DOI: 10.1111/jels.12311

ORIGINAL ARTICLE

JOURNAL OF EMPIRICAL LEGAL STUDIES

Intensified support for juvenile offenders on probation: Evidence from Germany

Christoph Engel[1] | **Sebastian J. Goerg**[2] | **Christian Traxler**[3]

[1]Max Planck Institute for Research on Collective Goods, Bonn, Germany

[2]Technical University of Munich, TUMCS for Biotechnology and Sustainability, Straubing, Germany

[3]Hertie School, Berlin, Germany

Correspondence
Christoph Engel, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, D 53113 Bonn, Germany.
Email: engel@coll.mpg.de

Abstract

This paper studies a probation program in Cologne, Germany. The program, which has a clear rehabilitative focus, offers intensified personal support to serious juvenile offenders over the first 6 months of their probation period. To evaluate the program's impact on recidivism, we draw on two research designs. Firstly, a small-scale randomized trial assigns offenders to probation with regular or intensified support. Secondly, a regression discontinuity design exploits a cutoff that defines program eligibility. The results suggest that the program reduces recidivism. The effect seems persistent over at least 3 years. Our evidence further indicates that the drop in recidivism is strongest among less severe offenders.

KEYWORDS

probation, randomized control trial, regression discontinuity design, recidivism, youth crime

INTRODUCTION

The distribution of crimes within Western populations is heavily skewed, with a large share of crimes committed by a fairly small number of individuals.[1] Hand in

[1]In 2016, in the US 3939.55 adults per 100,000 have been brought into formal contact with the criminal law system. In Germany, this number has been 2996.23. In the same year, in the US 670.3 persons per 100,000 have been in prison, while this rate was 78.22 in Germany (UNODC link, https://dataunodc.un.org/data/crime/Persons%20brought%20into%20formal%20contact).

hand with this pattern we typically observe very high recidivism rates: individuals who were convicted in the past are very likely to soon enter the criminal justice system again (Doleac, 2019). A crucial policy question, thus, concerns interventions that reduce recidivism.

Research has so far failed to identify promising strategies that effectively lower recidivism and coherently replicate in different contexts. On the contrary, the literature provides numerous null-results that fail to document any significant impact of prisoner reentry programs (Doleac et al., 2020), intensified supervision (Gendreau et al., 2000; Georgiou, 2014; Hyatt & Barnes, 2017), or intensified probation programs (Petersilia & Turner, 1993). The one noticeable exception is programs that prioritize treatment over deterrence, which have sometimes found to be effective (Lowenkamp et al., 2010).

The present paper provides evidence on a program with intensified support for young convicts in Germany. The program, which was independently developed by the regional Court of Cologne, offers a 6-month period with intensive support to high-risk youth criminals that are convicted to probation. The probation officers in this program work under a significantly reduced caseload, which allows for a swifter start of the probation period with a much higher contact frequency as compared to regular probation (RP). The program is particularly noteworthy for its courageous, almost grassroots-like approach. As they were concerned that juveniles might be put on a lasting criminal career when sent to prison, a single probation office has developed, on its own initiative, a program meant to bring them back on track, almost at the last minute. They had to do so with no additional resources, just relying on the enthusiasm of the probation officers, and on their collective sense of solidarity, as the leeway for the officers implementing intensive probation (IP) had to be provided by their colleagues shouldering an even higher workload in ordinary probation. Building on two research designs, we evaluate the program's impact on recidivism rates.

First, in cooperation with the local court and the probation office in Cologne, we conduct a small-scale randomized control trial (RCT). The trial randomly assigns serious juvenile offenders to probation with regular or "IP" (i.e., probation with intensified support). Second, we exploit that judges define program eligibility using a scorecard. Convicted offenders with a score below a certain cutoff are never assigned to IP. The cutoff allows us to implement a regression discontinuity design (RDD).

The results from the RCT indicate an imprecisely estimated decline in recidivism. Relative to a control group in RP, which has a 41% recidivism rate within 6 months after the start of the probation period, the cases assigned to IP have a 10 percentage point lower short-run recidivism. The effect slightly shrinks over time; however, a statistically insignificant but quantitatively meaningful 5–7 percentage points gap in recidivism remains during the second and third year after the start of the initial probation sentence. Our analysis further suggests that the treatment effect is mainly driven by a reduction in crime at the extensive rather than the intensive margin: the propensity to recidivate drops, but not the number of crimes.

The RDD delivers larger and more precisely estimated program effects. Mirroring the findings from the RCT, the decline in recidivism extends well beyond the initial 6-month period of the program. Note that the RDD identifies local average treatment effects (LATEs) at the lower end of the severity score of cases that are program eligible. Finding a larger LATE (as compared to the ATE from the RCT) suggests that the program might be more useful in preventing crimes among the relatively least problematic (but still fairly severe) offenders.

This is also one of the first studies that randomly assign an intervention within the German criminal justice system.[2] Germany is an interesting case in that, as a matter of judicial policy, incarceration rates are very low. As a matter of stated policy, retribution plays a very small role. Instead, resocialization and rehabilitative approaches are driving factors.

It is important to emphasize that both our RCT and the RDD exploit variation between probation with regular and intensified support. Our results are thus different but still complementary to a related strand of research which compares rehabilitative programs with harsher punishment or more or less severe forms of imprisonment (Bhuller et al., 2020; Lotti, 2020; Mastrobuoni & Terlizzese, 2019). This latter literature finds that harsher punishment can increase recidivism. Probation with intensified *support* also differs from programs with intensified *supervision* (Hyatt & Barnes, 2017), as the focus is not on increased supervision and drug screenings but on supporting convicts in avoiding circumstances that increase the likelihood of recidivism. Lastly, the context of the probation program differs from diversion programs preventing incarceration from being listed in criminal records and the stigma associated with it (Mueller-Smith & Schnepel, 2021). Also note that, in Germany, criminal records are not publicly accessible and are rarely considered in hiring processes.

The remainder of the paper is organized as follows. The next section discusses the literature on IP and puts the Cologne program into perspective. Sections "Research design" and "Data and samples" explain our research design and our data, respectively. Our results are presented in section "Results". After a critical discussion we conclude with a summary of our findings.

INTENSIVE PROBATION

Attempts at making probation more effective

All over the world, policymakers have explored intermediate sanctions that are less severe, and less costly, than incarceration, but more intense, and hopefully also more effective, than just releasing the convict on probation or parole (for overviews, see Cullen & Gendreau, 2000; Lipsey & Cullen, 2007; Lipsey &

[2]The only other RCT of which we are aware took place at about the same time, and tests the effect of electronic monitoring on recidivism (Meuer & Woessner, 2020). For earlier, non-experimental but quantitative evaluations of correctional programs in Germany, see Egg et al. (2000). For a hybrid between qualitative and quantitative analysis of socio-therapeutic prisons, see Lösel and Köferl (1989).

Wilson, 1998; National Research Council, 2007). Experiences have not been too encouraging, especially with respect to curbing recidivism.

The largest endeavor to assess the effectiveness of intermediate sanctions goes back to an initiative launched in Georgia in the 1980s (Erwin, 1986). All over the United States, policymakers were intrigued by the prospect of containing crime more effectively, and spending less money for the purpose, by using IP (for a review of the initiatives, see Petersilia & Turner, 1993). In 1986, the Bureau of Justice Assistance launched an initiative to evaluate the effectiveness of these programs and entrusted the Rand Corporation with the implementation. In multiple jurisdictions, more than 2000 convicts were evaluated (Petersilia & Turner, 1993, p. 292). Overall results were sobering. At no site convicts on IP were less often arrested than controls, they did not take more time before they first recidivated, they did not commit less serious offenses, and they were not less frequently convicted (Petersilia & Turner, 1993, pp. 310–312).

In all jurisdictions that participated in the evaluation exercise, convicts were randomly assigned to IP or ordinary probation (Petersilia, 1989). In the State of California, three counties participated (more detail from Petersilia & Turner, 1990a, 1990c, 1991). The most elaborate design was implemented in Ventura County. There, family and peer relationships were also evaluated. Convicts were put into the pool from which Rand randomly selected participants if they scored high enough on a risk assessment scale. The program targeted juveniles at medium risk of committing crime in the near future. For those treated the program took between 7 and 9 months. The program was evaluated a year later. However, only 57% of them participated in the interviews (Brank et al., 2008; Lane et al., 2005). There was no significant difference between treatment and control with respect to recidivism for ordinary crime, but those treated committed more technical violations (Lane et al., 2005; Petersilia & Turner, 1990a, 1991).

Contemporaneous experiments in Utah (Austin et al., 1990), Oregon (Petersilia & Turner, 1990b, 1993, 304 f.), with female drug offenders in San Francisco (Guydish et al., 2011), and with parolees in Texas (Turner & Petersilia, 1992, 1993, p. 307) did not find significant effects either. In Ohio, correlational evidence suggests a small positive effect (Lowenkamp et al., 2006). A clear positive effect was found in Philadelphia. Convicted juveniles were randomly assigned to either ordinary or intensive aftercare programs. Of 44 treated, 22 were rearrested within 9.9 months for misdemeanor or felony, while of 46 controls 34 were rearrested within 11.7 months. Eleven of the treated, and 19 of the controls, were rearrested for a felony (Sontheimer & Goodstein, 1993). Evaluations outside the United States did also not find a significant reduction in recidivism through various forms of IP. Programs failed in the United Kingdom (Folkard et al., 1974, 1976) and in Finland (Huttunen et al., 2014).

This sobering evidence resonates with the evaluation of other correctional interventions. A Californian program to assist prisoner reentry into society did not have a significant effect on recidivism (Farabee et al., 2014), nor did a

Dutch program assigning juveniles at risk to multisystemic therapy (Asscher, Deković, et al., 2014). Yet, results look brighter for other interventions. In general, behavioral/cognitive programs (Pearson et al., 2002) and programs aiming at the rehabilitation of adult offenders have been shown to be effective in reducing recidivism (Wilson et al., 2000), and interventions have successfully targeted truancy (Berg et al., 1977) and low performance in high school students (Rodriguez-Planas, 2012).

More recent evaluations of programs aimed at reducing recidivism without incarceration have sometimes had more success though. A meta-study of 43 reentry programs for incarcerated persons reports that 23 led to a significant reduction in recidivism, including a program on "intensive supervision (surveillance and treatment)", and a program on cognitive behavioral therapy for individuals classified as high or moderate risk (Bitney et al., 2017).[3] Programs have in particular been more successful if they were "treatment-oriented" (Drake et al., 2009, p. 184), prioritized "human service" over deterrence, and were executed in a spirit of integrity (Lowenkamp et al., 2010).[4]

Cologne's probation program with intensified support

Legal orders like the German one are skeptical about the benefits of incarceration. Many actors try to avoid prison sentences as long as possible. The German criminal code for juveniles (*Jugendgerichtsgesetz* [JGG]) adopts a hybrid approach between sanctioning, educating, and rehabilitating. Specialized juvenile courts have considerable discretion in defining what deems the appropriate reaction of the criminal justice system. One tool in their box is probation. Judges have considerable leeway in designing probation conditions. It is this discretion on which Cologne's probation project is built.

RP typically represents a relatively mild sanction with a low level of support and explicit or implicit supervision. Hence, most judges did not perceive RP as a suitable alternative to imprisonment for particularly severe youth crime. Yet, given concerns about the potential criminogenic effects from youth jails/prisons (Bayer et al., 2009; Stevenson, 2017) the courts try to keep juvenile offenders as long as possible out of prison. To accommodate juvenile offenders with relatively long criminal record at an early age, the Cologne regional court and the local probation office developed an alternative, "IP" program that offers intensified support (with the German term *ambulate Intensivbetreuung*). The 6-month program targets high-risk juvenile offenders who still qualify for a probation sentence but would be most likely incarcerated after any further criminal offense.

IP differs from RP in numerous dimensions. First, IP has a component of "swiftness" (Hawken & Kleiman, 2009): the first contact between the convicted

[3]Also see Wanner (2018).

[4]Also see Petersilia and Turner (1993, p. 315), reporting that offenders either employed during the year after reentry, or attending counseling sessions, performing community services, or paying restitution were 10%–20% less likely to recidivate. But they stress that these outcomes could result from selection.

offender and the assigned probation officer typically takes place within 1 or 2 weeks. Under RP, the first interaction with the probation officer might only occur after several weeks. Within our RCT sample introduced below, the probation office's e-documentation system reveals an average time gap between trial and a first personal meeting of 16 days in IP and 26 days in RP (two-sided *t*-test on mean, $p = 0.031$; $N = 49$).

Second, as indicated by its name, IP assures a more intensive and closer contact between officers and convicts. This point is documented in Figure 1. Juveniles in RP see their probation officer on average only once a month (they have on average 0.25 personal contacts per week). By contrast, convicts in the IP program have initially a weekly, personal meeting. If we include other forms of contact recorded in the probation office's e-documentation system (mostly phone calls but also text messages and letters), the difference in contact frequencies becomes even more pronounced: in IP, there are on average 3–4 weekly contacts over the first 4 months of probation; during the same time period, there are only 0.5–1.0 weekly contacts in RP (see bottom panel of Figure 1). The stark difference in interaction frequencies (in particular, in the frequency of personal meetings) shrinks over time but remains higher in IP than RP throughout the 6-month period of the program.[5]

The higher contact intensity clearly increases monitoring in IP. Probation officers are more closely following convicts' behavior and compliance with probation conditions. The main objective of IP, however, is not increased supervision but intensified support. Officers try to re-establish basic habits (starting with basic things like getting up at reasonably early hours) that facilitate (re-) integration. Officers try to make sure that convicts regularly attend school, go to job interviews or take part in anti-aggression therapy. In addition, IP officers provide help in many of these situations (e.g., they jointly go to appointments or support convicts in finding an apprenticeship or a housing opportunity).

When the program was started in 2006, the Cologne probation office did not get additional funds. It had to man the program with its regular workforce. Four probation officers were assigned to take over IP cases. IP officers work under a significantly reduced case load, typically handling 5 IP (plus 25 RP) cases at a time. Probation officers not assigned to the program shoulder a much higher caseload. Actually, they agreed to have their caseload further increased, to free up resources for the IP program. The probation officers not assigned to the program handle an average of 60 RP cases at a time.[6]

Based on anecdotal evidence, the judges at Cologne's juvenile court as well as the probation office perceived the program as a success. The common

[5]IP officially lasts for 6 months. Given that probation periods last longer (with a minimum of 12 months and modal probation period of 24 months), convicts return to regular probation (i.e., at a lower interaction frequency) but remain with the same probation officer. In practice, however, there is no exact cut after 6 months and the transition is more fluent (see Figure 1).

[6]Although there was some modest turnaround among officers before our sample period, the assignment to IP or RP remained constant during the evaluation. This implies that we will not be able to isolate officer specific effects.

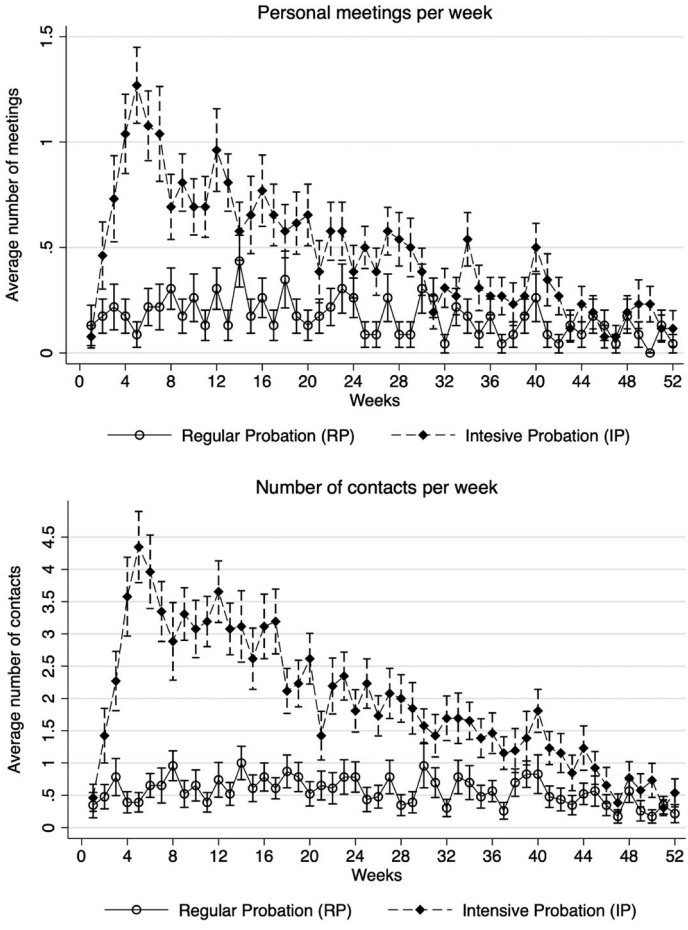

FIGURE 1 Contact between probation officers and offenders. The figure plots the weekly number of documented contacts between the probation officers and the convicted offenders for two groups who were randomly assigned to regular probation (RP; N = 26) or intensive probation (IP; N = 23). Bars represent SEs. The top panel shows personal, fact-to-face contact, the panel at the bottom includes also other forms of contact (mainly phone calls, but also text messages and letters). Data are based on the probation office's e-documentation system

impression was that the program would, at the very least, delay recidivism during the 6-month program period. In turn, this would contribute to a lower crime risk while aging and growing out of the high-crime youth period. The positive perception was also reflected in an excess demand for IP: juvenile courts asked for more slots in the IP program than could be offered by the probation office. This point will be reflected in our research strategy.

RESEARCH DESIGN

In 2010, the Cologne court approached us to evaluate the effectiveness of the program. In cooperation with the judges and the Cologne probation office, we developed an experimental research design. The core of the evaluation was a randomized controlled trial that built on judges' excess demand for available IP slots. In addition, we present evidence from an RDD that exploits a cutoff inherent in the definition of program eligibility.

Randomized treatment assignment

In close cooperation with the judges at the juvenile court and the probation office, we first developed a scorecard meant to evaluate juvenile convicts (see Appendix D). The scorecard captured the severity of the criminal history, the level of aggressiveness, drug and alcohol problems, as well as the convicts' family, housing, and schooling situation. The scorecard further included a set of "exclusion criteria." If a judge ticks a box that indicates at least one reason for exclusion (e.g., severe drug problems), a convict is not eligible for the program.

Judges agreed to filling out the scorecard for any juvenile convicted for a probation sentence during the evaluation period. Similar to Petersilia and Turner (1990c), we thus relied on judges' competence for defining which convicts are eligible for the program. Every convict with 13 or more points (and no exclusion reason, see above) on the scorecard was considered eligible for the program. Among this eligible population, we randomly assigned convicts to IP or RP.

Randomization took place in typically bi-weekly intervals. Before a randomization date, probation officers reported open IP slots to the research team. At the same time, the youth judges submitted scorecards for all new probation cases.[7] A member of the research team then entered the information from the scorecards into our database and, in the presence of a judge, used a simple randomization software to determine outcomes. The randomization procedure assured that all eligible convicts (who were in the pool at a given point of time) had an equal opportunity to be assigned to the program. An eligible convict would enter IP if (a) randomization assigned the convict to IP and if (b) at least one IP slot was available in this draw.[8] Otherwise, the juvenile convict would enter RP. In case of program assignment, the IP probation officer(s) would immediately be informed by the research team. Random treatment assignment was carried out over 1.5 years, from January 2011 till July 2012. In the light of scarce resources

[7]These cases could either be convictions from the past days, cases from trials where the hearing had already taken place, but the judge's written decision would be finalized in the following week.

[8]In principle, the randomization could have resulted in an outcome where, for example, one open IP slot remained "vacant" for (at least) a 2-week period. In practice, this never happened. In case of multiple open IP slots with different probation officers, the software randomly assigned convicts to different slots/officers.

(few IP slots) and the excess demand described above, judges were very supportive and approved the procedure.[9] The trial was officially approved by the Appellate Court (*Oberlandesgericht*) of Cologne and North-Rhine Westfalia's Ministry of Justice (Ordinance of the Ministry of October 18, 2010).

It was clear from the start that the number of observations would be small, as the probation office had not been given additional personnel for the program. Most importantly, the program does only make sense for offenders with a pronounced criminal record, and it does not make sense for offenders with severe, very likely insurmountable impediments, like serious drug addiction. Although the Cologne court district is one of the largest in the country, the number of eligible offenders per year is limited. It was therefore clear from the outset that the RCT would have low power; in the end, we had 30 treated and 27 untreated cases.

We had discussed these constraints with the court and the probation officers when planning the study. But the program had been running since 2006, and the probation officers were subjectively convinced that it was strikingly effective. Had this impression been supported by the data, we would have had a chance to establish a significant treatment effect, even with the small sample we could muster. Figure 2 uses simulation to show from which distributions the recidivism rates in the sample would have had to be chosen for treated (with IP) and untreated convicts (with ordinary probation), when allowing for a one-sided test (i.e., $\alpha = 0.1$; less recidivism with IP) and having an 80% probability ($\beta = 0.2$) to find the effect in the sample if it actually exists in the population. The critical parameter is the noise rate. As Figure 2 shows, at the noise rate that just meets the conventional criterion for power analysis, on average about half of the treated, but about three quarters of the untreated convicts would have to recidivate. Although an effect of this strength seemed ambitious, given the anecdotal evidence it did not ex ante appear impossible.[10]

Regression discontinuity design

The definition of program eligibility implied a discontinuity: Convicts with a score below 13 points had no chance to enter IP; convicts with a score of 13 or above had

[9]Judges made it clear that, in some exceptional cases, they might exclude convicts from the procedure and ad hoc assign them to IP. During the evaluation period, judges made use of this option in three cases. As discussed below, we will exclude these cases from our main analysis.

[10]The simulation proceeds as follows: We simulate a dataset with 30 treated and 27 untreated convicts. Whether the convict recidivates depends on the expression of a latent variable. It is mapped into recidivism if it is positive and into no recidivism otherwise. The latent variable is modeled as $1 - \tau + \varepsilon$, where τ is a dummy that is 1 if the convict is treated, and $\varepsilon \sim \mathcal{N}(0, \sigma)$. The simulation varies σ, that is, the SD of the normal distribution from which the error term is taken. For $\alpha = 0.1, \beta = 0.2$, with this data generating process $\sigma = 1.125$ is required. We find this threshold with the help of grid search, each time simulating 1000 datasets. Figure 2 simulates the data generating process, with $\sigma = 1.125$, and collects the resulting recidivism rates. The programs for implementing these simulations are available from the authors upon request.

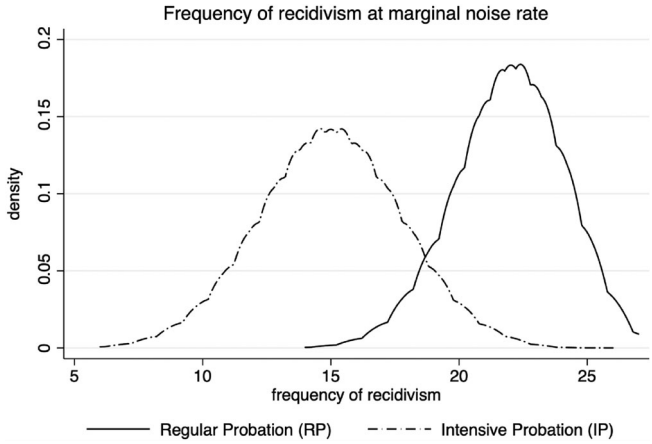

FIGURE 2 Distributions of recidivism required to establish a significant treatment effect with given sample size

a positive chance of treatment assignment. The cutoff rule thus implies a "fuzzy" treatment discontinuity: At the cutoff, the chance of getting into IP jumps from zero to (roughly) 50%.[11] Hence, even if the point score is plausibly correlated with the recidivism risk, the fact that the treatment (or rather: the chance of getting treated) changes discontinuously at the threshold, allows us to implement a fuzzy RDD (for background, see Lee & Lemieux, 2010).

Formally, we estimate reduced form equations of the structure,

$$Y_i = \mu + \tau D_i + f(Z_i) + e_i, \tag{1}$$

where Y_i is an outcome variable, D_i is a dummy indicating treatment eligibility, Z_i is the assignment variable (i.e., the score normalized around the cutoff such that $Z_i \geq 0$ implies $D_i = 1$), and f is a function of Z_i. Accounting for the discrete nature of the running variable and the sample size, we estimate $f(Z_i)$ parametrically. Our baseline specification will use two linear trends, separately defined for the range below ($Z_i < 0$) and above ($Z_i \geq 0$) the cutoff.[12]

Intuitively speaking, the RDD compares outcomes between juveniles that scored slightly below and those marginally above the cutoff. Accounting for the

[11]The fuzziness of the discontinuity is not due to non-compliance with the eligibility rule. Rather exactly because of the randomized control trial, treatment for those above the cutoff is only probabilistic. This point is also documented in Figure B1.

[12]The focus on linear specifications is motivated by model comparison based on the Akaike information criterion (AIC; Lee & Lemieux, 2010); models with linear trends tend to dominate models with quadratic terms and higher order polynomials in terms of minimizing the estimated information loss.

correlation between the point score Z_i and the outcome variable (i.e., conditional on f), the discontinuous jump in the chance of entering IP at the threshold then yields the reduced form estimate τ of the program's impact on an outcome.

In contrast to the RCT, which allows us to estimate an ATE for all program eligible offenders in our sample, the reduced form coefficient τ captures a LATE: The RD estimate is local, as it is identified from offenders with a point score Z_i around the eligibility cutoff. Put differently, the RD estimates will reflect the program's impact for less severe offenders, who are nevertheless severe enough to bring them close to program eligibility (as measured by the assignment score Z_i).

Note further that the coefficient τ must be interpreted in intention-to-treat (ITT) terms: Not every case above the eligibility cutoff enters IP. To estimate the (local) treatment effect on the treated (TOT), we use the discontinuity at the cutoff as an instrument. In particular, we will run two-stage least square regressions (2SLS). We first estimate

$$\mathrm{IP}_i = a + b\,D_i + f'(Z_i) + e'_i, \tag{2}$$

where IP_i is a dummy indicating that a case entered the program. The coefficient b then measures the discontinuity in the chance of entering the treatment at the cutoff. The function f' absorbs again any correlation between the point score Z_i and the left-hand side variable. Based on the first stage, we can then estimate

$$Y_i = \alpha + \beta\,\widehat{\mathrm{IP}}_i + \rho(Z_i) + \varepsilon'_i, \tag{3}$$

where $\widehat{\mathrm{IP}}_i$ is the instrumented indicator for program participation and ρ is again a parametrically estimated function of Z_i, which is allowed to differ on either side of the cutoff. The 2SLS estimate of β from this equation indicates the TOT effect of the program. The validity of the RDD is discussed in section "Data and samples."

Survey among probation officers

The last element of our research design consists of a survey among probation officers. This survey collected the officers' subjective evaluations of the convicted offenders at the start of the probation period and 6 months later. This allows us to observe within-case (and within probation officer) changes in the subjective evaluation of a case. We achieved a response rate of 80.7%, which hardly varied between treatment conditions (see Table 1).

TABLE 1 Structure of sample and data availability

		Cases with information from …	
	All cases	**e-Documentation**	**Officer survey**
Randomization period	171	152	46
RCT sample	57	49	46
IP ("treated")	30	26	25
RP ("control")	27	23	21
Nonqualified (below cutoff or exclusion criteria)	111	101	
Ad hoc assigned to IP (nonrandomized)	3	2	
Post-randomization period	38	—	
Total sample	209	—	

Note: The table illustrates the structure of the sample and data availability. The RCT sample is based on the 57 cases with randomized assignment to IP and RP. The RDD will explore the entire sample from the main study period (171 cases).
Abbreviations: IP, intensive probation; RCT, randomized control trial; RDD, regression discontinuity design; RP, regular probation.

We have asked officers to evaluate convicts on nine dimensions, mirroring those from the Judges' scorecards: their family background, their housing situation, school or apprenticeship, their social environment, alcohol and drug consumption, the ability to structure their day and to fulfill their daily duties, as well as their aggression level (for detail, see Appendix E). As subscales differ, for aggregation we normalize each subscale to the unit interval. A higher score indicates a higher degree of the respective social problem. For each convict, we generate an aggregated score that averages over the nine normalized subscales, separately for the beginning and the end of the first 6 months of probation (RP or IP).

DATA AND SAMPLES

Our analysis exploits data from multiple sources. First, we use information from the comprehensive scorecards discussed above. These data are augmented with case and offender specific information from the court's database. Second, we exploit data from the e-documentation platform used by Cologne's probation office. We mainly use these data to descriptively analyze contact frequencies between offenders and probation officers (see section "Intensive probation"). Third, we collect subjective evaluations among probation officers (see section "Survey among probation officers").

Our key outcome variables on recidivism are based on new criminal convictions after a convict has been put on probation. This information is obtained from the Federal Crime Register (the German *Bundeszentralregister*), which is highly protected by law.[13] The data allow us to observe the exact time of a crime that resulted in a conviction (before December 31, 2015). If there is an entry in the crime register, we also know the crime(s) for which the person has been convicted, and the sanction(s). Based on these data, we compute different recidivism measures that cover up to 3 years after the last convict started a probation period. In particular, we derive indicators of and count variables for re-convictions within certain time periods (based on the time between the initial probation sentence and the date of the crime that resulted in a later conviction). The data allow us to distinguish between reconvictions for violent, property, and other crimes.[14]

Overall, we collect data on 209 cases (see Table 1). During our main study period (all probation cases decided between January 2011 and June 2012), we observe 171 cases. Among these, 57 cases are treatment eligible, that is, cases with a point score above the cutoff and no exclusion criterion. These cases are randomly assigned to RP or IP.

A number of 111 cases are not treatment eligible and thus excluded from the randomization. These convicts either have a point score below the assignment cutoff and/or there are one or multiple reasons for exclusion (see above). In addition, there are three cases that are program eligible but ad hoc assigned to IP by the judges. As noted above (see footnote 9), these are exceptional cases where the respective judge would not have accepted an RP outcome. Our analysis of the RCT excludes these three cases. The RDD analysis, which exploits variation around the eligibility cutoff (rather than the actual treatment assignment), makes use of all 171 cases from the main study period.

For all 171 cases, we have detailed data on recidivism. For 86% from our core RCT sample (49 out of 57 cases), the e-documentation system allows us to observe interaction frequencies of convicts with probation officers during the probation period. These data served as the basis for Figure 1 (see above). For 81% of the RCT sample, we also have two data points (at the start and after 6 months of probation) from the subjective evaluation survey.[15]

[13]The authority keeping the register (*Bundesamt für Justiz*) has given us permission to receive this data. Permission was obtained after successful consultation of the authority with the Data Protection Commissioner of the Federal Republic of Germany (Ordinance of the *Bundesamt für Justiz* of May 30, 2014).

[14]Property crimes cover (non-violent) theft, fraud and blackmailing; violent crimes include violent and sexual assaults, robbery, restraint, and threat with gun/firearm. The residual category summarizes other crimes (e.g., vandalism, drug abuse, traffic violations).

[15]Probation officers sometimes failed to deliver an assessment within a reasonable time window around the end of the 6-month period. More generally, compliance with the survey protocol was lower in the RP sample. RAs where therefore instructed to focus on reminding/encouraging probation offers with RP cases from the RCT sample. Although this resulted in balanced return rates between treatment and control groups, it implied a very low return rate for cases beyond the RCT sample. Below we will therefore focus on the RCT sample.

Finally, we also collect some basic data for a 4-month period after the end of the randomization period. In this "post-study period," we merely ask judges to continue filling out the scorecards. We will briefly discuss these data below.

Randomization checks

Balancing checks for the 57 eligible cases are presented in Table 2. For none of the variables, do we find a significant difference between those assigned to IP and RP. The average age at the time of the randomization is around 18 years (with 90% of the data in a range between 16 and 21). The sample is almost exclusively male, with the two female convicts ending up in the control group.

The average in the overall score from the judges' scorecards (ranging from 13 to 28) is nearly identical between the two groups, with only minor differences in subcategories (e.g., alcohol or drug problems, or elevated aggression levels). Judges have used flexible tools for avoiding that the defendant goes to jail (according to §§ 27 and 57 JGG, the German juvenile courts law) in comparable fractions. In terms of crime categories, in the control group we observe a higher share of convicts that had committed violent crimes and fewer cases with property crimes. However, neither difference is statistically significant. Overall, the observable characteristics are consistent with random assignment of convicts to IP (treatment) and RP (control).

Validity of the RDD

The RDD provides a valid identification strategy as long as possible confounders (i.e., observed or unobserved individual characteristics that influence recidivism) change smoothly around the eligibility cutoff. In our context, one might be worried that this assumption could be violated. As all judges knew about the threshold, they could strategically have scored juveniles to place them either below or above the cutoff. In this case, ending up above or below the cutoff would not be as good as random but selected by judges, conditional on factors that are unobservable to the research team.

Given the high demand for IP slots from the pre-experimental period, we were mainly concerned about strategic *overrating* (i.e., selection into eligibility). In contrast, we expected that strategic *underrating* of convicts would be no issue, because judges could simply tick a box to indicate an exclusion criterion.[16] Despite this institutional feature, we detect some evidence suggesting that strategic underrating could nevertheless be an issue.

[16]Recall from above that judges would only have to tick one of the exclusion indicators in the scorecard to assure that a convict is excluded from the randomized program assignment (see Appendix D). Note further that these exclusion criteria would not alter the overall score.

TABLE 2 Summary statistics and balancing checks

Variable	RP ("Control")	IP ("Treated")	*p*-Value
Age	18.272	18.401	0.803
Male	0.926	1.000	0.134
Score[a]	17.630	17.567	0.946
Problematic peers[a]	0.333	0.300	0.792
Alcohol[a]	0.222	0.367	0.242
Drugs[a]	0.370	0.200	0.158
Agression-mid[a]	0.481	0.367	0.390
Agression-high[a]	0.185	0.300	0.323
§27 JGG	0.481	0.367	0.390
§57 JGG	0.148	0.100	0.588
No. of onvictions	1.500	1.241	0.497
Property crime	0.250	0.308	0.658
Violent crime	0.708	0.577	0.344

Note: The table presents the sample mean among treatment and control group for different variables. The third column includes the *p*-values from two-sided *t*-tests. Age is the age at the randomization date and Male indicates gender. Score is the overall point score from the scoreboard; the next variables—subcategories of the score card—indicate problematic peers, alcohol, or drug problems, respectively; intermediate or high aggression levels (3–4 or 5–6 points in this subcategory of the scorecard), respectively; § 27 JGG and § 57 JGG (the German juvenile courts law) are dummies capturing case specific information. The dummies indicate whether the juvenile court is empowered to be more flexible: According to § 27 JGG, the court may for the time being only declare that the defendant has violated the law, and define a period within which the defendant may be sent to prison, should later behavior call for that. According to § 57 JGG, the court may impose a prison sentence, but refrain from enforcing it for a defined period of time, and conditional on the convict not misbehaving. No. of convictions are the number of prior convictions (at the time of randomization); property crime and violent crime are dummies indicating for which crime an individual was convicted. Property crimes cover (nonviolent) theft, fraud, and blackmailing; violent crimes include violent and sexual assaults, robbery, restraint, and threat with gun/firearm. The residual category summarizes other crimes (e.g., vandalism, drug abuse, traffic violations). Sample is $N = 57$, except for the two crime category variables with $N = 50$.

Abbreviations: IP, intensive probation; JGG, *Jugendgerichtsgesetz*; RP, regular probation.

[a]These variables are based on information from the scorecards. All other variables are coded based on court data.

The left panel of Figure 3 illustrates the distribution of scorecard points for the 171 convicts during the 1.5-year randomization period. The figure shows some "heaping" below the cutoff: 14 convicts had exactly 12 points, whereas only 5 cases had 13 points. The dark-gray shade bars, which indicate cases where an exclusion criterion was ticked, suggest that exclusion occurred on either side of the cutoff.

One interpretation of this pattern is indeed a strategic underrating of some convicts (who would have otherwise scored just above the cutoff and would have thus been eligible for IP). However, the evidence does not coherently

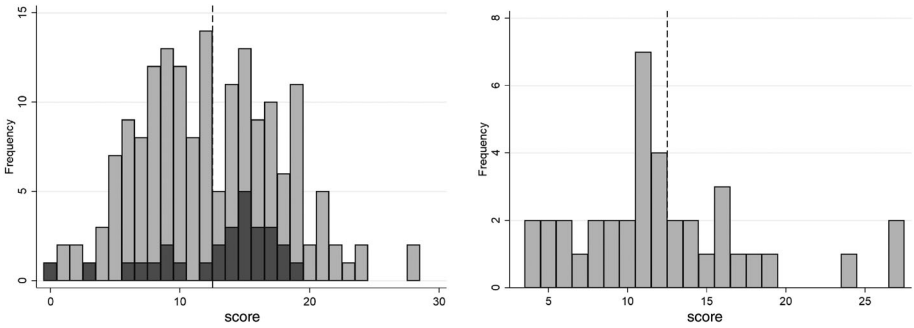

FIGURE 3 Distribution of scorecard points. The light-gray bars indicate the distribution of scores in the main sample ($N = 171$ from the main study period); the dark-gray bars indicate, within this sample, the cases where an exclusion criterion was ticked. The right panel illustrates the distribution in the post-randomization period ($N = 39$). The dashed black line indicates the cutoff

support this interpretation. First, we ran placebo estimates that explore whether any observable characteristics discontinuously change around the threshold. The tests, which are reported in the Appendix (see Table A1), do not indicate any discontinuities. The evidence suggests that observable characteristics are smooth around the cutoff (see Figure B1). These observations again speak against strategic scoring being the driver behind the point distribution from Figure 3.[17]

Second, we examined the point distribution from scorecards filled out by judges after the RCT ended. In this post-randomization period, judges continued to use the scorecards. The score, however, was no longer used to assign cases to randomization. Hence, there was no strategic reason to target a score just below 12 points. The right panel of Figure 3 nevertheless indicates heaping below the cutoff (with a seemingly above-normal number of cases with 11 and 12 points).[18]

One interpretation of the score distribution in the post-randomization data is that the heaping is, at least partially, driven by the mere design of the scorecard. The overall point score is the sum of 10 sub-scores. In most of these 10 dimensions, the modal score assigned by the judges was either 1 or 2 points. Hence, ending up with an overall score of 11 or 12 points just turned out to be very likely. Although this offers a possible alternative explanation, we ultimately

[17]Obviously, these tests (which are in the spirit of the balancing checks for the RCT from above) are of limited power and one can never rule out strategic scoring based on unobservables (i.e., information only available to a judge). However, to the extent that unobservables are correlated with some observable variables, the tests are at least indicative. The smoothness of covariates documented in Figure B1 is reassuring on this point, too. Note further that the use of background information obtained from the scorecards is problematic, as the different sub-scores must increase at the cutoff, by assumption. This is why Table A1 mainly focuses on other characteristics.
[18]It is also worth noting that the histogram does not indicate any clear missing mass on the right-hand side of the cutoff (see right panel of Figure 2).

cannot rule out strategic scoring. In our empirical analysis we will therefore complement our RD estimates with so called "donut estimates." The latter exclude the range just around the cutoff from the RDD (Barreca et al., 2011).

RESULTS

This section first presents the results from the randomized assignment to IP. Thereafter, we turn to the RDD.

Randomized treatment assignment

Subjective assessment by probation officers

We first consider the subjective assessment by the probation officers at the time of randomization and half a year later. Figure 4 presents these two scores. Within the control group, the average score improved from 0.58 when put on probation to 0.45, 6 months later.[19] According to the probation officer's scores, only three cases on RP deteriorated during this time period. Yet, the officers see more progress among convicts in the IP program. Among the treatment group, the average score drops from 0.65 to 0.32, 6 months later. The difference in differences is highly significant (Mann Whitney, $N = 46$, $p < 0.001$) and robust to excluding outliers (cases with the most extreme improvement/worsening). Recall, however, that all IP cases are handled by four probation officers. The pattern could therefore be driven by these officers being overly optimistic about the impact of their work. We thus turn to more objective data on recidivism.

Recidivism

Figure 5 compares cumulative recidivism rates between the treatment (IP) and the control (RP) groups. Consistently with the program's focus on the most problematic among young offenders, the figure documents very high recidivism rates. Within 10 months, every second convict committed another crime. After 3 years, this rate raises to more than 75%. However, the figure also suggests that IP has indeed a positive effect in terms of reducing recidivism.

During the first month, there is no noticeable difference. Between the second and the sixth months (the end of the "intensive" part of the probation period), recidivism rates in the IP group are between 10 and 13 percentage points below the rates observed for RP. In relative terms, this corresponds to a 20%–30%

[19]Recall that a lower score represents a "better," less problematic outcome (see section "Survey among probation officers").

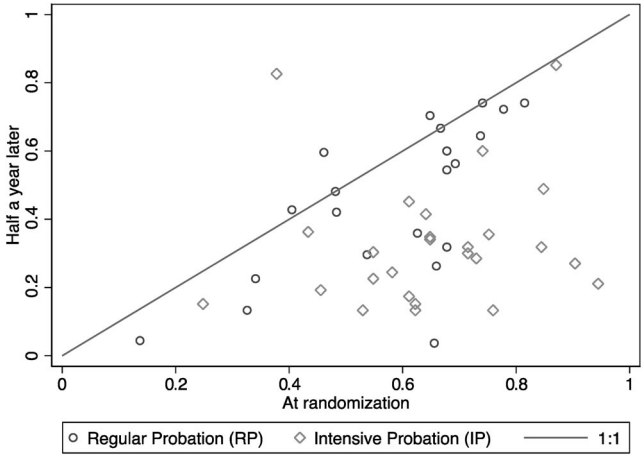

FIGURE 4 Subjective assessment by probation officers. The figure plots the probation officers' evaluation score at the beginning of the probation period and 6 months later. The score is normalized to unit interval (with a higher the score indicating a more problematic condition)

decline. However, the differences are statistically insignificant (with two-sided tests yielding p-values in the range of 0.250, $N = 57$).

After the end of IP, in particular between Months 6 and 10, the treatment gap closes again and becomes virtually zero. The evidence therefore suggests that IP seems to have a positive, stabilizing effect during the treatment period, but that the effect vanishes fairly quickly. The patterns in the second and third year, however, indicate again lower recidivism rates for IP (with statistically insignificant treatment differences in a range between 5 and 10 percentage points). Hence, conditional on not having re-committed a crime within the first year, the program might provide some beneficial mid-run effects, too. This pattern is also reflected in estimates that consider recidivism within intervals ranging from one to 36 months (see Figure B4).

Parametric estimates from duration models, which support our interpretation from above, are reported in Table A2. Considering re-offenses observed during the entire 3-year outcome window, the estimated hazard ratio is between 0.73 and 0.82 (Table A2, Columns 1 and 2), suggesting a roughly 20% lower hazard (i.e., probability of re-offending conditional on not having done so before) in IP as compared to RP. When we allow for time varying hazards, we observe—consistently with Figure 5—a stronger effect during the first 6 months (ratios between 0.50 and 0.67) with an opposing trend during the next 6 months (positive hazard ratios of 1.26 and 1.16; see Table A2, Columns 5 and 6, respectively). During the second year, however, hazard rates are consistently between 0.73 and 0.81. (In the third year, hazard rates are very close to unity.) Although none of these estimates is statistically significant, the effect sizes are meaningful

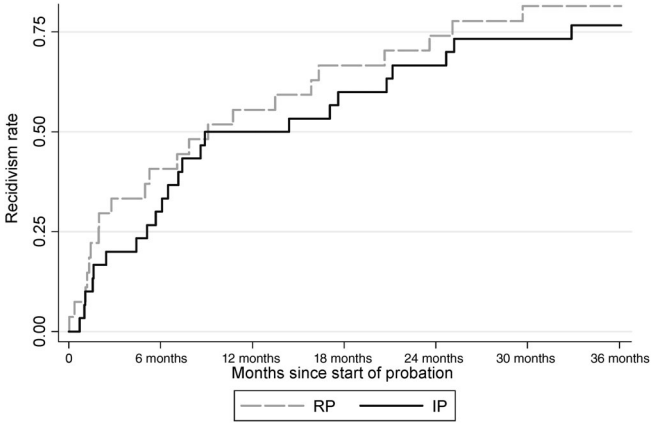

FIGURE 5 Recidivism over time. The figure plots the recidivism rates for the treatment (intensive probation) and control group (regular probation). N = 57

and, once more, point to both a short-run (during the 6 months of IP) and mid-run (during the second year after probation) effect from IP.

A complementary way to capture these time-varying effects is presented in Table 3. Here, we use linear probability models to estimate treatment effects on a binary recidivism indicator for time windows ranging from 3 months to 3 years. All estimates indicate a negative impact (a lower recidivism rate) for IP. Both unconditional and conditional treatment effects (without and with controls) are, in absolute terms, declining within the first year. During the second year, the absolute treatment difference raises again. For both, the short-run and mid-run recidivism, the estimates from some (but not all) models are significant at the 10% level.

In a next step, we replicate the analysis distinguishing between recidivism related to a violent crime or a property crime. The estimation results from Table 4 provide some indication of differential treatment effects on different crime types. For violent crimes, we observe a similar pattern as above: a relatively large effect of IP during the first 6 months (see Table 4, Panel A, Columns 1 and 2), which declines thereafter but survives throughout Years 2 and 3. For property crime, in contrast, there is no indication of any impact from IP. Both the short- and the mid-run effects on recidivism tend to be very close to zero (see Table 4, Panel B). This provides suggestive evidence that IP mainly works via preventing (or delaying) violent re-offenses.

The results from above offer weak evidence suggesting that IP lowers the risk of reoffending. To examine whether treatment effects at the extensive margin (i.e., reconviction: yes or no) are mirrored at the intensive margin, we study count

TABLE 3 LPM estimates: Recidivism rates I

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Recidivism	3 months		6 months		12 months	
IP	−0.133	−0.259*	−0.107	−0.207	−0.056	−0.077
	(0.119)	(0.144)	(0.129)	(0.197)	(0.135)	(0.192)
Control variables	N	Y	N	Y	N	Y
RP mean	0.333	0.333	0.407	0.407	0.556	0.556
Panel B						
Recidivism	1.5 years		2 years		3 years	
IP	−0.067	−0.259*	−0.074	−0.292*	−0.048	−0.154
	(0.130)	(0.144)	(0.123)	(0.162)	(0.109)	(0.159)
Control variables	N	Y	N	Y	N	Y
RP mean	0.667	0.667	0.741	0.741	0.815	0.815

Note: The table presents estimated treatment effects on recidivism (binary outcome), considering time frames between 3 months and 3 years. All estimates are based on linear probability models, with a sample of $N = 57$. Robust SEs are in parenthesis. RP mean presents the average recidivism rate in the control group with regular probation. Control variables in specifications (2), (4), and (6) include age, gender, dummies for the start of the probation period as well as indicators based on the different dimensions in the judges' scorecards (e.g., alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers).
Abbreviations: IP, intensive probation; LPM, linear probability model; N, no; RP, regular probation; Y, yes, the set of control variables is included in the specification.
* indicate significance at the 10% level.

variables that measure the number of crimes for which the individual has been convicted during different time windows. The estimates, which are reported in Table A3, are all negative (see Panel A), indicating a relatively small, statistically insignificant decline in the number of crimes.[20] The point estimates are not too different from the extensive margin drop in recidivism reported in Table 3 above. A similar picture—that is, small and insignificant estimates—emerges when we consider the severity of re-offenses (as captured by the imposed legal sanction; estimates not reported). Overall, the data on the number and severity of re-convictions provide no compelling evidence on intensive margin responses to IP. If at all, the program seems to alter crime outcomes mainly at the extensive margin.

RDD results

The results from the randomized program assignment provide weak evidence suggesting that IP reduced recidivism rates. However, the main limitation of the evidence from above is its restricted power associated with the small sample size.

[20]Very similar results are obtained if we estimate count data models (e.g., Poisson regressions).

TABLE 4 LPM estimates: Recidivism rates II (violent and property crimes)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: *Violent crimes*								
Recidivism	6 months		1 year		2 years		3 years	
IP	−0.156	−0.165	−0.093	−0.195	−0.100	−0.150	−0.107	−0.039
	(0.103)	(0.106)	(0.110)	(0.122)	(0.121)	(0.158)	(0.129)	(0.189)
Controls	N	Y	N	Y	N	Y	N	Y
RP mean	0.222	0.222	0.259	0.259	0.333	0.333	0.407	0.407
Panel B: *Property crimes*								
Recidivism	6 months		1 year		2 years		3 years	
IP	0.011	−0.056	0.033	0.061	0.056	−0.023	−0.026	−0.042
	(0.113)	(0.171)	(0.129)	(0.165)	(0.135)	(0.191)	(0.133)	(0.179)
Controls	N	Y	N	Y	N	Y	N	Y
RP mean	0.222	0.222	0.333	0.333	0.444	0.444	0.593	0.593

Note: The table presents estimated treatment effects on crime-specific recidivism indicators (binary outcome), that is, for violent (Panel A) and property crimes/re-offenses (Panel B), distinguishing a time frame between 6 months and 3 years. RP mean presents the average violent crime or property crime recidivism rate in the control group with regular probation. Control variables in specifications (2), (4), (6), and (8) include dummies for the start of the probation period as well as indicators based on the different dimensions in the judges' scorecards (e.g., alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers). All estimates are based on linear probability models, with a sample of $N = 57$. Robust SEs are in parenthesis.

Abbreviations: IP, intensive probation; LPM, linear probability model; N, no; RP, regular probation; Y, yes, the set of control variables is included in the specification.

* indicate significance at the 10% level.

To assess whether we find similar results when we explore a slightly larger sample, we now turn to the RDD. We assess if the discontinuity in treatment rates yields findings that are consistent with the results from the previous sub-section. Before doing so, however, it is important to recall that section "Randomized treatment assignment" reported ATEs. The RDD, in contrast, will yield LATEs, for cases that are on the edge to be program eligible. Put differently, the RDD effects are identified from the "less severe" convicts (among convicts with records sufficiently severe to qualify for the program in the first place)—those with a lower point score on the scorecard. With this important caveat in mind, let us first provide some graphical evidence.

Figure 6 visualizes our basic, reduced-form RD estimates (using linear functions of the running variable, which may differ on either side of the cut-off, for f from Equation 1). Note first that the linear fits indicate an intuitive, positive correlation between point score and recidivism: More problematic offenders (with a higher point score) are more likely to re-offend within a 1- or 3-year time period, respectively. Second, there seems to be a discontinuous drop in recidivism rates at the eligibility cutoff. Those marginally above the cutoff—who

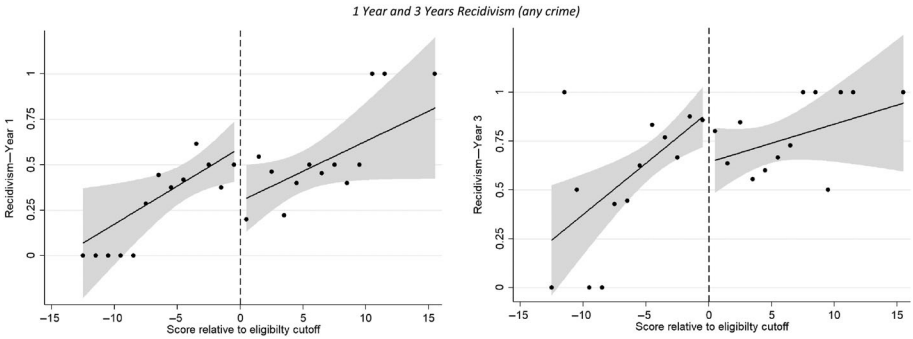

FIGURE 6 Discontinuity in recidivism at the intensive probation (IP) eligibility cutoff. The figure presents estimated discontinuities in 1-year (left) and 3-year recidivism rates (right sub-figure) at the IP eligibility cutoff (linear fits with 95% confidence intervals). The binned data present the average recidivism rate among convicts with a given point score (relative to the eligibility cutoff). N = 171

have an approximately 50% chance of entering probation with intensified support—display lower rates of recidivism (both, within a 1- and 3-year time period).[21]

The estimates from Table 5 confirm this discontinuity, indicating that the observed drops in recidivism are statistically significant. The coefficients suggest that reoffending declines between 10 (6 months) and around 30 percentage points (1–3 years period) once the IP eligibility cutoff is passed. The results further indicate that the point estimates hardly change when we add control variables. This is reassuring and, together with the smoothness of covariates around the cutoff (see Table A1 and Figures B1 and B2), provides support for the validity of the RDD.

The estimates suggest that the IPs' crime reducing effect builds up during the first year and remains roughly constant thereafter. This picture is confirmed in Figure 7, which offers a more detailed impression of the dynamics of the reduced form effects. The figure plots point estimates and confidence intervals of 36 RD estimations of Equation (1) with binary dependent variables Y_m, indicating a re-offense within m-months and $m = \{1, \ldots, 36\}$. The figure suggests that the effect builds up during the first year. For any period after 10 months, we detect a significantly negative effect. After the 12th month, the effect is fairly stable throughout the 3-year period. Hence, the RD estimates do not indicate any decay in the LATE.[22]

As noted above, the reduced form estimates must be interpreted as a local, ITT effect. To turn from an ITT to a TOT effect, Table 6 presents 2SLS estimates of Equation (3). Consistently with the roughly 50% discontinuity in the treatment propensity, the point estimates from the 2SLS in Panel A are roughly twice as large as

[21]In robustness exercises discussed below, which focus on more narrow bandwidths around the cutoff (and thus exclude outliers in terms of very high/low recidivism rates within a bin), we obtain similar results.

[22]The corresponding graph for the RCT estimates (Figure B4) is characterized by much larger confidence intervals. It is nevertheless worth noting that the LATE seems to follow a slightly different trend: Relative to the ATE, the effect seems to emerge less quickly during the 6 months of the core period of IP.

TABLE 5 Reduced form RD estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Recidivism	6 months		1 year		2 years		3 years	
Eligibility	−0.107	−0.069	−0.298**	−0.224	−0.343***	−0.324**	−0.273**	−0.267**
	(0.126)	(0.141)	(0.129)	(0.146)	(0.122)	(0.127)	(0.116)	(0.119)
Controls	N	Y	N	Y	N	Y	N	Y

Note: The table presents reduced form RD estimates (of Equation 1) at the IP-eligibility cutoff. All specifications control linearly for the correlation between point-score and outcomes (differentiating for the range below and above the cutoff). The dependent variables are recidivism binary indicators for different time periods. Specifications (2), (4), (6), and (8) include controls (time period fixed effects for the start of the probation period as well as indicators for alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers). Sample: $N = 171$. Robust SEs are in parenthesis.

Abbreviations: IP, intensive probation; N, no; RD, regression discontinuity; Y, yes, the set of control variables is included in the specification.

***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

those from the reduced form from Table 5.[23] Although some of these coefficients appear very large, note that (a) the estimates are relatively imprecise and the 95% confidence intervals typically overlap with the ATEs observed in the RCT. Alternatively, (b) the large point estimates could suggests that the program has a particularly pronounced impact on less severe offenders that are close to the eligibility cutoff: "Less problematic" cases could benefit more from IP.[24]

Table 6 further examines recidivism separately for violent and property crimes (see Panel B). We find statistically insignificant results for both of the two types of recidivism. For violent crimes, the point estimates meander around zero. For property crimes, the estimates are consistently negative. Note that this contrasts to the evidence from the RCT sample, where there were some indications for a decline in violent crimes. For the RDD, the drop in recidivism seems to stem from property crimes and minor (nonviolent and nonproperty) crimes. This difference again highlights that the LATE is identified for relatively less severe offenders (as compared to the ATE estimated from the RCT).

We also explored crime at the intensive margin (Panel C). The estimated discontinuities for the number of crimes tend to be negative, but are all statistically insignificant. For the crime count after 2 and 3 years, the estimates are quantitatively very similar to the estimates reported in Panel A. The latter observation suggests that the crime reduction is again driven by extensive rather than intensive margin responses.

[23]Figure B3 provides evidence on the treatment discontinuity (i.e., the first stage). Consistently with random treatment assignment, we observe a jump in the treatment rate of roughly 50%. In order to get to TOT effects, one therefore has to deflate our reduced form (ITT) estimates by 0.5.

[24]Consistently with this interpretation, we find larger average treatment effects if we split our RCT sample into low (vs. high) point score cases. Due to the limited sample size, however, this heterogeneity analysis comes with very large SEs.

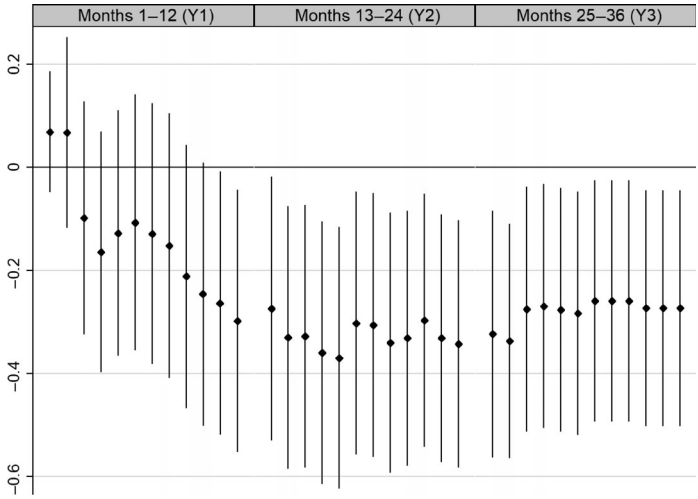

FIGURE 7 Reduced form RD Estimates over Time. NOTES: The figure presents the estimated coefficients and 95% confidence intervals from 36 different reduced form RD estimates (in the spirit of Table 5). Dependent variables are dummies indicating recidivism within m months, with m {1,…, 36}. N = 171

One concern with the RD analysis relates to the potential heaping of cases below the eligibility cutoff (see section "Validity of the RDD"). As pointed out above, any strategic sorting of cases would question the validity of the RDD. To assess the robustness of our results, we run so called "donut" RD estimates, which exclude a range around the cutoff (in particular, convicts with one point below and above the eligibility cutoff). Reduced form donut estimates, which are reported in Table A4, confirm the findings from above. In fact, the point estimates document a significantly negative drop in property crimes.[25]

If one is not concerned about the heaping, one might consider the opposite of a donut hole strategy and rather focus on a narrower range around the cutoff (i.e., excluding observations relatively far away from the cutoff). Doing so, one obtains similar but less precisely estimated results. In the light of the smaller number of observations, however, this might not be too surprising.

In further robustness exercises, we worked with a slightly larger set of control variables (that comes at the cost of a smaller sample due to missing entries for covariates). Results were hardly affected. However, our estimates are sensitive to the linearity assumption applied above: If one accounts for the link between the point score and outcomes using quadratic or higher order

[25]Similar (more imprecise) results are obtained in 2SLS donut estimates as well as when we impose a larger "donut hole" (with ±2 points around the cutoff). In the latter case, however, the sample size shrinks substantially and SEs increase accordingly.

TABLE 6 Two-stage least square regression (2SLS) RD estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: *Any crime*								
Recidivism	6 months		1 year		2 years		3 years	
IP (instrumented)	−0.152	−0.093	−0.603*	−0.511	−0.726**	−0.769**	−0.510*	−0.560*
	(0.292)	(0.348)	(0.342)	(0.401)	(0.341)	(0.372)	(0.307)	(0.332)
Panel B: *By crime type*								
	Violent crimes		*Violent crimes*		*Property crimes*		*Property crimes*	
Recidivism	1 year		3 years		1 year		3 years	
IP (instrumented)	0.138	0.059	0.009	−0.101	−0.197	−0.210	−0.199	−0.246
	(0.247)	(0.264)	(0.295)	(0.324)	(0.279)	(0.339)	(0.323)	(0.369)
Panel C: *Number of crimes*								
Count	6 months		1 year		2 years		3 years	
IP (instrumented)	−0.165	−0.130	0.114	0.298	−0.614	−0.542	−0.507	−0.542
	(0.408)	(0.499)	(0.612)	(0.732)	(0.806)	(0.891)	(1.058)	(1.183)
Controls	N	Y	N	Y	N	Y	N	Y
First stage *F*-stat.	19.745	19.125	19.745	19.125	19.745	19.125	19.745	19.125

Note: The table presents 2SLS RD estimates (see Equation 3), using the IP-eligibility cutoff as instrument. Every second specification includes a vector of control variables. Kleibergen-Paap *F*-statistics from the first stage regressions (see Equation 2, section "Research design") are provided at the bottom of the table. Note that the first stage is independent of the outcome variable and only varies with the inclusion of controls. The dependent variables in Panels A and B are recidivism indicators (binary outcomes) for different time periods. Panel A considers any recidivism, Panel B distinguishes between violent and property crimes/re-offenses. Panel C uses a count for (any) crimes as outcome variables. Sample: $N = 171$. Robust SEs are in parenthesis.

Abbreviations: IP, intensive probation; N, no; RD, regression disconitnuity; Y, yes, the set of control variables is included in the specification.

***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

polynomials, one obtains different point estimates and much larger SEs. Following Lee and Lemieux (2010), we computed the AIC to compare the fit of the linear and the quadratic models (see footnote 12). The comparison indicates that models with linear trends typically outperform models with quadratic trends (and, even more so, models that include higher order polynomials). The more we restrict the sample to cases around the cutoff, the stronger the linear model dominates.

DISCUSSIONS

Mechanisms and measurement

As discussed in section "Intensive probation", the IP program involves many different dimensions. In general, probation officers take over new cases more swiftly and can then devote more time to personal meetings and monitoring. Our data do not allow us to distinguish which program feature is most relevant in shaping the results reported above.

One concern might be that a higher contact intensity implies a stricter level of control and supervision in IP. This program feature could mechanically increase the chance of detecting further crimes or violations of probation conditions. In turn, this could increase the measured recidivism rates among the IP group, both at the extensive and intensive margin. There are at least three reasons why we do not think that this is an issue. First, there is a growing body of evidence suggesting that increased levels of supervision for probationers and parolees has no impact on recidivism (Georgiou, 2014; Hyatt & Barnes, 2017). This speaks against the relevance of this concern. Second, note that both the RCT and the RDD estimates point to persistent effects that remain stable during the second and third year after the start of IP, that is, long after the end of the period of intensive supervision. Finally, if the mechanism would nevertheless play a role, it would bias our results against finding a recidivism reducing program effect. Our estimates would then be lower bounds of "true" treatment effects.

Hawthorne effect

It was not possible to implement the RCT without the consent of probation officers and judges. Probation officers were thus aware of the study. Although officers did neither know about the RD design nor about who is in the relevant evaluation sample, one might nevertheless worry about a Hawthorne effect (Adair, 1984; Landsberger, 1958; McCambridge et al., 2014). It could be that some probation officers work extra hard as they know that they are evaluated. This could

lead to either an over- and/or under-estimation of the treatment effect (if either IP or RP officers put in relatively more effort after learning about the evaluation).

To assess whether the findings from our RCT are confounded by such a Hawthorne effect, we exploit the data on the scorecards for the post-randomization period. Recidivism rates (during a 6-, 12-, or 24-month period) for the 38 cases observed in this period are not statistically different from the rates in our RCT sample. Among the 38 filled scorecards 14 cases would have qualified for the IP program. Recidivism rates in this group are again statistically indistinguishable from our IP or our RP group (with p-values from Fisher's exact tests in the range between 0.690 and 0.863). Similar null results are obtained for the severity of later convictions, the number of days between the first and the next conviction, or the number of later convictions for violent and for property crime, respectively. Although these null-results do not rule out the presence of a Hawthorne effect, the analysis at least suggests that there has been no major structural break after the communicated end of our evaluation period.

Power

As noted above, one of the main limitations of our study is the limited sample size of the RCT. We had to live with the fact that the probation office did not have more manpower and that (fortunately) the large district of the regional Court of Cologne did not produce more eligible cases. As the evaluation was placing some administrative burden on the court (and the team of research assistants), we could not extend the evaluation period. In addition, a longer evaluation period would have increased the risk of changes over time in unobserved contextual variables. The limited sample size obviously impedes our ability to precisely identify small effects. Comparing our results with the ex ante expectations by the court, and with the subjective ex post evaluations by the probation officers, stresses the importance of rigorous evaluation, all the more so if a pilot project is meant to inform policy makers about the desirability of a roll out. We have a nuanced message for them. Although we do not document a strong, unqualified program effect, it is reassuring that the RDD finds a meaningful, positive effect of the program.

Cost–benefit analyses

IP reduces the total amount of cases a probation officer handles. Therefore, it is more expensive than RP. To see whether the potential benefits (in terms of lower recidivism rates) outweigh the cost, we performed a simple back-of-the-envelope cost–benefit analysis. Details of our calculations, which use our recidivism estimates and proxies for the direct victimization costs, are provided in Appendix C.

For our RCT sample, we estimate that the additional resource costs are higher than the reduction in victimization costs (both within 1 and 3 years after

the program). However, as the total social costs of recidivism are plausibly larger than the mere victimization costs (see Appendix C), the social benefits should compensate the costs. By contrast, the results from the RDD point to much larger social benefits. In fact, for the RDD sample the reduced victimization costs alone are roughly twice as large as the resource costs of the program. Given our small sample size, the results from this simple cost–benefit analysis should be treated with caution.[26]

External validity

We evaluated a pilot project by one regional court. Germany is a federation and the states ("Länder") have jurisdiction for organizing the administration, including correctional staff. Results from the city of Cologne do, therefore, not directly extrapolate to the entire country, or beyond. Yet, Cologne is situated in the state of *North Rhine-Westphalia*. The incidence of crime in this state is about the average of all of Germany. Cologne is the fourth biggest city in the country. In cities with more than 500,000 inhabitants, the incidence of crime is highest and comparable.[27] Although other parts of the country could certainly not just copy the program as is, it therefore stands to reason that the findings from Cologne are meaningful. The program has promise. But to safely establish, and evaluate, the program, more court districts would have to be involved.

Internationally, crime rates, the judicial and correctional systems, and the number of prison inmates differ widely.[28] The modest success of the Cologne program might well hinge on societal factors beyond the control of the judiciary, like the willingness of employers to give juvenile convicts a second chance. Differences in the institutional fabric, and in the personnel acting in the courts and correctional facilities, might make it more difficult in other countries to implement a comparable program. Yet, at the core of the program are interventions that do not require the societal, cultural, or institutional German context, like educating convicts to take responsibility for their lives and families, or helping them with professional training and job searches. The results of the study may, therefore, be read as a cautious sign of hope: When handled by zealous probation officers, and supported by the court administration, intensive intervention with the palpable intention to help the

[26]The cost–benefit analysis also neglects general equilibrium effects: If IP is perceived as the most likely "punishment" for those who are on the edge to imprisonment, this might imply lower deterrence.

[27]In 2019, in North Rhine-Westphalia per inhabitant 0.0685 crimes have been reported. The Land with the lowest incidence is Bavaria (0.0461 crimes per inhabitant and year), and the Land with the highest incidence is Berlin (0.141 crimes per inhabitant). In 2019, Cologne had 1.086 Mio inhabitants, wikidata.org. In cities with more than 500,000 inhabitants, in 2019, 0.284 crimes per inhabitant have been reported. All data from Polizeiliche Kriminalstatistik Bundesrepublik Deutschland Jahrbuch 2019 Band 1: Fälle, Aufklärung, Schaden, https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/PolizeilicheKriminalstatistik/2019/Jahrbuch/pks2019Jahrbuch1Faelle.pdf?__blob=publicationFile&v=3

[28]https://dataunodc.un.org/content/prison-population-regional-and-global-estimates

convicts leading a meaningful life free from crime has a chance to work. This is all the more reassuring as the positive effects are probably underestimated given the program focuses on juveniles for which the successful prevention of criminal careers yields much higher returns in the long run (Cohen & Piquero, 2009).

CONCLUDING SUMMARY

Against a widely shared pessimism of "nothing works" in reducing recidivism, our study provides experimental and quasi-experimental evidence that points to the effectiveness of a probation program with intensified support from Cologne, Germany. Data from an RCT and an RDD suggest that the program, which significantly increases the personal support that young probationers receive over a 6-month period, tends to reduce recidivism in the short- and the mid-run. The crime reducing effect seems to operate at the extensive rather than the intensive margin. The RDD, which documents a larger drop in recidivism, indicates that the decline in recidivism is more pronounced among less severe offenders. This limited, but measurable, success is all the more remarkable as the program is implemented with no additional resources and at the initiative of a local probation office (and continues to the present day). It shows how much can be achieved with the help of energetic and enthusiastic officers, supported by the solidarity of their colleagues who are willing to shoulder an even higher workload, to make the program possible. It is up to future work to isolate the specific mechanism that shapes this decline in crime.

ACKNOWLEDGMENTS

This project would not have been possible without the constructive cooperation and support of numerous people. We are particularly grateful to Michael Klein, Maren Sütterlin-Müsse, and the other judges at Cologne's youth court, Christian Schmitz-Justen, and Frank Czaja at the regional court of Cologne, as well as Martin Kuhnigk, Lucia Lennartz-Schweda, and the team at the probation office. Avista Assadi, Mimoun Berrissoun, Julia Möller, Anja Roesner, Melanie Röver, and Denise Wettley provided excellent research assistance. We also benefited from detailed and very constructive comments of three anonymous reviews.

REFERENCES

Adair, J. G. (1984). The Hawthorne effect. A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*(2), 334–345.

Asscher, J. J., Deković, M., Manders, W., van der Laan, P. H., Prins, P. J. M., van Arum, S., & Dutch MST Cost-Effectiveness Study Group. (2014). Sustainability of the effects of multisystemic therapy for juvenile delinquents in the Netherlands. Effects on delinquency and recidivism. *Journal of Experimental Criminology*, *10*(2), 227–243.

Austin, J., Joe, K., Krisberg, B., & Steele, P. A. (1990). The impact of juvenile court sanctions. A court that works. *Focus: The National Council on Crime and Delinquency*, 1–7.

Barreca, A. I., Guldi, M., Lindo, J. M., & Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *Quarterly Journal of Economics*, *126*(4), 2117–2123.

Bayer, P., Hjalmarsson, R., & Pozen, D. (2009). Building criminal capital behind bars. Peer effects in juvenile corrections. *Quarterly Journal of Economics*, *124*(1), 105–147.

Berg, I., Hullin, R., McGuire, R., & Tyrer, S. (1977). Truancy and the courts. Research note. *Journal of Child Psychology and Psychiatry*, *18*(4), 359–365.

Bhuller, M., Dahl, G. B., Løken, K. V., & Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, *128*(4), 1269–1324.

Bitney, K., Drake, E., Grice, J., Hirsch, M., & Lee, S. (2017). *The effectiveness of reentry programs for incarcerated persons*. http://www.wsipp.wa.gov/ReportFile/1667/Wsipp_The-Effectiveness-of-Reentry-Programs-for-Incarcerated-Persons-Findings-for-the-Washington-Statewide-Reentry-Council_Report.pdf

Brank, E., Lane, J., Turner, S., Fain, T., & Sehgal, A. (2008). An experimental juvenile probation program. Effects on parent and peer relationships. *Crime & Delinquency*, *54*(2), 193–224.

Cohen, M. A., & Piquero, A. R. (2009). New evidence on the monetary value of saving a high risk youth. *Journal of Quantitative Criminology*, *25*(1), 25–49.

Cullen, F. T., & Gendreau, P. (2000). Assessing correctional rehabilitation. Policy, practice, and prospects. *Criminal Justice*, *3*(1), 299–370.

Doleac, J. L. (2019). *Encouraging desistance from crime*. http://jenniferdoleac.com/wp-content/uploads/2019/02/Doleac_Desistance_Feb2019.pdf

Doleac, J. L., Temple, C., Pritchard, D., & Roberts, A. (2020). Which prisoner reentry programs work? Replicating and extending analyses of three RCTs. *International Review of Law and Economics*, *62*, 105902.

Drake, E. K., Aos, S., & Miller, M. G. (2009). Evidence-based public policy options to reduce crime and criminal justice costs. Implications in Washington state. *Victims and Offenders*, *4*(2), 170–196.

Egg, R., Pearson, F. S., Cleland, C. M., & Upton, D. S. (2000). Evaluations of correctional treatment programs in Germany: A review and meta-analysis. *Substance Use & Misuse*, *35*(12–14), 1967–2009.

Entorf, H. (2014). *Jenseits der Fallzahlen: Die Kriminalitätsentwicklung bei ökonomischer Bewertung der Schäden*. https://www.econstor.eu/handle/10419/104974

Erwin, B. S. (1986). Turning up the heat on probationers in Georgia. *Federal Probation*, *50*, 17–24.

Farabee, D., Zhang, S. X., & Wright, B. (2014). An experimental evaluation of a nationally recognized employment-focused offender reentry program. *Journal of Experimental Criminology*, *10*(3), 309–322.

Folkard, M. S., Fowles, A. J., McWilliams, B. C., Smith, D. D., Smith, D. E., & Roy Walmsley, G. (1974). *Impact, intensive matched probation and after-care treatment, Volume I. The design of the probation experiment and an interim evaluation*. Home Office.

Folkard, M. S., Smith, D. E., & Smith, D. D. (1976). *IMPACT: Intensive matched probation and after care treatment 2: The results of an experiment*. Home Office.

Gendreau, P., Goggin, C., Cullen, F. T., & Andrews, D. A. (2000). The effects of community sanctions and incarceration on recidivism. *Forum on Corrections Research*, *12*(2), 10–13.

Georgiou, G. (2014). Does increased post-release supervision of criminal offenders reduce recidivism? Evidence from a statewide quasi-experiment. *International Review of Law and Economics*, *37*, 221–243.

Guydish, J., Chan, M., Bostrom, A., Jessup, M. A., Davis, T. B., & Marsh, C. (2011). A randomized trial of probation case management for drug-involved women offenders. *Crime & Delinquency*, *57*(2), 167–198.

Hawken, A., & Kleiman, M. (2009). *Managing drug involved probationers with swift and certain sanctions: Evaluating Hawaii's HOPE*. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=252477

Huttunen, K., Kerr, S. P., & Mälkönen, V. (2014). *The effect of rehabilitative punishments on juvenile crime and labor market outcomes*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2492430

Hyatt, J. M., & Barnes, G. C. (2017). An experimental evaluation of the impact of intensive supervision on the recidivism of high-risk probationers. *Crime & Delinquency*, *63*(1), 3–38.

Landsberger, H. A. (1958). *Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry*. Ithaca: Cornell University.

Lane, J., Turner, S., Fain, T., & Sehgal, A. (2005). Evaluating an experimental intensive juvenile probation program. Supervision and official outcomes. *Crime & Delinquency*, *51*(1), 26–52.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation. A review of systematic reviews. *Annual Review of Law and Social Science*, *3*, 297–320.

Lipsey, M. W., & Wilson, D. B. (1998). *Effective intervention for serious juvenile offenders. A synthesis of research*. Sage Publications, Inc.

Lösel, F., & Köferl, P. (1989). Evaluation research on correctional treatment in Germany: A meta-analysis. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system. Psychological perspectives* (pp. 334–355). Springer.

Lotti, G. (2020). Tough on young offenders. Harmful or helpful? *Journal of Human Resources*. https://doi.org/10.3368/jhr.57.4.1017-9113R3

Lowenkamp, C. T., Flores, A. W., Holsinger, A. M., Makarios, M. D., & Latessa, E. J. (2010). Intensive supervision programs. Does program philosophy and the principles of effective intervention matter? *Journal of Criminal Justice*, *38*(4), 368–375.

Lowenkamp, C. T., Paler, J., Smith, P., & Latessa, E. J. (2006). Adhering to the risk and need principles. Does it matter for supervision-based programs. *Federal Probation*, *70*, 3–8.

Mastrobuoni, G., & Terlizzese, D. (2019). *Leave the door open? Prison conditions and recidivism*.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect. New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, *67*(3), 267–277.

Meuer, K., & Woessner, G. (2020). Does electronic monitoring as a means of release preparation reduce subsequent recidivism? A randomized controlled trial in Germany. *European Journal of Criminology*, *17*(5), 563–584.

Mueller-Smith, M., & Schnepel, K. T. (2021). Diversion in the criminal justice system. *Review of Economic Studies*, *88*, 883–936.

National Research Council. (2007). *Parole, desistance from crime, and community integration*. National Academies Press.

Pearson, F. S., Lipton, D. S., Cleland, C. M., & Yee, D. S. (2002). The effects of behavioral/cognitive-behavioral programs on recidivism. *Crime & Delinquency*, *48*(3), 476–496.

Petersilia, J. (1989). Implementing randomized experiments. Lessons from BJA's intensive supervision project. *Evaluation Review*, *13*(5), 435–458.

Petersilia, J., & Turner, S. (1990a). Comparing intensive and regular supervision for high-risk probationers. Early results from an experiment in California. *Crime & Delinquency*, *36*(1), 87–111.

Petersilia, J., & Turner, S. (1990b). *Diverting prisoners to intensive probation. Results of an Experiment in Oregon*. Rand.

Petersilia, J., & Turner, S. (1990c). *Intensive supervision for high-risk probationers. Findings from Three California experiments*. Rand.

Petersilia, J., & Turner, S. (1991). An evaluation of intensive probation in California. *Journal of Criminal Law and Criminology*, *82*, 610–658.

Petersilia, J., & Turner, S. (1993). Intensive probation and parole. *Crime and Justice*, *17*, 281–335.

PKS. (2018). *Police crime statistics*. Federal Republic of Germany Report 2018. https://www.bka.de/SharedDocs/Downloads/EN/Publications/PoliceCrimeStatistics/2018/pks2018_englisch.html?nn=113788

Rodriguez-Planas, N. (2012). Longer-term impacts of mentoring, educational services, and learning incentives. Evidence from a randomized trial in the United States. *American Economic Journal: Applied Economics*, *4*(4), 121–139.

Sontheimer, H., & Goodstein, L. (1993). An evaluation of juvenile intensive aftercare probation. Aftercare versus system response effects. *Justice Quarterly*, *10*(2), 197–227.

Stevenson, M. (2017). Breaking bad. Mechanisms of social influence and the path to criminality in juvenile jails. *Review of Economics and Statistics*, *99*(5), 824–838.

Turner, S., & Petersilia, J. (1992). Focusing on high-risk parolees. An experiment to reduce commitments to the Texas Department of Corrections. *Journal of Research in Crime and Delinquency*, *29*(1), 34–61.

Wanner, P. (2018). *Inventory of evidence-based, research-based, and promising programs for adult corrections*. http://www.wsipp.wa.gov/ReportFile/1681/Wsipp_Inventory-of-Evidence-Based-Research-Based-and-Promising-Programs-for-Adult-Corrections_Report.pdf

Wilson, D. B., Gallagher, C. A., & MacKenzie, D. L. (2000). A meta-analysis of corrections-based education, vocation, and work programs for adult offenders. *Journal of Research in Crime and Delinquency*, *37*(4), 347–368.

How to cite this article: Engel, C., Goerg, S. J., & Traxler, C. (2022). Intensified support for juvenile offenders on probation: Evidence from Germany. *Journal of Empirical Legal Studies*, *19*(2), 447–490. https://doi.org/10.1111/jels.12311

APPENDIX A: ADDITIONAL TABLES

TABLE A1 Placebo checks: Discontinuity in observables

	(1) §57 JGG	(2) §27 JGG	(3) Alcohol Addiction	(4) Drug Addiction	(5) Aggression High
Discontinuity	0.0112	−0.0663	0.0814	0.0813	−0.0002
	(0.887)	(0.613)	(0.494)	(0.509)	(0.997)
	(6) Problematic Peers	**(7) Age**	**(8) Gender**	**(9) Violent Crime**	**(10) Property Crime**
Discontinuity	0.0112	−0.0663	0.0814	0.0813	−0.0002
	(0.887)	(0.613)	(0.494)	(0.509)	(0.997)

Note: The table presents RDD estimates at the IP-eligibility cutoff considering observable characteristics. Each entry reports the discontinuity from a different estimation (i.e., a different outcome variable). All specifications control linearly for the correlation between point score and outcomes (differentiating for the range below and above the cutoff). Sample: $N = 171$ in specifications (1)–(8) and $N = 152$ in (9)–(10). Robust SEs are in parenthesis.
Abbreviations: IP, intensive probation; JGG, *Jugendgerichtsgesetz*; RDD, regression discontinuity design.

TABLE A2 Duration analysis: Time until first re-offense

	(1)	(2)	(3)	(4)	(5)	(6)
IP	0.824	0.729				
	[0.514]	[0.343]				
IP × Year 1			0.825	0.664		
			[0.594]	[0.333]		
IP × Months 0–6					0.669	0.493
					[0.365]	[0.214]
IP × Months 6–12					1.260	1.154
					[0.712]	[0.809]
IP × Year 2			0.729	0.806	0.729	0.814
			[0.607]	[0.737]	[0.607]	[0.749]
IP × Year 3			1.053	0.988	1.053	0.990
			[0.952]	[0.988]	[0.952]	[0.990]
Controls	No	Yes	No	Yes	No	Yes

Note: The table presents estimated hazard ratios from Cox proportional hazard models. Specifications (3)–(6) present time dependent hazard ratios. Specifications (2), (4), and (6) control for age, gender, and include dummies for the start of the probation period as well as indicators based on the different dimensions in the judges' scorecards (e.g., alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers). *p*-values, based on robust SEs, are presented in brackets. $N = 57$.
Abbreviation: IP, intensive probation.

TABLE A3 Number of convictions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: All crimes								
Number of crimes	6 months		1 year		2 years		3 years	
IP	−0.148	−0.298	−0.115	−0.178	−0.141	−0.630	−0.104	−0.581
	(0.159)	(0.254)	(0.249)	(0.383)	(0.309)	(0.483)	(0.427)	(0.654)
Controls	N	Y1	N	Y1	N	Y1	N	Y1
RP mean	0.481	0.481	0.815	0.815	1.407	1.407	2.037	2.037
Panel B: By crime type								
	Violent crimes		*Violent crimes*		*Property crimes*		*Property crimes*	
Number of crimes	1 year		3 years		1 year		3 years	
IP	−0.096	−0.223	−0.056	−0.062	−0.011	−0.025	−0.056	−0.238
	(0.137)	(0.176)	(0.238)	(0.367)	(0.176)	(0.237)	(0.244)	(0.342)
Controls	N	Y1	N	Y1	N	Y1	N	Y1
RP mean	0.296	0.296	0.630	0.630	0.444	0.444	0.852	0.852
Panel C: Smaller sample and extended controls								
	All Crimes		*All Crime*		*Violent Crimes*		*Violent Crimes*	
Number of crimes	1 year		3 years		1 year		3 years	
IP	−0.217	−0.279	−0.060	−0.211	−0.523	−0.701	−0.196	−0.257
	(0.276)	(0.261)	(0.383)	(0.376)	(0.722)	(0.702)	(0.142)	(0.164)
Controls	Y1	Y2	Y1	Y2	Y1	Y2	Y1	Y2
RP mean	0.913	0.913	2.130	2.130	0.304	0.304	0.696	0.696

Note: The table presents estimated treatment effects on the number of crimes during time windows between 6 months and 3 years. Panel A counts all crimes, Panels B and C distinguish between different types of crimes. All estimates are based on linear models. Panels A and B are based on the full sample ($N = 57$), Panel C explores a smaller sample for which the extended controls are available ($N = 50$). Robust SEs are in parenthesis. RP mean presents the average number of crimes in the control group with regular probation. The basic control variables (Y1) include age, gender, dummies for the start of the probation period as well as indicators based on the different dimensions in the judges' scorecards (e.g., alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers). The augmented set of controls (Y2) further includees indicators for the type of crime resulting in the probation sentence.

Abbreviations: IP, intensive probation; N, no; RP, regular probation; Y1 and Y2, limited availability of data, the set of controls differs between specifications.

TABLE A4 "Donut" RD estimates (reduced form)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: *Any crime*								
Recidivism	6 months		1 year		2 years		3 years	
Eligibility	−0.181	−0.160	−0.333**	−0.304*	−0.440***	−0.426***	−0.283*	−0.280*
	(0.150)	(0.155)	(0.154)	(0.167)	(0.150)	(0.157)	(0.148)	(0.161)
Controls	N	Y	N	Y	N	Y	N	Y
Panel B: *By crime type*								
	Violent crimes		*Violent crimes*		*Property crimes*		*Property crimes*	
Recidivism	1 year		3 years		1 year		3 years	
Eligibility	0.079	0.048	0.169	0.138	−0.296**	−0.286*	−0.025	−0.066
	(0.112)	(0.116)	(0.134)	(0.146)	(0.147)	(0.164)	(0.166)	(0.180)
Controls	N	Y	N	Y	N	Y	N	Y
Panel C: *Number of crimes*								
Count	6 months		1 year		2 years		3 years	
Eligibility	−0.265	−0.253	−0.220	−0.217	−0.521	−0.605	−0.641	−0.765
	(0.192)	(0.211)	(0.260)	(0.275)	(0.374)	(0.374)	(0.468)	(0.477)
Controls	N	Y	N	Y	N	Y	N	Y

Note: The table presents reduced form "donut" RDD estimates at the IP-eligibility cutoff. Panel A replicates the estimates from Table 4, excluding observations with a score one point below and above the cutoff (i.e., the "donut hole," 12 and 13 points on the scorecard). Sample size thus shrinks to $N = 151$. All specifications control linearly for the correlation between point score and outcomes (differentiating for the range below and above the cutoff). The dependent variables in Panels A and B are recidivism indicators (binary outcomes) for different time periods. Panel A considers any recidivism, Panel B distinguishes between violent and property crimes/re-offenses. Panel C uses a count for (any) crimes as outcome variables. Specifications (2), (4), (6), and (8) add further controls (time period fixed effects for the start of the probation period as well as indicators for alcohol or drug addiction; intermediate or high aggression levels; highly problematic peers). Robust SEs are in parenthesis.
Abbreviations: IP, intensive probation; N, no; RD, regression discontinuity; RDD, regression discontinuity design; Y, yes, the set of control variables is included in the specification.
***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

APPENDIX B: ADDITIONAL FIGURES

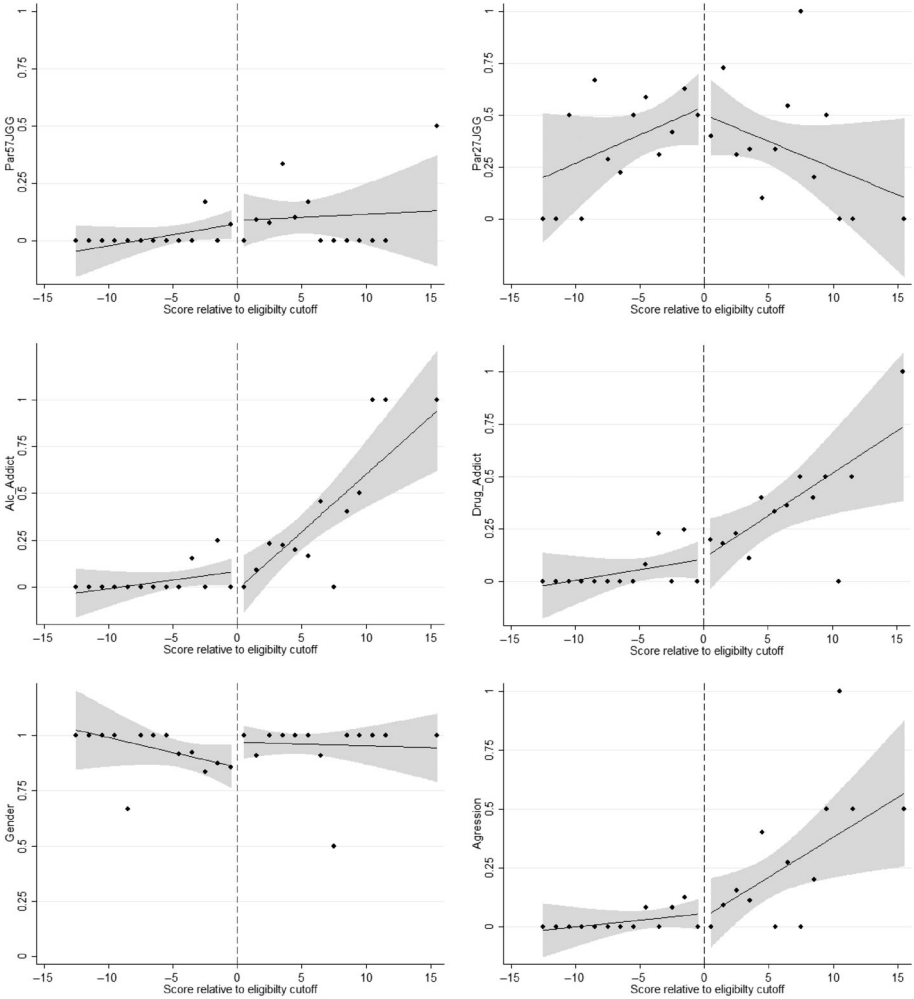

FIGURE B1 Smoothness of covariates around the IP-eligibility cutoff (1/2). The figure presents linear fits (with 95% confidence intervals, allowing for differential slopes on either side of the cutoff) for different (pre-treatment) covariates at the IP eligibility cutoff: § 27 JGG and § 57 JGG (the German juvenile courts law) are dummies capturing case specific information, indicating whether the juvenile court is empowered to be more flexible (see Table 2 for further details). The other variables indicate alcohol or drug addiction, gender (with male defined as 1), and high aggression level. The dots indicate average values for a given point score (relative to the eligibility cutoff). $N = 171$. IP, intensive probation, JGG, *Jugendgerichtsgesetz*

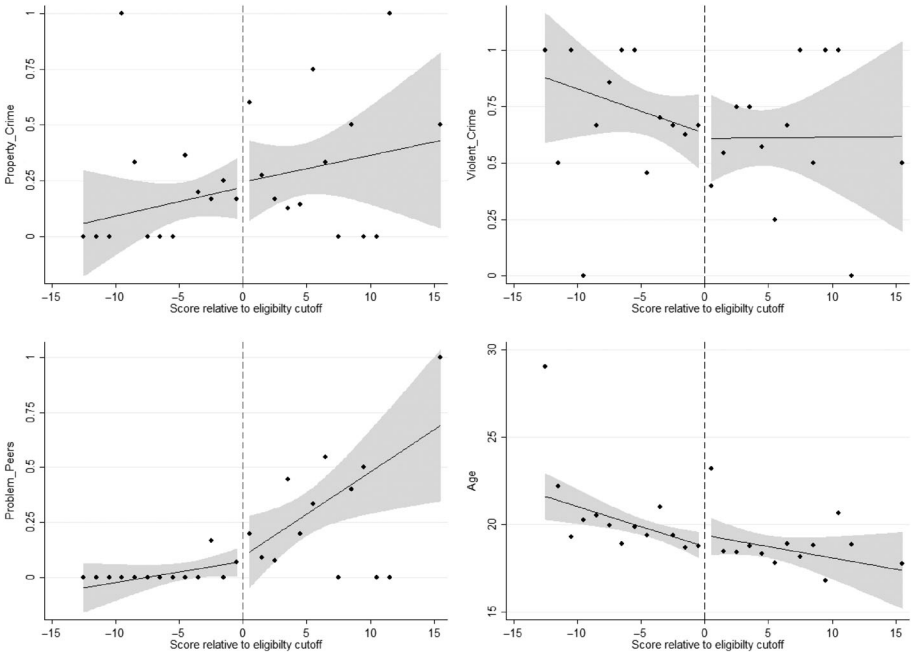

FIGURE B2 Smoothness of covariates around the IP-eligibility Cutoff (continued, 2/2). The figure presents linear fits (with 95%-confidence intervals, allowing for differential slopes on either side of the cutoff) for different (pre-treatment) covariates at the IP eligibility cutoff: convicted for property of violent crime (binary), presents of highly problematic peers, and age (in years). The dots indicate average values for a given point score (relative to the eligibility cutoff). $N = 171$

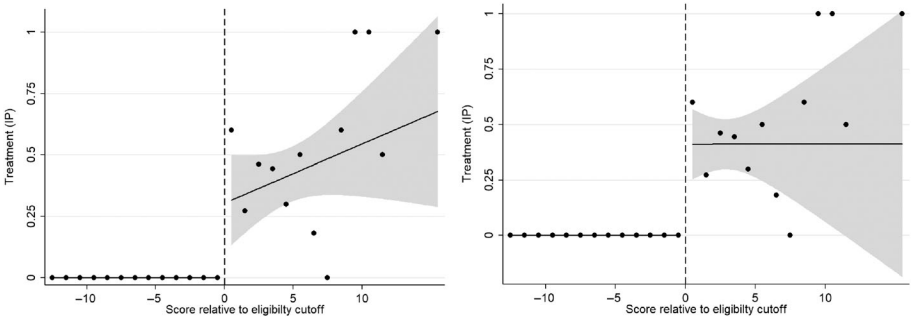

FIGURE B3 Treatment discontinuity at the IP-eligibility cutoff treatment rates: unweighted (left) and weighted fit (right panel). The figure presents estimated discontinuities in the treatment rates: the share of convicts which enter IP (linear fits with 95% confidence intervals). The dots present average treatment rates among convicts with a given point score (relative to the eligibility cutoff). The estimate in the right panel is weighted by the number of observations at a given point score. $N = 171$

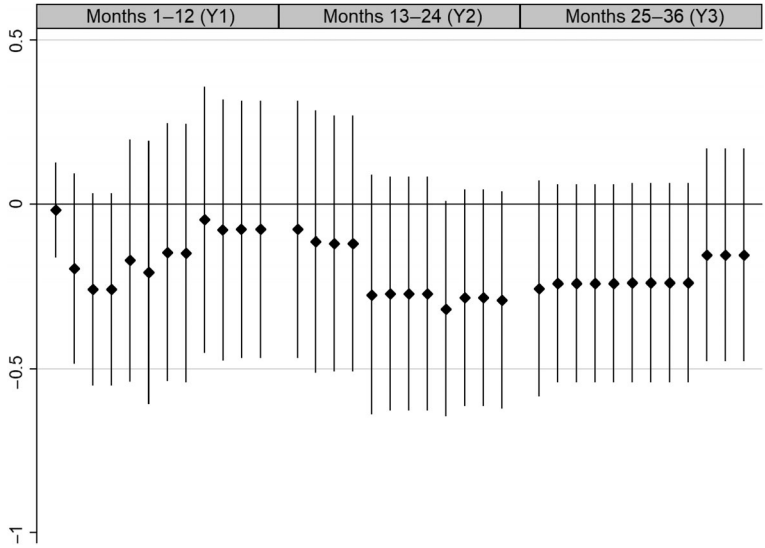

FIGURE B4 RCT: Estimated treatment effects on recidivism over time (binary). The figure presents point estimates and 95% confidence intervals from 36 different estimates of treatment effects on recidivism in the spirit of Table 3. The dependent variable is a dummy indicating a re-offense within m-months, with $m = \{1, \ldots, 36\}$. $N = 57$. RCT, randomized control trial

APPENDIX C: COST–BENEFIT ANALYSES

Let us first turn to the costs of the IP program. Probation officers in the IP program dedicate more time to the juvenile offenders, which reduces the total number of cases they administrate. The annual gross salary of probation officers is based on the tariff agreements for state employees (TV-L 10). For an officer with at least 10 years of experience it is 58,505 Euro. The work load for an officer who is only active in RP is 60 cases and for a probation officer who is active in IP it is 5 IP plus 25 RP cases (at any given month). One IP case is thus equivalent to 7 RP cases. Based on this conversion rate (and accounting for the fact that IP is limited to 6 months) we arrive at additional cost of 1 IP (over 1 RP) case of (6825–975)/2 = **2925 Euro**. Given that other administrative procedures (and costs) do not differ between IP and RP cases, this number gives us a good proxy for the marginal cost of offering one case of intensive rather than regular probation.

Next, we consider the program's benefits in terms of reducing recidivism. Unfortunately, good estimates for the social costs of different crimes in Germany are not available. What is available, though, are reasonable estimates for the victimization costs, which are calculated based on the German Police Crime Statistics, PKS (2018), and Entorf (2014). Some of these estimates are listed below:

Severe theft	2500 Euro	Robbery	8500 Euro
Theft	500 Euro	Aggrav. assault	31,500 Euro
Fraud	2300 Euro	Rape	92,000 Euro

Source: PKS (2018) and Entorf (2014).

Note that studies on the social costs of crime indicate that victimization costs only account for between 33% and 75% of the total direct costs of crime (i.e., including costs for the justice system and offenders' productivity costs). In addition, the willingness-to-pay for crime reductions is typically much higher than the total direct costs of crime (Cohen & Piquero, 2009). Considering only victimization costs, we therefore obtain a very conservative lower bound on the potential program benefits.

With these caveats in mind, we use the estimates from PKS (2018) and Entorf (2014) to compute the average victimization cost of re-offenses in the RCT and the RDD sample. In doing so, we account for the distribution of different types of crimes observed within a 1- and 3-year interval. For the RCT sample, we arrive at an *average* cost per re-offense (within 1 and 3 years) of 11,000 Euro and 12,300 Euro, respectively. In our RDD sample, these costs are smaller with 10,400 Euro and 11,900 Euro, respectively.[29] Based on these numbers, we now use the estimated program impact on recidivism after 1 and 3 years to approximate the gains from lower victimization costs.

RCT Estimates: The point estimates from Table 3 indicate that IP reduces recidivism by 7.7 percentage points within 1 year and by 15.4 percentage points within 3 years (Column 6, Panels A and B). Based on the proxies from above, these effects translate into a decline in victimization costs of 850 Euro within 1 year and roughly 1900 Euro after 3 years. These numbers are clearly below the marginal costs of the program (2925 Euro). However, if the total social costs of crime were at least 50% larger than the mere victimization costs (which seems plausible, see Cohen & Piquero, 2009), the program's benefits after 3 years clearly outweigh the costs (given reasonable discount factors). Note further that Table 4 provides suggestive evidence indicating that the decline in recidivism is mainly driven by a lower risk of violent (rather than property) crimes—in particular, within the first year. This would again imply higher program benefits.

RDD Estimates: The point estimates from the RDD reported in Table 6, Column (4) and (8) indicate 51.1 and 56.0 percentage points drop in recidivism within 1 and 3 years after the start of the probation period, respectively. This, in turn, yields a reduction in victimization costs of approximately 5300 and 6700 Euro, respectively. The estimated treatment effects thus indicate that the

[29]For the RCT sample, the victimization cost of an average violent crime is roughly 24,500 Euro and the average cost of an average property crime is roughly 1800 Euros. In our RDD sample, the corresponding numbers are 26,200 Euros for a violent crime and 2400 Euros for a property crime.

program benefits clearly outweigh its costs, with net benefits of around 2400 Euro after 1 year and 3700 Euro after 3 years. Note, again, that the social benefits would be larger as we only consider the decline in direct victimization costs.

APPENDIX D: JUDGES' SCORECARD

1. General Structural Shortcomings in the Living Environment: Family

No reliable family relationship	3
Participation in family life limited to eating and sleeping	2
Participation in family life in family with poor parenting skills	1
Reliable family relationship	0

2. General structural shortcomings in the living environment: Housing

Lives in a facility	3	2	1	0
Lives in personal apartment	3	2	1	0
Lives in hotel	3	2	1	0
Homeless	3	2	1	0

3. General structural shortcomings in the living environment: School/education

No participation in school/apprenticeship/training/therapy	3	
Poor participation in school/apprenticeship/training/therapy	2	1
Regular participation in school/apprenticeship/training/therapy	0	

4. Personal shortcomings: Structured day

Not maintaining a structured day	3	
Limited ability to maintain a structured day	2	1
Maintaining a well-structured day	0	

5. Personal shortcomings: Fulfillment of duties

Deficient fulfillment of duties	2
Limited fulfillment of duties	1
Good fulfillment of duties	0

6. Peers and social contact

Exclusively problematic contacts	5	4	3	2
Problematic and unproblematic contacts		3	2	1
Exclusively unproblematic contacts				0

7. Addiction problems: Alcohol

Alcohol consumption	3	2	1	0

8. Addiction problems: Drugs

Drug consumption	3	2	1	0

9. Aggression

Level of aggression	6	5	4	3	2	1	0

10. Willingness to actively participate in program

No	1
Yes	0

Exclusion rules

Shortcomings and obstacles with regard to understanding and susceptibility

Every single criterion leads to exclusion

	Yes	No
1. Proband speaks no German		
2. Substantial intellectually poor aptitude, resulting in a significant reduction in the ability to follow instructions		
3. Mental illness (medical diagnosis)		
4. Excessive use of hard drugs (junkie)		
5. Lack of problem awareness regarding own delinquency		
6. Lack of intention to change		
7. Current involvement in at least two active interventions (over-intervention)		

APPENDIX E: PROBATION OFFICERS' SCORECARD

1. General structural shortcomings in the living environment: Family

(Please *tick only once* per time)

	Beginning of probation	After 6 months
No reliable family relationship	3 ☐	3 ☐
Participation in family life limited to eating and sleeping	2 ☐	2 ☐
Participation in family life with poor parenting skills	1 ☐	1 ☐
Reliable family relationship	0 ☐	0 ☐

2. General structural shortcomings in the living environment: Housing

	Beginning of probation	After 6 months
Lives in facility	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐
Lives in personal apartment	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐
Lives in hotel	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐
Homeless	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐

3. General structural shortcomings in the living environment: Education

	Beginning of probation	After 6 months
No participation in school/apprenticeship/training/therapy	3 ☐	3 ☐
Poor participation in school/apprenticeship/training/therapy	2 ☐ 1 ☐	2 ☐ 1 ☐
Regular participation in school/apprenticeship/training/therapy	0 ☐	0 ☐

4. Personal competence shortcomings: Structured day

	Beginning of probation	After 6 months
No structured day	3 ☐	3 ☐
Limited ability to maintain a structured day	2 ☐ 1 ☐	2 ☐ 1 ☐
Well-structured day	0 ☐	0 ☐

5. Personal competence shortcomings: Fulfillment of duties

	Beginning of probation	After 6 months
Deficient fulfillment of duties	2 ☐	2 ☐
Limited fulfillment of duties	1 ☐	1 ☐
Good fulfillment of duties	0 ☐	0 ☐

6. Peers and social contacts

	Beginning of probation	After 6 months
Exclusively problematic contacts	5 ☐ 4 ☐ 3 ☐ 2 ☐	5 ☐ 4 ☐ 3 ☐ 2 ☐
Problematic and unproblematic contacts	3 ☐ 2 ☐ 1 ☐	3 ☐ 2 ☐ 1 ☐
Exclusively unproblematic contacts	0 ☐	0 ☐

7. Addiction problems: Alcohol

	Beginning of probation	After 6 months
Alcohol consumption	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐

8. Addiction problems: Drugs

	Beginning of probation	After 6 months
Drug consumption	3 ☐ 2 ☐ 1 ☐ 0 ☐	3 ☐ 2 ☐ 1 ☐ 0 ☐

9. Aggression

	Beginning of probation	After 6 months
Level of aggression	6 ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1 ☐ 0 ☐	6 ☐ 5 ☐ 4 ☐ 3 ☐ 2 ☐ 1 ☐ 0 ☐

10. Cooperation and willingness to accept support

Beginning of probation	After 6 months
(Fairly) good ☐	(Fairly) good ☐
(Fairly) bad ☐	(Fairly) bad ☐