Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program*

Yotam Shem-Tov Steven Raphael Alissa Skog

January 14, 2022

Abstract

This paper studies the effect of a restorative justice intervention targeted at youth ages 13 to 17 facing felony charges of medium severity (e.g., burglary, assault). Eligible youths were randomly assigned to participate in the Make-it-Right (MIR) restorative justice program or a control group where they faced standard criminal prosecution. We estimate the effects of MIR on the likelihood that a youth will be rearrested in the four years following randomization. Assignment to MIR reduces the probability of a rearrest within six months by 19 percentage points, a 44 percent reduction relative to the control group. Moreover, the reduction in recidivism persists even four years after randomization. Thus, our estimates show that restorative justice conferencing can reduce recidivism among youth charged with relatively serious offenses and can be an effective alternative to traditional criminal justice practices.

*Yotam Shem-Tov, Assistant Professor, Department of Economics, University of California, Los Angeles; shemtov@econ.ucla.edu. Steven Raphael, Professor and James D. Marver Chair in Public Policy, Goldman School of Public Policy University of California, Berkeley; and NBER; stevenraphael@berkeley.edu. Alissa Skog, Senior Research Associate, California Policy Lab, University of California, Berkeley; alissaskog@berkeley.edu. A pre-analysis plan for this project can be found at the Open Science Foundation, https://osf.io/3sb4u/. We are especially grateful to Katherine Miller, Maria McKee, Tara Regan Anderson, Mikaela Rabinowitz, and Jean Roland for the many insights they provided over the course of this project, as well as to the former District Attorney of San Francisco, George Gascon, who initiated Make-it-Right and advocated for the random allocation of individuals into the program which allowed for this study to be conducted. We are also thankful for invaluable information Lauren Brown and staff from Community Works and Huckleberry Youth shared about the principles of restorative justice and the program's implementation. The California Policy Lab provided continuous support and help throughout this project, we would especially like to thank April Chang, Johanna Lacoe, and Evan White. We thank Hadar Avivi, Annabelle Berrios, Gale Burford, Robynn Cox, Peng Ding, Sara Heller, Dmitri Koustas, Nicholas Li, Juliana Londoño-Vélez, Maxim Massenkoff, Conrad Miller, Andrew Penner, Deepak Premkumar, Heather Strang, and Santiago Tobón for helpful comments and discussions. We also thank seminar and conference participants at Harvard Kennedy School Social Inequality Seminar, University of British Columbia, WEAI, UCLA, NBER SI Crime, California Policy Lab, 14th All-California Labor Economics Conference, and LACEA Workshop on the Economics of Crime for numerous helpful comments and suggestions. Chandni Raja, Shivani Ghate, Logan Spencer provided outstanding research assistance. Any opinions and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the San Francisco District Attorney, San Francisco Juvenile Probation Department, Community Works, Huckleberry Youth, or the California Policy Lab. We thank Arnold Ventures, the University of California Office of the President Multicampus Research Programs and Initiatives, MRP-19-600774 and M21PR3278, The James Irvine Foundation, and the Bylo Chacon Foundation for their generous support. The views expressed here do not necessarily reflect the views of the funders. All errors should be attributed to the authors.

1 Introduction

The United States criminal justice policy has historically relied on sanctions to enforce compliance. While the objectives of criminal justice policy are complex and multifaceted, much policy debate and research focuses on the efficacy of sanctions (e.g., Kuziemko, 2013; Aizer and Doyle, 2015; Bhuller et al., 2020; Rose and Shem-Tov, 2021). Sanctions became increasingly punitive beginning in the late 1970s, leading to large increases in the correctional populations (Raphael and Stoll, 2013). These policies had disparate impacts and generated large racial disparities in incarceration rates (Neal and Rick, 2014; Lofstrom and Raphael, 2016). Recent years have seen an effort to dial back the severity of punishments and use alternatives to prosecution to reduce correctional population without impacting public safety (Mueller-Smith and Schnepel, 2020; Rose, 2020). Alternative programs for juveniles might be especially effective, given the strong correlation between age and criminal involvement (Hirschi and Gottfredson, 1983; Ulmer and Steffensmeier, 2014).

Restorative justice conferencing is an alternative that emphasizes accountability through repairing harm rather than imposing sanctions. While not always used instead of traditional prosecution, several countries and jurisdictions throughout the U.S. have experimented with this alternative prosecution path. Restorative justice programs typically involve a structured conference involving the victim, the accused, and supporters of both parties, followed by an agreement whereby the accused makes amends for the crime through a mutually agreed-upon set of actions. The process involves the individual charged with the offense taking responsibility for their actions and engaging in dialogue with the victim(s), as well as family and other community members, about the impact of their actions. The current evidence on the effectiveness of restorative justice programs in reducing recidivism is mixed (Wilson et al., 2018).

This paper studies "Make-it-Right" (MIR), a restorative justice conferencing program implemented in San Francisco for teenagers who would otherwise be charged with felony offenses of medium severity (e.g., burglary, theft, assault). Eligible cases were randomly assigned to either the MIR program (treatment group) or a regular felony prosecution (control group). Youth assigned to the treatment group were offered the opportunity to participate in the restorative justice program run by the organization Community Works (CW) instead of standard case processing and prosecution. Successful completion of the program results in formal charges never being filed.

The experiment included 143 youth, constituting 13 percent of all juveniles charged with a felony in San Francisco during the time period. Among the experimental sample, 99 were assigned to MIR, and 44 faced regular felony prosecution (control regime). Although the sample size is relatively small, the treatment effects are large enough to credibly conclude that MIR has large recidivism reducing effects along multiple dimensions that persist even four years after referral to the program. To conduct inference, we report p-values using both the standard methods based on asymptotic approximations as well as randomization inference (Fisher, 1935) which is finite-sample exact.

The program's target population is high-risk youth: 43 percent of control group members are rearrested within six months, and 75 percent are rearrested within four years of treatment assignment. We estimate the intent-to-treat (ITT) and treatment-on-the-treated (TOT) effects of the program on recidivism in the four years after randomization. We find that MIR has large crime-reducing effects. Juveniles assigned to MIR are less likely to be subsequently arrested for any reason, for felonies, or for more severe offenses. Correspondingly, after treatment, juveniles are also less likely to be convicted of a new offense. The effects of the program are large. Youths assigned to MIR are 19 percentage points less likely to be arrested for another offense in the six months after randomization. This estimate represents a 44 percent reduction relative to the rearrest rate among the control group of 43 percent. Moreover, the reduction persists years after randomization. Those assigned to MIR are 15 percentage points (20 percent) less likely to be rearrested within three years and 27 percentage points less likely (32 percent) after four years. Among those assigned to MIR, 81 percent enrolled in the program, and 53 percent completed it. Accounting for imperfect take-up of treatment magnifies the results. Two-stage least squares (2SLS) estimates of the TOT with respect to enrollment and completion are roughly 1.3 and 1.9 times larger than the ITT effects.

Next, we compare the recidivism rates among the MIR control group and the full population of juveniles charged with felony offenses in San Francisco between October 2013 to May 2019. Although eligibility to MIR was not random (e.g., there are restrictions based on criminal history, gang affiliation, and type of offense), the MIR control group's rearrest rate is similar to that of the entire population of juveniles with felony arrests in San Francisco. Thus, the MIR experimental population is not a selected low-risk population. The similarity in risk between the experimental population and the broad population of juveniles arrested for a felony offense suggests that our experimental estimates may be predictive of the effect of expanding the program to the wider pool of juvenile cases.

Finally, we investigate the mechanisms through which assignment to MIR impacts recidivism. MIR has two essential components: restorative justice conferencing itself and diversion from felony prosecution. As a diversion program, MIR affects case outcomes. For example, among youth assigned to MIR, only five percent of the cases result in a felony conviction relative to 20 percent in the control group. To distinguish between MIR's effects through restorative justice conferencing from any effects that are mediated through impacts on avoiding a felony conviction, we restrict attention to youth who have not been convicted of a felony offense. Individuals whose case results in a felony conviction are likely those of equal or higher risk of recidivism. Among the control group, 20 percent of cases will result in a felony conviction relative to five percent in the treatment group. Thus, removing these cases drops more individuals of potentially higher recidivism risk from the control group. Nevertheless, in this restricted sample, the impacts of MIR on recidivism are similar to those in the full and non-restricted experimental sample. One potential cause for these findings is that in California, arrests, charges, and convictions in juvenile court do not show

up in employment criminal history background checks.

The relatively muted impact of the diversion component of MIR suggests that fostering empathy and dialogue with the victim are a driving force behind MIR's large recidivism-reducing effects. Moreover, it supports the hypothesis that the interaction with the victim during the restorative justice conference is a transformative experience for the youth. Indeed, among youth who participated in conferencing, 95 percent fulfilled the agreement and completed the program.

Our results contribute to a broad literature across the social sciences on the determinants of criminal behavior. Previous studies document that an individual's criminal behavior is influenced by punishment severity and swiftness (Drago et al., 2009; Hjalmarsson, 2009; Abrams, 2012; Aizer and Doyle, 2015; Davidson et al., 2019; Eren and Mocan, 2021), the likelihood of being caught (Di Tella and Schargrodsky, 2004; Blanes i Vidal and Mastrobuoni, 2018), the economic rewards of the crime (Draca et al., 2019; Kirchmaier et al., 2020), neighborhoods and peers (Ludwig et al., 2001; Kling et al., 2005; Damm and Dustmann, 2014), and opportunities in the labor market (Bushway, 2004; Raphael, 2014; Yang, 2017; Agan and Starr, 2018; Schnepel, 2018; Britto et al., 2020; Doleac and Hansen, 2020; Cullen et al., 2021; Rose, 2021; Khanna et al., 2021).[1] However, little is known about whether or not criminal behavior can be influenced by less punitive interventions that use self reflection and empathy to directly change a youth's decision-making.

Two related studies investigate the effects of cognitive-behavioral therapy (CBT) on anti-social and criminal behavior. Heller et al. (2017) conduct a randomized controlled trail (RCT) among youth in Chicago and show that CBT changes a youth's decision-making, reduces criminal behavior, and increases high school graduation. Blattman et al. (2017) conduct an RCT in Liberia recruiting high-risk men aged 18-35. They find that CBT reduces anti-social behavior. However, the effects dissipate and are not significant after a year unless combined with a monetary grant. Our analysis complements these findings by showing that similar to CBT, interventions that appeal to one's sense of responsibility (and perhaps shame) and, unlike CBT, specifically encourage pro-social behaviors have the potential to reduce recidivism. Our findings are especially encouraging when considering that our sample is comprised of high-risk juveniles charged with felony offenses of mid-level severity (e.g., burglary, assault) and not those charged with only minor offenses. Through the lens of Becker (1968), both MIR and CBT-type interventions do not change the economic incentives individuals are facing (i.e., the costs and benefits from criminal behavior); rather, they influence the individual's decision making (i.e., preferences) while holding the setting fixed.

We also contribute to the growing literature that evaluates the efficacy of different diversion programs aimed at addressing the needs of those who become involved with the criminal justice system. Such diversion programs range from courts focused specifically on the treatment of defendants with severe mental illness (Cuellar et al., 2006; Seward et al., 2021), to drug courts that

[1] A growing recent literature also finds that parental criminal involvement can impact children's likelihood of interacting with the criminal justice system (e.g., Bhuller et al., 2018; Dobbie et al., 2018; Huttunen et al., 2019; Arteaga, 2020; Norris et al., 2020).

focus on compliance with substance use disorder treatment, sobriety, and recidivism (Mitchell et al., 2012), to specialty court focusing on the needs of specific populations such as veterans or domestic violence cases (Owens et al., 2021). Two recent studies use different sources of exogenous variation to identify the effects of felony diversion on recidivism. Augustine et al. (2021) exploit quasi-random assignment to judges and Mueller-Smith and Schnepel (2020) leverage natural experiments associated with shifts in diversion policy. Both studies find evidence that diversion programs reduce recidivism.

Although related, restorative justice conferencing is fundamentally different than common diversion programs as it presents an alternative model to addressing the harms caused by a criminal incident. For instance, the conference gives the victim an active and prominent role. Moreover, in our setting, we show that MIR's large recidivism-reducing impacts are not mediated by its effects on avoiding a felony conviction (through its diversion component). Thus, our findings suggest that fostering dialogue (and perhaps empathy) between the victim and the accused can have meaningful long-term effects on recidivism.

The remainder of this paper is organized as follows. Section 2 describes restorative justice practices and reviews the previous literature on such interventions. Section 3 describes the institutional setting, data sources, and summary statistics. Section 4 presents our empirical strategy. Section 5 reports our main findings and additional analyses regarding the external validity of the estimated effects of MIR. Section 6 discusses the potential mechanisms driving our results. Finally, Section 7 concludes.

2 Restorative Justice Conferencing: Theoretical Background and Empirical Evidence

In this section, we begin by briefly describing the theoretical considerations behind the current sentencing system in the U.S. and some of its drawbacks. We then explain what restorative justice is and its theoretical motivation. Lastly, we discuss the existing empirical evidence on the efficacy of restorative justice interventions in reducing recidivism.

2.1 The Current System of Punishment

Criminal justice policy and practice in the U.S. has historically been structured around several normative criteria. Consequentialist policy focuses on creating an enforcement and punishment architecture in service of deterrence, incapacitation, and rehabilitation. Policy levers devoted to increasing the likelihood of detection and the swiftness and/or severity of sanction are used to minimize criminal offending, often with an eye to the fiscal and social costs of exercising these policy levers. On the other hand, retributivist sentencing focuses on the moral content of the

criminal offense, the degree to which the offense violates community norms, and the punishment proportional to the degree of harm caused by the offense, broadly defined. Retributivist theory focuses on the actions and autonomy of the individual, and the reaffirmation of community values by delivering through sentencing the "just deserts" merited by the violation of community norms. These alternative normative approaches to criminal sanctioning are in some instances complementary, but are often in conflict. For example, lengthy prisons sentences that incarcerate people into advanced ages who commit serious violent offenses in their youth may be justified based on the gravity of the crime, but may not be justifiable in terms of deterrence and incapacitation (especially during the latter years of a sentence).

Despite their different objectives and at times prescriptions, these two normative approaches to criminal justice are firmly nested within policy approaches to criminal offending based almost entirely on sanctioning. Case adjudication centers around the actions and rights of the accused. Accountability, pursued in the name of either restributivist or consequentialist objectives, is enforced primarily through the severity of the sanctions issued.

Standard criminal case processing in the U.S. rarely addresses the underlying factors that contribute to one's propensity to criminally offend, or more generally, to have frequent interactions with law enforcement. For example, there is ample evidence of substantial overlap between individuals charged with and convicted of criminal offenses and individuals who have been victimized in the past (Sampson and Lauritsen, 1990; Berg et al., 2012). Moreover, the criminal justice system often fails to center the experiences and needs of crime victims. Criminal offenses are framed as crimes against the community or state, with the actual victim serving primarily as a witness. While convicted offenders can be ordered to pay some form of restitution and victims may seek redress through civil litigation beyond the criminal case, standard processing has very little restorative content from the victim's viewpoint beyond metaphorically making the accused pay for the harm they have caused.

2.2 Restorative Justice

Advocates of restorative justice argue for the direct involvement of the harmed and a fuller consideration of the factors contributing to the behavior of the person committing the harm. Such advocates argue that criminal cases can be adjudicated in a manner that builds greater empathy for the victim and appreciation of the consequences of one's actions while simultaneously improving the well-being and satisfaction of victims. As summarized by Umbreit and Armour (2011), adversarial criminal case processing focuses on establishing what laws have been broken, who broke them, and what punishment the person(s) responsible deserve(s). By contrast, restorative justice systems focus on establishing who has been harmed, what needs to occur to restore the welfare of the harmed, and who is obligated to make the restoration.

While restorative justice programming can take many forms, in criminal justice settings it

usually involves direct conferencing between the person responsible for the criminal offense and the person who has been harmed.[2] Restorative justice conferences can occur in addition to standard criminal processing (Angel et al., 2014) or in lieu of criminal processing (as in MIR). The process may involve a direct discussion between victim and perpetrator of a specific offense (McGarrell, 2001) or the conference may be between individuals who have committed and have been harmed by specific offenses but not involved in the same incident (Fulkerson, 2001). Conferencing requires the consent of both the harmed as well as the person who committed the offense, and typically requires that the offender acknowledge their culpability. To date, much experimentation has occurred in juvenile justice settings. These settings are seen as early opportunities for alternative case processing that may prevent or reduce adult criminal offending.[3] Moreover, while these methods have a longer history of application in juvenile settings and for less serious offenses, some countries have experimented with application to more serious criminal cases, including domestic violence (Strang et al., 2002).

A restorative justice conference usually begins with a statement by the person accused of the crime, followed by an opportunity for the harmed party to directly address the accused. Conferences also typically incorporate supporters of both the harmed and the accused, with everyone involved permitted to communicate openly about how the offense impacted their well-being. Supporters often include family members and friends; in juvenile justice settings the family support system almost always includes a parent or guardian. Conferences close with an agreement phase, whereby the victim and the person who committed the offense agree to a plan of action that the person will undertake to restore the welfare of the harmed. The actions can include writing formal letters of apology, paying restitution, agreeing to specific community service, and/or tailored actions of good faith.[4]

Restorative justice has its origins in a theory of shaming in criminal justice practice first articulated by Braithwaite (1989). Braithwaite posits two forms of shaming for those who commit harm. Stigmatic shaming permanently associates the criminal offense with the character of the individual who committed the offense and thus creates separation between the individual and society. By contrast, reintegrative shaming separates the offense from the individual, effectively condemning the action but offering the person a path back into the community through the acknowledgement

[2]Sherman and Strang (2007) discuss several alternative incarnations of restorative practices including direct conferencing, conferencing through third-party mediation, and the payment of restitution. The authors note the long history of restorative practices in response to criminal offense throughout the world, and the influence of the practice of indigenous cultures (e.g., the Maori in New Zealand, Native Americans in the United States) in informing conventional thinking about restorative justice.

[3]The U.K. has also experimented with adjunct restorative justice conferences in adult corrections (Shapland et al., 2008).

[4]The authors observed a presentation of the service provider for MIR, CW, to the San Francisco Sentencing Commission that highlighted a case where it was revealed during the conference that the youth who committed an offense was a talented muralist and the victim a lover of Disney characters. An element of the action plan from the conference was that the youth would paint a mural of the harmed party's favorite character in her apartment, a provision of the agreement on which the youth willfully and enthusiastically followed through.

of responsibility and efforts to make amends. While modern incarnations of restorative justice tend to emphasize restoration over shaming, most do require that the person who committed the offense take responsibility for their actions. Moreover, through direct communication during the conference, the person who committed the offense is typically confronted with the consequences of their actions not only for the victim, but often by the victim's family and friends, and even members of their own family. In theory, this process should not only engender greater empathy for the person harmed, but will also make the accused more mindful and aware of how their actions impact others. Beyond the impact on the person causing harm, many hypothesize that the ability to conference directly with the person causing harm and the efforts towards sincere restoration minimize the trauma suffered by victims and permit them to move on with their lives to a greater degree than would otherwise be possible.

2.3 The Empirical Evidence on the Efficacy of Restorative Justice Conferencing

While there have been only a handful of small-scale restorative justice programs in the U.S. juvenile justice system, restorative justice is a key component of juvenile justice in New Zealand (Ministry of Justice, 2004) as well as a standard juvenile diversion alternative used in Australia (Little et al., 2018) following a decade of experimentation (Strang et al., 2013). The U.K. has also experimented with restorative justice interventions, usually in addition to standard criminal case processing rather than as an alternative (Shapland et al., 2008).

The existing evaluation research focuses both on victim outcomes (usually measures of satisfaction, or post-incident indicators of trauma) as well as measures of recidivism among those who committed the offenses. The extant body of research generally finds that restorative justice programs are well received by victims and may help alleviate post-traumatic stress symptoms (McCold and Wachtel, 1998; McGarrell, 2001; Angel et al., 2014; Brooks, 2013; Sherman et al., 2015).[5]

In contrast to findings regarding the effects on victims, evidence on the impacts of restorative justice interventions on recidivism is less conclusive. In a relatively recent review, Livingstone et al. (2013) highlights there is "currently a lack of high-quality evidence regarding the effectiveness of restorative justice conferencing for young offenders" and more research is needed. The Australian experiments reviewed by Sherman et al. (2015) find some evidence of a reduction in repeat

[5] Fulkerson (2001) studies an offshoot of restorative justice conferences on domestic violence victims. The author investigate the effect of victim impact panels (VIPs) in domestic violence cases in Arkansas. VIPs permit crime victims to directly address offenders, usually as a group, in a safe and secure and environment that is not restricted by the formal procedures of a courtroom. Treatment subjects were asked to participate in a VIPs whereby a panel of victims were given the opportunity to address a panel of offenders, with the one restriction being that victims could not participate on a panel where the individual that victimized them was among the group of offenders being addressed. While the study reports that most participants believed that the panels where helpful (both among victims and offenders), no contrast is drawn with individuals assigned to the control group.

offending, especially for offenses that involved a victim. However, there is some weak evidence in sub-group analyses of increased offending, though the estimates are underpowered and appear to be exploratory. An important feature of these experiments is that randomization occurred only among offenders who agreed to participate, leading to a selected sample.

In the U.S., the evidence is mixed. First, in an observational study Brooks (2013) evaluates a restorative justice community conferencing program focused on juveniles in Baltimore, Maryland. Brooks uses a propensity score matching design to compare youth who participated in the program to observably similar individuals who did not. Interestingly, Brooks finds higher recidivism rates among the participants of the restorative justice program.

Second, the results of two RCTs from the 1990s are inconclusive and find no persistent reductions in recidivism. McGarrell (2001) and McGarrell and Hipple (2007) evaluate an RCT of a restorative justice program in Indiana focused on children (the average age is 12.5). They find meaningful decreases in recidivism within one year relative to a control group allocated to another diversion program. However, overtime the effects disappear and no impacts on recidivism are detectable in the longer-run (Jeong et al., 2012). McCold and Wachtel (1998) evaluate a family-group conference intervention in Pennsylvania and find no evidence of reductions in recidivism after one year. However, the take-up among those assigned to treatment was low (roughly 40 percent).

The two experimental studies by McGarrell (2001) and McCold and Wachtel (1998) are more related to our evaluation of MIR than the non-U.S. studies reviewed in Sherman et al. (2015). In these two RCTs, randomization was done without conditioning on the youth's willingness to participate (similar to MIR and unlike the non-U.S. studies), and they involve only juveniles who have not been convicted. However, they also differ from MIR in key aspects. In McCold and Wachtel (1998), the conferencing was conducted by *police officers* and involved only youth charged with infractions and misdemeanors (no felonies). In McGarrell (2001), youth were especially young and generally engaged in less severe offenses. Another key difference is that MIR has an agreement monitor that supports the youth in completing the agreement post-conferencing. The agreement monitor joins the restorative justice process at the conference and then meets with the youth regularly (at least once a week) post-conferencing to make sure they are on track to fulfill the agreement signed during the conference.

Our evaluation of the MIR program contributes to this literature in several ways. First, it provides credible evidence that restorative justice conferencing can cause lasting reductions in recidivism. Importantly, there are no observed or unobserved confounders to the intervention in our setting since assignment to the treatment and control groups was done at random. Moreover, the take-up was high; 81 percent of those assigned to MIR enrolled in the program, and 67 percent completed the program among those who enrolled. These rates are especially high given that the accused's consent to participate in MIR was not an eligibility requirement for inclusion in the randomization. Second, unlike other restorative justice experiments where youth assigned to the control group were funneled into various alternative diversion programs, control group members

under MIR faced traditional felony prosecution. Third, it highlights the potential importance of having an agreement monitor to check-in and support the youth after the restorative justice conference is over, and the onus of completing the agreement is on the youth.

A key difference between MIR and most non-U.S. experiments is that the accused is not required to consent to participate in the RCT—that is, consent is not an eligibility requirement for inclusion in the randomization. Therefore, the ITT estimates take into account any non-participation due to a lack of willingness on the accused's side. Moreover, effects for those who participated in MIR can also be recovered using 2SLS, as discussed below.[6]

Thus, our ITT estimates capture exactly the policy-relevant estimand of comparing assignment to a restorative justice intervention relative to regular felony prosecution. Lastly, we find economically meaningful and statistically significant estimates that MIR reduces recidivism both in the short-run and in the long run after four years from randomization. These findings are informative to the current debate on whether restorative justice conferencing can reduce recidivism or not.

3 Setting and Data

We begin by describing our data, the MIR program, and the process by which juveniles were randomized to either the MIR treatment group or a business-as-usual control group. We then provide summary statistics on the MIR experimental sample and compare it to the full population of juveniles charged with a felony offense in San Francisco during the same time period.

3.1 Data Sources

Our evaluation draws upon three data sources, all pertaining to criminal offenses occurring within San Francisco. First, we were provided programmatic information on all youth who were part of the MIR experiment. Programmatic data include information on whether the eligible youth was assigned to MIR or the control group and information on MIR program participation. For those who enrolled in MIR, we were provided information on key dates, including enrollment date, conference date, date of completion, and for those who did not complete the program, the date the case was sent back to the office of the San Francisco District Attorney (SFDA). Second, we were provided data on the universe of juvenile arrests for offenses occurring in San Francisco between October 2010 and November 2020. The data includes a description of the offense (by penal code), whether charges have been filed, demographics such as sex, race, date-of-birth, name, and age at the first criminal incident.

[6] In MIR, the consent of the victim was a requirement for eligibility to be randomized. However, in conversations with the juvenile prosecutor in charge, we have been told that *all* the victims consented when asked if the case can be eligible for MIR. While all victims consented for the juvenile to participate in the MIR experiment, not all agreed to take part in the restorative justice conferencing themselves. In cases in which the victim is not able to participate, surrogate actors take their role.

These records allow us to construct measures of criminal history for the youth participating in the RCT. Moreover, we report summary statistics for all the juveniles for whom the prosecutor decided to file felony charges (similar to our experimental sample) during the experiment's time period. Third, we were provided with administrative records on all adults arrested and presented for a charging decision in San Francisco between October 2010 and November 2020. Given that some of the youth in the experiment turn 18 shortly after the arrest for the offense associated with their inclusion in the experiment, the adult arrests records are essential to construct accurate recidivism outcomes.

3.2 The Make-it-Right Program

The SFDA piloted the MIR program at the end of 2013. The program diverts youth who have been arrested for certain felony offenses to a restorative justice conferencing program. Conferencing involves facilitated community-based conversation between the involved minor, their family, the person harmed, and a community representative, leading to an agreed-upon plan for addressing that harm. Eligible youth are randomized to either the treatment or control groups after the juvenile prosecutor reaches the decision to file charges but before charges are formally filed. Thus, all the individuals in the control group will face criminal charges, and youth assigned to MIR but who did not complete the program will also automatically face criminal charges. The SFDA does not file criminal charges (or any other sanction) against youth who successfully complete MIR.

Eligibility criteria. The program is targeted at juveniles 13 to 17 years of age charged with medium-severity felony offenses such as the unlawful taking of a vehicle, grand theft, burglary, or assault. The sample of eligible youth, therefore, includes only those for which the prosecutor decided to file felony charges. To be referred to the program, the youth must be a resident of San Francisco County or Northern Alameda County, and must have no prior 707(b) arrests or sustained petitions.[7] In addition, the youth must not have caused moderate or significant injuries to the victim, cannot be affiliated with a gang, cannot have used a weapon, and the minor cannot be currently under probation supervision or in detention. Thus, by construction, youth randomized into MIR will not be a part of any criminal justice proceedings while in the program.[8]

MIR was intended to be a relatively small pilot: SFDA expected to enroll no more than 25 individuals per year. In practice, the number of eligible youth and enrollees was lower than expected. One reason for this is the steady reduction in juvenile crime in San Francisco, which affected the overall volume of cases. In total, the experiment lasted 5.5 years and included 99

[7] 707(b) offenses are those that would count as a strike under California's three-strikes for juveniles 16 and over, for example, violent felonies such as kidnapping or robbery.

[8] These requirements ensure that an eligible youth does not have any other pending cases; otherwise, they would be either on probation supervision or in detention. Moreover, if the youth is rearrested for a new offense, then either their participation in MIR will be stopped, or the new offense will be merged with the original one for which the youth was assigned to MIR.

participants in the restorative justice treatment group and 44 participants in the control group when it concluded in May 2019, constituting 13 percent of the juveniles charged with a felony during that period.

Treatment Assignment Process Timeline. First, the prosecutor is presented with an arrest and decides whether to proceed with a prosecution (and on what charges) or not. Second, among cases in which the prosecutor decided to file felony charges, eligibility for the MIR program is assessed (see details above). Third, among eligible cases, allocation to MIR or the control group is done at random (see details below). Fourth, individuals assigned to MIR are contacted by CW staff who describes the program to them (and their families).

Randomization Procedure. The SFDA's Office designed and led the implementation of the MIR RCT; the research team was not involved in devising nor implementing/monitoring the randomization process. The SFDA conducted the randomization of individuals to the treatment and control groups at the case level, which corresponds to individuals except for cases involving co-defendants. The randomization process was designed to separate the actual treatment assignment from the prosecutor in charge of the program. Specifically, once a case was deemed eligible, it was sent to a paralegal not involved with the program and unrelated to the juvenile prosecutor, who consulted a prepared list of assignments, selected the next available assignment, and then communicated the results back to the prosecutor. We confirmed with the juvenile prosecutor that she never overrode the allocation communicated to her by the paralegal.[9]

The MIR program. Table 1 provides a concise description of the MIR program, its activities, and each of its steps. Once assigned to treatment, CW, a nonprofit organization located in Oakland, California, specializing in restorative justice, assesses the youth's ability to participate. Unsuitable cases are referred back to the SFDA for felony prosecution. Importantly, no sanctions are imposed on youth who do not enroll in MIR, and the reason for unsuitability is not disclosed to the SFDA. An essential requirement for participation is demonstrating reflection and accountability for one's actions. Minors and their parents may decline to participate, effectively opting for the case to be referred back to the SFDA for prosecution. For cases that proceed to conferencing, CW conducts the pre-conference planning involving the youth (referred to as the responsible party), the victim (harmed party), and any other individuals who will take part in the conferencing (such as parents and/or supporters of the harmed and responsible parties). Moreover, CW also mediates the conference. Appendix Table A.1 illustrates how a restorative justice conference looks. Post-conference case management and compliance monitoring is managed by the Huckleberry Youth's

[9]Initially, the SFDA randomly assigned 50 percent of eligible individuals to MIR. However, shortly after commencing the experiment, the SFDA discovered that the number of youth eligible for MIR was lower than initial estimates as overall juvenile crime and incarceration in the county was decreasing. The assignment probability to treatment was thus altered to 70 percent in May 2014 (six months after the pilot date). As we discuss below, including cohort fixed effects yields almost identical estimates. Moreover, removing the observations from the earlier periods all together also yields similar results.

Community Assessment and Resource Center. Youth who fail to follow through with the program have their cases referred back to SFDA for prosecution.

Juveniles in the control group continue through the traditional juvenile justice process. They are charged and prosecuted in juvenile court, with the minor being supervised by the JPD during the process. While some minors are detained, most are supervised in the community using electronic monitoring or day reporting centers.

Figure 1 depicts the flow of cases through the treatment and control arms. In total, 143 participated in the experiment, with 99 (69.2 percent of study subjects) assigned to the treatment group and 44 assigned to the control group (30.8 percent). Youth assigned to MIR either enroll in the program or are deemed unsuitable (e.g., when refusing to assume responsibility for their share in the incident or unable to participate due parental refusal). The take-up rate was especially high, with 80.8 percent of those assigned to MIR enrolling in the program. This contrasts with earlier restorative juvenile justice experiments in the U.S. that we review above. The higher take-up rate may reflect the fact the MIR enrolls youth charged with relatively serious offenses (all felonies), unlike the eligibility of lower-level charges in prior experiments. Moreover, the alternative to participating in conferencing is not another diversion program but felony prosecution and potentially severe sanctions. The process of enrolling into MIR was relatively quick. On average, juveniles waited 15 days, and the median waiting time was 21 days.

Among those enrolled in MIR, 66.7 percent completed the program, and as a result, formal charges were never filed against them. The average and the median duration of the program (time between enrollment and completion) were six months. Overall, 52.5 percent of those assigned to MIR completed the program. Thus, while most youth assigned to MIR completed the program, a non-negligible portion did not. Multiple reasons can lead to this result. For example, the accused and the victim did not reach an agreement during conferencing, or the youth did not fulfill the agreed contract with the victim.

In what follows, we present the final evaluation of the MIR program that makes use of the full sample of juveniles who took part in the experiment.[10]

3.3 Summary Statistics

Table 2 presents summary statistics for the different groupings of juveniles randomized under the MIR experiment and for all juveniles charged with a felony offense between October 2013 and May 2019. The first and second columns present average characteristics for youth assigned to the control and treatment groups. The third column reports the characteristics of those assigned to

[10]In an early analysis of the program we study, Huntington et al. (2017) provide a preliminary outcome evaluation of the MIR program. This earlier phase of the intervention was severely under-powered, and the analysis focused largely on reviewing comparable experiments, thinking through the decision of whether to continue the experiment, and offered very preliminary outcome comparisons. Our analysis was completed independently of this effort, based on separately procured data extracts, and a completely independent process of data pre-processing and analysis.

MIR and who also enrolled in the program. The final column describes all the juveniles charged with a felony offense during the experiment (most of which were not eligible for MIR).

Comparisons of the averages for the control (Column (1)) and treatment (Column (2)) groups reveal that random assignment yielded balance on observable characteristics. The square brackets ([·]) in Column (2) report p-values for the null hypothesis that the averages in Columns (1) and (2) are equal. While there are a few differences, an F-test of the overall significance of a regression of a treatment dummy on the covariate list yields a p-value of 0.789. Moreover, except for one covariate, all the individual p-values are not significant at the five percent level, and only three are significant at the ten percent level. Roughly 90 percent of youth in the study are male, approximately half are Black, and one-third are Hispanic. Average age at arrest is 16, though age at first arrest is lower for both groups (14.75 for the control group and 15.2 for the treatment group). Around one-fourth of the treatment and control groups have a prior arrest in San Francisco. However, prior arrests for felony offenses are less common (of those with a prior arrest, 13.6 percent among the control group and 33.3 percent in the treatment group have a prior felony arrest). Moreover, the average number of prior arrests is low among both the control and treatment groups, 0.18 and 0.36, respectively. Importantly adjusting for any difference in covariates does not change any of our estimates as we discuss in Section 5.3.[11]

Next, we describe the type of offenses youth in MIR are typically facing. The most common charge is felony theft (64 percent of the control group and 66 percent of the treatment group), followed by burglary (approximately 40 percent) and felony assault (roughly 14 percent).[12]

The final four rows present means for a predicted recidivism (defined as a subsequent arrest in San Francisco). To generate these predicted values, we use auxiliary data on the entire sample of youth charged with a felony offense between October 2013 and May 2019, exclusive of youth who participated in the RCT. We estimate an OLS model where the dependent variable is an indicator for a rearrest within t months of the focal incident. The explanatory variables include all of the covariates presented in Table 2. We then use the estimated parameters from this model to generate predicted recidivism probabilities for each youth in the MIR treatment and control groups. Appendix B describes in more detail the construction of the predicted recidivism variables. Roughly 36 percent of the treatment and control groups are predicted to be arrested within six months, with predicted felony recidivism rates of about 17 percent for both groups.

Comparison of the means in Columns (2) and (3) suggest that youth assigned to the MIR program and who also enrolled are indistinguishable from youth who do not take up the treatment (as can be seen from the p-value of the joint F-test presented at the bottom of Table 2). Moreover, as summary measures, the predicted recidivism probabilities are almost identical in Columns (2) and (3).

[11]Note that any observable imbalances are not systematic; for example, individuals in the control group are less likely to have a prior felony arrest; however, they have more prior arrests on average.

[12]The charge proportion sum to greater than one since a given case is often associated with more than a single eligible charge.

Relative to the broader population of juveniles charged with felony offenses, described in Column (4), the MIR youth are more likely to be male, somewhat less likely to be Black or Hispanic, are less likely to have a prior arrest, and have a first arrest occurring at a slightly older age (approximately 15 vs. 14.5). There are notable differences in the charge distribution, reflecting the MIR eligibility criteria. Specifically, over a third of non-MIR juveniles are arrested for severe person offenses (2 percent for homicide, 1.4 percent for a sex offense, and 38.7 percent for robbery) compared with roughly two percent of MIR youth arrested for robbery. We also observe a higher proportion of all JPD referrals involving a weapons offense relative to MIR youth.

To give more context on our setting and the individuals who interact during the restorative justice conferencing, Appendix Table A.1 presents the demographic characteristics of the victim (harmed party) and compares them to those of the accused (responsible party).[13] A few patterns are worth noting. First, the average age of the victim is 35, roughly double the average age of the average accused youth. Second, the racial composition of the two groups is meaningfully different, and in 79 percent of the incidents, the harmed party is of another race/ethnicity than the responsible party. While most of the youth are Black and Hispanic, most of the individuals in the harmed group are White and Asian. Females are also overly represented in the harmed (victim) group, 41 relative to 11 percent among the responsible youth.

4 Empirical Strategy

Our empirical approach is straightforward given that individuals were randomized between assignment to MIR (treatment) or business-as-usual prosecution (control). Comparisons of mean outcomes among units assigned to treatment and control regimes are sufficient to identify the ITT effect of being assigned to MIR relative to felony prosecution.

We begin by estimating the ITT effect using Equation (1) which describes the relationship between assignment to MIR and any rearrest:

$$Y_{it} = \gamma_0 + \gamma_1(\text{Assigned MIR})_i + \epsilon_{it} \tag{1}$$

where $Y_{it} \in \{0, 1\}$ indicates whether individual i rearrested within t months from the date of randomization. The individuals in our sample have been randomized on different dates, and hence we observe fewer individuals for longer time horizons. To exploit all available information, we complement Equation (1) with Kaplan-Meier estimates of the failure function (defining a failure as a new arrest).

In addition to the ITT analyses, we also estimate the TOT effect of participation in MIR, by using the random assignment to MIR as an instrumental variable for whether or not a youth

[13]The information on the demographic characteristics of the victims in the MIR experimental sample (although partial) was provided to us by CW and was collected as part of the restorative justice conferencing process.

enrolled into the program. Equations (2) and (3) describe our 2SLS estimator:

$$Y_{it} = \beta_0 + \beta_1(\text{Enrolled MIR})_i + \eta_{it} \tag{2}$$

$$(\text{Enrolled MIR})_i = \alpha + \alpha_1(\text{Assigned MIR})_i + \xi_i \tag{3}$$

The 2SLS estimator from Equations (2) and (3) identifies the TOT under the LATE framework assumptions (Imbens and Angrist, 1994; Angrist et al., 1996), as there is only one-sided non-compliance in our setting (Bloom, 1984). In our setting, these assumptions are likely to hold. Monotonicity must hold since all the individuals in the control group did not participate in MIR. Exclusion is also likely to hold as assignment to MIR is unlikely to impact recidivism except through participation in MIR. If a youth does not complete MIR, the prosecutor does not have any information about the reason. CW emphasizes that they do not share any information with the prosecutor that can potentially be used against the youth, including the circumstances that lead a juvenile not to participate or complete MIR. Moreover, enrollment or completion of MIR has no bearing on decisions in future prosecutions. The prosecutor does not take into account whether a youth completed MIR or not when making future charging decisions. Lastly, to quantify the magnitude of the estimated ITT and TOT effects, we compare them to the control group complier mean (Katz et al., 2001).

Throughout our analysis, we report two types of p-values. The first is based on cluster-robust standard errors, clustered at the case level. The second p-value is based on randomization inferences using random permutations of cases to placebo MIR and control regimes to generate the sampling distribution under the null hypothesis (as was suggested by Young, 2019). This procedure is known as randomization, Fisherian, or permutation inference and was first proposed by Fisher (1935). As was advocated by Chung and Romano (2013), we use as our test statistic the standardized t-statistic.[14]

Finally, in our pre-analysis plan we specified that we will evaluate only one-sided hypothesis tests pertaining to whether the MIR program reduced the likelihood of recidivism or not. This choice was aimed to maximize statistical power given our small sample size.[15] Pre-analysis plans have been mentioned in the literature on research transparency and reproducibility as a tool to

[14]Note that randomization inference yields valid p-values for two null hypotheses. First, it provides finite-sample exact inference on the sharp null of no treatment effect on all individuals. Second, when the sharp null hypothesis is incorrect (e.g., some units are impacted positively while others negatively), randomization inference still provides asymptotically valid inference on the null hypothesis of no average treatment effect, also known as the Neyman null hypothesis. Thus, regardless of whether the sharp null hypothesis is correct or not, randomization inference provides valid asymptotic inference on the null of no average treatment effect. For the latter interpretation to hold, a standardized statistic (e.g., t-statistic) must be used as the test statistic in the randomization inference procedure. The second interpretation of randomization inference tests was formalized only recently by Li and Ding (2017), Wu and Ding (2020), and Zhao and Ding (2021).

[15]We submitted the pre-analysis plan before looking at the data and without knowing the exact sample size. However, we knew its general range.

increase statistical power by pre-specifying one-sided hypothesis (e.g., Olken, 2015).[16] Given our pre-specification choice of focusing on one-sided hypothesis tests, we are unable in this study to test whether the program increases the likelihood of an arrest among participants. However, this is unlikely since the signs on all our estimated effects point towards reductions in recidivism. Although we find no evidence that MIR increases recidivism, this finding is not universal. As we noted above, in some instances, prior experimental results from Australia (Sherman et al., 2015) revealed situations where restorative justice programs increased offending among some individuals.

5 Main Results

We begin with ITT estimates by visualizing the recidivism patterns that we observe in the data among individuals who have been randomly assigned to MIR and those assigned to the control regime. Figure 2 presents Kaplan-Meier estimates of the failure functions depicting the relationship between the probability of being rearrested at least once and the number of days since randomization took place over a four-year period.[17] For both groups, the likelihood of rearrest is quite high, with nearly half of the control group arrested within six months of randomization and over 70 percent rearrested by the end of the four-year period. However, we observe markedly lower rearrest rates among youth assigned to the treatment group. The difference in the percentage ever arrested on a new offense reaches roughly 20 percentage points within six months. It then fluctuates around this level for the remainder of the observation period (though we should note, the number of observable individuals for the latter periods narrows to roughly half of our sample).

Next, we examine the effects of MIR on measures of recidivism that focus on more severe interactions. Appendix Figure A.2 shows similar results when measuring recidivism as future felony arrests, future arrests for offenses that are at least as severe as the original charges for which the youth was arrested, or future arrests that result in a conviction. These findings show that MIR causes reductions in recidivism across a variety of different measures, reducing both the likelihood of any future arrests as well as the likelihood of arrests for relatively serious offenses.

We perform formal hypothesis tests for equality of the two cumulative failure functions using the standard non-parametric Peto-Peto-Prentice test (see Klein and Moeschberger (2006) for a textbook description) and calculate p-values in two ways. First, using a non-parametric inference

[16] Another example is Christensen and Miguel (2018) who advocated the use of pre-analysis plans and mentioned as one of their advantages the fact that they can allow researchers to specify their interest in one-sided hypothesis in advance and, by doing so, increase the accuracy of statistical tests, "PAP [Pre-Analysis-Plans] bind the hands of researchers and greatly limit specification searching, allowing them to take full advantage of the power of their statistical tests (even making one-sided tests reasonable)."

[17] Note, the last randomization occurred in October 2019, and we observe recidivism data through (including) November 2020. Hence, we have at least 14 months post-randomization for all youth, but obviously longer periods for youth randomized into the program in the early years. Figure 2 presents Kaplan-Meier estimates of the failure function which are based on one minus the product of period-specific survival probabilities (through a given time interval), implicitly assuming that the recidivism hazard function is stable across cohorts defined by randomization date. We examine and discuss in more detail differences in recidivism across cohorts in Section 5.3.3.

procedure that is based on asymptotic approximations (as is standard practice). Second, we use randomization inference and calculate the finite sample distribution of the test statistic from 1,000 random placebo permutations of the treatment assignment. The top of Figure 2 reports the one-tailed p-value from these tests. Both tests reject the null hypothesis of equal failure functions, with the p-value from the randomization inference ($p = 0.017$) roughly 2.5 times the value based on the standard non-parametric asymptotic inference ($p = 0.0071$).[18]

Table 3 presents our principal results. Each column presents model estimates where the dependent variable is an indicator variable equal to one if the youth is rearrested within the time period indicated by the column heading (e.g., within 6 months, within 12 months and so on). Panel (a) presents our 2SLS estimates of the TOT effect of participating in MIR on subsequent rearrests. Along with each estimate, we report standard errors clustered at the case level (in parentheses), the p-value from a one-tailed test of the null hypothesis of a non-negative impact on recidivism against the alternative of a negative (i.e., crime reducing) impact based on the standard error (in curly brackets), and the p-value from a comparable one-tailed test using randomization inference with 1,000 placebo permutations of the treatment assignment (in square brackets). Panel (b) presents the ITT effects along with the comparable standard error and p-value calculations to those presented for the TOT estimates. Panel (c) presents the first stage results for the effect of assignment to MIR on participation. Given that there is no crossing over from control to treatment in the data, these first stage effect estimates are the proportion of youth who are assigned to MIR and who comply and enroll in the program. Finally, at the bottom of the table we present averages of the dependent variable outcome for the control group and for compliers under the control regime.

Before describing the effect sizes, we should note the high rates of rearrest observed among the control group. For example, 43 percent are rearrested within six months of random assignment, 63 percent are rearrested within two years, and 75 percent are rearrested within four years. The patterns in these average control outcomes highlight the fact that MIR impacts rearrest rates for youth who are at a very high risk of future contact with the criminal justice system.

A second pattern to note is the high take-up rate. While the take-up rate varies slightly across the time periods analyzed, it never falls below 73 percent (Panel (c) of Table 3).[19] Again, this stands in contrast to the juvenile interventions reviewed above where take-up rates are considerably lower and we speculate that the relatively high take-up rate for MIR is likely driven by the fact that the alternative is a felony prosecution rather than a diversion program.

The reduced form effects (i.e., the ITT effect estimates) indicate that assignment to MIR reduces the likelihood of rearrest by 18.9 percentage points within the first six months, 18.4 percent-

[18] One reason for the fact that randomization inference yields larger p-values relative to the standard inference formulas is that the Peto-Peto-Prentice test p-value does not take into account clustering at the case level in the assignment to treatment. As we show in Table 3, once clustering is taken into account, randomization inference and regular cluster-robust standard errors yield similar p-values.

[19] Note, differences in sample size across models are the result of differences in the post-randomization time horizon in which the youth is observed based on their date of randomization.

age point within the first year, and 14.4 percentage points within the first two years. Relative to the control complier means, these effect sizes imply a 44 percent, 33 percent and 24 percent reduction in recidivism, respectively. The overall effect sizes hold up if we measure rearrest within three years of randomization (14.7 percentage points equivalent to 20 percent of the control-complier mean) and widens for four years (26.7 percentage points equivalent to 30 percent of the control-complier mean).

The final column models recidivism during the period 12 to 48 months post-randomization. This additional outcome allows us to assess whether referral to MIR has longer term effects on behavior beyond the immediate reduction in the likelihood of new arrest recidivism following the conference. Within 12 months, 99 percent of those assigned to MIR completed or failed the program. The average and median duration of MIR is 189 days. The estimate in Column (6) indicates that assignment to MIR reduces recidivism between years one and four post-randomization by 27 percentage points (equivalent to 37 percent of the control-complier mean). Regarding inference, the p-values from the one-tailed tests indicate significance at the five percent (or less) level of confidence for the six-month, 12-month, and 48-month outcomes, as well as for the outcome measuring recidivism one to four years after randomization. The p-values based on the randomization tests are generally larger, though we still observe significance at the five percent level for the six and 12-month outcomes and significance at the ten percent level for the 48-month outcome.

Turning to 2SLS estimates, the TOT effects are generally 1.3 to 1.4 times larger than the ITT effects. Relative to the control-complier means, the TOT estimates range from 30 percent of the recidivism rate for the 24-month outcome to 54 percent of recidivism occurring within six months and 51 percent of recidivism occurring in years one through four. The p-values of the TOT estimates are significant at the five percent level for all of the outcomes except the 24-month outcome (with a p-value of 0.0759) and the 36 month outcome (with a p-value of 0.0979). The p-values from the randomization tests align with the regular p-values.

Finally, we discuss whether non-compliance (being assigned to MIR and not enrolling) is related to recidivism propensities. Appendix Figure A.3 presents the cumulative failure functions for youth assigned to the control group and those assigned to the treatment group but did not enroll in MIR (the "never-takers"). The two curves are similar to each other, and a test for equality of the two curves fails to reject the null hypothesis that the two are equivalent ($p = 0.9485$). Moreover, the average rearrest rates at the bottom of Table 3 show that for all the outcomes, the control group members and control compliers are quite close to one another, suggesting that non-compliance is unrelated to recidivism propensities.

5.1 The Effect of Enrollment vs. Completion of MIR

Our focus so far was on the ITT effect estimates of assignment to MIR and 2SLS estimates that re-scale the ITT by the share of individuals who enrolled into the program (Equations (2) and

(3)). This section discusses an alternative 2SLS specification that re-scales the ITT based on the share of youths who completed the MIR program rather than the share who enrolled in it.

According to the principles of restorative justice, meeting with the harmed individual and completing the contract agreed with them can be transformative for the youth. A key question is whether completion of the program is needed for it to have an effect. The summary statistics in Figure 1 suggest that the ITT recidivism reducing effect of assignment to MIR is driven by the individuals who completed the program. Specifically, youth assigned to MIR but who did not enroll have rearrest rates within one year of 57.9 percent. This is similar to the 57.7 percent rate of those who enrolled but did not complete the program. These rates are also similar to the 56.8 percent rearrest rate among the control units. However, among youth who completed MIR, the rearrest rate is much lower (19.2 percent). These summary statistics suggest that it is plausible that program completion is particularly important.

Table 4 reports 2SLS estimates of the TOT using an indicator for completing MIR as the treatment of interest. The TOT effects are larger than when using enrollment since the first stage (i.e., effect on completion) is smaller. Columns (1) and (2) report that completion of MIR reduces the likelihood of a rearrest within six and 12 months by about 36 percentage points. Relative to the control-complier means, these are reductions of 76 and 65 percent. The impacts on rearrests along longer time horizons are also large and range from 38 to 56 percent.

What factors lead individuals who enrolled to complete the MIR program? Appendix Table A.2 compares the characteristics of individuals who completed the program relative to those who did not (conditional on enrolling in MIR). A few patterns are noteworthy. First, in terms of the youths' demographic characteristics, sex and age do not predict completion. However, Black youth are less likely to complete the program than non-Black (mainly Hispanic) youth. Specifically, individuals who completed the program are 35 percentage points less likely to be Black. Second, there is suggestive evidence that individuals who are less likely to recidivate are completing the program with higher likelihoods. They have lower predicted recidivism, fewer prior arrests, and were older at their first arrest. Third, while youth who completed the program are generally less likely to recidivate based on observables, they can still be individuals charged with serious offenses. For example, youth arrested for an assault offense (the more severe type of crime in our sample) are more likely to complete the program than not.[20]

5.2 External Validity of the Estimated Effects

As discussed above, youth eligible to participate in the RCT are a non-random sample (e.g., individuals affiliated with a gang are not eligible). However, one measure of external validity, arguably the most important one, is recidivism rates. Figure 3 presents comparisons of the rearrest

[20]Individuals who participate in a restorative justice conference almost always complete the program. Only three individuals out of the 28 who did not complete the program took part in a conference.

rates of the MIR control group against the broader population of juveniles in San Francisco charged with a felony offense. Interestingly, the empirical failure function of all juveniles accused of a felony is similar to that of the control group, and we cannot reject the null hypothesis that the two curves are equal.[21]

A potential explanation for the findings in Figure 3 is that some of the common types of offenses eligible for MIR have especially high rearrest rates. Appendix Figure A.4 reports Kaplan-Meier estimates of the likelihood of a future arrest by the type of offense a juvenile is accused of committing. Burglary defendants have the highest rearrest rate for both any rearrest and felony rearrests. Moreover, among rubbery defendants, which is the most common offense category that is usually not eligible for MIR, the rearrest rates are similar to those of defendants charged with theft and assault. Thus, juveniles in the experimental sample are more commonly charged with Burglary (40 percent), the offense type with the highest rearrest rate, relative to the general population (14 percent).

Our results suggest that the effects estimates presented in the previous sections are likely predictive of the impacts of MIR in the broader population of youth who are typically charged with felonies in San Francisco.

5.3 Robustness Analyses

5.3.1 Statistical Power

The MIR experiment includes 143 individuals, which is a relatively small sample size. In this section, we discuss the concerns in an experiment that is potentially underpowered and their implications to our setting.

The key concern in an underpowered experiment is concluding that the treatment had no effect while in fact the statistical tests lacked sufficient power to detect small treatment effects due to insufficient sample size. However, in our case, the treatment effects are large enough to reject the null hypothesis of no treatment effect (or recidivism increasing impacts). Moreover, we see consistent patterns when examining a variety of different measures of reoffending, such as any new arrest, new arrests that lead to a conviction, new arrests for felony offenses, or new arrests for offenses of equal or higher severity.

Gelman and Carlin (2014) mention two other potential concerns. The first is a sign error, estimating that the treatment has a positive effect while the true effect is negative. In our case, this will mean concluding that MIR reduces recidivism when it actually increases it. A sign error is improbable in our setting. Following the procedure proposed by Gelman and Carlin (2014), we estimate the probability of a sign error in the effects of enrollment to MIR to be 0.00002 and

[21] While the two groups have similar rearrest likelihoods, the experimental sample has a lower predicted likelihood of recidivism (see Table 2). These findings indicate that the MIR experimental sample might be negatively selected relative to the general population along certain dimensions.

0.00003 for rearrests within one and four years, for example.

The second potential error is in exaggerating the magnitude of the treatment effect. Reassuringly, we estimate that the scope of this possible error is also limited. Again, following the procedure proposed by Gelman and Carlin (2014), we estimate an average potential exaggeration ratio of 1.2 in the effect of enrollment to MIR on rearrests within one and four years. In other words, on average, our estimates might indicate that the impact of enrollment to MIR causes a reduction of 23.4 percentage points while the true effect is a reduction of 19.5 percentage points. Thus, although small, our sample size and estimated effects are sufficient to draw firm conclusions on the effectiveness of MIR and rule out potential inference errors.

5.3.2 Covariate Adjustment

To test the robustness of our difference-in-means comparisons to any finite sample imbalances in covariates, we also present ITT and TOT effects estimates that adjust for any imbalances in the predicted likelihood of a future arrest. To limit researcher degrees of freedom in deciding how to adjust for covariates (e.g., which controls to includes in the model or not), we pre-specified our procedure for conducting covariate adjustment with an eye on parsimony. The challenge is that including all the covariates in the regression model can meaningfully reduce the degrees of freedom since there can be many relevant variables.[22]In this approach, we adjust only for a single summary index that can be viewed as a dimension reduction of the relevant information from all the covariates into a single factor. Specifically, we calculate the predicted likelihood of a future arrest using data on all the youth charged with a felony offenses who did not participate in the experiment (i.e., Column (4) of Table 2). We use this auxiliary data to estimate a simple OLS model where the dependent variable is an indicator of a future arrest and the controls include sex, race/ethnicity, number of past arrests, number of past felony arrests, fixed effects for the category of the referral offense, age at arrest, and age at first arrest. We then use the estimated model coefficients (estimated only using the auxiliary data) and the covariates of study participants to generate the predicted likelihood of a future arrest and add this summary index to our base specification. Appendix Figure B.1 shows that the predicted future arrest index is highly correlated with the likelihood of a new arrest observed in the experimental sample, and we cannot reject a coefficient of one. In Appendix B, we present a more detailed description of this covariate adjustment procedure.[23]

[22]For example, controlling for the broad category of the current offense, age and other demographics, number of past offenses and the types of past offenses can result in a model with more than 30 covariates. Adding 30-plus covariates to the model in Equation (1) will have a meaningful impact on the number of degrees of freedom. To avoid needing to choose which covariates to include in the model, we opted for the option of controlling for a single index that summarizes the predictive ability of all the covariates as we describe in detail in Appendix B.

[23]This is the procedure used to calculate the predicted likelihood of new arrest averages presented at the bottom of Table 2 and Appendix Table A.2. The idea of using auxiliary observational data to improve the accuracy of experimental estimates has been proposed in other studies (e.g., Gagnon-Bartsch et al., 2020).

Appendix Table A.3 reports the results. The table is structured similarly to Table 3 but also includes the coefficient on the predicted recidivism index, which is a weighted average of the pre-treatment covariates. The weights are determined by the degree to which each covariate is predictive of the outcome in the auxiliary observational data on the full population of juveniles charged with a felony offense in San Francisco. The results are very close to those from the models omitting the predicted recidivism index control variable. Thus, any finite sample imbalances in covariates do not impact our treatment effect estimates.

5.3.3 Differences in Effects Across Cohorts

The analyses above are based on individuals who have been assigned to MIR between October 2013 and May 2019. Our long-run (i.e., four-year) effects on recidivism are estimated using the earlier cohorts for which we have a longer time horizon to measure rearrests. While there is variation in the time horizon that we observe a youth for post-randomization, for all the youth in our sample, we observe recidivism within at least one year from randomization. Next, we focus on rearrests occurring within six months or one year, for which we have a balanced sample, and examine differences across cohorts. Appendix Table A.4 reports 2SLS estimates with and without cohort fixed effects. The estimated effects are almost identical with and without the cohort fixed effects. Moreover, the first stage coefficient also does not change by the inclusion of the cohort fixed effects. Columns (3) and (6) examine an even richer specification that allows the effect of MIR to vary by cohort. The estimates are the largest for the 2016-2017 cohort. To formally test for cross-cohort differences, we conduct a joint F-test. We cannot reject the null hypothesis that the program's impacts are the same in all three cohorts ($p = 0.55$ and $p = 0.7$ for rearrests within six months and one year). Thus, in our balanced sample, the effects of MIR are similar across cohorts.

Lastly, in Appendix Table A.4, the coefficients on the cohort indicators are negative, indicating that, on average, the rearrest rate is lower in the more recent cohorts. In Appendix Figure A.5, we compare the rearrest rates of youth assigned to MIR in different time horizons (cohorts) and the control group. The figure shows that the later cohorts assigned to MIR generally have lower rearrest rates than the earlier cohorts. Thus, our findings suggest that the long-run effects would not be smaller if we observed four-year rearrest rates for all of our samples and not only among the earlier cohorts.

5.3.4 Robustness to Including or Not Including Arrests Due to Probation Violations

Our main measure of recidivism includes all rearrests including those that are the result of probation violations. As MIR can impact the case processing outcomes, using measures that include technical probation violation might classify similar behavior by the youth as recidivism depending on whether the individual is on probation. Moreover, whether one is on probation is impacted by

MIR given the higher conviction rate for control group members. Therefore, to show our results are robust to the way recidivism is measured, we also report effect estimates using only arrests that are caused by a new criminal incident. Appendix Figure A.6 shows that when including only arrests for new criminal incidents, MIR has large and statistically significant recidivism-reducing effects that are similar to those documented in Figure 2. Moreover, the 2SLS and ITT estimates of the impacts of MIR (Appendix Table A.5) are similar to those reported in Table 3. These analyses confirm that the estimated effects of MIR are not sensitive to the decision of whether or not to measure recidivism using all rearrests or using only rearrests for new criminal incidents.

5.3.5 The COVID-19 Period

In this section we present a key robustness check pertaining to the overlap of the MIR observation period and the COVID epidemic. Recall that study subjects are randomized into MIR through October 2019 and our observations period extends through the end of 2020. Hence, many of the youth in the study have later observation periods that overlap with the stringent stay-at-home orders in place in California (and the Bay Area in particular) .

Appendix Figure A.7 presents Kaplan-Meier estimates of the failure functions by treatment group where we have truncated the observation period for all youth to end at March 15, 2020.[24] Note, this truncation causes us to lose sample, especially for the observation periods beyond 16 months post-initial arrest. Nonetheless, the patterns we observe here are similar to what we observe for the failure functions using un-truncated observation periods. Large disparities in the failure functions open up soon after the initial arrest and persist throughout the observation period. Both inference strategies reject the null hypothesis of equal failure functions for the treatment and control groups when we focus on arrests for any new offenses. Similar patterns also emerge when examining effects on more severe interactions such as felony rearrest, rearrests for new offenses that are as severe as the original offense, or rearrests that lead to a conviction.

6 Mechanisms

In this section, we discuss the potential mechanisms by which MIR may be impacting recidivism. One hypothesis is that restorative justice conferencing is a transformative event that causes youth to change their behavior. Furthermore, the meeting with the victim and the support the youth receives throughout the process from the conference coordinator and the agreement monitor can cause a lasting change in the youth's behavior. For example, one of the program participants describes "In conference, we talked about how the person was harmed and how I harmed myself. I started to see things differently when I had to write a letter to myself about all the harm I

[24] On March 16, 2020 San Francisco and five other Bay Area counties enacted a strict shelter-in-place orders that greatly reduced social interactions outside of the home and closed all in-person instruction in public schools throughout the region.

had done." Indeed, 95 percent of the youth who participated in a restorative justice conference completed the MIR program. Moreover, interviews and discussions with CW staff (e.g., conference coordinators) reinforce the hypothesis that restorative justice conferencing is an emotionally impactful event for the youth that can potentially cause lasting effects on their behavior.

In addition to being a restorative justice conferencing program, MIR is also a diversion program. It diverts youth after the prosecutor decided to file charges but before charges have been formally filed. As a diversion program MIR influences case outcomes. Table 5 compares the conviction and case outcomes of individuals assigned to MIR relative to the control group. Panel (a) shows that among youth assigned to MIR, only five percent of the cases result in a felony conviction relative to 20 percent in the control group. Thus, assignment to MIR reduces the likelihood of being convicted of a felony offense by 15 percentage points. However, unlike criminal records in the adult system, in California, employers generally cannot use juvenile arrests and proceedings (e.g., charges, convictions) in employment decisions (Assembly Bill No. 1843) and this information does not appear in criminal history background checks. This restriction limits the degree to which interactions with the criminal justice system as juveniles can impact one's labor market options.

Previous studies found that among adults, a felony conviction can increase recidivism and negatively impact labor market outcomes (e.g., Mueller-Smith and Schnepel, 2020; Augustine et al., 2021). Next, we examine whether the impacts of MIR can be explained by the case not resulting in a felony conviction. To do this, we restrict the sample only to individuals whose case did not result in a felony conviction and compare the rearrest rate of youth assigned to MIR relative to the control group. Importantly, restricting the sample only to cases that did not result in a felony conviction biases us, if anything, against finding any effects of MIR. Individuals whose case results in a felony conviction are likely of higher risk of recidivism (at least not of lower risk), and 20 percent of the control group will have a felony conviction relative to only five percent of the treatment group. Indeed, Appendix Figure A.8 shows that among youth charged with a felony offense, those who are eventually convicted of a felony are meaningfully more likely to be rearrested in the future for any offense as well as more severe offenses such as felonies. Thus, removing these cases drops more individuals of potentially higher recidivism risk from the control group. Figure 4 shows that even among this sub-sample, youth participants assigned to MIR had a meaningfully lower likelihood of being rearrested both in the first few months as well as after four years from randomization. Moreover, restricting the sample even further to individuals whose case resulted in neither a felony nor a misdemeanor conviction yields similar results (see Appendix Figure A.9).

The results in Figure 4 show that even with this sample restriction to cases without a felony conviction, the MIR participants still have lower rearrest rates, indicating that the effects of MIR likely operate through its restorative justice conferencing component.[25] Furthermore, the ITT

[25] Agan et al. (2020) emphasize the importance of having any criminal history, even for non-violent misdemeanor offenses, on the likelihood of recidivism. MIR diverts youth before charging but after an arrest took place. Thus, both individuals in the control and treatment groups will have a criminal record of the initial arrest. In addition,

effects of MIR on the likelihood of being rearrested range from 14 to 27 percentage points, which is meaningfully larger than the effects of MIR on the case not resulting in a felony conviction. Thus, the primary channel for MIR's recidivism-reducing effects is its conferencing component.

Another channel through which case outcomes can impact recidivism is the likelihood a youth will be removed from their home. However, Panel (b) of Table 5 shows there is no difference in home removal rates between individuals assigned to MIR and those in the control group suggesting that differential home removal rates also cannot explain our findings. There is also no evidence of any type of differential "incapacitation" between the two groups.[26]

Having established that restorative justice conferencing is the primary mediating channel for MIR's effects, we next discuss the potential channels through which it may be altering behavior. One possibility is that the program changes an individual's perspective, making them more aware of their impacts on others, increasing empathy for those who are a victim of a criminal offense, and rendering the participant more deliberative, considerate, and self-reflective. Such an interpretation is consistent with the findings in past research that restorative justice conferencing appears to be most effective when the offense involves an identifiable victim (Sherman and Strang, 2007). Moreover, the shame associated with facing a crime victim during the conference might provide a more powerful deterrent than typical sanctions meted out through prosecution. Notably, a face-to-face meeting with the victim in which the youth listens to the harmful impacts of their actions on others is an emotional event that can influence an individual in multiple ways.

7 Concluding Remarks

The MIR project randomized youth charged with select felonies to either prosecution as usual or referral to a restorative justice alternative. Take-up rates for those assigned to the MIR treatment were relatively high, especially compared to past restorative justice experiments. Assignment to MIR causes a large reduction in the likelihood of future arrest, both in the immediate aftermath of the initial arrest as well as in the four years following. We observe this for arrests overall and for more severe felony arrests.

Two aspects of MIR deserve further discussion when thinking on its welfare implications and possible expansions. First, what are the financial costs of the MIR program relative to standard case processing? Second, to what degree are the large effects related to the types of offenses eligible for participation in MIR.

the record of the arrest will still be visible to prosecutors in future cases (if the youth is rearrested). However, as we mentioned above, in California, juvenile arrests and proceedings (e.g., charges, convictions) do not appear in employment criminal history background checks.

[26] Panel (b) of Table 5 also shows that assignment to MIR impacts the likelihood that youth will be under both informal probation and juvenile probation supervision. Both types of supervision are for a term of six months. They do not include any conditions that increase the likelihood of detecting criminal behavior. For example, there are no drug tests. Moreover, in many cases, youth undergo a needs assessment at Huckleberry Youth both when under probation supervision and at the agreement monitoring stage of the MIR program.

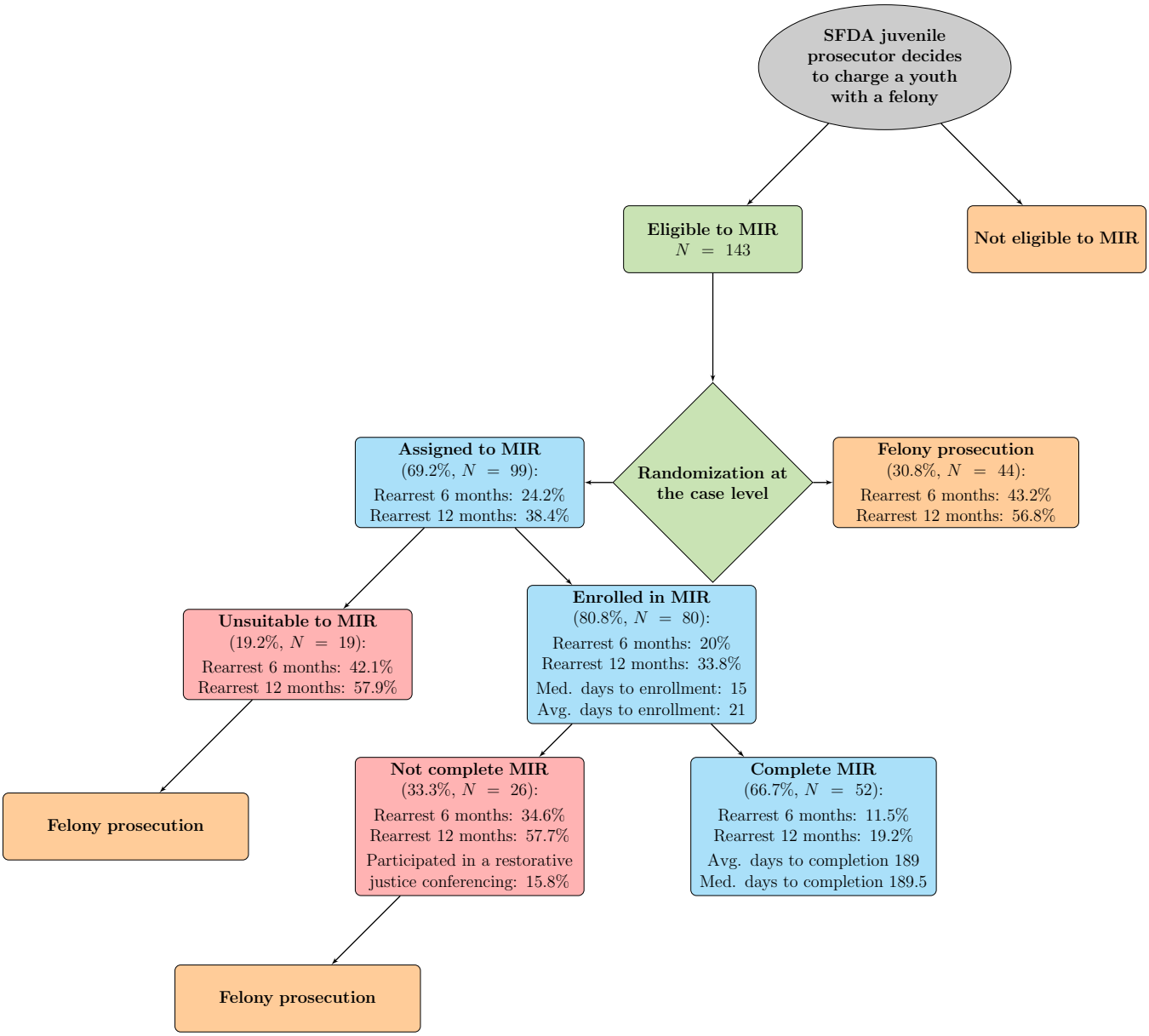
Costs and benefits. We do not have exact estimates of the costs of the MIR program relative a regular felony prosecution. However, in our interviews with employees of the district attorney and juvenile probation department, we have been told that the cost of allocating youth to MIR is not higher and might be even lower than the cost of juvenile criminal proceedings. Therefore, as long as the program does not increase recidivism, it helps to reduce the number of individuals involved in criminal proceedings with no additional costs. Our estimates show that the program even has large recidivism reducing effects, highlighting its potential gains.

Offense severity. The MIR program treated youth charged with more serious offenses that are typically not eligible for similar restorative justice or diversion programs. It may be the case that there is simply more opportunity to reduce the likelihood of future arrests even for juveniles charged with serious offenses than generally understood or that interventions targeted at less serious offenses may simply widen the net of the criminal justice system and apply an invasive intervention to instances that do not merit such an intervention. Given the relatively small scale of the experiment, we did not evaluate whether effect size varies by offense severity.

To conclude, our findings show that restorative justice conferencing can reduce recidivism among youth charged with relatively serious offenses and be an effective alternative to traditional criminal justice practices.

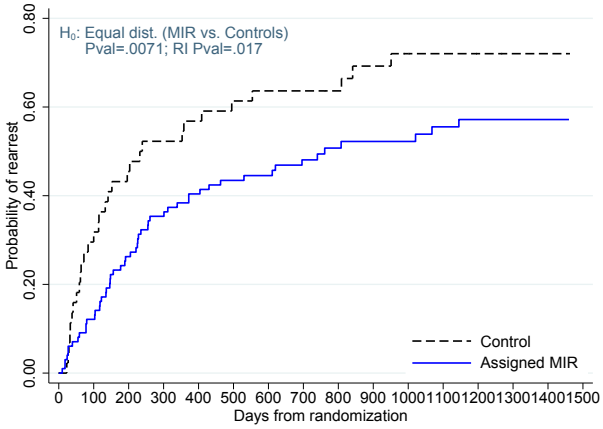
Figures

Figure 1: Make-it-Right Assignment, Enrollment, and Completion Process and Distribution



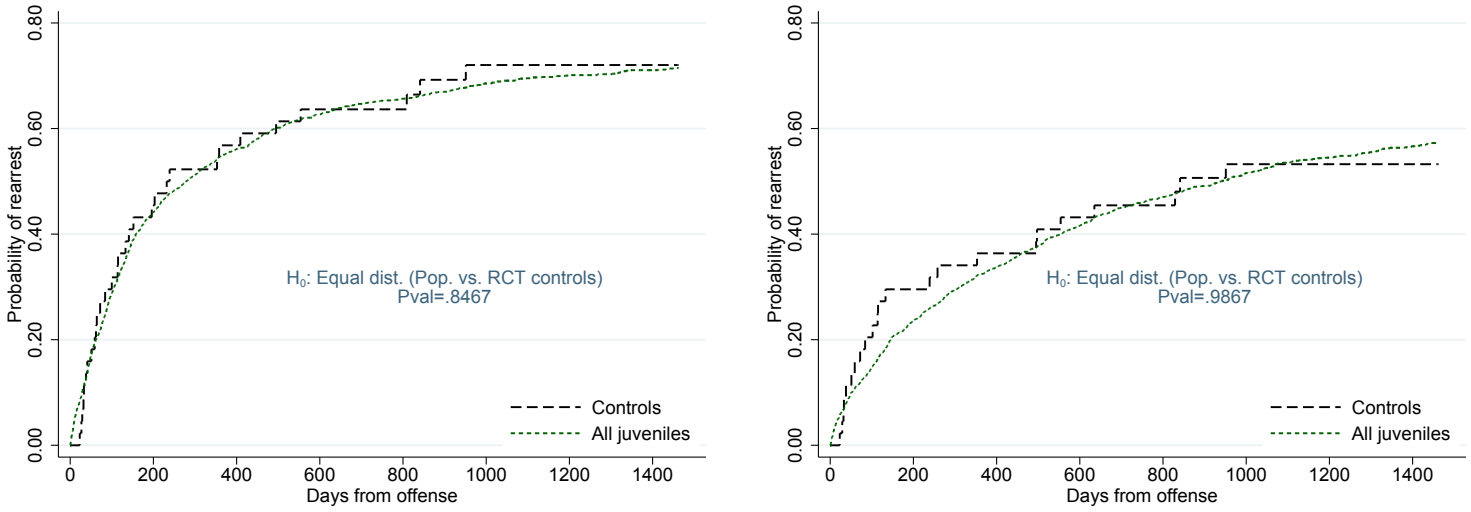
Notes: This figure depicts the process through which youths are assigned to the Make-it-Right (MIR) program. The process starts with the San Francisco District Attorney (SFDA) office receiving a reference / charges from the police or probation department regarding a youth. The SFDA juvenile prosecutor then decides whether or not to charge the youth with a felony offense. Our starting point (grey circle) is when the juvenile prosecutor decided to charge the youth with a felony offense. Then eligibility to MIR is assessed. If eligible, the youth is randomized to either MIR or to the control group which faces a traditional criminal prosecution for a felony offense. After being assigned to MIR the suitability of the youth to the program is being assessed by Community Works (CW) who administrates the MIR restorative justice conferencing program. Among the individuals assigned to MIR, 80.8% will be found suitable by CW and continue to enroll in MIR and 19.2% will be determined by CW as unsuitable and will face a standard criminal prosecution for a felony offense. Among the youths who enrolled into MIR, 66.7% will continue to complete the program, and as a result, formal charges were never filed against them. Another 33.3% will not succeed to complete the program and will face standard prosecution for their felony offense. CW administrates the restorative justice conferencing and the preparation towards it. This includes enrollment and unsuitability decisions. Following the conference, Huckleberry Youth provides the post-conference monitoring and support to the youth. The agreement monitor from Huckleberry Youth is responsible on making the determination whether a youth completed the agreement that was agreed upon in conferencing and thus completed the program. Importantly, youth who complete the program will not face criminal charge or any other sanction or supervision. The figure also reports the rearrest rates within six months and 12 months from the date in which youth are randomized into the different treatment regimes.

Figure 2: Rearrest Rate of Juveniles Randomly Assigned To Make-it-Right Relative to the Experimental Control Group



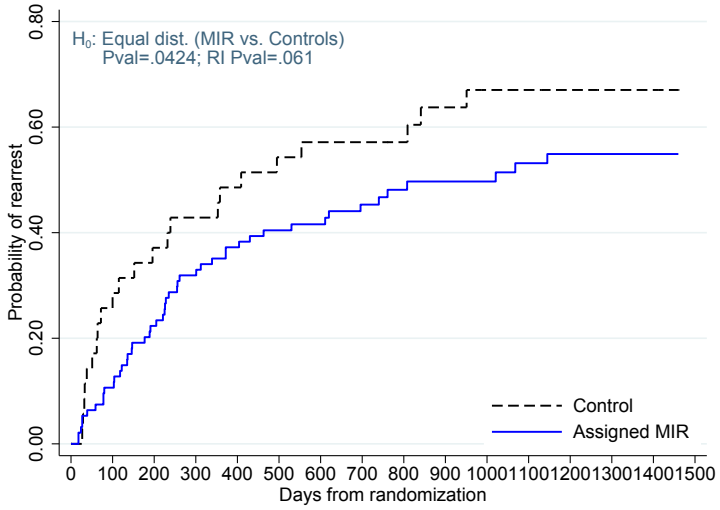
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization between assignment to Make-it-Right (MIR) and the control group which faces traditional criminal prosecution. The figure plots Kaplan-Meier estimates for any rearrest. We report p-values from an hypothesis test for whether the failure functions are the same among individuals assigned to MIR and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "RI Pval" which is based on randomization inference using 1,000 simulations in which we randomly assigned cases to MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that would to conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

Figure 3: A Comparison Between the Make-it-Right Experimental Sample and the Full Population of Juveniles Charged with a Felony Offense in San Francisco



Notes: This figure plots Kaplan-Meier estimates of the failure function of being rearrested for any offense and for felony offenses. It compares the rearrest rates of the experimental control group (dashed black line) and the full populations of youth charged with a felony offense in San Francisco (most of which are not eligible for MIR). The dotted green line reports Kaplan-Meier estimates of the rearrest rates of the full population of youth charged with a felony offense in San Francisco. We report p-values from an hypothesis test for whether the failure functions are the same or not. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). The p-values in these hypothesis tests are two-sided since we did not pre-specified in the pre-analysis plan how they will be conducted.

Figure 4: The Effects of Make-it-Right Among the Sub-Sample of Individuals Who Did Not Incur a Felony Conviction in Their Case Outcomes



Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization into eligibility for Make-it-Right (MIR) or the control group which faces traditional criminal prosecution. In this figure (unlike Figure 2), we restrict the sample to individuals whose case did not result in a felony conviction. As can be seen from the figure, even among this restricted population the individuals who have been assigned to MIR have meaningfully lower rearrest rates. We report p-values from an hypothesis test for whether the failure functions are the same among MIR participants and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "RI Pval" which is based on randomization inference using 1,000 simulations in which we randomly assigned cases to MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that we would conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

Tables

Table 1: Description of the Make-it-Right Restorative Justice Conferencing Program

Stage	Description/Examples of Activities
Suitability Assessment	CW coordinator has initial meeting(s) with the youth (responsible party) and his/her family to determine if they are suitable for restorative justice community conferencing. The youth must agree to participate and demonstrate reflection and accountability to self, family, community, and person harmed. The main question that the coordinator asks when determining suitability is whether they feel confident putting the responsible youth in front of the person harmed. A youth who is unwilling to take responsibility will be deemed unsuitable at this point.
Pre-Conference (post-enrollment)	The CW coordinator holds several pre-conference meetings with the **responsible youth** and their family/support person to prepare them for the conference, typically between three to four. The youth must finalize an **apology letter** ahead of the conference. This is a reflective apology, for example, how do I feel about my actions now? If I had to do it over again, what would I do differently? What would I like the victim to know? What can or will I do to make up for what I did? The conference coordinator also conducts preparation meetings with the **harmed party**. These meetings aim to set expectations from the conference and for understanding the limitations the youth is facing.
Conference	The conference begins with the youth (responsible party) reading the **apology letter** to the harmed party. Next there is a **roundtable discussion** on how to address the four quadrants of the harm to: self, victim, family, and community. The conference results in a **consensus-based plan of action** (i.e., an agreement) for the youth's accountability and to prevent the youth from engaging in future criminal activity. The agreement's objective is to restore welfare by addressing the **four quadrants of the harm**: self, victim, family, and community. **All parties sign the agreement**. This concludes the formal involvement of the harmed party. Multiple conferences can be held until the plan is developed—if no plan is developed, the youth is referred back to the SFDA for prosecution.
Examples of Agreement Activities	**Academic**: Tutoring sessions, meet high school attendance requirements, make a plan for college/technical school application. **Employment**: Make a goal to create a resume and apply to a certain number of jobs. **Reflection Writing**: Journaling, poems, and/or essays reflecting on opportunities for self-improvement. **Yoga**: Attend a set number of yoga classes. **Anger management**: Attend/complete a set number of anger management sessions. **Restitution**: Identify amount to provide to harmed party and/or community. **Chores**: Keeping one's room clean, taking out the trash, helping with dinner, etc. Goal is to help the youth engage more in family life. **Family Systems Therapy**: Counseling sessions with identified family members. **Community Service**: Youth repairs harm done to the community by performing a set number of community service hours at a local organization.
Agreement Implementation	After the conference, the Huckleberry Youth agreement monitor debriefs with the youth. They finalize the details of the restorative plan (i.e., the agreement) and set target completion dates. Youth and agreement monitor meet on a weekly basis to review the youth's progress toward completion of the plan. The meetings are not only about making sure the youth is on track to finish the plan; they also discuss other issues that the youth is facing and develop a plan to address them.

Notes: This table summarizes the activities and sequence of events of the Make-it-Right (MIR) program. The initials CW refer to Community Works which is a non-profit community organization that specializes in restorative justice conferencing and administrates the MIR program in San Francisco. The initials SFDA refer to the San Francisco District Attorney.

Table 2: Summary Statistics of Make-It-Right Experimental Sample and the Full Sample of Juveniles Charged With Felony Offenses in San Francisco

	(1) Feloy prosecution (control)	(2) Assigned MIR (treatment)	(3) Enrolled to MIR (compliers)	(4) All juveniles charged with a felony
Demographics:				
Male	0.909	0.889	0.900	0.802
		[0.756]		
Black	0.500	0.531	0.487	0.607
		[0.788]		
Hispanic	0.318	0.323	0.359	0.239
		[0.966]		
Age	16.023	16.091	16.113	16.124
		[0.814]		
Criminal history:				
Any past arrests	0.318	0.434	0.425	0.568
		[0.215]		
Number of past arrests	0.773	0.616	0.588	1.949
		[0.583]		
Any past felony arrests	0.136	0.333	0.350	0.459
		[0.020]		
Number of past felony arrests	0.182	0.364	0.362	0.943
		[0.089]		
Age at first criminal offense	14.750	15.198	15.269	14.554
		[0.091]		
Type of most severe offense:				
Homicide/Manslaughter	0.000	0.000	0.000	0.020
		[.]		
Sex offense	0.000	0.000	0.000	0.014
		[.]		
Roberry	0.000	0.030	0.000	0.387
		[0.087]		
Assault	0.159	0.131	0.138	0.291
		[0.690]		
Burglary	0.318	0.434	0.487	0.142
		[0.245]		
Theft	0.636	0.657	0.713	0.218
		[0.833]		
Drug	0.000	0.000	0.000	0.061
		[.]		
Weapons	0.000	0.020	0.025	0.131
		[0.165]		
Other	0.205	0.293	0.287	0.472
		[0.447]		
Predicted recidivism:				
Pred. recidivism 6 months	0.372	0.362	0.363	0.427
		[0.519]		
Pred. recidivism 12 months	0.470	0.459	0.459	0.548
		[0.428]		
Pred. felony recidivism 6 months	0.164	0.170	0.169	0.222
		[0.757]		
Pred. felony recidivism 12 months	0.257	0.255	0.254	0.325
		[0.934]		
Joint F-test of MIR assignment on covariates p-value	0.789			
Joint F-test of MIR compliers and never-takers covariates p-value	0.914			
Number of observations	44	99	80	1531
Number of individuals	44	99	80	1094

Notes: The table reports summary statistics (means) of the individuals randomly assigned to face standard felony prosecution, the control group, (Column (1)), to Make-it-Right (MIR), the treatment group, (Column (2)), the compliers—those assigned to MIR and who also enrolled into the program (Column (3)), and the full population of juveniles charged with felony offenses between October 2013 and May 2019. The square parenthesis in Column (2) report p-values for whether the difference in each characteristic between Columns (1) and (2) is different than zero. The mean characteristics of compliers in Column (3) are calculated using the standard formula from Abadie (2003). Specifically, using a 2SLS regression of a covariate interacted with an indicator for enrollment into MIR (i.e., $\text{MIR}_i \cdot X_i$) as the outcome, an indicator for MIR enrollment as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up rate is about 80 percent. The joint F-tests at the bottom of the table are based on randomization inference (Fisher, 1935) using 1,000 random placebo permutations of treatment assignment.

Table 3: The Effects of Assignment (ITT) to and Participation (TOT) in Make-it-Right on the Likelihood of Being Arrested in the Subsequent Four Years

	(1) 6 months	(2) 12 months	(3) 24 months	(4) 36 months	(5) 48 months	(6) 12-48 months
Panel (a)			*2SLS*			
Participated in MIR (treated)	-0.234	-0.228	-0.184	-0.196	-0.363	-0.368
	(0.103)	(0.111)	(0.128)	(0.151)	(0.165)	(0.199)
	{0.0120}	{0.0211}	{0.0759}	{0.0979}	{0.0157}	{0.0344}
	[0.0030]	[0.0050]	[0.0722]	[0.1150]	[0.0230]	[0.0250]
Panel (b)			*Reduced form*			
Assigned to MIR (ITT)	-0.189	-0.184	-0.144	-0.147	-0.267	-0.270
	(0.084)	(0.092)	(0.103)	(0.118)	(0.133)	(0.154)
	{0.0132}	{0.0237}	{0.0830}	{0.1092}	{0.0249}	{0.0423}
	[0.0140]	[0.0410]	[0.1130]	[0.1680]	[0.0850]	[0.1040]
Panel (c)						
First-Stage coefficient	0.808	0.808	0.781	0.750	0.736	0.736
	(.0463)	(.0463)	(.0558)	(.0676)	(.0832)	(.0832)
Rearest rate among controls	0.432	0.568	0.632	0.750	0.833	0.667
Rearest rate among compliers controls	0.434	0.566	0.606	0.745	0.876	0.726
Includes controls	No	No	No	No	No	No
Number of observations	143	143	120	100	71	71

Notes: The table reports estimates of the effects of Make-it-Right (MIR) on the likelihood of a future arrest. Each cell in the table reports four numbers: the point estimate, standard error clustered at the case level, a one-sided p-value using the cluster-robust standard errors, and a one-sided p-value using randomization inference (Fisher, 1935) based on 1,000 random permutations. The compliers rearrest rates under the control regime (bottom of the table) are calculated using the standard formulas from Imbens and Rubin (1997) and Abadie (2003). Specifically, using a 2SLS regression of the outcome interacted with an indicator for enrollment into MIR (i.e., $(1 - \text{MIR}_i) \cdot \text{Rearrest}_i$) as the outcome, an indicator for not enrolling into MIR (i.e., $(1 - \text{MIR}_i)$) as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up rate is about 75% and is reported at the bottom of the table (i.e., the First-Stage coefficient). The number of observations changes across the columns because the sample in each of the regressions is restricted to individuals that are observed at least the mentioned time horizon (e.g., 48 months in Column (5)) after the date of randomization.

Table 4: The Effects of Completing Make-it-Right on the Likelihood of Being Arrested in the Subsequent Four Years

	(1) 6 months	(2) 12 months	(3) 24 months	(4) 36 months	(5) 48 months	(6) 12-48 months
Panel (a)			*2SLS*			
Completed MIR (treated)	-0.361	-0.351	-0.268	-0.286	-0.525	-0.531
	(0.159)	(0.166)	(0.176)	(0.205)	(0.214)	(0.276)
	{0.0125}	{0.0178}	{0.0653}	{0.0837}	{0.0084}	{0.0295}
	[0.0010]	[0.0101]	[0.1180]	[0.1690]	[0.0460]	[0.0380]
Panel (b)						
First-Stage coefficient	0.525	0.525	0.537	0.515	0.509	0.509
	(.0726)	(.0726)	(.0825)	(.0974)	(.1231)	(.1231)
Rearest rate among controls	0.432	0.568	0.632	0.750	0.833	0.667
Rearest rate among compliers controls	0.476	0.543	0.541	0.743	0.932	0.827
Includes controls	No	No	No	No	No	No
Number of observations	143	143	120	100	71	71

Notes: The table reports estimates of the effects of Make-it-Right (MIR) completion on the likelihood of a future arrest. The key difference between the estimates in this table relative to Table 3 is that here the treatment is defined as *completion* rather than enrollment in MIR. Here the key assumption is that assignment to MIR impacts recidivism only by impacting the likelihood that an individual will complete the MIR program. In other words, participation in the program without completion has no effect on recidivism. Each cell in the table reports four numbers: the point estimate, standard error clustered at the case level, a one-sided p-value using the cluster-robust standard errors, and a one-sided p-value using randomization inference (Fisher, 1935) based on 1,000 random permutations. The compliers rearrest rates under the control regime (bottom of the table) are calculated using the standard formulas from Imbens and Rubin (1997) and Abadie (2003). Specifically, using a 2SLS regression of the outcome interacted with an indicator for enrollment into MIR (i.e., $(1 - \text{MIR}_i) \cdot \text{Rearrest}_i$) as the outcome, an indicator for not enrolling into MIR (i.e., $(1 - \text{MIR}_i)$) as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up/completion rate is about 50 percent and is reported at the bottom of the table (i.e., the First-Stage coefficient). The number of observations changes across the columns because the sample in each of the regressions is restricted to individuals that are observed at least the mentioned time horizon (e.g., 48 months in Column (5)) after the date of randomization.

Table 5: A Comparison of Convictions and Case Dispositions Between Individuals Assigned to Make-it-Right Relative to Those Assigned to the Control Group

	(1) Control	(2) Assigned MIR
Panel (a) Convicted offenses:		
Convicted of misdemeanor or felony	0.386	0.152
Convicted of misdemeanor	0.205	0.121
Convicted of felony	0.205	0.0505
Panel (b) Disposition (least to most severe):		
Charges dismissed	0.364	0.737
Informal probation	0.364	0.121
Probation	0.0455	0.0101
Wardship probation	0.136	0.0707
Out of home placement	0.0227	0.0202
Number of individuals	44	99

Notes: The table reports information on convictions and dispositions for the individuals in our experimental sample. These case outcomes are for the focal criminal incident that lead the individual to be in our experimental sample. We do not include charging, conviction, or disposition information from any future criminal incidents. Note that individuals who complete MIR will not have criminal charges filed against them and will not face any other criminal sanctions. All probation types (informal probation, probation, and wardship probation) are for a period of six months. Informal probation is levied on individuals who have not been convicted of a crime and acts as a diversion from further prosecution. Both "Probation" and "Wardship probation" are analogous to probation supervision in the adult criminal justice system. The difference is that wardship probation imposes relatively more restrictive.

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 2003, *113* (2), 231–263.

Abrams, David S, "Estimating the deterrent effect of incarceration using sentencing enhancements," *American Economic Journal: Applied Economics*, 2012, *4* (4), 32–56.

Agan, Amanda and Sonja Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 2018, *133* (1), 191–235.

_ , Jennifer Doleac, and Anna Harvey, *Misdemeanor Prosecution*, College Station, TX: Texas AM Working Paper, 2020.

Aizer, Anna and Joseph J. Doyle, "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *The Quarterly Journal of Economics*, 2015, *130* (2), 759–803.

Angel, Caroline M., Lawrence W. Sherman, Heather Strang, Barak Ariel, Sarah Bennett, Nova Inkpem, Anne Keane, and Therese S. Richmond, "Short-Term Effects of Restorative Justice Conferences on Post-Traumatic Stress Symptoms Among Robbery and Burglary Victims: A Randomized Controlled Trial," *Journal of Experimental Criminology*, 2014, *10*, 291–307.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Arteaga, Carolina, "Parental Incarceration and Children's Educational Attainment," 2020. Working paper.

Augustine, Elsa, Johanna Lacoe, Alissa Skog, and Steven Raphael, *The Impact of Felony Diversion in San Francisco*, Berkeley, CA: University of California Berkeley Working Paper, 2021.

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 1968, *76* (2), 169–217.

Berg, Mark T., Eric A. Stewart, Christopher J. Schreck, and Ronald L. Simmons, "The victim-offender overlap in context: Examining the role of neighborhood street culture," *Criminology*, 2012, *50*, 359–390.

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad, "Intergenerational Effects of Incarceration," *AEA Papers and Proceedings*, 2018, *108*, 234–40.

_ , _ , _ , and _ , "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 2020, *128* (4), 1269–1324.

Blattman, Christopher, Julian C Jamison, and Margaret Sheridan, "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia," *American Economic Review*, 2017, *107* (4), 1165–1206.

Bloom, Howard S, "Accounting for no-shows in experimental evaluation designs," *Evaluation review*, 1984, *8* (2), 225–246.

Braithwaite, John, *Crime, Shame, and Reintegration*, Cambridge, U.K.: Cambridge University Press, 1989.

Britto, Diogo, Paolo Pinotti, and Breno Sampaio, "The effect of job loss and unemployment insurance on crime in Brazil," 2020.

Brooks, Alison, *Moving forward: Two approaches to repairing the harm through restorative justice*, American University, 2013. PhD dissertation.

Bushway, Shawn D., "Labor Market Effects of Permitting Employer Access to Criminal History Records," *Journal of Contemporary Criminal Justice*, 2004, *20* (3), 276–291.

Christensen, Garret and Edward Miguel, "Transparency, reproducibility, and the credibility of economics research," *Journal of Economic Literature*, 2018, *56* (3), 920–80.

Chung, EunYi and Joseph P Romano, "Exact and asymptotically robust permutation tests," *The Annals of Statistics*, 2013, *41* (2), 484–507.

Cuellar, Allison E., Larkin S. McReynolds, and Gail A. Wasserman, "A Cure For Crime: Can Mental Health Treatment Diversion Reduce Crime Among Youth," *Journal of Policy Analysis and Management*, 2006, *25* (1), 197–214.

Cullen, Zoë, Will Dobbie, and Mitchell Hoffman, "Increasing the Demand for Workers with a Criminal Record," 2021.

Damm, Anna Piil and Christian Dustmann, "Does growing up in a high crime neighborhood affect youth criminal behavior?," *American Economic Review*, 2014, *104* (6), 1806–32.

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

Dobbie, Will, Hans Grönqvist, Susan Niknami, Mårten Palme, and Mikael Priks, "The intergenerational effects of parental incarceration," Technical Report, National Bureau of Economic Research 2018.

Doleac, Jennifer L and Benjamin Hansen, "The unintended consequences of "ban the box": Statistical discrimination and employment outcomes when criminal histories are hidden," *Journal of Labor Economics*, 2020, *38* (2), 321–374.

Draca, Mirko, Theodore Koutmeridis, and Stephen Machin, "The changing returns to crime: do criminals respond to prices?," *The Review of Economic Studies*, 2019, *86* (3), 1228–1257.

Drago, Francesco, Roberto Galbiati, and Pietro Vertova, "The Deterrent Effects of Prison: Evidence from a Natural Experiment," *Journal of Political Economy*, 2009, *117* (2), 257–280.

Eren, Ozkan and Naci Mocan, "Juvenile Punishment, High School Graduation, and Adult Crime: Evidence from Idiosyncratic Judge Harshness," *Review of Economics and Statistics*, 2021, *103* (1), 34–47.

Fisher, RA, "The design of experiments.," *The design of experiments.*, 1935, (1nd Ed).

Fulkerson, Andrew, "The use of victim impact panels in domestic violence cases: A Restorative Justice Approach," *Contemporary Justice Review*, 2001, *4*, 355–368.

Gagnon-Bartsch, Johann A, Adam C Sales, Edward Wu, Anthony F Botelho, Luke W Miratrix, and Neil T Heffernan, "Precise Unbiased Estimation in Randomized Experiments using Auxiliary Observational Data," 2020.

Gelman, Andrew and John Carlin, "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors," *Perspectives on Psychological Science*, 2014, *9* (6), 641–651.

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack, "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago," *The Quarterly Journal of Economics*, 10 2017, *132* (1), 1–54.

Hirschi, Travis and Michael Gottfredson, "Age and the explanation of crime," *American journal of sociology*, 1983, *89* (3), 552–584.

Hjalmarsson, Randi, "Juvenile Jails: A Path to the Straight and Narrow or to Hardened Criminality?," *Journal of Law and Economics*, 2009, *52* (4), 779–809.

Huntington, Margaret, Lisa Quan, Sarah Riley, and Richard Zarrella, *Make It Right: An Evaluation of a Youth Restorative Justice Program in San Francisco*, University of California, Berkeley: Goldman School of Public Policy, 2017.

Huttunen, K., M. Kaila, M. Kaila, T. Kosonen, and E. Nix, "Shared Punishment? The Impact of Incarcerating Fathers on Child Outcomes," 2019. Working paper.

i Vidal, Jordi Blanes and Giovanni Mastrobuoni, "Police patrols and crime," 2018.

Imbens, Guido and Donald Rubin, "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 1997, *64* (4), 555–574.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Jeong, Seokjin, Edmund F McGarrell, and Natalie Kroovand Hipple, "Long-term impact of family group conferences on re-offending: The Indianapolis restorative justice experiment," *Journal of Experimental Criminology*, 2012, *8* (4), 369–385.

Katz, Lawrence F, Jeffrey R Kling, and Jeffrey B Liebman, "Moving to opportunity in Boston: Early results of a randomized mobility experiment," *The Quarterly Journal of Economics*, 2001, *116* (2), 607–654.

Khanna, Gaurav, Carlos Medina, Anant Nyshadham, Christian Posso, and Jorge Tamayo, "Job Loss, Credit, and Crime in Colombia," *American Economic Review: Insights*, 2021, *3* (1), 97–114.

Kirchmaier, Tom, Stephen Machin, Matteo Sandi, and Robert Witt, "Prices, policing and policy: the dynamics of crime booms and busts," *Journal of the European Economic Association*, 2020, *18* (2), 1040–1077.

Klein, John P and Melvin L Moeschberger, *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media, 2006.

Kling, Jeffrey R, Jens Ludwig, and Lawrence F Katz, "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," *The Quarterly Journal of Economics*, 2005, *120* (1), 87–130.

Kuziemko, Ilyana, "How should inmates be released from prison? An assessment of parole versus fixed-sentence regimes," *The Quarterly Journal of Economics*, 2013, *128* (1), 371–424.

Li, Xinran and Peng Ding, "General forms of finite population central limit theorems with applications to causal inference," *Journal of the American Statistical Association*, 2017, *112* (520), 1759–1769.

Little, Simon, Anne Stewart, and Nicole Ryan, "Restorative Justice Conferencing: Not a Panacea for the Overrepresentation of Australia's Indigenous Youth in the Criminal Justice System," *International Journal of Offender Therapy and Comparative Criminolog*, 2018, *62* (13), 4067–4090.

Livingstone, Nuala, Geraldine Macdonald, and Nicola Carr, "Restorative justice conferencing for reducing recidivism in young offenders (aged 7 to 21)," *Cochrane Library*, 2013, (2).

Lofstrom, Magnus and Steven Raphael, "Crime, the Criminal Justice System, and Socioeconomic Inequality," *The Journal of Economic Perspectives*, 2016, *30* (2), 103–126.

Ludwig, Jens, Greg J Duncan, and Paul Hirschfield, "Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment," *The Quarterly Journal of Economics*, 2001, *116* (2), 655–679.

McCold, Paul and Benjamin Wachtel, *Restorative Policing Experiment: The Bethlehem Pennsylvania Police Family Group Conferencing Project*, Pipersville, PA: Community Service Foundation, 1998.

McGarrell, Edmund F, *Restorative Justice Conferences as an Early Response to Young Offenders*, Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice, 2001.

_ **and Natalie Kroovand Hipple**, "Family group conferencing and re-offending among first-time juvenile offenders: The Indianapolis experiment," *Justice Quarterly*, 2007, *24* (2), 221–246.

Ministry of Justice, *Restorative Justice: Best Practice in New Zealand*, Wellington, New Zealand: Ministry of Justice, 2004.

Mitchell, Ojmarrh, David B. Wilson, Amy Eggers, and Doris L. MacKenzie, "Assessing the Effectiveness of Drug Courts on Recidividsm: A Meta-Analytical Review of Traditional and Non-Traditional Drug Courts," *Journal of Criminal Justice*, 2012, *40* (1), 60–71.

Mueller-Smith, Michael and Kevin T. Schnepel, "Diversion in the Criminal Justice System," *Review of Economic Studies*, 2020, *rdaa030, https://doi.org/10.1093/restud/rdaa030.*

Neal, Derek and Armin Rick, "The prison boom and the lack of black progress after Smith and Welch," Technical Report, National Bureau of Economic Research 2014.

Norris, Samuel, Matthew Pecenco, and Jeffrey Weave, "The Effects of Parental and Sibling Incarceration: Evidence from Ohio," 2020. Working paper.

Olken, Benjamin A, "Promises and perils of pre-analysis plans," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Owens, Emily, Aria Golestani, and Kerri Raissian, "Specialization in Criminal Courts: Decision Making, Recidivism, and Re-victimization in Domestic Violence Courts in Tennessee," 2021. Working paper.

Raphael, Steven, "The New Scarlet Letter? Negotiating the U.S. Labor Market with a Criminal Record," 2014. Report.

_ **and Michael A. Stoll**, *Why Are So Many Americans in Prison*, New York, NY: Russell Sage Foundation, 2013.

Rose, E, "Who gets a second chance? effectiveness and equity in supervision of criminal offenders," *The Quarterly Journal of Economics*, 2020.

Rose, Evan K, "Does banning the box help ex-offenders get jobs? Evaluating the effects of a prominent example," *Journal of Labor Economics*, 2021, *39* (1), 79–113.

_ **and Yotam Shem-Tov**, "How does incarceration affect reoffending? Estimating the dose-response function," *Journal of Political Economy*, 2021, *129* (12), 3302–3356.

Sampson, Robert J. and Janet L. Lauritsen, "Deviant lifestyles, proximity to crime and the offender- victim link in personal violence," *Journal of Research on Crime and Delinquency*, 1990, *27*, 110–139.

Schnepel, Kevin T, "Good jobs and recidivism," *The Economic Journal*, 2018, *128* (608), 447–469.

Seward, Jonathan, Vivian S. Vigliotti, and Scott Cunningham, "Social Workers and Suicidality in Jail: Evidence from Travis County's Mental Health Court," 2021.

Shapland, Joanna, Anne Atkinson, Helen Atkinson, James Dignan, Lucy Edwards, Jeremy Hibbert, Marie Howes, Jennifer Johnstone, Gwen Robinson, and Angela Sorsby, "Does Restorative Justice Affect Reconviction: The Fourth Report from the Evaluation of Three Scheme," Technical Report, Ministry of Justice Research Series 10/08, United Kingdom 2008.

Sherman, Lawrence and Heather Strang, "Restorative Justice: The Evidence," Technical Report, The Smith Institute, London, United Kingdom 2007.

Sherman, Lawrence W., Heather Strang, Geoffrey Barnes, Daniel J. Woods, Sarah Bennett, Nova Inkpen, Dorothy Newbury-Birch, Meredith Rossner, Caroline Angel, Malcolm Mearns, and Molly Slothower, "Twelve Experiments in Restorative Justice: the Jerry Lee Program of Randomized Trials of Restorative Justice Conferences," *Journal of Experimental Criminology*, 2015, *11*, 501–540.

Strang, Heather, John Braithwaite et al., *Restorative justice and family violence*, Cambridge University Press, 2002.

_ , Lawrence W Sherman, Evan Mayo-Wilson, Daniel Woods, and Barak Ariel, "Restorative justice conferencing (RJC) using face-to-face meetings of offenders and victims: Effects on offender recidivism and victim satisfaction. A systematic review," *Campbell Systematic Reviews*, 2013, *9* (1), 1–59.

Tella, Rafael Di and Ernesto Schargrodsky, "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack," *American Economic Review*, 2004, *94* (1), 115–133.

Ulmer, Jeffrey Todd and Darrell J Steffensmeier, "The age and crime relationship: Social variation, social explanations," in "The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality," SAGE Publications Inc., 2014, pp. 377–396.

Umbreit, Mark S. and Marilyn Peterson Armour, "Restorative Justice and Dialogues: Impact, Opportunities, and Challenges in the Global Community," *Washington University Journal of Law Policy*, 2011, *36*, 65–89.

Wilson, David B, Ajima Olaghere, and Catherine S Kimbrell, *Effectiveness of restorative justice principles in juvenile justice: A meta-analysis* 2018.

Wu, Jason and Peng Ding, "Randomization tests for weak null hypotheses in randomized experiments," *Journal of the American Statistical Association*, 2020, pp. 1–16.

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

Young, Alwyn, "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *The Quarterly Journal of Economics*, 2019, *134* (2), 557–598.

Zhao, Anqi and Peng Ding, "Covariate-adjusted Fisher randomization tests for the average treatment effect," *Journal of Econometrics*, 2021.

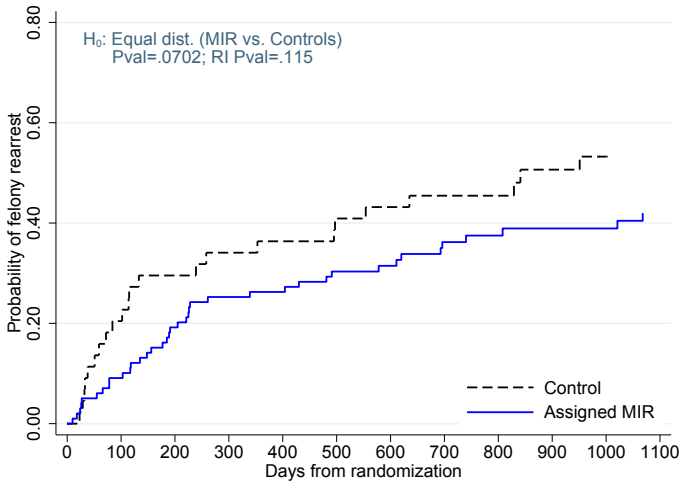
A Additional Figures and Tables

Figure A.1: Illustration of a Make-it-Right restorative justice conference

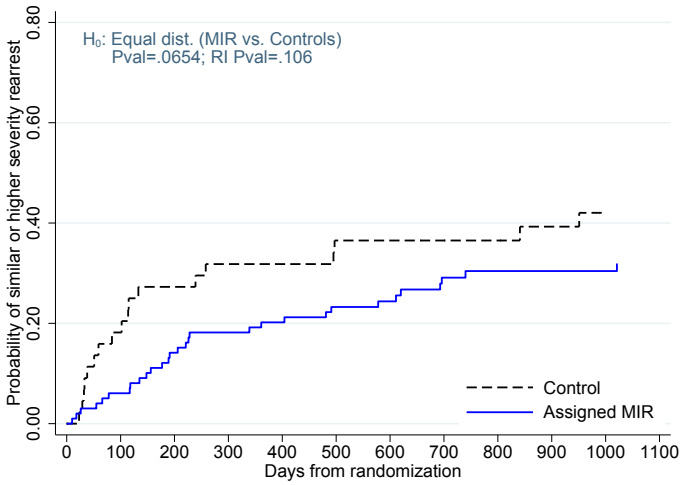


Notes: This figure illustrates how a Make-it-Right restorative justice conference looks like. The picture is from Community Works's website, http://communityworkswest.org/restorative-justice-circles/.

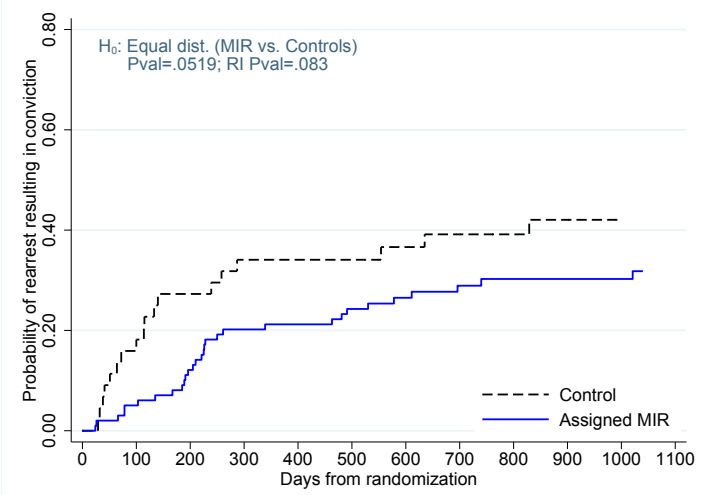
Figure A.2: A Comparison of Recidivism Rates Between Juveniles Randomly Assigned To Make-it-Right Relative to the Experimental Control Group Along Different Measures of the Severity of Reoffending



(a) Any felony rearrest from randomization



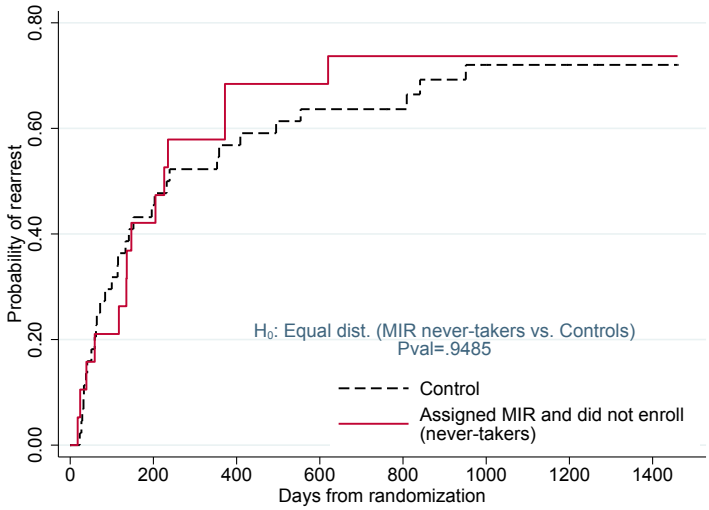
(b) Any rearrest from randomization for offenses at least as severe as the original offense



(c) Any rearrest from randomization that lead to a conviction

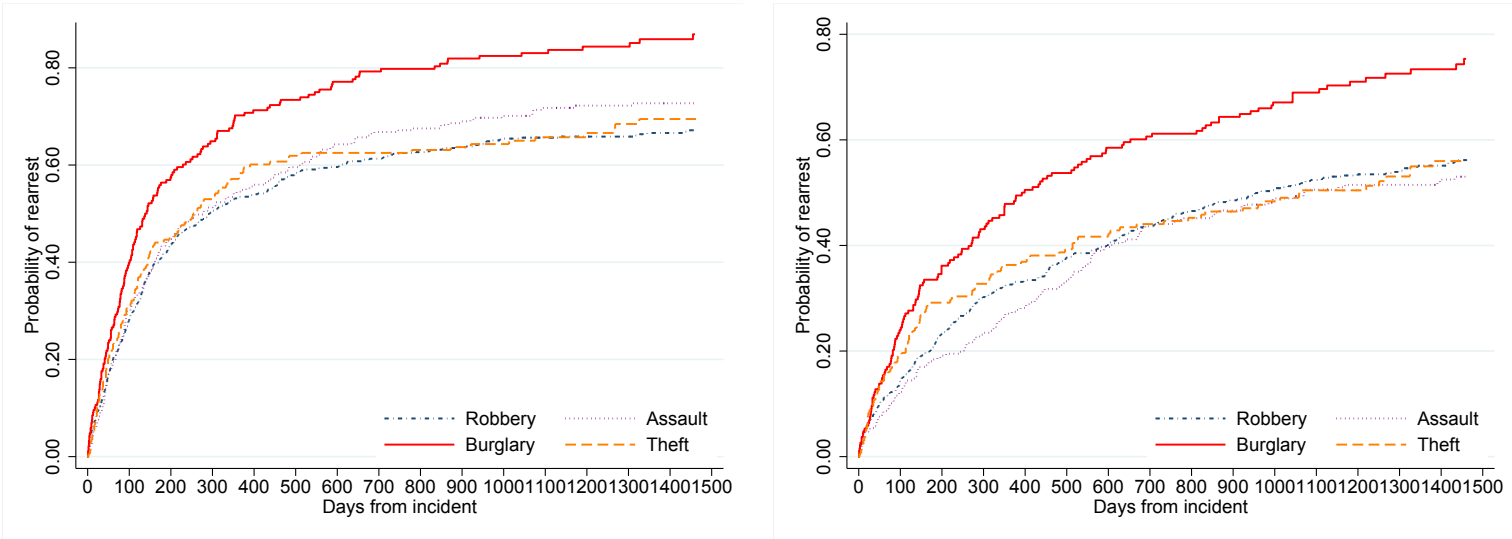
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization between assignment to Make-it-Right (MIR) and the control group which faces traditional criminal prosecution. Panel (a) plots Kaplan-Meier estimates for felony rearrest. Panel (b) for rearrests for offenses that as severe as the focal offense for which the youth appears in our experimental sample. Panel (c) only counts rearrests that resulted in a conviction. In all plots we report p-values from an hypothesis test for whether the failure functions are the same among MIR participants and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "PI Pval" which is based on permutation inference (Fisher, 1935) using 1,000 simulations in which we randomly assigned cases to MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that we would conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

Figure A.3: A Comparison of Recidivism Rates Between the Control Group and Individuals Assigned to Make-it-Right But Who Did Not Enroll ("Never-Takers")



Notes: This figure plots Kaplan-Meier estimates of the failure function of being rearrested. The figure compares compares between the failure curves of the experimental control group (dashed black line) and of youth randomly assigned to Make-it-Right (MIR) but who did not enroll into the MIR program, i.e., the "never-takers" (solid red line). The p-value is from an hypothesis test for whether the failure functions are the same or not. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). The p-values in these hypothesis test are two-sided since we did not pre-specify in the pre-analysis plan how they would be conducted.

Figure A.4: A Comparison of Rearrest Rates by the Type of Felony Offense a Juvenile Is Charged

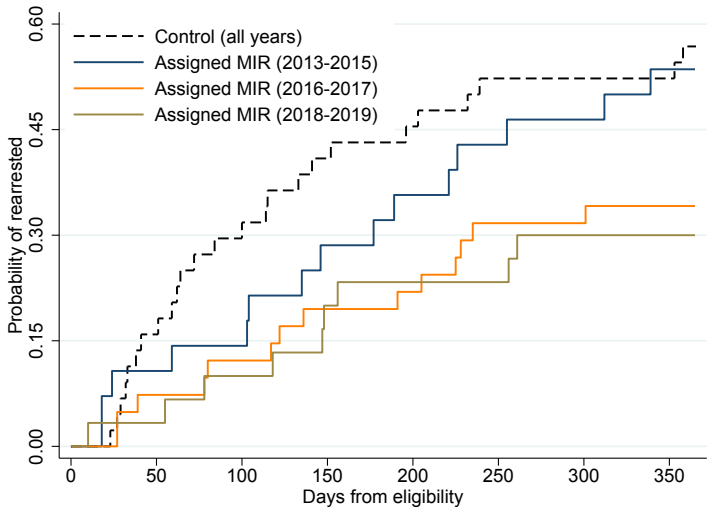


(a) A new arrest for any offense

A new arrest for a felony offense

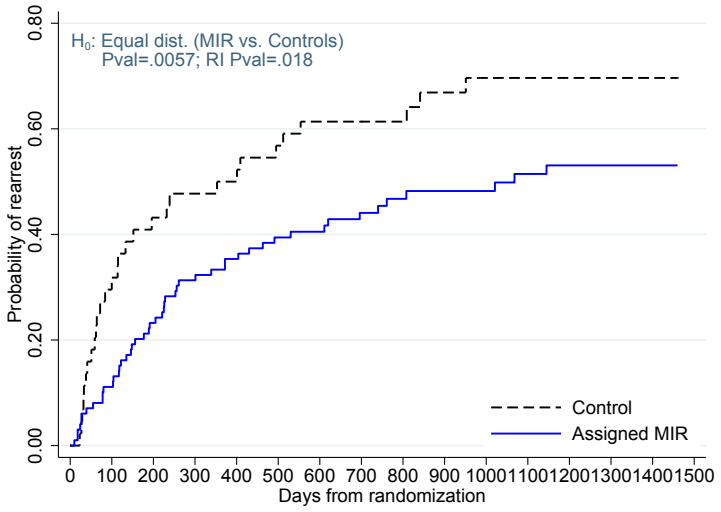
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date the focal offense took place. Each of the figures plots four failure function curves, one for each of the four most common types of felony offenses in San Francisco.

Figure A.5: Rearrest Rates of Juveniles Assigned To Make-it-Right in Different Time Periods Relative to the Experimental Control Group



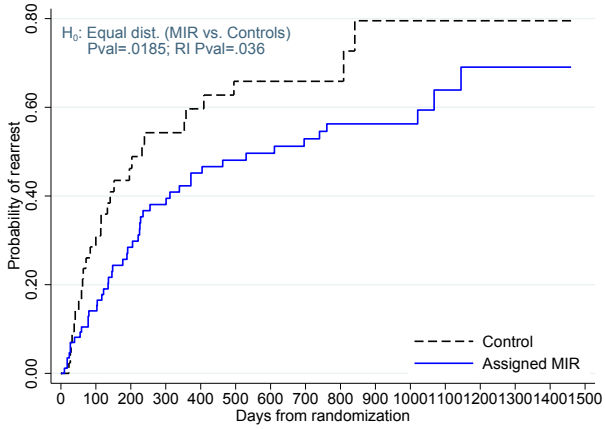
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within one year from the date of randomization into eligibility for Make-it-Right (MIR) or the control group which faces traditional criminal prosecution. We observe at least one year post-randomization for all the individuals in our sample. Thus, when comparing the rearrest rates within one year we do not need to drop any observations. However, when examining the likelihood of a rearrest in a longer time horizon the sample decreases.

Figure A.6: Recidivism Rates of Juveniles Randomly Assigned To Make-it-Right Relative to the Experimental Control Group Using Only Rearrests for New Criminal Incidents

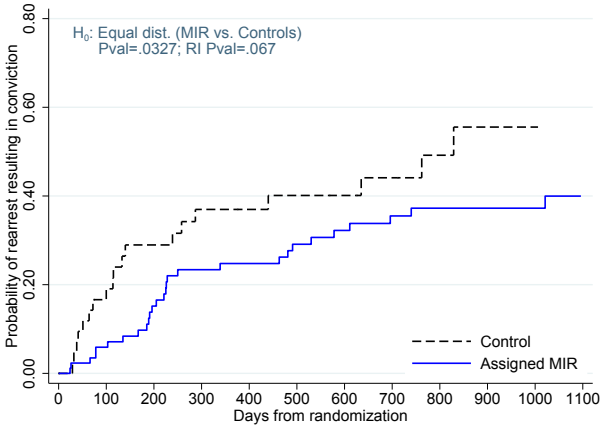


Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization between assignment to Make-it-Right (MIR) and the control group which faces traditional criminal prosecution. Recidivism is measured using only rearrests for new criminal incidents. Specifically, rearrests for probation or warrants violations will not be included in this recidivism measure. We report p-values from an hypothesis test for whether the failure functions are the same among MIR participants and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "RI Pval" which is based on randomization inference (Fisher, 1935) using 1,000 simulations in which we randomly assigned cases to MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that we would conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

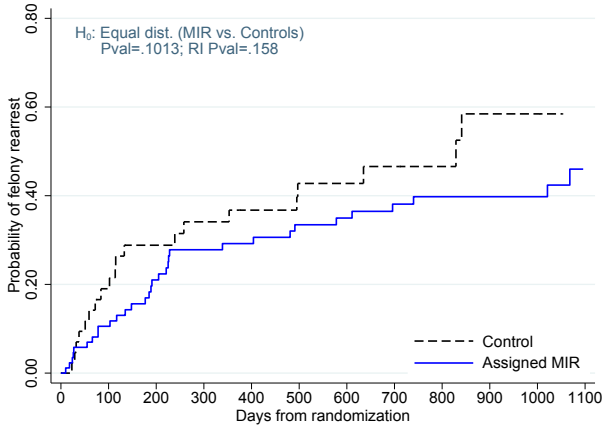
Figure A.7: Rearrest Rates of Juveniles Randomly Assigned to Make-it-Right Relative to the Experimental Control Group When Not Including Any Reoffending That Took Place After March 15, 2020 Which Is the Beginning of COVID-19 Restrictions in California



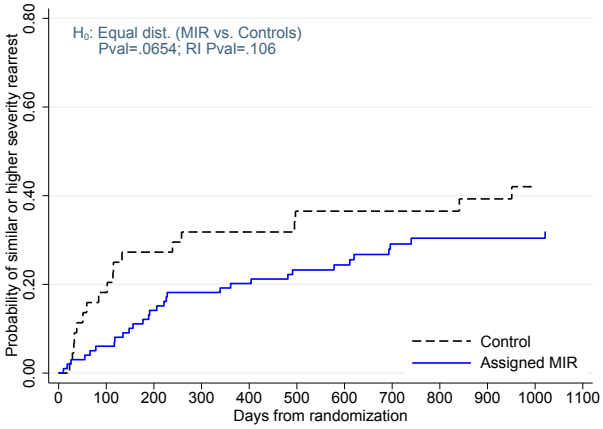
(a) Any rearrest from randomization



(b) A rearrest from randomization that leads to a conviction



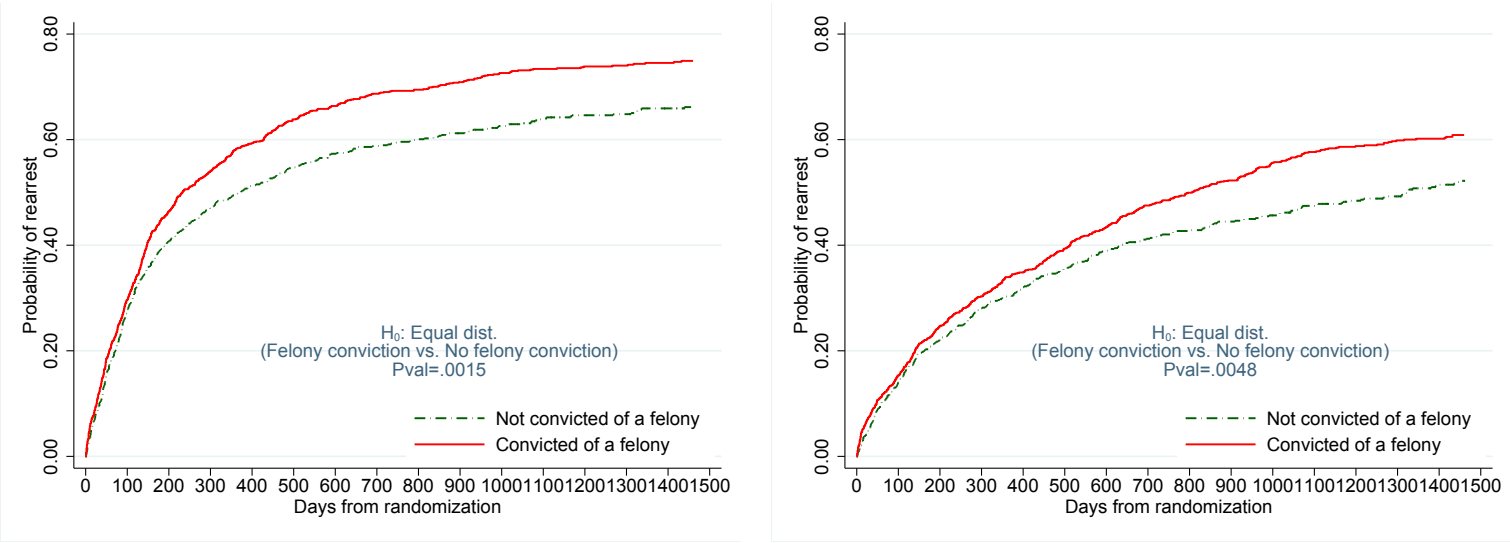
(c) Any felony rearrest from randomization



(d) Any rearrest from randomization for offenses at least as severe as the original offense

Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization into eligibility for Make-it-Right (MIR) or the control group which faces regular criminal prosecution. Panel (a) plots Kaplan-Meier estimates for any rearrest and Panel (b) only counts rearrests that lead to a conviction. Panel (c) plots rearrests for felony offenses. Lastly, Panel (d) presents failure function including only rearrests for offenses that are as severe as the original offense. In all plots we report p-values from an hypothesis test for whether the failure functions are the same among MIR participants and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "RI Pval" which is based on randomization inference (Fisher, 1935) using 1,000 simulations in which we randomly assigned cases to placebo MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier estimates of the failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that we would conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

Figure A.8: A Comparison of Recidivism Rates in the Full Population of Juveniles Charged with Felony Offenses in San Francisco Between Individuals Whose Case Resulted in a Felony Conviction Relative to Those Who Did Not

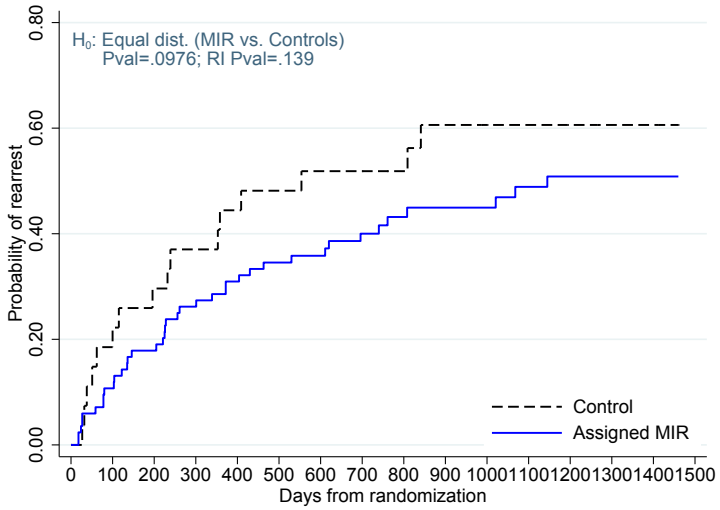


(a) A new arrest for any offense

A new arrest for a felony offense

Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date the offense took place. Each of the figures plots two failure function curves, one for individuals whose case resulted in a felony conviction (solid red line) and another for the individuals whose case did not result in a felony conviction (dashed green line). We report p-values from an hypothesis test for whether the failure functions are the same or not. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). The p-values in these hypothesis tests are two-sided since we did not pre-specified in the pre-analysis plan how they will be conducted.

Figure A.9: The Effects of Make-it-Right Among the Sub-Sample of Individuals Who Did Not Incur Neither a Felony Nor a Misdemeanor Conviction in Their Case Outcomes



Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization into eligibility for Make-it-Right (MIR) or the control group which faces traditional criminal prosecution. In this figure (unlike Figures 2 and 4), we restrict the sample to individuals whose case did not result in neither a felony nor a misdemeanor conviction. As can be seen from the figure, even among this restricted population the individuals who have been assigned to MIR have meaningfully lower rearrest rates. We report p-values from an hypothesis test for whether the failure functions are the same among MIR participants and controls. We use the Peto–Peto–Prentice test for equality of failure functions (for a detailed description see Klein and Moeschberger, 2006). We report two types of p-values: "Pval" which is based on standard variance formulas and "RI Pval" which is based on randomization inference (Fisher, 1935) using 1,000 simulations in which we randomly assigned cases to MIR and controls and calculated the distribution of the test statistic under the null of no treatment effect. Both p-values are one-sided, rejecting the null hypothesis that MIR is not preferable to the control regime when the MIR Kaplan-Meier failure curve dominates that of the control units. We pre-specified in the pre-analysis plan that we would conduct only one-sided tests that reject the null when MIR is preferable to the control regime in reducing recidivism.

Table A.1: Summary Statistics of the Demographic Composition of the Victim (Harmed Party) and the Youth (Responsible Party) in the Make-it-Right Experimental Sample

	(1) Victim (harmed party)	(2) Youth (responsible party)
Age	35.60	16.09
Sex:		
Male	0.585	0.889
Victim and youth of same sex	0.523	.
Missing sex	0.343	0
Race/ethnicity:		
Black	0.0820	0.531
Hispanic	0.148	0.323
White	0.443	0.0729
Asian	0.328	0.0938
Victim and youth of same race	0.213	.
Missing race	0.384	0

Notes: The table reports summary statistics (means) of the demographic characteristics of the youth (responsible party) who have been assigned to Make-it-Right (MIR) and the victim (harmed party) of the related criminal incidents. The demographic information for the youth comes from the administrative records provided to us by the San Francisco District Office and the San Francisco Department of Juvenile Probation. The demographic information for the victims comes from Community Works, which is the non-profit organization that implements MIR. As a result, we observe demographic information for only a subset of the victims.

Table A.2: Summary Statistics Comparing Individuals Who Completed Make-it-Right to Those Who Did Not Complete Make-it-Right Among the Youth Who Enrolled Into the Program

	(1) Did not complete MIR	(2) Completed MIR
Demographics:		
Male	0.893	0.904
Black	0.714	0.360
Hispanic	0.286	0.400
Age	15.86	16.25
Criminal history:		
Any past arrests	0.321	0.212
Age at first criminal offense	14.93	15.46
Assault	0.107	0.154
Burglary	0.429	0.519
Theft	0.714	0.712
Weapons	0.0357	0.0192
Other	0.107	0.385
Predicted recidivism:		
Pred. recidivism 6 months	0.323	0.271
Pred. recidivism 12 months	0.450	0.400
Pred. felony recidivism 6 months	0.208	0.163
Pred. felony recidivism 12 months	0.329	0.278
Number of individuals	28	52

Notes: The table reports summary statistics (means) of the characteristics of individuals who enrolled into the to the Make-it-Right (MIR) program and did not complete it (Column (1)) and of individuals who enrolled and completed the program (Column (2)).

Table A.3: The Effects of Assignment (ITT) to and Participation (TOT) in MIR on the Likelihood of Being Arrested in the Subsequent Four Years When Including Controls

	(1) 6 months	(2) 12 months	(3) 24 months	(4) 36 months	(5) 48 months	(6) 12-48 months
Panel (a)			*2SLS*			
Participated in MIR (treated)	-0.228	-0.212	-0.168	-0.170	-0.339	-0.345
	(0.102)	(0.108)	(0.126)	(0.143)	(0.168)	(0.199)
	{0.0136}	{0.0259}	{0.0922}	{0.1193}	{0.0239}	{0.0437}
	[0.0000]	[0.0110]	[0.0560]	[0.1280]	[0.0390]	[0.0310]
Predicted Y(0)	0.519	1.153	0.600	0.987	0.466	0.633
	(0.383)	(0.525)	(0.394)	(0.434)	(0.448)	(0.464)
Panel (b)			*Reduced form*			
Assigned to MIR (ITT)	-0.184	-0.171	-0.130	-0.126	-0.246	-0.254
	(0.084)	(0.089)	(0.100)	(0.111)	(0.134)	(0.155)
	{0.0151}	{0.0287}	{0.0993}	{0.1295}	{0.0353}	{0.0532}
	[0.0090]	[0.0380]	[0.1240]	[0.1590]	[0.0710]	[0.0830]
Predicted Y(0)	0.494	1.161	0.649	1.048	0.540	0.639
	(0.415)	(0.555)	(0.404)	(0.444)	(0.437)	(0.440)
Panel (c)						
First-Stage coefficient	0.809	0.808	0.774	0.743	0.727	0.735
	(.0461)	(.046)	(.0556)	(.0665)	(.0847)	(.0832)
Rearest rate among controls	0.432	0.568	0.632	0.750	0.833	0.667
Rearest rate among compliers controls	0.434	0.566	0.606	0.745	0.876	0.726
Includes controls	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	143	143	120	100	71	71

Notes: The table reports estimates of the effects of Make-It-Right (MIR) on the likelihood of a future arrest. Predicted $Y_i(0)$ is a summary index of the covariates based on how predictive they are on the outcome in the auxiliary observational data, in which none of the individuals where assigned to MIR. The construction of predicted $Y(0)$ based on the covariates is described in Section 4 and more formally in Appendix B. The only difference between this table and Table 3 is the inclusion of predicted $Y_i(0)$ in the regression specifications. Each cell in the table reports four numbers: the point estimate, standard error clustered at the case level, a one-sided p-value using the cluster-robust standard errors, and a one-sided p-value using randomization inference (Fisher, 1935) based on 1,000 random permutations. The compliers rearrest rates under the control regime (bottom of the table) are calculated using the standard formulas from Imbens and Rubin (1997) and Abadie (2003). Specifically, using a 2SLS regression of the outcome interacted with an indicator for enrollment into MIR (i.e., $(1 - \text{MIR}_i) \cdot \text{Rearrest}_i$) as the outcome, an indicator for not enrolling into MIR (i.e., $(1 - \text{MIR}_i)$) as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up rate is about 75% and is reported at the bottom of the table (i.e., the First-Stage coefficient). The number of observations changes across the columns because the sample in each of the regressions is restricted to individuals that are observed at least the mentioned time horizon (e.g., 48 months in Column (5)) after the date of randomization.

Table A.4: 2SLS Estimates of the Effects of Enrollment Into Make-it-Right on Rearrest Within Six Months and One Year for a Balanced Sample With and Without Cohort Fixed Effects

	(1) 6 months	(2) 6 months	(3) 6 months	(4) 12 months	(5) 12 months	(6) 12 months
Participated in MIR (treated)	-0.234	-0.233		-0.228	-0.212	
	(0.103)	(0.104)		(0.111)	(0.111)	
	{0.0120}	{0.0134}		{0.0211}	{0.0292}	
	[0.0000]	[0.0020]		[0.0080]	[0.0140]	
Participated in MIR (2013-2015 cohort)			-0.105			-0.175
			(0.215)			(0.205)
			{0.3136}			{0.1982}
			[0.2530]			[0.1150]
Participated in MIR (2016-2017 cohort)			-0.368			-0.314
			(0.157)			(0.181)
			{0.0103}			{0.0422}
			[0.0050]			[0.0670]
Participated in MIR (2018-2019 cohort)			-0.179			-0.107
			(0.171)			(0.173)
			{0.1485}			{0.2683}
			[0.0768]			[0.1800]
2016-2017 cohort		-0.059	0.074		-0.159	-0.088
		(0.088)	(0.172)		(0.096)	(0.172)
2018-2019 cohort		-0.024	0.000		-0.212	-0.267
		(0.099)	(0.196)		(0.104)	(0.187)
First-Stage coefficient	0.808	0.798		0.808	0.798	
	(.0463)	(.0509)		(.0463)	(.0509)	
Joint F-test of equal treatment effects in all cohorts			.551			.704
			[.912]			[.768]
Rearest rate among controls	0.432	0.432	0.432	0.568	0.568	0.568
Rearest rate among compliers controls	0.434	0.434	0.434	0.566	0.566	0.566
Includes controls	143	143	143	143	143	143

Notes: The table reports estimates of the effects of enrollment into Make-It-Right (MIR) on the likelihood of a future arrest. Importantly, the sample size does not change across columns as we restrict attention to shorter time horizons in which a balanced sample is observed. Each of the cells in the table reporting the effects of MIR includes four numbers: the point estimate, standard error clustered at the case level, a one-sided p-value using the cluster-robust standard errors, and a one-sided p-value using randomization inference (Fisher, 1935) based on 1,000 random permutations. In Columns (1) and (4), no covariates are included in the model. In Columns (2) and (5), fixed effects for the time period/cohort in which the incident took place. The omitted category is the 2013-2015 cohort, i.e., the cases for which we observe the longest follow-up period. Lastly, in Columns (3) and (6), we allow the effects of the MIR program to vary across cohorts. At the bottom of the table, we report a joint F-test for whether the effects of MIR are the same in all cohorts, and we cannot reject the null hypothesis of equality. In the square parenthesis, we also report a p-value for the joint F-test based on randomization inference. The compliers rearrest rates under the control regime (bottom of the table) are calculated using the standard formulas from Imbens and Rubin (1997) and Abadie (2003). Specifically, using a 2SLS regression of the outcome interacted with an indicator for enrollment into MIR (i.e., $(1 - \text{MIR}_i) \cdot \text{Rearrest}_i$) as the outcome, an indicator for not enrolling into MIR (i.e., $(1-\text{MIR}_i)$) as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up rate is about 80% and is reported at the bottom of the table (i.e., the First-Stage coefficient). Moreover, the take-up rate is not affected by whether or not cohort fixed effects are included in the model.

Table A.5: The Effects of Assignment (ITT) to and Participation (TOT) in Make-it-Right on the Likelihood of Being Arrested for a New Criminal Incident in the Subsequent Four Years

	(1) 6 months	(2) 12 months	(3) 24 months	(4) 36 months	(5) 48 months	(6) 12-48 months
Panel (a)			*2SLS*			
Participated in MIR (treated)	-0.244	-0.206	-0.197	-0.213	-0.440	-0.368
	(0.102)	(0.113)	(0.127)	(0.150)	(0.158)	(0.199)
	{0.0094}	{0.0351}	{0.0620}	{0.0791}	{0.0035}	{0.0344}
	[0.0000]	[0.0040]	[0.0401]	[0.0840]	[0.0020]	[0.0250]
Panel (b)			*Reduced form*			
Assigned to MIR (ITT)	-0.197	-0.167	-0.154	-0.160	-0.324	-0.270
	(0.083)	(0.093)	(0.102)	(0.118)	(0.127)	(0.154)
	{0.0097}	{0.0373}	{0.0681}	{0.0892}	{0.0069}	{0.0423}
	[0.0100]	[0.0520]	[0.0910]	[0.1360]	[0.0310]	[0.1040]
Panel (c)						
First-Stage coefficient	0.808	0.808	0.781	0.750	0.736	0.736
	(.0463)	(.0463)	(.0558)	(.0676)	(.0832)	(.0832)
Rearest rate among controls	0.409	0.500	0.605	0.719	0.833	0.667
Rearest rate among compliers controls	0.444	0.519	0.588	0.723	0.902	0.726
Includes controls	No	No	No	No	No	No
Number of observations	143	143	120	100	71	71

Notes: The table reports estimates of the effects of Make-it-Right (MIR) on the likelihood of a future arrest. Each cell in the table reports four number: the point estimate, standard error clustered at the case level, a one-sided p-value using the cluster-robust standard errors, and a one-sided p-value using randomization inference (Fisher, 1935) based on 1,000 random permutations. The only difference between this table and Table 3 is that here only rearrests for new criminal incidents are used. Specifically, rearrests for probation or warrant violations will not be included. The compliers rearrest rates under the control regime (bottom of the table) are calculated using the standard formulas from Imbens and Rubin (1997) and Abadie (2003). Specifically, using a 2SLS regression of the outcome interacted with an indicator for enrollment into MIR (i.e., $(1 - \text{MIR}_i) \cdot \text{Rearrest}_i$) as the outcome, an indicator for not enrolling into MIR (i.e., $(1 - \text{MIR}_i)$) as the endogenous treatment, and instrumenting using an indicator for whether the youth was randomly assigned to control or MIR. Note that not all individuals assigned to MIR took-up the program. The take-up rate is about 75% and is reported at the bottom of the table (i.e., the First-Stage coefficient). The number of observations changes across the columns because the sample in each of the regressions is restricted to individuals that are observed at least the mentioned time horizon (e.g., 48 months in Column (5)) after the date of randomization.

B Covariates Adjustment in an RCT Using Auxiliary Observational Data

We observe two samples. The first is an experimental sample in which individuals where randomly assigned to either Make-it-Right (MIR) or the control group. Denote by Z_i assignment to MIR (the treatment of interest), let Y_i^s be the outcome of interest, and denote by X_i^s a vector of pre-treatment covariates. The second sample is much larger, however, the treatment (Z_i) was not assigned to any of the individuals in this sample. Denote by Y_i^p and X_i^p the outcome and observable characteristics of individuals in the larger auxiliary sample. In our setting, this sample includes all juveniles not eligible for MIR who have been charged with felony offense between October 2013 and May 2019.

In the experimental sample, assignment to MIR is done at random, however, due to the small number of observations there can still be some imbalances between the observable and unobservable characteristics of the individuals who were assigned to the treatment and control groups. Specifically, the bias term can be expressed as:

$$\begin{aligned} &\mathbb{E}\left[Y_i^s|Z_i=1\right]-\mathbb{E}\left[Y_i^s|Z_i=0\right]= \\ &\mathbb{E}\left[Y_i^s(1)-Y_i^s(0)|Z_i=1\right]+\mathbb{E}\left[Y_i^z(0)|Z_i=1\right]-\mathbb{E}\left[Y_i^s(0)|Z_i=0\right] \end{aligned} \tag{B.1}$$

It is clear from Equation (B.1) that if we observed $Y_i^s(0)$ among both the control and treated units than we could control for it and correct for any potential finite sample imbalances.

It is common practice to use Ordinary-Least-Square (OLS) regression and estimate the following specification:

$$Y_i^s=\alpha Z_i+X_i^{s\prime}\beta^s+e_i^s \tag{B.2}$$

this model corrects for potential imbalances in observables between the treatment and control groups and with flexible/saturated enough controls it is completely non-parametric (i.e., it does not require making any parametric assumption such as linearity of the conditional expectation

function). Another motivation for controlling for X_i^s is increasing the statistical precision by improving the explanatory power of the OLS model.

The big challenge in Equation (B.2) is that increasing the dimensionality of X, i.e., including a greater number of covariates, entails a reduction in the number of degrees of freedom. Reducing the degrees of freedom can be costly in experiments with small sample sizes as is the case in our setting of the MIR program. Moreover, including only a subset of the potential Xs raises the question of which covariates to include and adds another "researcher degree of freedom". To overcome this problem, we use the auxiliary data on all juveniles charged with a felony offense between October 2013 and May 2019 in San Francisco.

We begin by using auxiliary observational data to derive an estimator of $\widehat{X_i^{s\prime}\beta^s}$. In both the experimental and observational samples, we observe the same vector of observable characteristics. We estimate the following OLS specification in the auxiliary data:

$$Y_i^p = X_i^{p\prime}\beta^p + e_i^p \tag{B.3}$$

next we use the estimated coefficient $\hat{\beta}^p$ to form our estimator of $X_i^{'\beta^s}$:

$$\hat{Y}_{0i} \equiv X_i^{s\prime}\hat{\beta}^p \tag{B.4}$$

and now we can estimate the following model:

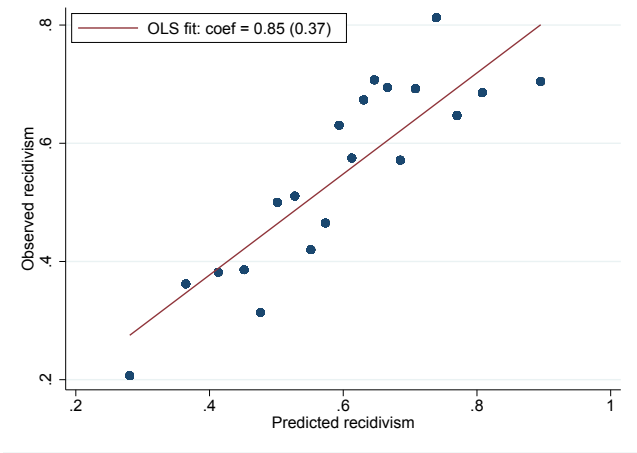
$$Y_i^s = \alpha Z_i + \gamma \hat{Y}_{0i} + \nu_i^s \tag{B.5}$$

where $\nu_i^s = (\gamma \hat{Y}_{0i} - X_i^{s\prime}\beta^s) + e_i^s$. Note that, if $\hat{\beta}^p \approx \beta^s$, then $\nu_i^s = e_i^s$ and the specification in Equation (B.5) yields the same results as the one in Equation (B.2) while using only one degree of freedom since only a *single* covariate is included in the model.

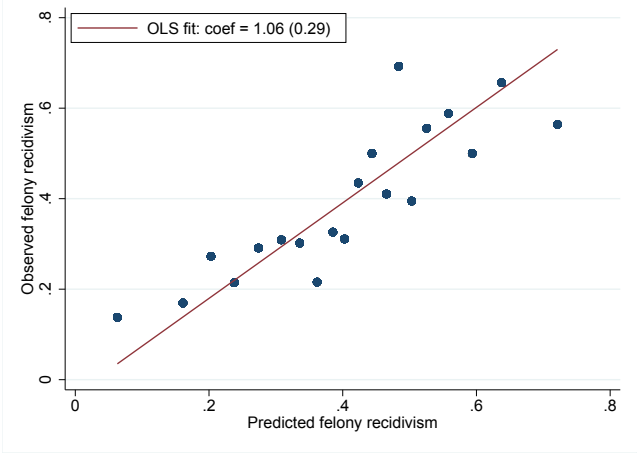
To validate that our predicted recidivism index ($X_i^{s\prime}\hat{\beta}^p$) is predictive of recidivism in the experimental sample, we examine the relationship between Y_i^s and $X_i^{s\prime}\hat{\beta}^p$. Figure B.1, depicts the relationship between the predicted and observed recidivism in the experimental sample. It is clear

that our predicted recidivism index is predictive of observed recidivism and the correlation is close to one. To obtain more power, we aggregated observed and predicted recidivism from multiple time horizons of 6, 12, 18, 24, 30, 36, 42, and 48 months from randomization.

Figure B.1: The Correlation Between Observed and Predicted Recidivism in the Experimental Sample



(a) Any rearrest



(b) Felony rearrest