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

Yotam Shem-Tov   Steven Raphael   Alissa Skog

August 3, 2023

Abstract

This paper studies the effect of a restorative justice intervention targeted at 143 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, Researcher, 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 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, Johanna Lacoe, and Santiago Tobón for helpful comments and discussions. We also thank many seminar and conference participants. Chandni Raja, Shivani Ghate, and 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, the Regents of the University of California, the California Policy Lab, or CPL's advisory board. We thank Arnold Ventures, the University of California Office of the President Multicampus Research Programs and Initiatives: M21PR3278, The James Irvine Foundation, and the Woven 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.

Historically, criminal justice policy in the United States has relied on sanctions to enforce compliance, with much policy debate and research focusing on sanction efficacy (e.g., Kuziemko, 2013; Aizer and Doyle, 2015; Bhuller et al., 2020; Rose and Shem-Tov, 2021). Restorative justice conferencing is an alternative that emphasizes accountability through repairing harm rather than imposing sanctions. While restorative justice has limited implementation in the U.S., it is a key component of juvenile justice in New Zealand (Ministry of Justice, 2004) and Australia (Little et al., 2018; Strang et al., 2013). Restorative justice programs typically involve a structured conference of the victim and the person accused, leading to a formal agreement through which the accused takes responsibility for their actions and commits to making amends. The current evidence regarding the effectiveness of restorative justice programs in reducing recidivism is mixed (Wilson et al., 2018).

This paper studies the "Make-it-Right" (MIR) program, a restorative justice conferencing intervention implemented by the San Francisco District Attorney (SFDA). The program targets teenagers who would otherwise face felony charges. Eligible cases were randomly assigned to either a treatment group where they were given the opportunity to participate in MIR, or a control group subjected to regular 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 study 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 caused a large reduction in recidivism that persists up to at least 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 that 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 83 percent are rearrested within four years of treatment assignment. Despite the eligibility restrictions for MIR (e.g., restrictions based on criminal history, gang affiliation, and offense), the MIR control group's rearrest rate is similar to that of the entire population of juveniles charged with a felony in San Francisco. Hence, the pilot program did not cherry pick easy-to-serve youth.

We find that MIR substantially reduced future arrests. Youths assigned to MIR were 19 percentage points less likely to be rearrested within six months of randomization (a 44 percent reduction). Moreover, the effects persisted years after randomization. Those assigned to MIR were 15 percentage points (20 percent) less likely to be rearrested within three years and 27 percentage points (32 percent) less likely after four years. Juveniles assigned to MIR were also less likely to be subsequently arrested for both new misdemeanor as well as new felony offenses. MIR youth were also less likely to be convicted for a future offense. Among those assigned to MIR, 81 percent enrolled, while 53 percent completed the program. Accounting for imperfect take-up of treatment makes the effects larger by roughly 1.23 times.

While there is a broad literature on the determinants of criminal behavior and its responses to various incentives (e.g., punishment severity or the economic rewards of the crime), little is known about whether criminal behavior can be influenced by less punitive interventions that use self-reflection and foster empathy to directly change a youth's decision-making.

Recent work shows that cognitive-behavioral therapy (CBT) changes an individual's decision making, reduces criminal behavior, and increases years of schooling (Heller et al., 2017; Blattman et al., 2017, 2022). Our analysis complements these findings by highlighting alternative levers that may impact decision making and offending among juveniles at high risk for arrest. Similar to CBT, restorative justice aims to foster greater deliberation and reduce impulsive behavior that may cause harm to others. However, while CBT aims to slow down choices and provide heuristic tools to deescalate risky situations and avoid quick reaction, restorative justice interventions go further and appeal to one's sense of responsibility to self and to others (and perhaps to one's sense of shame associated with causing harm) and explicitly aims to foster empathy for crime victims.

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 (Cuellar et al., 2006; Mitchell et al., 2012; Seward et al., 2021; 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. (2022) 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. Consistent with this evidence, Agan et al. (2021) finds that even for non-violent misdemeanor offenses, not being prosecuted leads to lasting reductions in recidivism.

Although related, restorative justice conferencing is fundamentally different than standard diversion programs as it presents an alternative model for addressing the harms caused by a criminal incident. The conference gives the victim an active and prominent role. Our findings suggest that fostering dialogue (and perhaps empathy) between the victim and the accused can lead to meaningful long-term reductions in recidivism.

1 Restorative Justice: Background and Evidence

Restorative justice has its origins in a theory of shaming first articulated by Braithwaite (1989). Braithwaite posits two forms: (1) stigmatic shaming that associates the offense with the offender, and (2) reintegrative shaming that condemns the action, but offers the person a path back into the community. Modern programs emphasize restoration over shaming and all require that the person who commits harm take responsibility for their actions and confront the consequences for the victim. The process is intended to engender greater empathy for the person harmed as well as to make the accused more mindful of the impact of their actions. Existing research generally finds that restorative justice minimizes the trauma suffered by victims (McCold and Wachtel, 1998; McGarrell, 2001; Angel et al., 2014; Sherman et al., 2015).

The evidence on the impact of restorative justice interventions on recidivism is mixed. The Australian experiments reviewed by Sherman et al. (2015) find some evidence of a reduction in repeat offending, especially for offenses that involved a victim. However, there is also some evidence in sub-group analyses of increased offending. An important feature of these experiments is that randomization occurred only among offenders who agreed to participate, leading to a selected sample. Evidence from the U.S. and Canada is similarly inconclusive (Bonta et al., 2002; Brooks, 2013).

The best evidence in the U.S. comes from two randomized control trials (RCT) from the 1990s. McGarrell (2001) and McGarrell and Hipple (2007) evaluate an RCT of a restorative justice program in Indiana focused on children arrested for a first-time offense (average age is 12.5). While they find meaningful decreases in recidivism within one year relative to a control group assigned to other diversion programs, the effects are short-lived and fade out over time (Jeong et al., 2012). McCold and

Wachtel (1998) evaluate a family-group conference intervention in Pennsylvania and find low take-up (42 percent) and no evidence of reductions in recidivism after one year.[1]

These experimental studies are more relevant to our evaluation of MIR than the non-U.S. studies reviewed in Sherman et al. (2015). Similar to the MIR intervention, randomization was done without conditioning on the youth's willingness to participate. These RCTs were also small in scale, relying on roughly 300 study participants. Aside from these similarities, the intervention we study differs along a number of likely relevant dimension. In McCold and Wachtel (1998), conferencing was conducted by *police officers*, the take-up rate was low, and it involved youth charged with infractions and misdemeanors. In McGarrell (2001), youth were especially young and engaged in less severe offenses. Furthermore, the alternative to restorative justice was one of 23 potential alternative diversion programs, making it difficult to determine the program's effects as the counterfactual regime is unclear. The interventions in these two studies were considerably less intensive, both in terms of pre-conference preparation and follow-up measures after the conferences. Lastly, both McCold and Wachtel (1998) and McGarrell (2001) took place in the 1990s in a meaningfully different environment and criminal justice system.

2 The Make-it-Right Program

The SFDA piloted MIR at the end of 2013. The pilot program diverted youth arrested for certain felony offenses to a restorative justice conferencing program. Conferencing involves a 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 were randomized to either the treatment or control groups after the juvenile division prosecutor decided

[1] In 2017, a group of master's students from the Goldman School of Public Policy at UC Berkeley wrote an internal report for the San Francisco District Attorney's Office on the MIR program. The report focused largely on reviewing comparable experiments and offered very preliminary and underpowered outcome comparisons. Our analysis was completed independently, based on separately procured data extracts, and a completely independent procedure of data pre-processing and analysis. We read the report only after completing our empirical analysis. It is hard to compare the estimates between the two studies for several reasons. First, the data used in Huntington et al. (2017) contained misspelling mistakes in the names of youth leading to inaccuracies in the calculation of rearrest rates. Second, the rearrest rates did not include incidents in the adult system.

to file charges, but before charges were formally filed. Thus, all the individuals in the control group faced criminal charges. Youth assigned to MIR who did not complete the program also automatically faced criminal charges. The SFDA does not file criminal charges (or impose any other sanction) against youth who successfully complete MIR.

Eligibility criteria. MIR targets juveniles 13 to 17 years of age charged with medium-severity felony offenses such as vehicle theft, grand theft, burglary, or assault. Youth were only eligible if the prosecutor determines the case is chargeable and would, in lieu of the program, prosecute the case. In addition, the eligibility criteria include the requirement that the youth must reside in San Francisco or Northern Alameda County. Additionally, they must not have any prior arrests or sustained petitions for juvenile offenses that would qualify as a strike under California's Three-Strikes law. The youth must not have caused significant injuries to the victim, must not be affiliated with a gang, cannot have used a weapon during the commission of the offense, and cannot be under probation supervision or in detention at the time of offense.[2]

MIR was intended to be a relatively small pilot: the SFDA expected to enroll no more than 25 individuals per year. However, the number of eligible youth and enrollees was lower even than expected, likely due to the steady reduction in juvenile crime in San Francisco. The experiment lasted 5.5 years and included 99 treatment group and 44 control group members when it concluded in May 2019. The experimental subjects constituted 13 percent of juveniles charged with a felony during this period.

Randomization Procedure. Randomization occurred at the case level, which corresponds to individuals except for cases involving co-defendants.[3] The process was designed to separate the actual treatment assignment from the decision to file

[2] These requirements ensure that eligible youth are not involved in any other pending cases and, therefore, are not on probation or in detention. Additionally, if a youth is arrested for a new offense, their participation in MIR will be halted, or the new offense will be merged with the original one for which they were assigned to MIR.

[3] Based on our interviews with the juvenile prosecutor, if one of the youths agrees to participate in MIR and the other does not, then one continues to MIR while the other faces regular felony prosecution. The choice of conducting the randomization at the case level was due to fairness consideration rather than a constraint that the entire case needs to be in only the treatment or the control regimes. Among the 143 participants in MIR, 13 percent of cases involved multiple individuals and youth from 115 unique criminal cases participated in the experiment.

charges. Once a case was deemed eligible (effectively after a charging decision was made), it was sent to a paralegal not involved with the program and unrelated to the juvenile prosecutor. The paralegal, through consultation with the SFDA policy director, consulted a prepared list of assignments, selected the next available assignment, and then communicated the results back to the prosecutor.[4] We confirmed with the juvenile division prosecutor that she never overrode the allocation communicated to her by the paralegal.[5]

The MIR program. Table 1 provides a concise description of the MIR program, its activities, and each of its steps. Once assigned to treatment, Community Works (CW) assesses the youth's ability to participate. Unsuitable cases are referred back to the SFDA for felony prosecution. Importantly, youth face no additional punishment for failing to enroll, and the reason for unsuitability is never disclosed to the SFDA. An essential requirement for participation is demonstrating reflection and accountability for one's actions.[6] Minors and their parents may decline to participate, effectively opting for standard case adjudication.

The harmed party must provide their consent before the responsible party can be offered MIR. Over the course of the experimental pilot *all* victims consented. However, not all victims agreed to take part in the restorative justice conferencing themselves. In such cases, surrogate actors filled this role.[7] Even if the victim does

[4] The randomization was carried out using referral lists of ten assignments (e.g., treatment, control, control, treatment, etc.) created using Excel's randomization function. The SFDA's Director of Policy generated the list of assignments and a second person (the Executive Secretary) consulted the list to assign treatment (MIR) or control (felony prosecution). When a new case was eligible, a paralegal would contact the Executive Secretary. If it was the first case in that randomization list, the Executive Secretary would check if the first assignment in the list (i.e., row #1) was assigned treatment or control and then relay the assignment to the paralegal. The process would continue as such; she would go down the list until the tenth referral came in. Then the Policy Director would generate a new randomization list, and the process would start again from line one. The Policy Director and the Executive Secretary were the only people with access to the randomization list.

[5] Initially, the SFDA randomly assigned 50 percent of eligible individuals to MIR. In May 2014 (six months after the pilot date), the assignment probability to treatment was increased to 70 percent due to a lower-than-expected number of eligible cases. As we discuss in the online appendix, including cohort fixed effects yields almost identical estimates. Moreover, removing the observations from the earlier periods also yields similar results.

[6] The requirement that the youth will recognize the harm in their actions and be willing to assume responsibility for them is essential to prevent revictimization of the victim during the restorative justice conferencing.

[7] During the pre-conference planning sessions, the youth is informed that a surrogate would take part in place of the actual victim. In order to explain to the youth the effects their acts had on the victim, a surrogate is typically either someone who has experienced a similar type of crime in the

not participate in the conference, they can still express how the incident has affected them by sending an impact statement or talking with the surrogate or the conference coordinator, who then relays the information to the youth.

In cases that proceed to conferencing, CW conducts 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 conference (such as parents and/or supporters of the harmed and responsible parties). Moreover, CW also mediates the conference. Pre-conference preparations are time-intensive. The conference coordinator spends about five hours preparing the youth and two to three hours with the victim and family members of the youth. In addition, the day prior to the conference, there is an hour preparatory meeting with each conference participant. The conference itself lasts about two hours. MIR has a more intensive process than other restorative conferencing models studied in past experiments in the U.S. For example, McCold and Wachtel (1998) report that the average conference lasted 43 minutes. Appendix Figure B.1 depicts the physical settings of a typical restorative justice conference. Post-conference case management and compliance monitoring are managed by the Huckleberry Youth's Community Assessment and Resource Center. Youth who fail to follow through with the program have their cases referred back to the SFDA for prosecution.

Juveniles in the control group continued through the traditional juvenile justice process. They were charged and prosecuted in juvenile court and supervised by the Juvenile Probation Department (JPD) during the process. While some minors were detained, most were 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 youths 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 enrolled in the program or were deemed unsuitable (e.g., when refusing to assume responsibility for their share in the incident or unable to participate due to parental refusal). The take-up rate was high, especially relative to previous U.S. experiments, with 80.8 percent of those assigned to MIR enrolling in the program. The higher take-up rate may reflect the fact the MIR enrolled youth charged with relatively serious offenses (all felonies) and that the alternative

past or a family member or close friend of the actual victim.

to participation is felony prosecution. The accused youth's suitability and willingness to participate in the program were the main causes of the imperfect take-up. In our sample, factors related to the victims did not lead to reduced take-up.

Among those enrolled in MIR, 66.7 percent completed the program (53 percent of those assigned to MIR). The average and the median duration of the program (time between enrollment and completion) was six months.

3 Data and Summary Statistics

Our evaluation draws upon three data sources. First, we were provided programmatic information on all youth who were part of the experiment. This information set included group assignment, and for those in the treatment group who enrolled in MIR, key dates associated with enrollment, program participation, completion, or failure to complete. Second, we received data on the universe of juvenile arrests and the associated dispositions occurring in San Francisco between October 2010 and November 2020. Third, we received comparable data for adult arrests, permitting us to measure recidivism that occurs after turning 18.

Table 2 presents summary statistics for juveniles randomized under the MIR experiment and for all juveniles charged with a felony offense between October 2013 and May 2019. Columns (1) and (2) present average characteristics for youth assigned to the control and treatment groups. Column (3) reports the characteristics of those assigned to MIR and who also enrolled in the program. The final column describes all the juveniles charged with a felony offense during the time period of 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 generally yielded balance on observable characteristics. There are, however, a few instances of imbalances. Juveniles assigned to treatment were more likely to have a past felony arrest, for example. The square brackets ([$\cdot$]) in Column (2) report p-values for the null hypothesis that the averages in Columns (1) and (2) are equal. A joint F-test of the null hypothesis that all the differences are zero yields a p-value of 0.789. Moreover, except for one covariate, all the individual p-values indiciate that the observed differences in means are not significant at the five-percent level, and only three are significant at the ten-percent level. Importantly adjusting for any difference in covariates does not change any of our estimates as we discuss in Appendix A.2.

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).

Relative to the broader population of juveniles facing felony charges (Column (4)), MIR youth were more likely to be male and somewhat less likely to be Black or Hispanic. Moreover, they were 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 that reflect the MIR eligibility criteria. Over a third of non-MIR juveniles were arrested for serious person offenses (e.g., 38.7 percent for robbery) compared with roughly three percent of the youth assigned to the MIR treatment group. The proportion of all juveniles charged with an offense involving a weapon was also a higher relative to MIR youth. For MIR youth, the most common charge during the experiment was 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).[8]

Victim and accused characteristics. Appendix Table B.1 presents the demographic characteristics of the harmed parties and compares them to those of the accused.[9] The average age of the harmed party was 35, roughly double the average age of the accused youth. In 79 percent of the incidents, the harmed party was of another race/ethnicity than the responsible party. While most of the youth in the study were Black or Hispanic, most of the individuals in the harmed group were White or Asian. Females were also overly represented amongst the harmed parties, constituting 41 percent relative to 11 percent among the responsible youth.

[8] The charge proportions may sum to more than one because a single case can involve multiple charges.

[9] Victim data (although partial) was collected by CW as part of the restorative justice conferencing process, and provided to the research team for this analysis.

4 The Impacts of the MIR Program on Recidivism

We first present estimates of the effects of MIR on recidivism by comparing the rearrest rates across the treated and control groups. We then discuss the external validity of our results and various robustness analyses.

Throughout our analysis, we report two types of p-values. First, using the standard methods based on asymptotic approximations. Second, using randomization inferences (as was proposed by Young, 2019) with 1,000 permutations of cases to placebo MIR and control regimes to generate the sampling distribution under the null hypothesis. Following Chung and Romano (2013), we use as our test statistic the standardized t-statistic.

Finally, we specified in our pre-analysis plan that we will evaluate only one-sided hypothesis tests pertaining to whether the MIR program reduced the likelihood of recidivism. This choice was aimed to maximize statistical power given our small sample size. We submitted the pre-analysis plan before looking at the data and without knowing the exact sample size. Pre-analysis plans have been mentioned in the literature on research transparency and reproducibility as a tool to increase statistical power by pre-specifying one-sided hypotheses (e.g., Olken, 2015).[10] Despite this pre-specification choice, in our main results we report p-values from two-sided hypotheses tests.

4.1 Results

We begin by visualizing the difference in recidivism patterns between treatment and control group members. 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 for a four-year period.[11] Nearly half of the control group was rearrested within six months and over 70 percent were rearrested within four years. Rearrest rates among youth assigned to MIR were markedly

[10] 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 hypotheses in advance.

[11] The last randomization occurred in October 2019, and we observe recidivism data through November 2020. Hence, we have at least 14 months post-randomization for all youth, but longer periods for youth randomized in the early years. The empirical failure function implicitly assumes that the recidivism hazard is stable across cohorts defined by randomization date. We show the robustness of our results to this assumption in Appendix A.4.

lower. The difference in the percentage rearrested reaches roughly 20 percentage points within six months. It then fluctuates around this level for the remainder of the observation period.

Appendix Figure B.2 presents similar failure functions for future felony arrests, future arrests for offenses that were at least as severe as the original charges, and future arrests that resulted in a conviction. Assignment to MIR caused reductions in recidivism across all of these alternative measures.

We perform formal hypothesis tests for equality of the two cumulative failure functions using the standard Peto-Peto-Prentice test (Klein and Moeschberger, 2006). We reject the null hypothesis that the two failure functions are equal at the 5% significance level, with the p-value from the randomization inference ($p = 0.03$) roughly twice the value based on the standard non-parametric asymptotic inference ($p = 0.014$).[12]

Table 3 presents our main results for rearrests up to one year from treatment assignment. For each estimate, we report standard errors clustered at the case level (in parentheses), and the p-value from a two-sided hypothesis test using randomization inference (in angle brackets). To quantify the magnitude of the estimated effects, we compare them to the mean outcome in the control group and to the control complier mean (CCM),[13] both are reported at the bottom of the table. Column (1) reports the effect of assignment to MIR on the likelihood of enrolling in the program. As is evident from Figure 1, take-up was high. Youth assigned to MIR had an 81 percent likelihood of participating in the program.

Assignment to MIR reduced the likelihood of rearrest by 18.9 and 18.4 percentage points within the first six months and one year, respectively (Columns (2) and (3)). Relative to the CCM, these effect sizes imply a 44 and 32 percent reduction in recidivism, respectively. Turning to 2SLS estimates, the TOT effects are about 1.23 times larger than the ITT effects (Columns (4) and (5)).[14] Relative to the CCM, the TOT estimates are at least 40 percent of the recidivism occurring within the first

[12] This difference likely reflects the fact that the Peto-Peto-Prentice test does not take into account clustering at the case level.

[13] Comparing effect estimates to the CCM is commonly done in settings with non-compliance (e.g., Katz et al., 2001; Heller et al., 2017).

[14] The exclusion restriction required for 2SLS estimates to identify the TOT is likely to hold as referral 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.

year among the control group.

Appendix Table B.3 extends the results in Table 3 and reports effects over a longer-time period, up to four years from program referral. The overall effect sizes hold up over time. The effect on rearrest within three years of randomization was 14.7 percentage points (equivalent to 20 percent of the CCM), and the program effectiveness increased when examining impacts after four years (26.7 percentage points equivalent to 30 percent of the CCM). The final column reports effect estimates when measuring recidivism only during the period of one to four years post-randomization. This additional outcome allows us to assess whether referral to MIR impacted behavior beyond the period of program participation. Within one year, 99 percent of those assigned to MIR completed or failed to complete the program. The estimate in Column (6) indicates that assignment to MIR reduced recidivism between years one and four by 27 percentage points (equivalent to 37 percent of the CCM). While the sample is smaller when examining recidivism within longer periods than one year, in Appendix A, we examine the impact of changes in sample composition over time and find no evidence that they drive any of our longer-run effects for rearrests within four years or rearrests between one to four years.

Finally, we examine whether non-compliance (being assigned to MIR and not enrolling) is related to recidivism propensities. Appendix Figure B.3 presents the cumulative failure functions for youth assigned to the control group and those who were assigned to the treatment group but who 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 the control group members and control compliers are quite close to one another, suggesting that non-compliance is unrelated to recidivism propensities.

4.2 External Validity of the Estimated Effects

MIR restricted eligibility to youth charged with medium-severity felonies. One way to explore the generalizability of our findings to the broader population of youth facing felony charges is by comparing the rearrest rates of MIR control group members to those of the broader population. Figure 3 presents this comparison. Interestingly, the empirical failure function of all juveniles accused of a felony is remarkably similar to that of the control group, and we cannot reject the null hypothesis that the two curves are equal.

A potential explanation for the finding is that some of the common types of offenses eligible for MIR have especially high rearrest rates. Appendix Figure B.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. Moreover, among robbery 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.

Scaling up MIR might also involve more or fewer cases with an actual victim instead of a surrogate. The rearrest rates among youth who completed the program are similar regardless of whether the actual victim or a surrogate attended the conference (Appendix Table B.2). Understanding the causal effect of using a surrogate rather than the actual victim is an important question for future research.

4.3 Robustness Analyses

Appendix A reports results from additional robustness analyses. We first discuss potential concerns that might arise in an experiment with a relatively small sample size and the report results from the tests proposed by Gelman and Carlin (2014) to highlight that our findings are not sensitive to common concerns. Second, we show that our results are robust to adjusting for observable covariates. Third, we show that our estimates are not sensitive to including or excluding arrests due to probation violations. Fourth, we discuss how changes in sample composition over time might impact our long-run estimates (e.g., rearrest after four years) and provide evidence that the long-run estimated effects are not driven by changes in sample composition and are capturing valid causal effects. Moreover, we also show that our estimates are robust to the inclusion of various time-cohort controls. Lastly, we show our results are also robust to excluding data from the period of COVID-19 (i.e., truncating the observation period for all youth to end on March 15, 2020).

5 Concluding Remarks

While our evaluation of the MIR program provides clear evidence that restorative justice conferencing can lead to persistent reductions in rearrests, there are key questions for future research pertaining to how restorative justice operates and whether eligibility can be expanded.

Offense severity. Understanding and quantifying heterogeneity in the effects of restorative justice conferencing is important for our understanding of how such interventions operate and for scaling up efforts. The MIR pilot program treated youth charged with more serious offenses that are typically not eligible for diversion. It may be the case that there is simply more opportunity to reduce the likelihood of future arrests for juveniles charged with serious offenses than is generally understood. Moreover, interventions targeted at less serious offenses may simply widen the net of the criminal justice system and apply an unnecessarily invasive intervention.

Can MIR be as effective for even more severe offenses? The answer is unknown. Given the relatively small scale of the experiment, we did not evaluate whether effect size varies by offense severity. The victim's willingness to participate might also vary with the nature of the offense. Assessing victim willingness to participate both in terms of approving the alternative process as well as in actual conferencing (as opposed to delegating a surrogate) is also an important topic for future research. A related question concerns the effect of a victim's refusal to participate on the youth. It is unclear whether being denied access to a restorative justice process due to victim unwillingness leads to worse outcomes. For example, being rejected by the victim might harden the youth and diminish empathy for others.

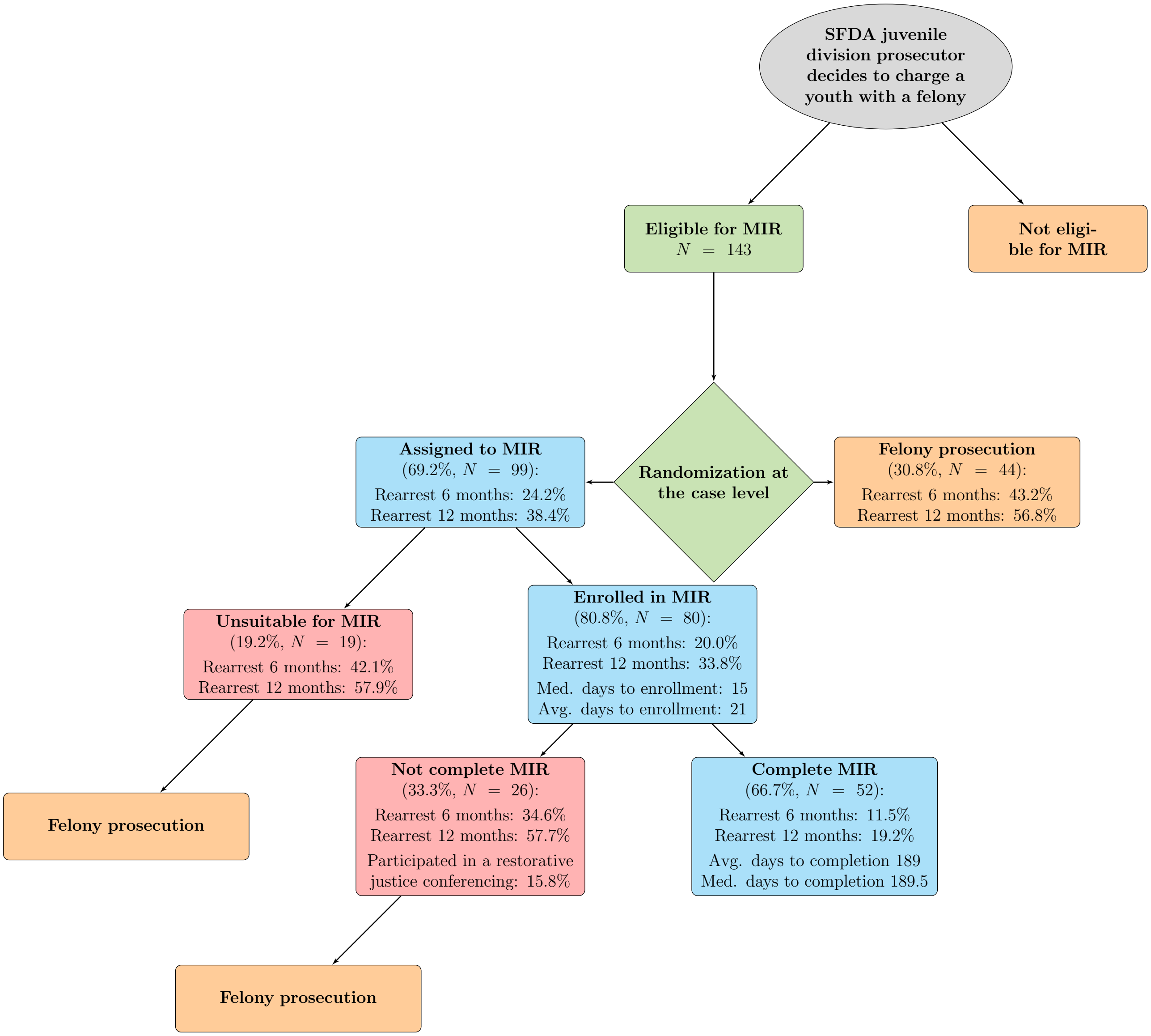
Separating the effects of diversion from those of restorative justice conferencing. MIR is both a restorative justice conferencing program and a diversion program. It diverts youth after the prosecutor decides to file charges, but before charges have been formally filed. As a diversion program, MIR influences case outcomes, for example, by reducing the likelihood of getting a felony conviction. An important and open question is what are the effects of restorative justice conferencing separately from its diversion component. Would MIR be as effective among youth who are not prosecuted? Among youth who are convicted and sentenced to a specific punishment? In a previous working paper draft (Shem-Tov et al., 2021), we presented some preliminary and suggestive analyses on this topic. However, these questions are very much open and should be addressed by future studies that are better designed and powered to answer them.

Separating the effects of conferencing with the victim from the impact of post-conference servicing. As we noted in the discussion of MIR, the intervention requires a substantial investment of time both before and after conference. Table 2 lists activities that may be included in the formal agreement between the harmed party and the youth. These activities could include financial restitution payments, commitments to attend school and contribute to household tasks, and engagement in programming intended to reduce future offending.[15] Notably, a separate service provider oversees the implementation of the agreement, meeting frequently with participating youth, and has the authority to refer non-compliant youth back to the SFDA for prosecution. Since all youth assigned to the treatment group who successfully complete a conference receive the same post-conference monitoring, we can not estimate the independent effects of the conference and the post-conference monitoring with our study design. This is an interesting question to address in future research.

[15] Many of the youth in the control group also get referred to various programs to reduce recidivism with the same service provider. Hence, it is unlikely that the programs available to the treatment group influenced our results.

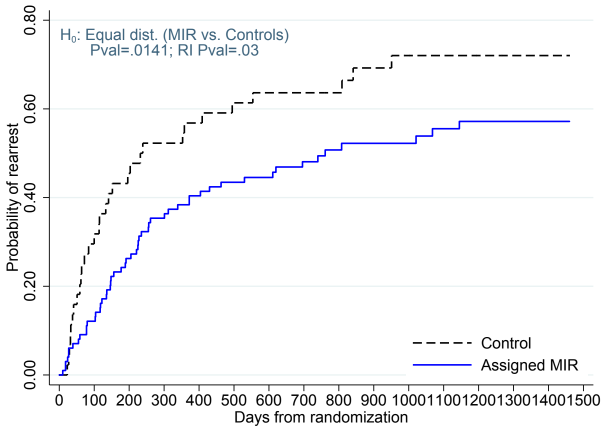
Figures and Tables

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


SFDA juvenile division prosecutor decides to charge a youth with a felony
Eligible for MIR $N = 143$
Not eligible for MIR
Randomization at the case level
Assigned to MIR (69.2%, $N = 99$): Rearrest 6 months: 24.2% Rearrest 12 months: 38.4%
Felony prosecution (30.8%, $N = 44$): Rearrest 6 months: 43.2% Rearrest 12 months: 56.8%
Unsuitable for MIR (19.2%, $N = 19$): Rearrest 6 months: 42.1% Rearrest 12 months: 57.9%
Enrolled in MIR (80.8%, $N = 80$): Rearrest 6 months: 20.0% Rearrest 12 months: 33.8% Med. days to enrollment: 15 Avg. days to enrollment: 21
Felony prosecution
Not complete MIR (33.3%, $N = 26$): Rearrest 6 months: 34.6% Rearrest 12 months: 57.7% Participated in a restorative justice conferencing: 15.8%
Complete MIR (66.7%, $N = 52$): Rearrest 6 months: 11.5% Rearrest 12 months: 19.2% Avg. days to completion 189 Med. days to completion 189.5
Felony prosecution

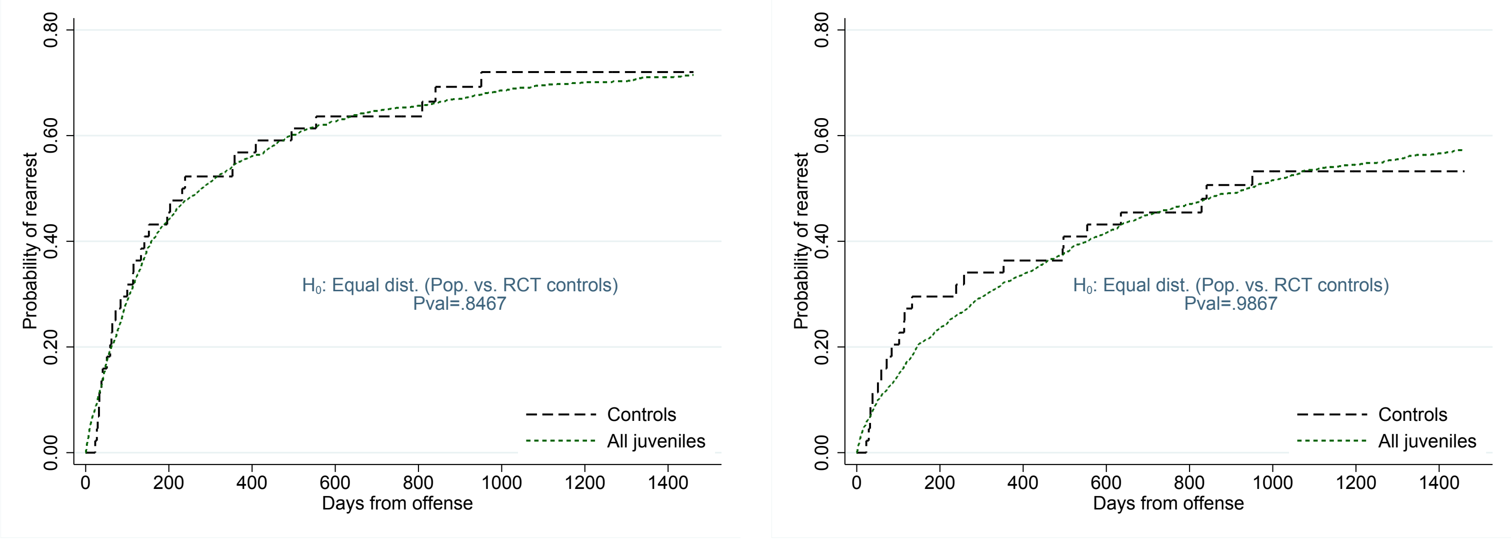
Notes: This figure depicts the process through which youths are assigned to the Make-it-Right (MIR) program.

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 reported p-values are from two-sided hypothesis tests.

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



(a) Future arrest

(b) Future felony arrest

Notes: This figure plots Kaplan-Meier estimates of the failure function of being rearrested for any offense (Panel (a)) and for any felony offense (Panel (b)). 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 reported p-values in these hypothesis tests are two-sided.

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

Stage	Description/Examples of Activities
Suitability Assessment	CW coordinator holds 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 - typically between three to four - pre-conference meetings with the **responsible youth** and their family/support person to prepare them for the conference. 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 harmed party 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 for the conference and help the harmed party understand 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 to apply to college or technical school. **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.

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

	(1)	(2)	(3)	(4)
	Feloy prosecution	Assigned MIR	Enrolled to MIR	All juveniles
	(control)	(treatment)	(compliers)	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
		[.]		
Robbery	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 average characteristics of compliers in Column (3) are calculated using the standard formula from Abadie (2003).

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

	First-Stage	Reduced-form		2SLS	
	(1)	(2) 6 months	(3) 12 months	(4) 6 months	(5) 12 months
Assigned to MIR (ITT)	0.808 (0.046) ⟨0.0000⟩	-0.189 (0.084) ⟨0.0300⟩	-0.184 (0.092) ⟨0.0630⟩		
Enrolled in MIR (TOT)				-0.234 (0.103) ⟨0.0010⟩	-0.228 (0.111) ⟨0.0070⟩
Rearest rate among controls		0.432	0.568	0.432	0.568
Rearest rate among compliers controls		0.434	0.566	0.434	0.566
Includes controls	No	No	No	No	No
Number of observations	143	143	143	143	143

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 three numbers: the point estimate, standard error clustered at the case level, and a p-value from a two-sided hypothesis using randomization inference (Fisher, 1935) based on 1,000 random placebo permutations of assignments to MIR. 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 81% and is reported at the bottom of the table (i.e., the First-Stage coefficient).

References

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

Agan, Amanda Y, Jennifer L Doleac, and Anna Harvey, "Misdemeanor prosecution," Technical Report, National Bureau of Economic Research 2021.

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.

Augustine, Elsa, Johanna Lacoe, Steven Raphael, and Alissa Skog, *The Impact of Felony Diversion in San Francisco*, Vol. 41 2022.

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad, "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 2020, *128* (4), 1269–1324.

Blattman, Chris, Sebastian Chaskel, Julian Jamison, and Margaret Sheridan, "Cognitive behavior therapy reduces crime and violence over 10 years: Experimental evidence," 2022. Working paper.

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.

Bonta, James, Suzanne Wallace-Capretta, Jennifer Rooney, and Kevin Mcanoy, "An Outcome Evaluation of Restorative Justice Alternative to Incarceration," *Contemporary Justice Review*, 2002, *5* (4), 319–338.

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

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

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.

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

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.

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.

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

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.

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

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.

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.

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.*

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.

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

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

Shem-Tov, Yotam, Steven Raphael, and Alissa Skog, "Can Restorative Justice Conferencing Reduce Recidivism? Evidence From the Make-it-Right Program," Technical Report, National Bureau of Economic Research 2021.

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, 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.

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

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.

Online Appendix

A Robustness Analyses

In this appendix, we discuss various additional analyses to demonstrate the robustness of our results to different decisions and show that the common concerns about experiments with relatively small sample sizes do not apply in this case.

A.1 Statistical Power

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

A.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. Appendix C describes in detail the covariate adjustment procedure.[16]

Appendix Table B.4 reports the results. The table is structured similarly to Appendix Table B.3 but also includes the coefficient on the predicted recidivism index, which is a weighted average of the pre-treatment covariates. The results are very close to those without any covariate adjustment. Thus, any finite sample imbalances in covariates do not impact our treatment effect estimates.

A.3 Robustness to the Inclusion of Arrests Due to Probation Violations

Our main measure of recidivism includes all rearrests including those that are the result of probation violations. To show our results are robust to the way recidivism is defined, we also report effect estimates using only arrests for a new criminal incident. Figure B.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 (Table B.6) are similar to those reported in Table B.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.

A.4 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

[16] This is the procedure used to calculate the predicted likelihood of new arrest averages presented at the bottom of Table 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).

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. Table B.5 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.

Last, Figure B.5 compares 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.

A.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-19 pandemic. 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).

Figure B.7 presents Kaplan-Meier estimates of the failure functions by treatment group where we have truncated the observation period for all youth to end on March 15, 2020.[17] Note, this truncation causes us to lose sample, especially for the observation periods beyond 16 months post-initial arrest. Nonetheless, the patterns

[17] 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.

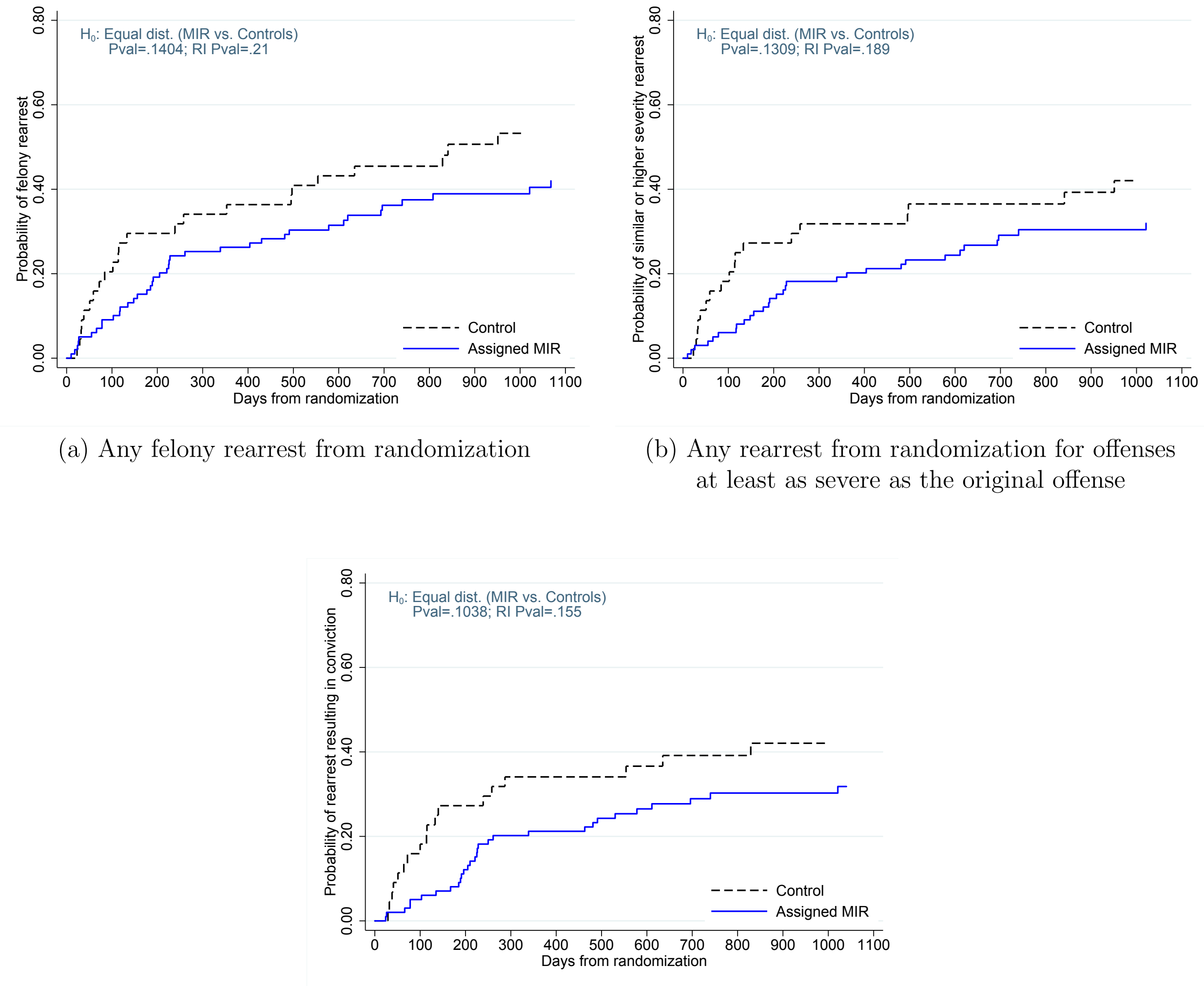
we observe here are similar to what we observe for the failure functions using untruncated observation periods. Large disparities in the failure functions open up soon after the initial arrest and persist throughout the observation period. Both inference strategies reject the null hypothesis of equal 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.

B Additional Figures and Tables

Figure B.1: Picture of a Make-it-Right Restorative Justice cCnference

Notes: This picture shows a typical Make-it-Right restorative justice conference, with people gathered in a circle of chairs. The picture is from Community Works's website, http://communityworkswest.org/restorative-justice-circles/.

Figure B.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



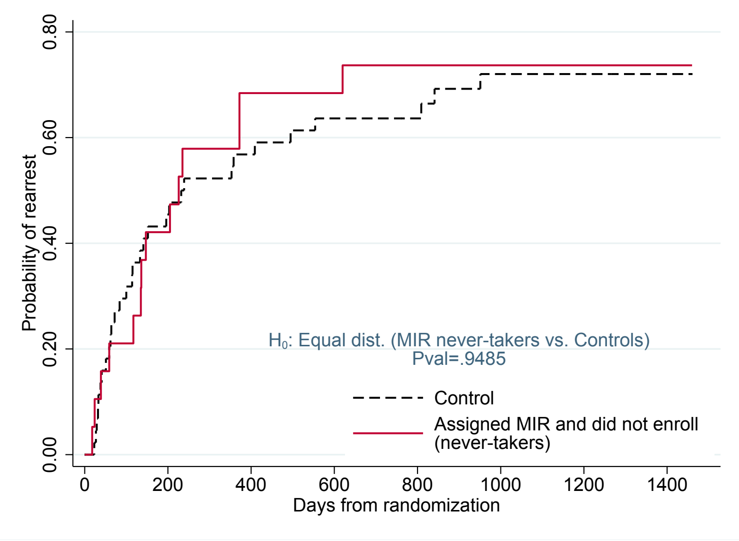
(a) Any felony rearrest from randomization

(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

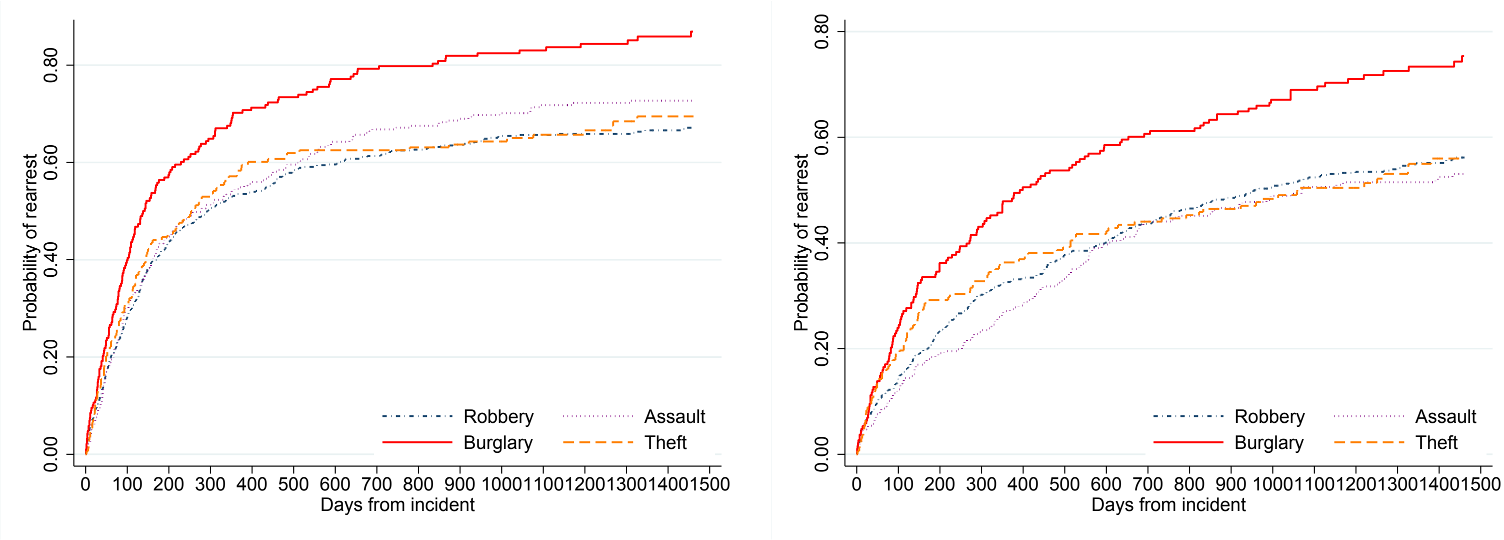
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date of randomization between assignment to Make-it-Right (MIR) and the control group which faces traditional criminal prosecution. Panel (a) plots Kaplan-Meier estimates for felony rearrest. Panel (b) for rearrests for offenses that are 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 a 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 from two-sided hypothesis tests.

Figure B.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 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 a 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-specify in the pre-analysis plan how they would be conducted.

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

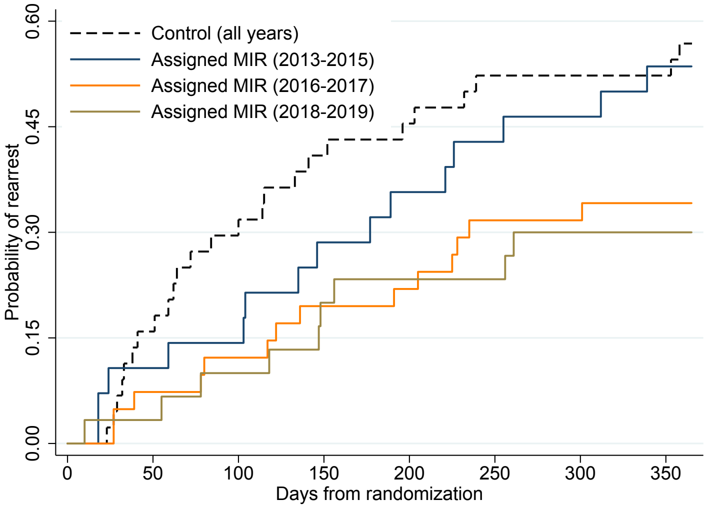


(a) A new arrest for any offense

(b) A new arrest for a felony offense

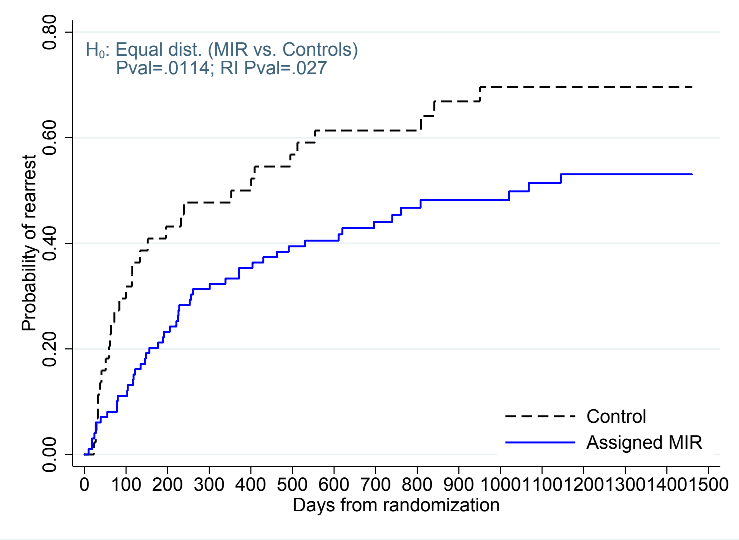
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within four years from the date the focal offense took place. Each of the figures plots four failure function curves, one for each of the four most common types of felony offenses in San Francisco.

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



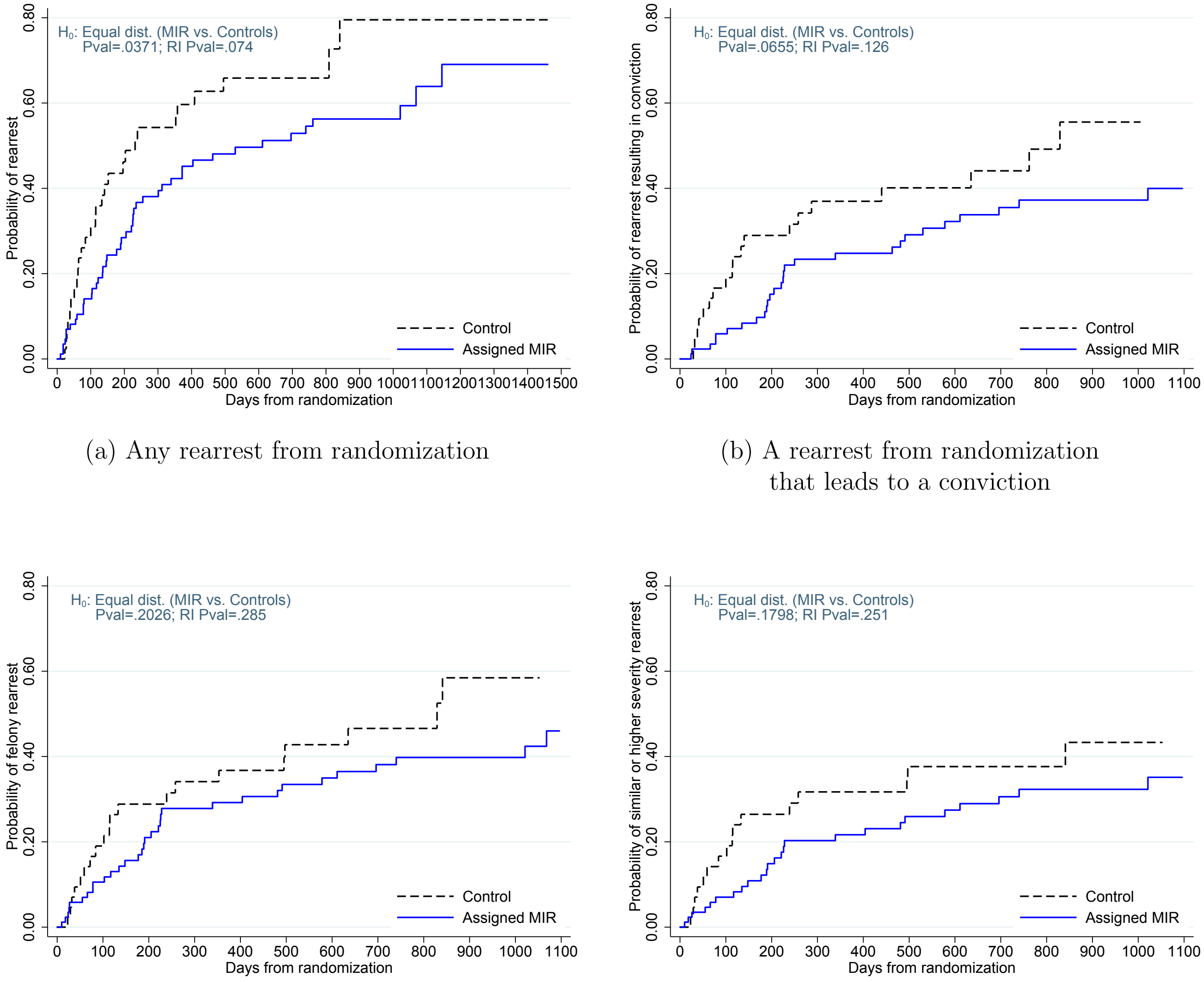
Notes: This figure plots Kaplan-Meier estimates of the failure function for being rearrested within one year from the date of randomization into eligibility for Make-it-Right (MIR) or the control group which faces traditional criminal prosecution. We observe at least one-year post-randomization for all 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. The data series "Control (all years)" includes only individuals in the experimental sample that is used when calculating the causal effects of MIR on rearrests.

Figure B.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 a 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 from two-sided hypothesis tests.

Figure B.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 (when COVID-19 restrictions began being implemented 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. Two-sided p-values can be obtained by multiplying the above ones by two.

Table B.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)	(2)
	Victim	Youth
	(harmed party)	(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 B.2: Summary Statistics of Youth Who Completed the MIR Program With Surrogate Victim Relative to the Actual Victim

	(1) Surrogate victim	(2) Actual victim
Rearrested within one year	0.190	0.227
Demographics:		
Male	0.857	0.909
Black	0.476	0.364
Hispanic	0.333	0.409
Age	16.38	16.09
Criminal history:		
Any past arrests	0.524	0.318
Any past felony arrests	0.476	0.227
Age at first criminal offense	15.62	15.14
Type of most severe offense:		
Assault	0.143	0.227
Burglary	0.524	0.455
Theft	0.714	0.682
Number of individuals	21	22

Notes: The table reports summary statistics (means) of one-year rearrest rates, demographic characteristics, criminal history, and offense types of youth who completed the Make-it-Right program with a surrogate victim (Column (1)) relative to the actual victim (Column (2)).

Table B.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)
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)
Panel (c)						
First-Stage coefficient	0.808	0.808	0.781	0.750	0.736	0.736
	(.0463)	(.0463)	(.0558)	(.0676)	(.0832)	(.0832)
Rearrest rate among controls	0.432	0.568	0.632	0.750	0.833	0.667
Rearrest rate among compliers controls	0.434	0.566	0.606	0.745	0.876	0.726
Includes control variables	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. Two-sided p-values can be obtained by multiplying the above ones by two. 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 B.4: 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)
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)
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)
Rearrest rate among controls	0.432	0.568	0.632	0.750	0.833	0.667
Rearrest rate among compliers controls	0.434	0.566	0.606	0.745	0.876	0.726
Includes control variables	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 Appendix C. The only difference between this table and Table B.3 is the inclusion of predicted $Y_i(0)$ in the regression specifications. Each cell in the point estimate and standard error clustered at the case level. 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).

Table B.5: 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)	(2)	(3)	(4)	(5)	(6)
	6 months	6 months	6 months	12 months	12 months	12 months
Participated in MIR (treated)	-0.234	-0.233		-0.228	-0.212	
	(0.103)	(0.104)		(0.111)	(0.111)	
Participated in MIR (2013-2015 cohort)			-0.105			-0.175
			(0.215)			(0.205)
Participated in MIR (2016-2017 cohort)			-0.368			-0.314
			(0.157)			(0.181)
Participated in MIR (2018-2019 cohort)			-0.179			-0.107
			(0.171)			(0.173)
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
			[.907]			[.764]
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 reports the point estimate of the effects of MIR and associated standard error clustered at the case level. 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.

Table B.6: 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)
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)
Panel (c)						
First-Stage coefficient	0.808	0.808	0.781	0.750	0.736	0.736
	(.0463)	(.0463)	(.0558)	(.0676)	(.0832)	(.0832)
Rearrest rate among controls	0.409	0.500	0.605	0.719	0.833	0.667
Rearrest rate among compliers controls	0.444	0.519	0.588	0.723	0.902	0.726
Includes control variables	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 the point estimate and standard error clustered at the case level. The only difference between this table and Table B.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).

C Covariates Adjustment in an RCT Using Auxiliary Observational Data

We observe two samples. The first is an experimental sample in which individuals were 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(s) 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:

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

It is clear from Equation (C.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:

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

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 (C.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 $X_i^{s\prime}\beta^s$. 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:

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

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

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

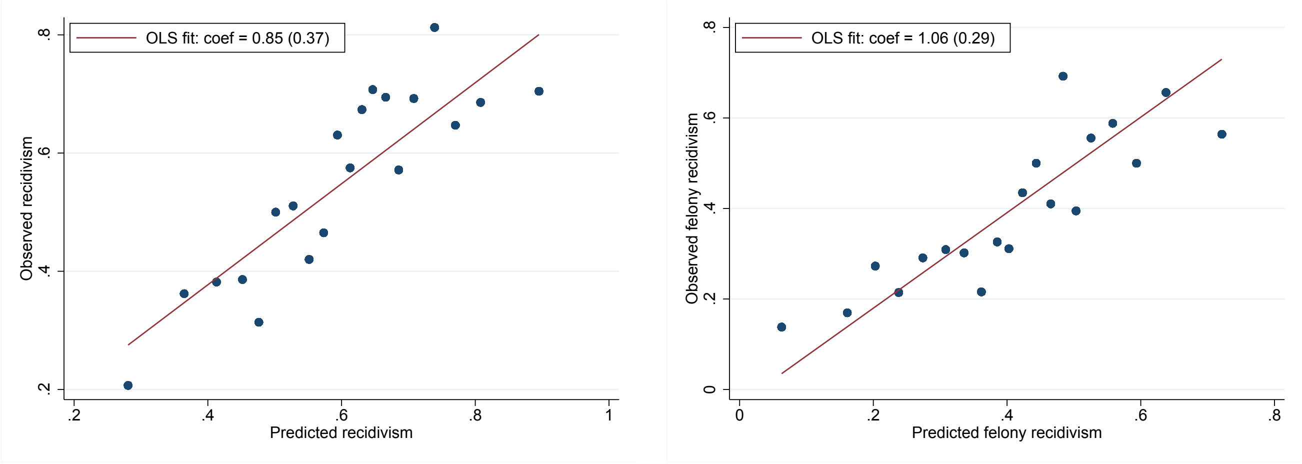
and now we can estimate the following model:

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

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 (C.5) yields the same results as the one in Equation (C.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 C.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 C.1: The Correlation Between Observed and Predicted Recidivism in the Experimental Sample



(a) Any rearrest

(b) Felony rearrest