Econometrica Supplementary Material

SUPPLEMENT TO "CAN RESTORATIVE JUSTICE CONFERENCING REDUCE RECIDIVISM? EVIDENCE FROM THE MAKE-IT-RIGHT PROGRAM"
(*Econometrica*, Vol. 92, No. 1, January 2024, 61–78)

YOTAM SHEM-TOV
Department of Economics, UC Los Angeles

STEVEN RAPHAEL
Goldman School of Public Policy, UC Berkeley and NBER

ALISSA SKOG
California Policy Lab, UC Berkeley

APPENDIX A: ROBUSTNESS ANALYSES

IN THIS APPENDIX, WE DISCUSS additional analyses that 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) mentioned 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.

Yotam Shem-Tov: shemtov@econ.ucla.edu
Steven Raphael: stevenraphael@berkeley.edu
Alissa Skog: alissaskog@berkeley.edu

https://doi.org/10.3982/ECTA20996

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 effect 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 include 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.[1]

Appendix Table B.IV reports the results. The table is structured similarly to Appendix Table B.III 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. Appendix 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 (Appendix Table B.VI) are similar to those reported in Appendix Table B.III. 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 are estimated using the earlier cohorts for which we have a longer time horizon to measure rearrests. While there is variation in the time horizon that we observe a youth for post-randomization, for all the youth in our sample, we observe recidivism within at least one year from randomization. Next, we focus on rearrests occurring within six months or one year, for which we have a balanced sample, and examine differences across cohorts. Appendix Table B.V reports 2SLS estimates with and without cohort fixed effects. The estimated effects are almost identical with and without the cohort fixed effects. Moreover, the first-stage coefficient also does not change by the inclusion of the cohort fixed effects. Columns (3) and (6) examine an even richer specification that allows the effect of MIR to vary by cohort. The estimates are the largest for the 2016–2017 cohort. To formally test for cross-cohort differences, we conduct a joint F-test. We cannot reject the null hypothesis that the program's impacts are the same in all three cohorts ($p = 0.55$ and $p = 0.7$ for rearrests within six months and one year). Thus, in our balanced sample, the effects of MIR are similar across cohorts.

Lastly, Appendix 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

[1] This is the procedure used to calculate the predicted likelihood of new arrest averages presented at the bottom of Table II. The idea of using auxiliary observational data to improve the accuracy of experimental estimates has been proposed in other studies (e.g., Gagnon-Bartsch, Sales, Wu, Botelho, Miratrix, and Heffernan (2020)).

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 observation period extends through the end of 2020. Hence, many of the youth in the study have later observation periods that overlap with the stringent stay-at-home orders in place in California (and the Bay Area in particular).

Appendix Figure 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.[2] Note, this truncation causes us to lose sample, especially for the observation periods beyond 16 months post-initial arrest. Nonetheless, the patterns we observe here are similar to what we observe for the failure functions using un-truncated observation periods. Large disparities in the failure functions open up soon after the initial arrest and persist throughout the observation period. Both inference strategies reject the null hypothesis of equal failure functions for the treatment and control groups when we focus on arrests for any new offenses. Similar patterns also emerge when examining effects on more severe interactions such as felony rearrest, rearrests for new offenses that are as severe as the original offense, or rearrests that lead to a conviction.

APPENDIX B: ADDITIONAL FIGURES AND TABLES

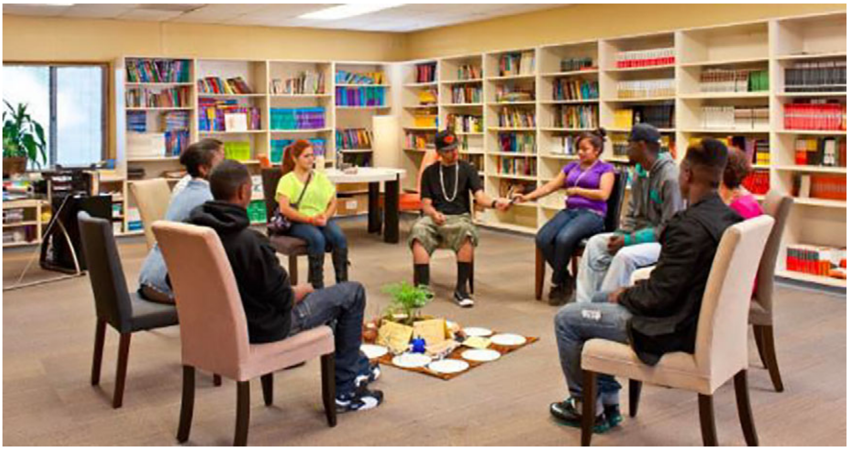
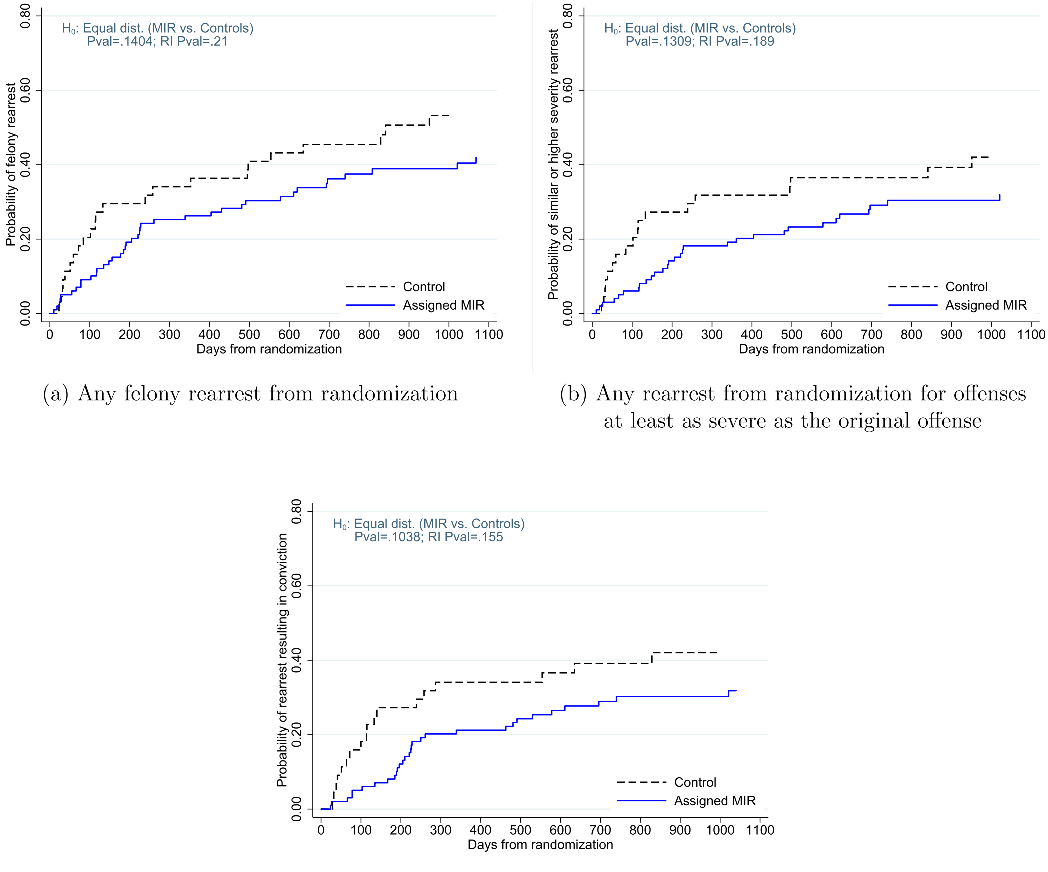


FIGURE B.1.—Picture of a Make-it-Right restorative justice conference. *Notes*: This picture shows a typical Make-it-Right restorative justice conference, with people gathered in a circle of chairs. The picture is from Community Works's website, http://communityworkswest.org/restorative-justice-circles/.

[2] On March 16, 2020, San Francisco and five other Bay Area counties enacted 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.



(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

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. *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 1000 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.

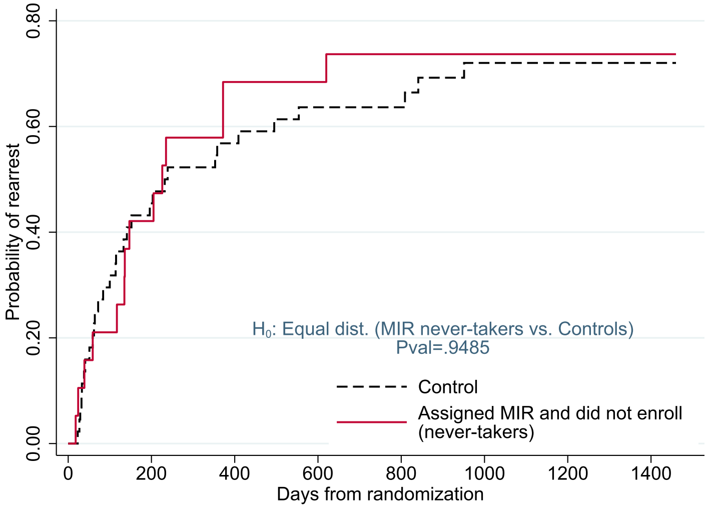
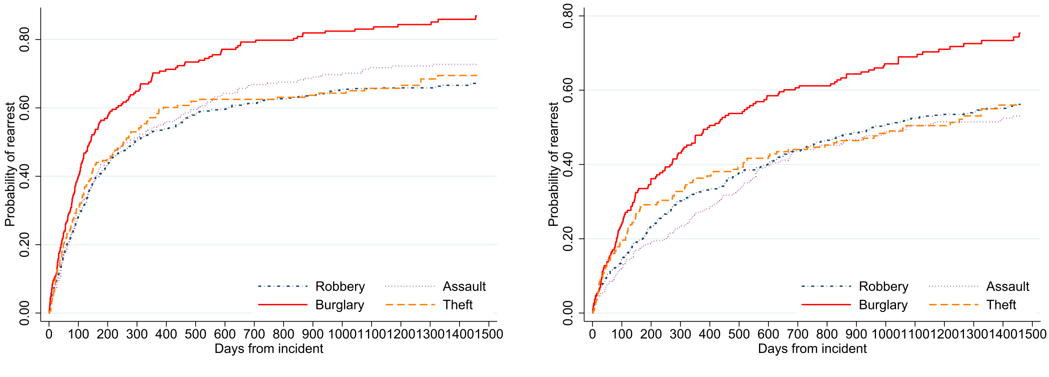


FIGURE B.3.—A comparison of recidivism rates between the control group and individuals assigned to Make-it-Right but who did not enroll ("Never-Takers"). *Notes*: This figure plots Kaplan–Meier estimates of the failure function of being rearrested. The figure compares between the failure curves of the 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, that is, 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 are from two-sided hypothesis tests.



(a) A new arrest for any offense

(b) A new arrest for a felony offense

FIGURE B.4.—A comparison of rearrest rates by the type of felony offense a juvenile is charged with. *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.

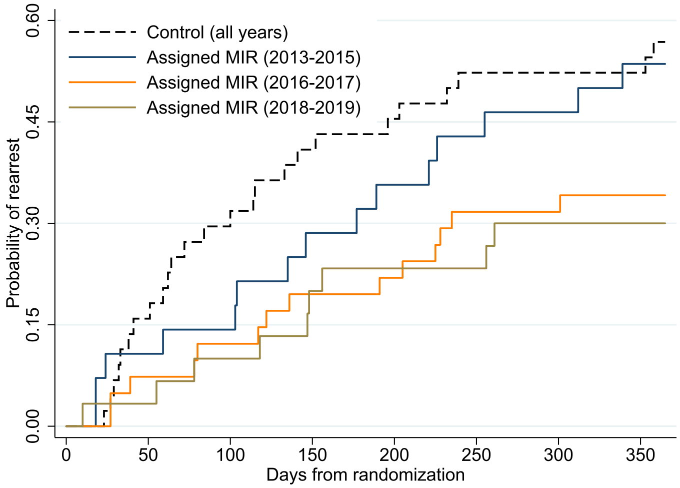


FIGURE B.5.—Rearrest rates of juveniles assigned to Make-it-Right in different time periods relative to the experimental control group. *Notes*: This figure plots Kaplan–Meier estimates of the failure function for being rearrested within one year from the date of randomization into eligibility for Make-it-Right (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.

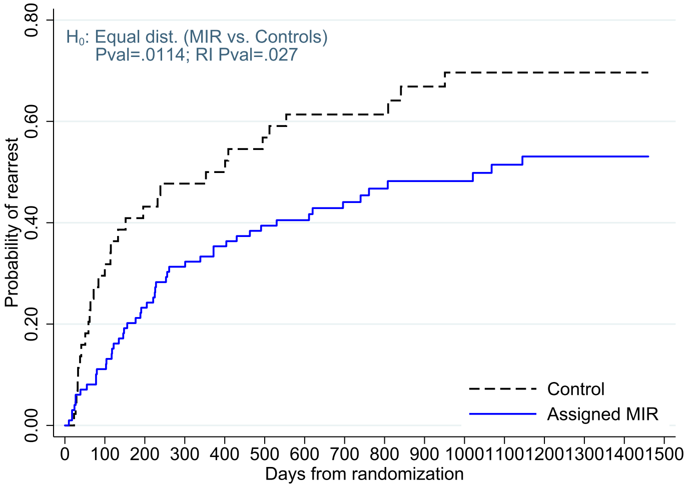
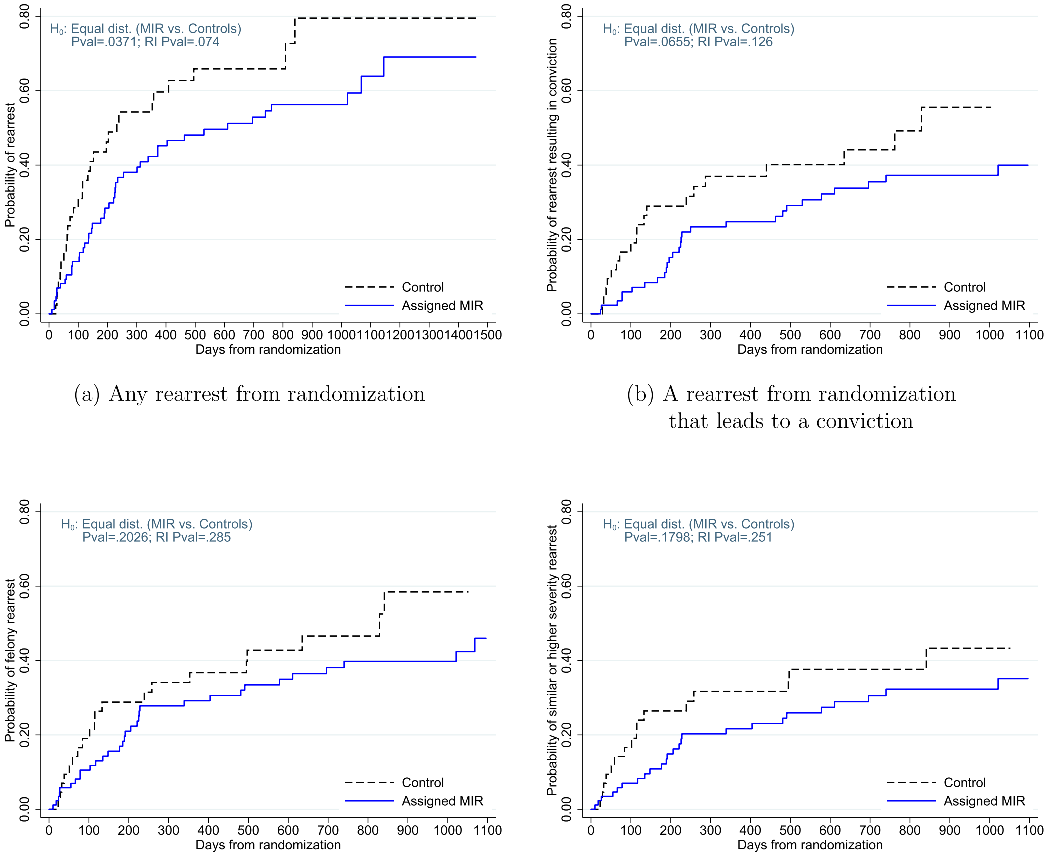


FIGURE B.6.—Recidivism rates of juveniles randomly assigned to Make-it-Right relative to the experimental control group using only rearrests for new criminal incidents. *Notes*: This figure plots Kaplan–Meier estimates of the failure function for being rearrested within four years from the date of randomization 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 1000 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.



(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

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). *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 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 1000 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 from two-sided hypothesis tests.

TABLE B.I

SUMMARY STATISTICS OF THE DEMOGRAPHIC COMPOSITION OF THE VICTIM (HARMED PARTY) AND THE YOUTH (RESPONSIBLE PARTY) IN THE MAKE-IT-RIGHT EXPERIMENTAL SAMPLE.

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

Note: 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.II

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

Note: 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.III

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
	(0.0463)	(0.0463)	(0.0558)	(0.0676)	(0.0832)	(0.0832)
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

Note: 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 1000 random permutations. Two-sided p-values can be obtained by multiplying the above ones by 2. 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.IV

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
	(0.0461)	(0.046)	(0.0556)	(0.0665)	(0.0847)	(0.0832)
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

Note: 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 were 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.III 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.V

2SLS ESTIMATES OF THE EFFECTS OF ENROLLMENT INTO MAKE-IT-RIGHT ON REARREST WITHIN SIX MONTHS AND ONE YEAR FOR A BALANCED SAMPLE WITH AND WITHOUT COHORT FIXED EFFECTS.

	(1) 6 Months	(2) 6 Months	(3) 6 Months	(4) 12 Months	(5) 12 Months	(6) 12 Months
Participated in MIR (treated)	−0.234	−0.233		−0.228	−0.212	
	(0.103)	(0.104)		(0.111)	(0.111)	
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	
	(0.0463)	(0.0509)		(0.0463)	(0.0509)	
Joint F-test of equal treatment effects in all cohorts			0.551			0.704
			[0.907]			[0.764]
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

Note: 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, that is, 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 brackets, we also report a p-value for the joint F-test based on randomization inference.

TABLE B.VI

The effects of assignment (ITT) to and participation (TOT) in Make-it-Right on the likelihood of beingaArrested 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
	(0.0463)	(0.0463)	(0.0558)	(0.0676)	(0.0832)	(0.0832)
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

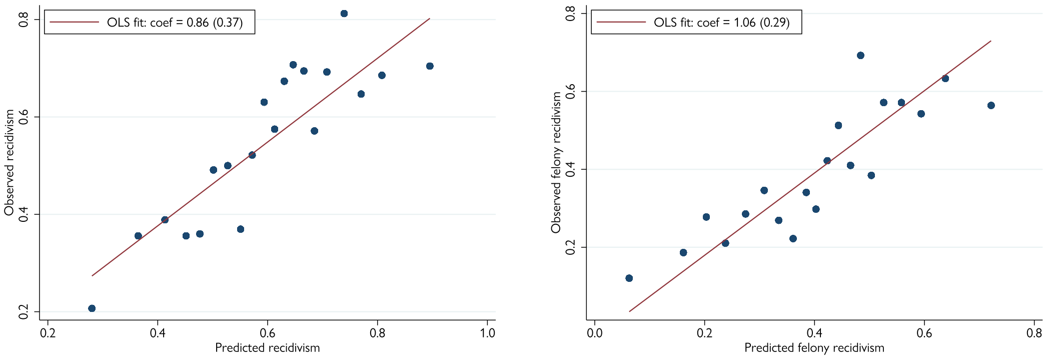
Note: The table reports estimates of the effects of Make-it-Right (MIR) on the likelihood of a future arrest. Each cell in the table reports the point estimate and standard error clustered at the case level. The only difference between this table and Table B.III is that here only rearrests for new criminal 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).

APPENDIX 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

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



(a) Any rearrest (b) Felony rearrest

FIGURE C.1.—The correlation between observed and predicted recidivism in the experimental sample.

It is clear from Equation (C.1) that if we observed $Y_i^s(0)$ among both the control and treated units, then 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 nonparametric (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, that is, 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 X's 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:

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

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

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

and now we can estimate the following model:

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

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 (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 $X_i^{s\prime}\hat{\beta}^p$. Appendix 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 1. 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.

REFERENCES

ABADIE, ALBERTO (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of econometrics*, 113 (2), 231–263. [9,10,12]

FISHER, RONALD (1935): *The Design of Experiments* (First Ed.). Oliver & Boyd: Edinburgh & London. [4,6,7, 9]

GAGNON-BARTSCH, JOHANN A., ADAM C. SALES, EDWARD WU, ANTHONY F. BOTELHO, LUKE W. MIRATRIX, AND NEIL T. HEFFERNAN (2020): "Precise Unbiased Estimation in Randomized Experiments Using Auxiliary Observational Data." [2]

GELMAN, ANDREW, AND JOHN CARLIN (2014): "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9 (6), 641–651. [1]

IMBENS, GUIDO, AND DONALD RUBIN (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64 (4), 555–574. [9,10,12]

KLEIN, JOHN P., AND MELVIN L. MOESCHBERGER (2006): *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media. [4-7]

Co-editor Oriana Bandiera handled this manuscript.

Manuscript received 20 July, 2022; final version accepted 16 September, 2023; available online 25 September, 2023.