paso Police Department. 2022. "El Paso Police Department Procedures Manual." Manual courtesy of El Paso Open Records.

El Paso Times. 2020. "Girl's suicide fuels mother's campaign for better mental health from El Paso police." Available at `https://www.elpasotimes.com`.

EPPD. 2018. "El Paso Police Department Crisis Intervention Team." Presentation courtesy of El Paso Open Records.

Frank, Richard G., and Thomas G. McGuire. 2010. "Mental Health Treatment and Criminal Justice Outcomes." In Controlling Crime: Strategies and Tradeoffs.

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."

Hjalmarsson, Randi, and Matthew J. Lindquist. 2013. "The Origins of Intergenerational Associations in Crime: Lessons from Swedish Adoption Data." Working paper.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." Econometrica, 62.

Jácome, Elisa. 2020. "Mental Health and Criminal Involvement: Evidence from Losing Medicaid Eligibility." Working paper.

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

King County Prosecuting Attorney's Office. 2019. "Filing and Disposition Standards." Section 18.II.B.I.

Krumholz, Sam. 2019. "The Effect of District Attorneys on Local Criminal Justice Outcomes." Working paper.

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.

Owens, Emily. forthcoming. "The Economics of Policing." In The Economics of Risky Behavior. , ed. Dave Marcotte and Klaus Zimmerman. Springer Nature.

Peterson, Jillian, and James Densley. 2018. "Is Crisis Intervention Team (CIT) training evidence-based practice? A systematic review." Journal of Crime and Justice, 41.

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

Rehavi, M. Marit, and Sonja B. Starr. 2014. "Racial Disparity in Federal Criminal Sentences." Journal of Political Economy, 122.

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

Shi, Lan. 2009. "The limit of oversight in policing: Evidence from the 2001 Cincinnati riot." Journal of Public Economics, 93.

Sloan, CarlyWill. 2022. "Racial Bias by Prosecutors: Evidence from Random Assignment." Working paper.

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/`.

Tuttle, Cody. 2021. “Racial Disparities in Federal Sentencing: Evidence from Drug Mandatory Minimums.” Working paper.

Washington Post. 2022. “Police Shootings Database.” Available at `https://www.washingtonpost.com/graphics/investigations/police-shootings-database/`.

Washington State Liquor and Cannabis Board. 2021. “I-502 Implementation.” Available at `lcb.wa.gov/mj-education/know-the-law`.

Yang, Crystal S. 2016. “Resource Constraints and the Criminal Justice System: Evidence from Judicial Vacancies.” American Economic Journal: Economic Policy, 8.

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

APPENDIX A

CHAPTER 1: PRISONER REENTRY PROGRAMS

Table A.1: DYT: Additional Outcomes

VARIABLES	Failed Drug Test	Any Arrest	Arrest for Violation of Probation	Arrest for Technical Violation of Probation	Completed Probation	Referral to/ Enrollment in Drug Treatment	% Drug Tests Failed	Missed Appointment with Probation Officer	Absconded
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table A.2: Aftercare: Summary Statistics (full sample)

	All	Oxford House	Theraputic Community	Control	OH: Difference from Control	TC: Difference from Control
	(1)	(2)	(3)	(4)	(5)	(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. Reprinted with permission from Doleac et al. (2020).

Table A.3: Aftercare: Main Outcomes (full sample)

	Our Results								
	Original Covariates			Amended Covariates			TOT effects (participated 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									
	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table A.4: Aftercare: Additional Outcomes

	Our Results														
	Original Covariates					Ammended Covariates					TOT effects (partic. 30+ days)				
	Days of Alcohol Use	Days of Drug Use	Illegal Income	Legal Issues	Psych. Hosp.	Days of Alcohol Use	Days of Drug Use	Illegal Income	Legal Issues	Psych. Hosp.	Days of Alcohol Use	Days of Drug Use	Illegal Income	Legal Issues	Psych. Hosp.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table A.5: 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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

Table A.6: MCORP: Additional Outcomes

	Original Results		**Our Results**			
			Original Covariates		Amended Covariates	
		Tech. Violation		Tech. Violation		Tech. Violation
	Reconviction	Revocation	Reconviction	Revocation	Reconviction	Revocation
	(1)	(2)	(3)	(4)	(5)	(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 assignment 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: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reprinted with permission from Doleac et al. (2020).

A.1 Matching Analysis for the MCORP Study

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).[1] We show results based on both methods.

In Figure A.1, 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.[2]

Figure A.2 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.[3]

Tables A.7 and A.8 display the covariate balance summary of raw data next to those using the PSM and IPW methods. Specifically, Table A.7 displays the number of observations pre- and post-matching, and Table A.8 displays the standardized differences in means between the two groups pre- and post-matching (we exclude post-randomization variables here). Comparing columns (1)

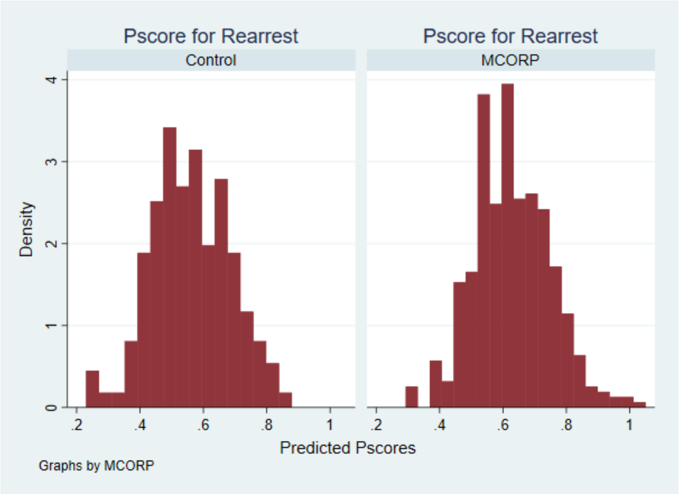
[1] PSM estimators impute the missing potential outcomes for each subject by using an average of the outcomes of similar subjects that received the other treatment level. The treatment effect is computed by taking the average of the difference between the observed and potential outcomes for each subject.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

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

[3] 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 (2) of Table A.8 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.[4] 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.

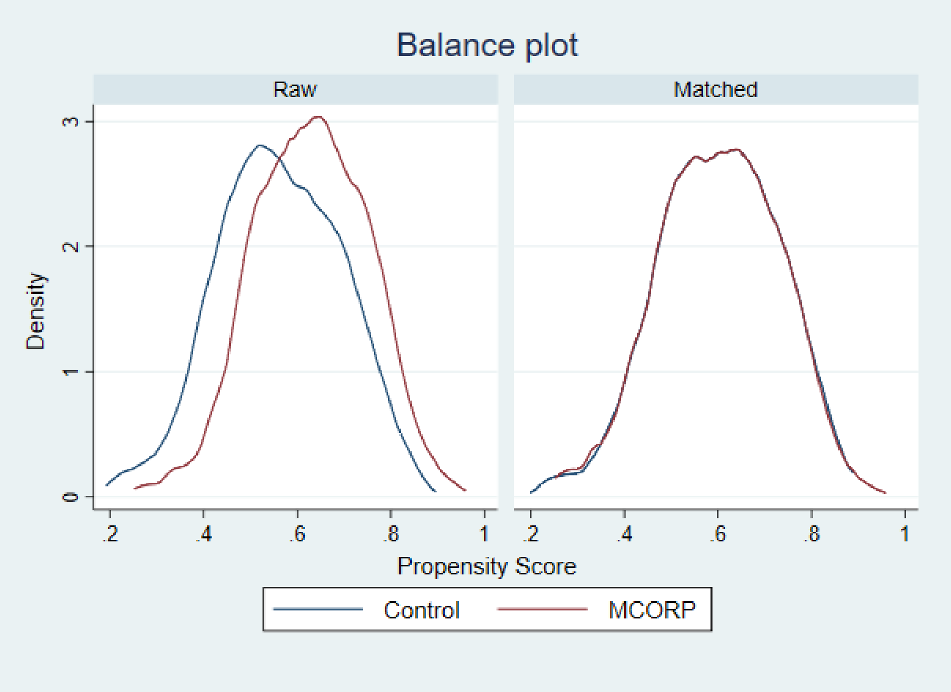
Figure A.1: MCORP: Histogram of Propensity Scores for Rearrest as an Outcome



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. Reprinted with permission from Doleac et al. (2020).

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

Figure A.2: 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. Reprinted with permission from Doleac et al. (2020).

Table A.7: MCORP: Number of Observations for PSM and IPW

	Raw	PSM (matched)	IPW (Weighted)
	(1)	(2)	(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. Reprinted with permission from Doleac et al. (2020).

Table A.8: MCORP: Covariate Balance Summary: Standardized Differences

	Raw	PSM	IPW
	(1)	(2)	(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. Reprinted with permission from Doleac et al. (2020).

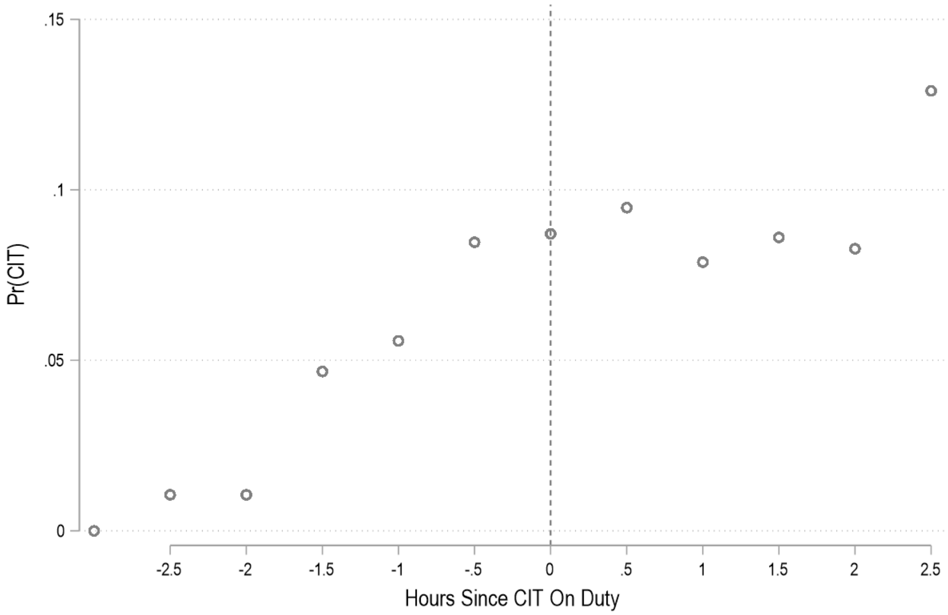
APPENDIX B

CHAPTER 3: CRISIS INTERVENTION TEAMS

B.1 Alternative Cutoff

B.1.1 Figures

Figure B.1: 12am Cutoff: No Discontinuity in the Probability of a CIT Unit Response Across Call Time (First Stage)

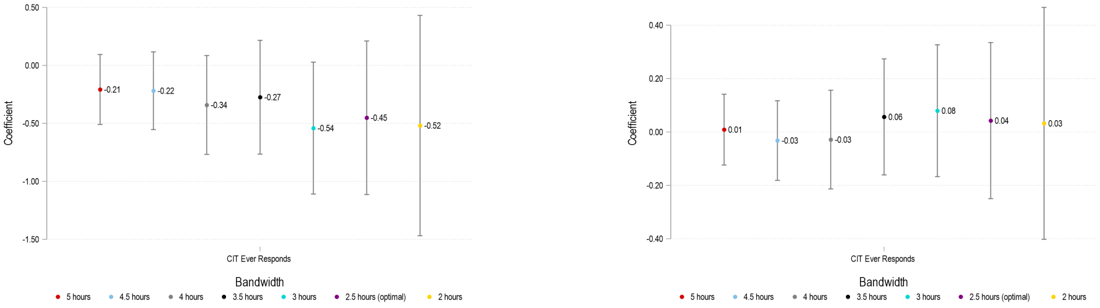


Notes: This figure plots the raw data for all calls against call time. Each marker represents the average likelihood of receiving a CIT unit response (y-axis) during a thirty-minute interval (x-axis). Time $t = 0$ corresponds to 12am. This figure uses a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico et al. (2020).

B.2 Robustness

B.2.1 Figures

Figure B.2: Robustness of LATEs to Varying Bandwidths

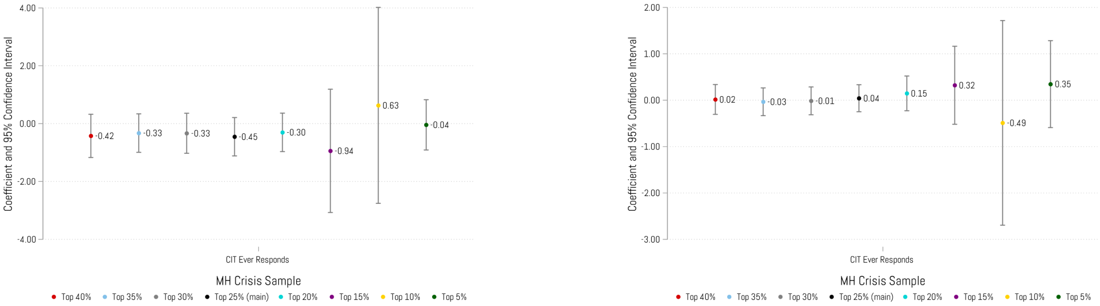


(a) Arrest

(b) Use of Force

Notes: Each figure plots the estimated coefficients and 95% confidence intervals from regressions of a given outcome (arrest or use of force) on the probability that CIT ever responds to a call (using call time as an instrument for a CIT unit response) using varying bandwidths. These estimates come from our preferred specification, which controls for exogenous call characteristics and includes year and month fixed effects. Controls include indicators for weekend, call priority, and call event type. Standard errors are clustered by call time. The optimal bandwidth of 2.5 hours was determined by following the optimal bandwidth selection methods outlined by Calonico et al. (2020). In each regression we limit our analysis sample to calls who are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%.

Figure B.3: Robustness of LATEs to Alternate MH Crisis Samples



(a) Arrest

(b) Use of Force

Notes: Each figure plots the estimated coefficients and 95% confidence intervals from regressions of a given outcome (arrest or use of force) on the probability that a CIT unit ever responds to a call (using call time as an instrument for a CIT unit response) using alternate MH crisis samples. These estimates come from our preferred specification, which controls for exogenous call characteristics and includes year and month fixed effects. Controls include indicators for weekend, call priority, and call event type. Standard errors are clustered by call time. The optimal bandwidth of 2.5 hours was determined by following the optimal bandwidth selection methods outlined by Calonico et al. (2020). In each regression we limit our analysis sample to calls who are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT units are in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top x percent. In our main analysis, we define our MH crisis calls as those with predicted likelihoods of a CIT unit response in the top 25%. In this figure we also show the LATEs for calls with predicted likelihoods in the top 40%, top 35%, top 30%, top 20%, top 15%, top 10%, and top 5% (excluding suicides).

B.2.2 Tables

Table B.1: Robustness in Estimating the Effect of a CIT Unit Response on Arrest: Bandwidth Selection

	BW=2	BW=2.5 (OPT.)	BW=3	BW=3.5	BW=4	BW=4.5	BW=5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: CIT Ever Responds (First Stage)							
CIT on Duty	0.0265**	0.0341***	0.0373***	0.0384***	0.0417***	0.0491***	0.0520***
	(0.0103)	(0.00914)	(0.00853)	(0.00803)	(0.00758)	(0.00709)	(0.00678)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.00	0.00
Treatment Effect (%)	285.67	434.09	520.39	614.76	760.22	1030.05	1247.88
B: Arrest Made when CIT on Duty (Reduced Form)							
CIT on Duty	-0.0137	-0.0154	-0.0202**	-0.0105	-0.0142	-0.0107	-0.0108
	(0.0118)	(0.0108)	(0.00981)	(0.00937)	(0.00871)	(0.00826)	(0.00788)
Control Mean	0.05	0.05	0.04	0.05	0.05	0.05	0.06
Treatment Effect (%)	-30.51	-33.44	-45.90	-21.25	-28.64	-20.10	-19.33
C: Arrest Made when CIT Responds (LATE)							
CIT Ever Responds	-0.519	-0.451	-0.541*	-0.274	-0.342	-0.219	-0.208
	(0.485)	(0.338)	(0.290)	(0.250)	(0.217)	(0.171)	(0.154)
Control Mean	0.03	0.03	0.03	0.04	0.04	0.04	0.04
Treatment Effect (%)	-1483.72	-1291.18	-1614.84	-776.63	-939.93	-577.40	-531.01
Complier Mean	0.37	0.37	0.33	0.40	0.37	0.42	0.45
Complier Treatment Effect (%)	-141.98	-123.62	-165.89	-67.88	-92.93	-51.43	-45.68
Observations	3,667	4,571	5,462	6,409	7,387	8,498	9,589
Controls	X	X	X	X	X	X	X
Year & Month FE	X	X	X	X	X	X	X

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in arrest from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in arrest from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). In each column we vary the bandwidth by 30 minutes. Column (2) is the optimal bandwidth used in the main analysis. All estimates come from our main specification, which includes indicators for weekend, call priority, and call event type, as well as year and month fixed effects. Standard errors are clustered by call time. We limit our analysis sample to calls who are most likely to get CIT. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top 25%. Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

Table B.2: Robustness in Estimating the Effect of a CIT Unit Response on Use of Force: Bandwidth Selection

	BW=2	BW=2.5 (OPT.)	BW=3	BW=3.5	BW=4	BW=4.5	BW=5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A: CIT Ever Responds (First Stage)							
CIT on Duty	0.0265**	0.0341***	0.0373***	0.0384***	0.0417***	0.0491***	0.0520***
	(0.0103)	(0.00914)	(0.00853)	(0.00803)	(0.00758)	(0.00709)	(0.00678)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.00	0.00
Treatment Effect (%)	285.67	434.09	520.39	614.76	760.22	1030.05	1247.88
B: Force Used when CIT on Duty (Reduced Form)							
CIT on Duty	0.000852	0.00145	0.00297	0.00217	-0.00119	-0.00157	0.000466
	(0.00590)	(0.00512)	(0.00470)	(0.00427)	(0.00392)	(0.00372)	(0.00352)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Treatment Effect (%)	12.87	23.48	41.42	30.70	-19.10	-24.68	6.45
C: Force Used when CIT Responds (LATE)							
CIT Ever Responds	0.0322	0.0425	0.0796	0.0566	-0.0285	-0.0320	0.00896
	(0.222)	(0.149)	(0.126)	(0.111)	(0.0943)	(0.0761)	(0.0676)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Treatment Effect (%)	514.67	775.19	1484.98	1081.97	-527.89	-575.99	157.80
Complier Mean	0.02	0.03	0.06	0.05	0.03	0.03	0.04
Complier Treatment Effect (%)	163.18	158.34	142.52	106.48	-109.06	-122.33	20.28
Observations	3,667	4,571	5,462	6,409	7,387	8,498	9,589
Controls	X	X	X	X	X	X	X
Year & Month FE	X	X	X	X	X	X	X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in use of force from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in use of force from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). In each column we vary the bandwidth by 30 minutes. Column (2) is the optimal bandwidth used in the main analysis. All estimates come from our main specification, which includes indicators for weekend, call priority priority, and call event type, as well as year and month fixed effects. Standard errors are clustered by call time. We limit our analysis sample to calls who are most likely to get CIT. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT response in the top 25%. Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

Table B.3: Robustness in Estimating the Effect of a CIT Unit Response on Arrest: Alternate MH Crisis Samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Top 40%	Top 35%	Top 30%	Top 25% (Main)	Top 20%	Top 15%	Top 10%	Top 5%
A: CIT Ever Responds (First Stage)								
CIT on Duty	0.0247***	0.0288***	0.0305***	0.0341***	0.0306***	0.0123	-0.00658	0.0254
	(0.00698)	(0.00756)	(0.00838)	(0.00914)	(0.0104)	(0.0114)	(0.0142)	(0.0294)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Treatment Effect (%)	340.82	372.37	347.62	434.09	327.56	106.58	-47.28	121.24
B: Arrest Made when CIT on Duty (Reduced Form)								
CIT on Duty	-0.0104	-0.00941	-0.0102	-0.0154	-0.00922	-0.0116	-0.00415	-0.00104
	(0.00902)	(0.00956)	(0.0105)	(0.0108)	(0.00993)	(0.00812)	(0.00711)	(0.0114)
Control Mean	0.05	0.05	0.05	0.05	0.03	0.02	0.01	0.00
Treatment Effect (%)	-21.87	-19.35	-19.70	-33.44	-31.47	-73.98	-46.87	-29.69
C: Arrest Made when CIT Responds (LATE)								
CIT Ever Responds	-0.422	-0.326	-0.334	-0.451	-0.302	-0.942	0.631	-0.0408
	(0.381)	(0.340)	(0.353)	(0.338)	(0.339)	(1.087)	(1.728)	(0.443)
Control Mean	0.04	0.04	0.04	0.03	0.02	0.01	0.01	0.00
Treatment Effect (%)	-1115.55	-839.54	-835.51	-1291.18	-1266.30	-10236.48	12249.90	-931.97
Complier Mean	0.39	0.38	0.43	0.37	0.18	0.20	0.11	-0.02
Complier Treatment Effect (%)	-108.76	-86.07	-77.24	-123.62	-169.61	-465.53	552.53	227.89
Observations	6,826	6,067	5,318	4,571	3,826	3,074	2,041	726
Controls	X	X	X	X	X	X	X	X
Year & Month FE	X	X	X	X	X	X	X	X

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in arrest from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in arrest from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). In each column, we limit our analysis sample to calls that are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top x percent. Column (4) is the MH crisis sample used in the main analysis. All estimates come from our main specification, which includes indicators for weekend, call priority priority, and call event type, as well as year and month fixed effects. Standard errors are clustered by call time. All samples include calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico et al. (2020). Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

Table B.4: Robustness in Estimating the Effect of a CIT Unit Response on Use of Force: Alternate MH Crisis Samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Top 40%	Top 35%	Top 30%	Top 25% (Main)	Top 20%	Top 15%	Top 10%	Top 5%
A: CIT Ever Responds (First Stage)								
CIT on Duty	0.0247***	0.0288***	0.0305***	0.0341***	0.0306***	0.0123	-0.00658	0.0254
	(0.00698)	(0.00756)	(0.00838)	(0.00914)	(0.0104)	(0.0114)	(0.0142)	(0.0294)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Treatment Effect (%)	340.82	372.37	347.62	434.09	327.56	106.58	-47.28	121.24
B: Force Used when CIT on Duty (Reduced Form)								
CIT on Duty	0.000399	-0.000979	-0.000433	0.00145	0.00451	0.00397	0.00321	0.00883
	(0.00408)	(0.00440)	(0.00467)	(0.00512)	(0.00579)	(0.00378)	(0.00266)	(0.00712)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
Treatment Effect (%)	5.23	-12.65	-5.55	23.48	75.14	240.87	254.15	.
C: Force Used when CIT Responds (LATE)								
CIT Ever Responds	0.0162	-0.0340	-0.0142	0.0425	0.147	0.323	-0.489	0.347
	(0.164)	(0.153)	(0.153)	(0.149)	(0.191)	(0.429)	(1.125)	(0.479)
Control Mean	0.01	0.01	0.01	0.01	0.01	0.00	0.00	0.00
Treatment Effect (%)	287.26	-583.57	-241.80	775.19	2832.25	11840.84	-31631.64	23774.26
Complier Mean	0.08	0.07	0.07	0.03	0.03	-0.03	-0.01	-0.04
Complier Treatment Effect (%)	21.12	-46.89	-20.27	158.34	528.84	-1151.11	8221.31	-894.37
Observations	6,826	6,067	5,318	4,571	3,826	3,074	2,041	726
Controls	X	X	X	X	X	X	X	X
Year & Month FE	X	X	X	X	X	X	X	X

Standard errors in parentheses
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Notes: Panel A displays the estimated discontinuity in a CIT response from a CIT shift change (first stage). Panel B displays the estimated discontinuity in use of force from a CIT shift change (reduced form). Panel C displays the estimated discontinuity in use of force from a CIT unit response using a CIT shift change as an instrument for the likelihood of a CIT unit response (local average treatment effect). In each column, we limit our analysis sample to calls that are most likely to get a CIT unit. In this exercise, we aim to examine the calls ex-ante most likely to receive a CIT unit response. To do this, we predict the likelihood of receiving a CIT unit response when CIT is in full operation (9am-11pm) based on the following exogenous covariates: indicators for call priority, call event type, day of month, and day of week, as well as month and year fixed effects. Robust standard errors are used. We then define a subsample of calls with a predicted CIT unit response in the top x percent. Column (4) is the MH crisis sample used in the main analysis. All estimates come from our main specification, which includes indicators for weekend, call priority priority, and call event type, as well as year and month fixed effects. Standard errors are clustered by call time. All samples include calls that occur within a bandwidth of 2.5 hours, which was determined by following the optimal bandwidth selection methods outlined by Calonico et al. (2020). Control means are calculated by estimating the average for calls in the hours prior to a CIT shift change (i.e., calls between 5:30am and 7:59am). Treatment effects are calculated by dividing the estimate by the control mean and multiplying by 100. Since the LATE is the average treatment effect (ATE) for compliers, we calculate complier means following the work of Angrist et al. (1996), Abadie (2003), Dahl et al. (2014), Dobbie & Yang (2018), and Agan et al. (2021). Complier treatment effects are calculated by dividing the estimate by the complier mean and multiplying by 100.

B.3 Technical Appendix

B.3.1 Understanding Compliers

Our treatment is binary: CIT either responds to a call or doesn't. Thus, treatment D is equal to 0 or 1. Furthermore, following Imbens and Angrist (1994) and Abadie (2003) we can classify potential treatments as $D_{z=0} = 0$ and $D_{z=1} = 1$, where z is the value of the instrument (here, $z = 1$ when CIT is on duty and $z = 0$ when CIT is off duty). Per Angrist, Imbens and Rubin (1996) we can divide our study population into subpopulations based on these potential treatments. For compliers, $D_0 = 0$ and $D_1 = 1$ (or $D_1 > D_0$). Compliers receive treatment (CIT response) when the instrument is turned on (CIT is on duty), and do not receive treatment when the instrument is turned off (CIT is off duty). For always-takers, $D_0 = D_1 = 1$. This group receives a CIT response no matter what. While it would seem impossible for calls to receive a CIT unit response when CIT is off duty, in our setting CIT may follow up with non-active calls (some of which may occur outside CIT shift hours). Too, there is a significant lag in CIT units going off duty at midnight, resulting in calls during off-hours receiving a CIT unit response. As a result, we have a small proportion of always-takers. For never-takers, $D_0 = D_1 = 0$. This group never receives a CIT response no matter what. The last group is defiers: $D_0 = 1$ and $D_1 = 0$ (or $D_1 < D_0$). Under the monotonicity assumption, the sign of the first stage is the same for everyone: in our case, when a CIT shift begins, some calls have a higher likelihood of receiving a CIT unit response. Some calls will not be affected by CIT coming on duty, but it will never be the case that calls will have a lower probability of receiving a CIT unit response when a CIT shift begins (i.e., defiers do not exist). Thus, our population can be divided into three subpopulations: compliers, always-takers, and never-takers.

Following the work of Angrist, Imbens and Rubin (1996), Abadie (2003), Dahl, Kostol and Mogstad (2014), Arnold, Dobbie and Yang (2018), and Agan, Doleac and Harvey (2021), we can calculate the share of each subpopulation. Let D_i represent CIT receipt for call i. For calls that receive a CIT response $D_i = 1$; for calls that do not receive a CIT response $D_i = 0$. Additionally, let z_i be the value of the instrument for call i such that $z_i = 0$ when CIT is off duty and $z_i = 1$ when CIT is on duty. The share of always-takers is given by:

$$\pi_a \equiv P[(D_i = 1|z_i = 0) = (D_i = 1|z_i = 1) = 1] \tag{B.1}$$

The share of never-takers is given by:

$$\pi_n \equiv P[(D_i = 1|z_i = 0) = (D_i = 1|z_i = 1) = 0] \tag{B.2}$$

The share of compliers is given by:

$$\begin{aligned}\pi_c \equiv& P(D_i = 1|z_i = 0) = 0 \wedge P(D_i = 1|z_i = 1) = 1 \\ =& P(D_i = 1|z_i = 1) - P(D_i = 1|z_i = 0) \\ =& P[(D_i = 1|z_i = 1) > (D_i = 1|z_i = 0)] = P(D_1 > D_0)\end{aligned} \tag{B.3}$$

We calculate the share of never-takers by counting the number of calls that do not receive a CIT unit when CIT is on duty and dividing that by the total number of calls that occur when CIT is on duty. For our analysis sample (calls with predicted likelihoods of a CIT unit response in the top 25% and that occur between 5:30am and 10:30am) $\pi_n = 0.964$. Likewise, we calculate the share of always-takers and compliers by counting the number of calls that receive CIT when CIT is on duty and dividing that by the total number of calls that occur when CIT is on duty. For our analysis sample $\pi_a + \pi_c = 0.036$. Finally, we calculate the share of always-takers by counting the number of calls that receive CIT when CIT is off duty and dividing that by the total number of calls that occur when CIT is off duty. For our analysis sample $\pi_a = 0.005$. Thus, $\pi_c = 0.036 - 0.005 = 0.031$.

To better understand our LATE, we can calculate average outcomes for complier calls that do not receive a CIT unit response when CIT is off duty: $E[Y_i(0)|D_1 > D_0]$. Calls that do not receive a CIT unit when CIT is off duty are a mixture of never-takers and compliers. Average outcomes for these calls are represented by:

$$E[Y_i|D_i = 0, z_i = 0] = \frac{\pi_c}{\pi_c + \pi_n} E[Y_i(0)|D_1 > D_0] + \frac{\pi_n}{\pi_c + \pi_n} E[Y_i(0)|D_1 = D_0 = 0] \quad \text{(B.4)}$$

Calls that do not receive CIT when CIT is on duty are never-takers. Average outcomes for these calls are represented by:

$$E[Y_i|D_i = 0, z_i = 1] = E[Y_i(0)|D_1 = D_0 = 0] \quad \text{(B.5)}$$

By combining these two equations, we can calculate the average outcomes for complier calls when CIT is off duty:

$$E[Y_i(0)|D_1 > D_0] = \frac{\pi_c + \pi_n}{\pi_c} E[Y_i|D_i = 0, z_i = 0] + \frac{\pi_n}{\pi_c} E[Y_i|D_i = 0, z_i = 1] \quad \text{(B.6)}$$

The first expectation term in this equation is the expected outcome for calls that do not receive a CIT unit when CIT is off duty. For arrests this equals 0.05, and for use of force this equals 0.006. The second expectation term in this equation is the expected outcome for calls that do not receive CIT when CIT is on duty. For arrests this equals 0.03, and for use of force this equals 0.005.