Does increased post-release supervision of criminal offenders reduce recidivism? Evidence from a statewide quasi-experiment

Georgios Georgiou

Visiting Assistant Professor of Economics, Colgate University*†

Abstract

Approximately 4.8 million offenders are subject to community supervision in the United States. This paper examines whether a program that assigned different supervision levels based on a risk assessment instrument, had any effect on offenders' recidivism rates. Using a large statewide sample of adult offenders in Washington State and a regression discontinuity design, I compare offenders whose risk characteristics are similar but who received different levels of post-release supervision. I find that offenders who received more supervision were not less likely to reoffend. The result holds for high-risk and low-risk offenders and for various types of recidivism.

JEL classification: C26, K42

Keywords: regression discontinuity, instrumental variables, community supervision, recidivism

*Colgate University, Economics Department, Persson Hall, 13 Oak Drive, Hamilton, NY 13346, USA. Tel.: +1 831 4197091. Email: ggeorgiou@colgate.edu.

†I would like to thank my advisors Donald Wittman and Carlos Dobkin for their invaluable help. I am also grateful to Elizabeth Drake and Laura Harmon of the Washington State Institute for Public Policy for providing me with the data sets used in this paper, and to David Daniels of the Washington Department of Corrections for helping me understand the state's community custody programs. I benefited from comments and suggestions from Julian Caballero, David Kaun, Jean Paul Rabanal, Sean Tanoos, and an anonymous referee. I thank them all wholeheartedly.

1 Introduction

In the United States at yearend 2011, 1 in 50 adults were under community supervision—that is, approximately 4.8 million offenders (Bureau of Justice Statistics, 2012). They were either sentenced directly into the community as probationers (around 3,970,000) or released from prison as parolees (around 843,000). This paper seeks to answer the following question: Does a higher level of supervision reduce recidivism rates?

This question is not new in the literature. Under the name Intensive Supervision Program or Intensive Supervision Probation/Parole (in both cases ISP), increased supervision has been studied since the 1960s for both theoretical and practical reasons (Clear and Hardyman, 1990; Petersilia and Turner, 1990). Increased supervision entails either an increased frequency of contact between the authorities and the supervised offender, a decrease in the caseload (number of supervised offenders) for each probation officer, or an increase in the frequency of other supervisory methods, such as drug tests.

At first it was the rehabilitative ideal that gave impetus to the effort of devising effective treatment in the community programs. However, after Martinson's "what works" study (1974), which included a special section on various failed programs, the effort came to a halt; it was revived in the 1980s, when prison overcrowding became a primary source of concern. More recently, contributions from Petersilia (1999, 2009) and Travis (2005) generated a renewed interest in the topic that is focused more on prisoner reentry, the parole side of the community supervision spectrum. However, the balance of the evidence suggests that increased supervision does not improve outcomes.

This study contributes to the research by presenting a quasi-experimental evaluation of a statewide differential supervision program that was based on a risk assessment instrument. Most evaluations of intensive supervision programs thus far have employed small samples and have usually been limited to a specific target group of offenders. By using a sample of considerable size (more than 50,000 observations) and adult offenders of various risk levels, as measured by a risk instrument, the present study expands the scope of previous work.

In particular, the study examines the effectiveness of community supervision by evaluating empirically a treatment program that was initiated in the state of Washington in 1999 and was in force until 2012. The program details are presented in Section 2; briefly, the Washington authorities provided different levels of supervision to offenders who posed different levels of risk for reoffending. Reoffending risk was measured by way of an "actuarial" instrument that generated a "risk score" for each individual offender and allowed the authorities to rank the offenders based on their risk. Using a regression discontinuity design, I compare individuals who had similar risk characteristics, but who received different levels of supervision as determined by their risk scores. This approach provides quasi-experimental evidence on whether extra supervision is effective in reducing recidivism.

As noted, the existing literature does not support the effectiveness of increased supervision.

Randomized controlled trials conducted at fourteen sites across the United States in the late 1980s (involving 2,000 offenders overall) showed that "[i]ntensive supervision probation did not decrease the frequency or seriousness of new arrests" (Petersilia and Turner, 1993, p. 281). This large randomized experiment has been documented by several studies that found similar outcomes for specific categories of offenders as well, such as drug offenders (Petersilia, Turner, and Deschenes, 1992). Some years later, an overall evaluation of intensive supervision and other programs (such as electronic monitoring and home confinement) by Cullen, Wright, and Applegate (1996) found negative results as well. Barnoski (2003) in a study that was conducted at twelve sites in the state of Washington in the late 1990s, involving 1,688 juvenile offenders, found that intensive supervision did not succeed in reducing their recidivism rates. Moreover, Sherman, Gottfredson, MacKenzie, Eck, Reuter, and Bushway (1997) and MacKenzie (2002), reviewing the literature existing at the time, note that there is no evidence that increased supervision reduces recidivism rates. In line with the previous authors, Harris, Gingerich, and Whittaker (2004) showed that differential supervision of offenders, based on a case management classification system, did not produce the desired results either. Finally, meta-analyses by Farrington and Welsh (2005), Bonta, Rugge, Scott, Bourgon, and Yessine (2008), Gill (2010), and Drake (2011) reached similar conclusions.[1]

The present study adds to this literature a quasi-experimental approach to the issue by using the technique of regression discontinuity in the wide setting that was described above. The method has been increasingly popular in the applied economics literature since the late 1990s and has been employed in a variety of settings (Imbens and Lemieux, 2008; Lee and Lemieux, 2010). Specifically, in the economics of crime, regression discontinuity studies have examined a number of different treatments, such as sanction severity and length (Lee and McCrary, 2005; Kuziemko, 2007), and prison security levels (Berk and de Leeuw, 1999; Chen and Shapiro, 2004). Hjalmarsson (2009), in particular, using data from the state of Washington and employing discontinuities in the state's sentencing guidelines, studied the effect of incarceration on post-release behavior of juveniles.

Moreover, Berk, Barnes, Ahlman, and Kurtz (2010) showed that regression discontinuity can actually replicate the results of randomized controlled trials. In fact, the setting that the authors

[1] However, Gill (2010, p. 38) notes that she does not believe that these "results should be taken as conclusive evidence that intensive probation supervision is a failed intervention." In fact, the literature documents intensive supervision programs that had favorable outcomes. However, most of them have a treatment element besides the control and surveillance aspect of supervision. Among them are programs in Georgia (Erwin, 1986) and New Jersey (Pearson, 1988) in the early to mid-1980s that spurred renewed interest on the subject. Additionally, Sherman, Gottfredson, MacKenzie, Eck, Reuter, and Bushway (1997) mention a number of programs from the late 1980s and early 1990s, which were based on increased treatment and showed promising results. More recently, successful interventions have been based on the so-called "responsivity" principle, which also takes into account the learning abilities of the offenders. Such programs with positive results were reported in Maryland (Taxman, Yancey, and Bilanin, 2006), where the method followed was the development of an individualized, treatment oriented, supervision plan, Hawaii (Hawken and Kleiman, 2009), where the focus was on personal responsibility and immediate sanctioning in case of failure, as well as in Oklahoma, Iowa (Jalbert, Rhodes, Kane, Clawson, Bogue, Flygare, Kling, and Guevara, 2011) and Canada (Bonta, Bourgon, Rugge, Scott, Yessine, Gutierrez, and Li, 2011). In 2012, Washington State actually switched to a program similar to Hawaii's (Drake and Aos, 2012).

employed was an experiment that involved a decrease in supervision intensity for low-risk offenders. The results showed that the estimates generated by the regression discontinuity design were effectively identical to those of the randomized trial, providing further justification for the use of regression discontinuity in the present study. Finally, Jalbert, Rhodes, Kane, Clawson, Bogue, Flygare, Kling, and Guevara (2011) used the method to estimate the effect of reduced caseload for probation officers on recidivism of medium- and high-risk offenders in agencies that supplement surveillance with other therapeutic strategies. The objects of their analysis were one county in Iowa and four districts in Colorado.

My results on the effectiveness of the differential community supervision program, in short, are the following. I find evidence that supervision for higher-risk offenders is indeed more intensive. Depending on the specification and the level of risk, offenders assessed as higher-risk receive 20 minutes to 2.4 hours more supervision per month than similar lower-risk offenders, when the average time of supervision per offender is 4.7 hours per month. However, I also find that increased supervision did not reduce recidivism in a substantial or statistically significant way. All my estimates for the effect of increased hours of supervision on recidivism are very close to zero and not significant. The result holds for various types of recidivism, namely felony and misdemeanor, property, drug, and violent recidivism. Moreover, the inclusion of a number of control variables as well as other robustness checks does not change the baseline estimates, and thus provides additional support for the validity of the result.

The policy conclusions of this paper may be extended beyond this negative outcome. In a similar vein with an already existing line of research (Barnes, Ahlman, Gill, Sherman, Kurtz, and Malvestuto, 2010), this study provides support for the idea that for a very specific and identifiable group of offenders, decreased supervision intensity will not necessarily bring about increased recidivism levels. In this respect, resources allocated to supervising those offenders could be used more efficiently elsewhere, thereby increasing social welfare. It should be stressed, however, that given the regression discontinuity setting of the study, this policy implication only applies within certain limitations, which are explained in detail in Sections 4 and 5.

The paper proceeds as follows. Section 2 presents the features of the differential supervision program implemented in Washington State. Section 3 describes the data sets used. Section 4 gives an account of the empirical method used in this paper. Section 5 presents and discusses the results, and Section 6 concludes.

2 Description of differential supervision in Washington

The task of administering the correctional system in Washington State falls to the Department of Corrections (DOC), while research and scientific support is provided by the Washington State Institute for Public Policy (WSIPP). Legislation passed in 1999 (the Offender Accountability Act

(OAA)) required the DOC to determine the proper level of supervision for an offender once he or she is released into the community after serving time in prison or being sentenced directly into the community (on probation). The OAA specified the criteria that determined supervision intensity as: risk of reoffending and seriousness of the offense committed (or "harm done").

Reoffending risk was measured using the actuarial instrument Level of Service Inventory - Revised or LSI-R. This is a popular risk assessment instrument whose qualities and mechanics have been analyzed in detail by numerous studies (e.g., Andrews and Bonta (1995); Gendreau, Little, and Goggin (1996); Gendreau, Goggin, and Smith (2002)). Briefly, the instrument consists of 54 items, each one addressing a specific risk factor. The presence of a risk factor is marked with a 1 and its absence with a 0, the total score of an offender being the sum of 1s. The scale generated runs from 0 through 54, with higher numbers representing higher risk and lower numbers lower risk. The instrument is administered as a questionnaire that offenders complete periodically, most importantly around the time that they will become at-risk to the community—that is, around the time of their release. Their answers are corroborated by official documentation and other methods, where possible.

The LSI-R is built into the system of supervision allocation (Risk Management Identification system (RMI)) provided for by the OAA. Three cut-off scores on the 55-point scale separate the four levels of supervision of the RMI system (RMA, RMB, RMC, and RMD), where the first one corresponds to the highest-risk offenders and the last one to the lowest-risk. The three cut-off scores are 24, 32, and 41. If an offender's score exceeds a cut-off score, he or she is moved to the next supervision category.

The fact that the LSI-R is administered as a questionnaire raises the issue of possible manipulation, since the authorities know the cut-off scores for each category. Manipulation in such cases, originating not with the agents (offenders) but with the authorities, would be what is known in the literature as "complete" (McCrary, 2008) or "precise" manipulation (Lee and Lemieux, 2010) and could potentially undermine the validity of the regression discontinuity design.[2] In that case Lee and Lemieux (2010, p. 299) note that the variation in treatment generated by the regression discontinuity design would no longer be "as good as randomly assigned."

Indeed, in Fig. 1, which presents the distribution of LSI-R scores, note three large jumps. Note also that these jumps occur exactly at the three cut-off points: 24, 32, and 41. This is not a coincidence. Georgiou (2012) showed that the authorities in Washington State used a sophisticated system of instrument manipulation by adding points to the scores of offenders that they thought should be supervised more intensively than their LSI-R score would otherwise warrant. In other words, as the authorities administered the LSI-R questionnaire, if they saw that the LSI-R score

[2] A similar situation susceptible to complete manipulation is discussed by Jacob and Lefgren (2004), who study the effect of summer school and grade retention on academic achievement. In that study, treatment assignment was a discontinuous function of students' past test performance. However, the grades on these tests were generated by teachers, who already knew the points of discontinuity and therefore were, potentially, able to manipulate them.

of an offender would not be high enough to place him or her in a sufficiently intensive supervision category (according to their subjective evaluation), then they would add a few points to the offender's score. This manipulated score would place the offender over (or, as Fig. 1 shows, exactly on) a threshold and thus justify his/her assignment to a more intensive supervision level.

Since the LSI-R score is the variable according to which treatment levels are assigned for the purposes of the regression discontinuity design, its discontinuous distribution on the thresholds constitutes a threat to the validity of the design. However, Georgiou (2012) showed that the manipulated LSI-R distribution depicted in Fig. 1 can be made smoother (see Fig. 2) by removing the items in the questionnaire that the authorities used to add points (the "manipulated" items) to an offender's score, thereby restoring the validity of the design.[3]

Fig. 3 shows that exceeding any of the three cut-off scores considerably raises the probability of being upgraded to the next supervision level. The reason that there is an increase in the probability and not certainty of an upgrade is that the seriousness of the offense committed also plays a role in the decision of the authorities (Aos, 2003).[4] However, as shown in Fig. 3, the LSI-R score is a crucial parameter that the authorities follow in allocating supervision levels.[5]

According to the latest data, in 2005 the annual cost budgeted by the DOC per offender for community supervision was \$5,500 for RMA and RMB offenders, \$1,249 for RMC, and \$505 for RMD.[6] At the same time, the number of hours of supervision budgeted per offender per month were 9.2 hours for RMA, 7.6 for RMB, 5.4 for RMC, and 1.6 for RMD. According to DOC officials, the same number of hours were budgeted in 2011.[7] However, it must be noted that these are budgeted hours; they do not represent the exact amount of supervision an offender receives, but rather the average supervision time allocated to an offender in a specific risk category. This is a limitation of

[3] Georgiou (2012) demonstrated how the manipulation was conducted. Overall, 7 out of the 54 items were used most often by the authorities to add points to the offenders' scores. These were identified as the "manipulated" items. Most of these items had a subjective content (e.g., "poor attitude toward sentence"), which made the manipulation harder to detect or challenge. The study also showed that the manipulation was performed in a professionally impeccable manner. In fact, the manipulated scores predict the time it will take an offender to reoffend at least as well as the non-manipulated, "corrected" scores that the author generated.

[4] Therefore, an offender who has committed a serious offense may be placed in a high supervision category even though his or her LSI-R score is not high enough to justify such a placement. Conversely, an offender who has committed a non-serious offense may be placed in a lower supervision category even though his or her score would normally indicate a higher one.

[5] For example, the probability that an offender is classified at the low-risk RMD category is about 73 percent if his or her score is 23, but falls to roughly 6 percent if his or her score is 24. Conversely, the probability that an offender is placed at the RMC category is approximately 4 percent if his or her LSI-R score is 23, but increases to almost 78 percent if his or her score is 24.

[6] WSIPP obtained these numbers from personal communication with DOC staff (Aos and Barnoski, 2005).

[7] Data from 2002 show that budgeted supervision time was 10.6 hours for RMA, 9.5 hours for RMB, 3.6 hours for RMC, and 0.8 hours for RMD (Department of Corrections, Briefing Document, House Criminal Justice and Corrections Committee, December 6, 2002). Unfortunately I have not been able to verify the exact date that the number of budgeted hours changed from the 2002 numbers to the 2005 numbers. However, again according to personal communication with DOC officials, I was apprised that the most reasonable switch date could be assumed to be the beginning of the 2003 fiscal year, which was July 1, 2003. I used this date in the analysis.

this study.

The specific conditions of community supervision are determined initially by the court and subsequently by the DOC. DOC policies further specify the operation of community supervision and the specific duties of supervision officers. Specifically, officers are required to develop an Offender Supervision Plan that outlines the objectives for each offender, reporting and contact requirements, as well as any specific treatment programs that the offender may be subject to (DOC Policy 380.200). Unless waived by the court, conditions of an offender's supervision include: reporting and being available for contact with the assigned officer, pursuing approved education or employment, and refraining from possessing or consuming controlled substances (Revised Code of Washington (RCW) 9.94A.703(2)). At the discretion of the court, conditions that may be further imposed include remaining within a geographic area, not consuming alcohol, and refraining from contacts with the victim (RCW 9.94A.703(3)). The DOC may impose additional conditions that are reasonably related to the crime of conviction, the offender's risk, and the safety of the community (RCW 9.94A.704(2)(a) and DOC Policy 390.600).

The officers ensure that the offenders comply with these conditions. According to personal communication with DOC staff, the officers check on the offenders in various ways: meeting with them in and outside the office (on field visits), contacting persons associated with the offenders (collateral contacts), performing urinalysis (for drug offenders), and ensuring that offenders have a job, go to school, or have a place to live. As expected, the higher the risk category of the offender (and, accordingly, the more hours budgeted for his or her supervision), the more checks are performed.[8]

Washington has adopted a determinate sentencing system, and thus there is no discretionary parole for incarcerated offenders. Community supervision is imposed on offenders who are either released from prison or sentenced directly into the community on probation. The length of supervision ranges from less than a year to three years, depending on the nature of the offense committed.[9] However, I do not have information on the length of the community supervision term imposed on each individual offender. So even though offenders are followed for recidivism events for a period of three years, I do not know the exact amount of time they were subject to supervision. According to the law, high-risk offenders should be supervised for the whole three-year term, and low-risk offenders for shorter terms. But because I do not know the exact length of the supervision, for my baseline specification I assume it to be the same for all offenders and spanning the entire three-year

[8]With respect to monthly contacts for example, low-risk RMD offenders were required to contact the DOC if they changed their address, place of employment, or contact information. RMC offenders were required to contact the DOC at least once a month. Accordingly, higher-risk offenders classified as RMB and RMA, were required to be contacted more times per month (see also the Attachment to DOC Policy 380.200, which outlines minimum contact standards under the current regime).

[9]According to RCW 9.94A.701, the increments, depending on the severity of the offense, are: three years, eighteen months, and one year. According to RCW 9.94A.702 community supervision terms shorter than a year may be imposed on offenders who were sentenced to one year or less.

period. This assumption could influence my estimates of the effect of supervision on recidivism by understating the difference in the treatment levels, since higher-risk offenders will be supervised for a longer period than lower-risk offenders. This is why Section 5 provides several robustness checks to test whether the baseline estimates are sensitive to different lengths of supervision.

As early as 2003, the officials of WSIPP proposed modifications to the LSI-R tool in order to improve its predictive power. Finally, in August 2008 the authorities discontinued the use of the LSI-R instrument and replaced it with another instrument designed specifically for the needs of Washington State (Barnoski and Drake, 2007; Drake and Barnoski, 2009).

The overall effectiveness of allocating supervision based on risk characteristics as determined by the OAA was assessed by WSIPP in two reports, an interim and a final one (Aos and Barnoski, 2005; Drake, Aos, and Barnoski, 2010). Although the interim report (2005) presented optimistic preliminary results, the final report (2010) presented a much more ambiguous picture of the effectiveness of the entire project.

In its final report (2010, p. 5), WSIPP notes that the ideal way to evaluate the OAA would be by comparing offenders randomly assigned to supervision under the OAA with those that were not, but "[s]ince the OAA was simultaneously implemented statewide, . . . , random assignment was not possible." As an alternative, I believe that the regression discontinuity design of this paper is a valid way to sidestep this obstacle and obtain a quasi-experimental evaluation of the OAA's effectiveness.

3 Data

All of the data used in this study were provided by WSIPP. Specifically, WSIPP made available to me data on the LSI-R score dating from 1999 to 2008. The data consist of 437,335 unique scores for 110,421 individual offenders (many individual offenders have been evaluated by way of the LSI-R questionnaire multiple times). WSIPP also provided data on the risk classification of offenders according to the RMI system. The data consist of 244,602 unique classifications for 92,358 individual offenders dating from 2000 to 2008. As with the LSI-R score, many offenders were categorized multiple times. Finally, WSIPP provided demographic, criminal history, and recidivism data for offender cohorts from 1990 to 2004, consisting of 303,190 unique cases for 197,119 individual offenders.

After merging the three data sets, following WSIPP's precedent (Barnoski and Aos, 2003), I restricted my sample to cases in which the LSI-R was administered within 90 days from the day an offender became at-risk to the community. If there was more than one LSI-R interview within that period, I used the one closest to the release date. Therefore the data set is organized by offense, and each offense corresponds to only one interview. The resulting data set contains 51,957 offenses committed by 47,154 individual offenders, spanning the period July 2000 to September 2004.

Table 1 presents summary statistics on offenders' gender, race, and age, as well as the nature of the most serious offense each one was currently convicted of (in seven broad categories) and the type of sentence (imprisonment or community sentence). The sample consists mostly of male, white, adult (young and early middle-age) offenders who served sentences in the community; almost one-third of the offenses are drug related. Table 1 presents also similar information for the four risk categories. Overall, the lower-risk categories, RMC and RMD, represent about 70 percent of the sample, and the higher-risk categories, RMA and RMB, the remaining 30 percent. With respect to supervision intensity, the table shows that average time of supervision for the entire sample is 4.7 hours per month. The risk categories exhibit considerable variance, with offenders assigned to the highest-risk category (RMA) receiving 10.2 hours of supervision and those assigned to the lowest-risk category (RMD) receiving 1 hour.

Table 2 presents recidivism summary statistics for the sample. A recidivism event is any conviction for a felony or misdemeanor offense committed by an offender within a 36-month period after becoming at-risk in the community (Drake, Aos, and Barnoski, 2010).[10] A 12-month adjudication period is also allowed. Therefore technical violations of community custody conditions do not constitute recidivism events for the purposes of this study.[11] Overall, the sample has a general recidivism rate of almost 50 percent. More specifically, the three highest supervision categories (RMA, RMB, RMC) display a recidivism rate of 55-58 percent, while the lowest supervision category (RMD) has a recidivism rate of 32 percent.

4 Empirical methodology

In this study I use the results obtained in Georgiou (2012) with respect to the items of the LSI-R questionnaire that were manipulated. As noted in Section 2, that paper identified the manipulated items and produced a new "corrected" distribution of the LSI-R scores by removing those items from the calculation of the index. This new distribution (Fig. 2) is smoother since all manipulation has been eliminated, and thus allows me to study similar individuals with similar LSI-R scores. I can therefore proceed to my current task, which is to assess how large an effect, if any, increased supervision has on recidivism.

More specifically, I want to examine whether the difference in the treatment received by the four groups of offenders (RMA, RMB, RMC, and RMD) had a bearing on their respective recidivism rates. As already noted, the treatment is the level of supervision (measured in hours) an offender receives once he or she is released (at-risk) into the community. The question is how to find a

[10] Even though the literature seems to be in favor of arrests as an outcome variable (Petersilia and Turner, 1990), the state of Washington lacks a statewide arrests' record, and thus the use of conviction data is preferable (Barnoski, 1997).

[11] The lack of information on technical violations constitutes another limitation of the present study. I will address this issue at length in Section 5.

causal effect between the hours of supervision and the recidivism rate.

The answer that is proposed in this study is the regression discontinuity design. Having the offenders' risk scores ordered by the LSI-R instrument allows me to use this econometric technique. In particular, I compare individuals who have very similar scores on the LSI-R index, but who, because of the cut-off scores, have been classified in different risk categories and therefore treated differently. The regression discontinuity allows me to identify individuals with virtually identical risk characteristics and thereby reach a conclusion about the possible causal effect of hours of supervision on the recidivism rate.

According to the relevant literature, a necessary condition for a valid regression discontinuity design is the absence of discontinuities in the distribution of the assignment variable (in this case the LSI-R instrument) close to the thresholds (McCrary, 2008). This requirement has been met after the correction of the LSI-R scores as shown in Fig. 2.

In addition, the validity of the design relies on the absence of discontinuities in the distribution of observed characteristics at the cut-off points (Imbens and Lemieux, 2008; Angrist and Pischke, 2009; Lee and Lemieux, 2010), as well as on the assumption that the unobserved characteristics do not demonstrate such discontinuities at the cut-offs. Table 3 provides weak but corroborating evidence that there are dissimilarities between offenders around the three thresholds for the manipulated distribution of LSI-R scores depicted in Fig. 1. The distribution patterns for different observed characteristics are presented graphically in Fig. 12 of Appendix A. Overall, Table 3 shows that for five out of the 42 regressions run, the characteristics are not comparable across the thresholds.

With respect to the specific characteristics that were found to be dissimilar, note that four out of the five concern a discontinuity around threshold 32. Three of them are crime characteristics, namely, drug offenses, assaults, and robberies. The most pronounced dissimilarity is with respect to offenders that were imprisoned prior to being placed under supervision (as opposed to being placed under supervision without prior imprisonment). The probability that an ex-prisoner has a score that is over 32 is almost 5 percentage points higher than having a score that is under 32. Finally, note that black offenders are 3 percentage points more likely to be over threshold 41 than under it.[12]

Table 4 performs the same function but for the corrected LSI-R distribution that corresponds to Fig. 2. The table shows that, after the removal of the manipulated items and the recalculation of the LSI-R scores, there is only one regression for which the characteristics across the thresholds are not comparable, namely crimes related to weapons for the high threshold (35 in the corrected distribution). The corresponding graphical presentation is given by Fig. 13 in Appendix A.[13]

[12] Based on these results, one cannot safely conclude that the authorities systematically targeted specific crime or other characteristics when they added points in order to manipulate an offender's LSI-R score.

[13] Figs. 12 and 13 of Appendix A show the distribution of all the observed characteristics that are known to me. The three dashed lines in each panel correspond to the cut-off points of the manipulated and the corrected distribution respectively. Regression lines around the cut-offs have been omitted for simplicity. However, the regression

The risk instrument defines, not fully but still strongly, the risk category an offender is assigned to. This is known in the relevant literature as "fuzzy regression discontinuity" (Imbens and Lemieux, 2008; Angrist and Pischke, 2009). Unlike the "sharp" design of Berk, Barnes, Ahlman, and Kurtz (2010), where assignment to treatment was a "deterministic and discontinuous function" of a risk score scale (Angrist and Pischke, 2009, p. 251), in this study the risk score is not the sole criterion for treatment assignment. However, in Fig. 3 I present evidence that it is a crucial criterion, or in the words of Hahn, Todd, and Van der Klaauw (2001, p. 202), "the probability of receiving treatment... is discontinuous" at the three thresholds of the risk scale.

As indicated above, the seriousness of the offense committed also plays a role in determining the intensity of supervision (RMI level). In Fig. 3, note that not all offenders who have a score that is below the threshold score of 24 are characterized as RMD, and conversely, there are offenders who are characterized as RMD even though their score is higher than 24.[14] Therefore there are offenders with scores under threshold 24 who are getting more supervision than their score would warrant, and vice versa. Obviously, this pattern creates a more blurry or "fuzzy" picture than the "sharp" design as to the assignment of treatment levels. The fuzzy regression discontinuity design addresses this problem by actually calculating the magnitude of difference in treatment for offenders whose scores are near the thresholds. In order for the design to be valid, that difference has to be strong and statistically significant. In Section 5, I show that this is the case for this study.

Given that I am interested in examining the effect of supervision on recidivism, the simplest specification that could capture the relationship is:

$$recidivism_i = \beta_0 + \beta_1 supervisionhours_i + \epsilon_i. \tag{1}$$

This is a binary response model, where $recidivism_i$ is the outcome dummy variable that takes the value 1 if an offender has reoffended after being at-risk in the community and 0 otherwise. The variable $supervisionhours_i$ has four possible levels that correspond to the risk category to which an offender has been assigned. Then the estimate, $\hat{\beta}_1$, will measure the effect on recidivism of receiving one additional hour of supervision as a result of being assigned to a category for higher-risk offenders.

The problem with this specification is that it is likely to result in estimates that are not consistent because there may be other unobservable characteristics correlated with hours of supervision, and which also affect recidivism (Dobkin, Gil, and Marion, 2010; Anderson, Dobkin, and Gross, 2012). The goal of my analysis is to tackle this problem by coming back to $\hat{\beta}_1$ through the channel of the fuzzy regression discontinuity design, which is equivalent to an instrumental variables approach (Hahn, Todd, and Van der Klaauw, 2001; Angrist and Pischke, 2009).

discontinuity coefficients are reported in Tables 3 and 4 of the body of the text.

[14] The same observation holds for thresholds 32 and 41.

The first stage of my instrumental variables model is to regress the hours of supervision on the risk score variable. I run this regression three separate times, one for each of the three thresholds of the manipulated LSI-R distribution, and I repeat the process with the corrected distribution. For simplicity, I present here a generic form of the regression equation pertaining to any of the thresholds; therefore the term *threshold* below refers to the numbers 24, 32, or 41, (manipulated scores) or 21, 27, or 35 (corrected scores) depending on the threshold used in the regression.

Most importantly, each regression is focused around its respective threshold. Specifically, I use 5 points below each threshold and 6 points above it (including the threshold itself). This window of observations applies to both the first stage and the reduced form of my model, for both the manipulated and the corrected data.[15]

The exact generic specification for the first stage is the following:

$$supervisionhours_i = \alpha_0 + \alpha_1 rd_i + \alpha_2(score_i - threshold) + \alpha_3 rd_i(score_i - threshold) + u_i, \quad (2)$$

where rd_i is the regression discontinuity dummy defined as an indicator variable that takes the value 1 if an offender's risk score is on or above the respective threshold and 0 otherwise ($rd_i = 1\{score_i \geq threshold\}$). The variable $score_i$ represents the risk score of an offender on the LSI-R grid. A linear polynomial of the risk score concludes the list of the right-hand side variables. The estimate $\hat{\alpha}_1$ measures the degree of association between an offender's score on the LSI-R instrument and his or her hours of supervision. In order for my instrument to be valid, this association needs to be strong and statistically significant.

The reduced form of my model will identify the effect of having a score that exceeds the threshold on recidivism. The relevant specification is the following:

$$recidivism_i = \pi_0 + \pi_1 rd_i + \pi_2(score_i - threshold) + \pi_3 rd_i(score_i - threshold) + v_i, \quad (3)$$

where all the variables have already been defined. Then the IV estimator is just $\hat{\beta}_{1IV} = \hat{\pi}_1/\hat{\alpha}_1$. Therefore, with the help of the regression discontinuity and the instrumental variables technique, I can get an answer to my initial question: What is the effect of increased supervision on recidivism rates?

The discreteness of the assignment variable (the LSI-R score) brings up certain additional issues that need to be tackled. First, following the literature (Lee and Card, 2008), since I do not observe individuals that were not treated on the thresholds, I construct a counterfactual value for them with respect to both the outcome variable (recidivism) and the treatment variable (hours of supervision).

[15] Therefore, to use the terminology of Angrist and Lavy (1999), my "discontinuity samples" (the range of observations around the discontinuity points) are the following: When I work with the manipulated scores, for threshold 24, I use observations in the range 19-29, for threshold 32 observations in the range 27-37, and for threshold 41 observations in the range 36-46. When I work with the corrected scores, for threshold 21, I use observations in the range 16-26, for threshold 27 observations in the range 22-32, and for threshold 35 observations in the range 30-40.

This is done by linearly extrapolating these variables based on data from the preceding discrete values of the LSI-R score.[16] Second, again following the literature (Lee and Lemieux, 2010), I use the means of the treatment variable (hours of supervision) and the outcome variable (recidivism) for each level of the assignment variable (LSI-R score) to graph the first stage and the reduced form of my model. The discreteness of the LSI-R score simplifies the problem of bandwidth selection for the purposes of this study.

At this stage, it should be noted that the normal limitation of such a regression discontinuity design is that it identifies treatment effects locally in a twofold sense (Angrist and Pischke, 2009). First, the treatment effects refer only to those offenders who actually received increased supervision when their LSI-R score exceeded the relevant threshold (the "compliers") and not those who did not receive it. This is an implication of the fuzziness of the design. Second, the treatment effects are only applicable to those offenders with an LSI-R score that is close to the thresholds and not to offenders with a score that is far away from them.

In the analysis below, I also explore alternative specifications, using the richness of the information available in the data set. As far as recidivism is concerned, I can identify the effects of supervision on different levels of offense, such as felony or misdemeanor, violent or not, as well as on different types of crime, such as drug or property. With regard to the right-hand side variables, apart from the simple specifications presented above in Eqs. (1)-(3), I also refine the analysis by adding control variables, measuring different characteristics of the offenders, namely gender, age, race, type of crime, type of sentence (prison or community), and number of prior adult felony adjudications. Furthermore, I examine whether extra supervision had an effect on specific population groups. I accomplish this by running my model exclusively for offenders that belong to a particular demographic, age, or other group.

Finally, I cross-check the validity of my estimates by testing alternative windows around the cut-offs. As noted, the baseline specification includes 5 points below and 6 above the thresholds (including the threshold itself). For robustness purposes, I both shrink and widen this window to verify that my estimates are not sensitive to such specification modifications.

[16] For example, in the first panel of Fig. 6, the solid dots are the means of the hours of supervision for each level of the LSI-R score. The two regression lines fit the micro data for the same levels of the LSI-R score. I used the predicted values for scores 19-23 to create a counterfactual value of the hours of supervision that offenders with a score of 24 would have received, if 24 had not been the cut-off point. Then the regression discontinuity estimate measures the difference between this counterfactual value for offenders with a score of 24 and the actual predicted value. Graphically, this difference is the vertical distance between the two fitted lines in the three panels of Fig. 6. The numerical value of the regression discontinuity (RD) estimates and their standard errors (SE) are reported above each graph.

5 Results and discussion

The short answer to the issue of whether the differential supervision program used in Washington worked is "no." The detailed answer is presented in Tables 5 and 6 and Figs. 4-11 and further explained in this section.

The first and simplest way of obtaining the IV estimator for the effect of hours of supervision on recidivism is to run Eqs. (2) and (3), the first stage and the reduced form of the model respectively. The results for different recidivism outcomes are presented in Table 5 for each of the three thresholds. Moreover, I use both the manipulated (first three columns) and the corrected LSI-R scores that were obtained after cleansing the instrument of the contaminated items (next three columns). The use of the manipulated scores, even though they do not allow me to capture individuals with similar risk characteristics around the cut-off points, is warranted because they represent the original risk decisions made by the Washington State authorities. Therefore they can serve as a benchmark, to which the estimates from the corrected scores can be compared.

Starting the analysis from the manipulated scores, the first section of Table 5 presents the RD estimates for the first stage of the model—that is, the effect of the risk score on hours of supervision. For all three thresholds, the instrument (RD dummy) is strong and statistically significant. Indeed, offenders above each threshold receive additional hours of supervision. This effect is also captured graphically in panel (a) of Fig. 4, and also in Fig. 6, which focuses around the thresholds. The first and especially the third threshold entail a more significant increase in the hours of supervision (1.6 and 2.4 more hours of supervision respectively) than the middle threshold (around 50 minutes more supervision).

The second section of Table 5 reports the RD estimates for the reduced form. The first row of this section refers to any kind of recidivism, and the subsequent rows further differentiate the effect for different types of recidivism. Having a score that exceeds any of the three thresholds is not associated with a higher or lower recidivism rate of any type. The RD estimates for general type recidivism are also depicted in panel (b) of Fig. 4 as well as in Fig. 7.

Finally, the third section of Table 5 gives the IV estimates, which capture the effect of hours of supervision on recidivism rates. Here, the IV estimates for all thresholds and types of recidivism are very close to zero, and none are statistically significant. This result holds throughout the rest of the analysis, as I perform several robustness checks.

Turning to the corrected scores that are reported in the rightmost three columns of Table 5, the RD estimates for the first stage are now attenuated, as can also be seen in panel (a) of Fig. 5 and in Fig. 8, but they are still strongly significant. Their sizes for the first, second, and third thresholds are 35 minutes, 20 minutes, and 70 minutes more time of supervision respectively. The attenuation can be attributed to the fact that, after the removal of the manipulated LSI-R items and the recalculation of the scores, the new thresholds do not perfectly capture the change in the amount of supervision.[17] The reduced-form RD estimates for general recidivism are still close to zero and non-significant (panel (b) of Fig. 5 and Fig. 9). As a result, the IV estimates for the corrected scores also show no effect of extra supervision on recidivism.

For the other types of recidivism, except for two cases, the situation is generally similar. The IV estimates reveal again that, except for property felony offenses and misdemeanor offenses, there is no effect of hours of supervision on recidivism. For the former case, the results show that crossing the last threshold and getting more supervision reduces recidivism by 2.6 percent, while for the latter case, surprisingly, crossing the same threshold increases recidivism by the same amount. Given that I ran a large number of regressions for many different outcomes and found no overall effect of supervision on recidivism, the significance of these two conflicting results should not be overstated.

In Table 6 I simplify the analysis even more by not differentiating the effects of supervision by threshold. Specifically, I ask the following question: What is the effect of increased hours of supervision if an offender has an LSI-R score that is greater than *any* threshold? The observations that were used for each of the three threshold-specific regressions (5 points below each threshold and 6 above) were pooled in a single data set for the manipulated scores, and in another data set for the corrected scores. The first column presents the estimates for the manipulated scores and the second column the estimates for the corrected scores. Now the regression discontinuity dummy takes the value 1 if the score is greater than or equal to *any* of the three thresholds. I call this specification, the "combined-threshold" analysis.

The difference in the first stage between manipulated and corrected scores observed in the threshold-specific regressions persists here too; 1.5 more hours of supervision, according to the manipulated scores, compared to 40 minutes more, according to the corrected scores. This is displayed graphically in the difference between panel (a) of Fig. 10 and panel (a) of Fig. 11. The IV estimates for both manipulated and corrected scores again show no statistically significant effect of extra hours of supervision on any type of recidivism. This is my preferred specification because it summarizes clearly and succinctly the absence of any effect of increased supervision on recidivism for any of the three thresholds.

Appendix A provides several robustness checks of the previous results. Specifically, Appendix Tables 7-9 indicate that the results are robust to the inclusion of control variables. The controls are powerful predictors of recidivism but they do not affect the size of the IV estimates. In addition, in Appendix Figs. 14-16 and Tables 10 and 11, I deal with the problem of the attenuated first stage in the corrected scores by removing observations that correspond to the 2 points around the thresholds. The first stage is indeed strengthened, but the IV estimates remain close to zero and not significant.

[17] Appendix A and, specifically, Figs. 14-16 and Tables 10 and 11 offer a solution that addresses this issue.

Appendix Table 12 deals with the issue of the length of the supervision term. As indicated in Section 2, for the baseline specification the assumption was made that all offenders are supervised for the entire three-year term during which they are followed for recidivism events. To test whether this assumption influences the effect of supervision on recidivism, four shorter supervision terms are introduced, namely two years, one and a half years, one year, and six months. For example, when the supervision period is two years, the recidivism event must have occurred within that period. The shorter the term, the more likely it is that all the offenders in the relevant sample are actively under supervision and that supervision has not lapsed for a certain number of them. The table uses recidivism as the outcome variable and provides evidence that even if shorter terms are assumed, the IV estimates remain largely unchanged and always insignificant.

Appendix Table 13 provides evidence on whether supervision was effective for particular demographic, age, or crime groups. The outcome variable is recidivism, and the IV estimates for both the manipulated and the corrected scores are presented. The only statistically significant result is for black offenders above the high threshold of the corrected scores, indicating a recidivism reduction of 9.3 percent. The concerns expressed above regarding the reliability of such isolated significant estimates apply in this case as well. However, this shred of positive evidence for supervision should not go completely unnoted. The rest of the estimates remain generally close to zero and not significant, thereby reinforcing the result obtained for the whole sample.

Finally, Table 14 of Appendix A tests the sensitivity of the results to different regression windows around the thresholds. The baseline specification is 5 points around the thresholds and in the table I both shrink (to 3 and 4 points around the thresholds) and widen (to 6 and 7 points around the thresholds) this baseline window.[18] The IV estimates do not change substantially from the baseline specification of 5 points that was presented in the body of the text. Specifically, the estimates still remain around zero and not significant both for the manipulated and the corrected data.

As indicated in Section 3, the lack of information on technical violations is a source of concern for the validity of the estimates presented in this study. It is a known fact in the literature that intensive supervision increases the number of technical violations detected for the group that is being supervised intensively (see, e.g., MacKenzie (2002); Gill (2010)). Therefore, offenders who are supervised more intensively are more likely to be found in violation of community custody conditions and sent to jail or prison, where they are in no position to reoffend. This might have a confounding effect on the results reached by the present study.

There are several reasons I believe that the inclusion of technical violations as an outcome variable would not change the size and significance of the estimates derived. First, Drake and Aos (2012) actually evaluated the effect of technical violations on felony recidivism in Washington from

[18] It should be noted, though, that widening the window was only possible for the manipulated scores. Widening for the corrected scores was impossible because the window would overlap with the next threshold. For example, a 6-point window above threshold 21 would end at score 27 which is the threshold for the next supervision category.

2002 to 2008. They report that on average only 4 percent of the offenders under active supervision are incarcerated for a technical violation and receive an average confinement term of 30 days.[19] Second, Petersilia and Turner (1990, 1993) showed that technical violations are not negatively correlated with arrest rates. Most important, Drake and Aos (2012) demonstrated, specifically for Washington, that the use of confinement does not reduce felony recidivism.[20] Finally, the issue at hand is related to the length of the supervision term, which was discussed above. The shorter the supervision term, the more unlikely it is for a technical violation to occur for any given offense. Therefore, even if estimates for longer supervision terms (such as my baseline specification of a three-year term) suffer from the omission of technical violations, estimates for terms as short as six months should not suffer at all, or at least, as much from that omission. However, Table 12 showed that shorter term specifications generated results that were equivalent to the baseline specification. This is another indication that the omission of technical violations should not have influenced the size of the estimates.

Another concern related to the above analysis could be the so-called "surveillance effect." According to this argument, it could be that more surveillance actually reduces recidivism, but also reveals the commission of crimes that would otherwise go undetected (Drake, Aos, and Barnoski, 2010). The two effects of surveillance go in opposite directions and a potential *ceteris paribus* recidivism reduction might be concealed by the greater number of crimes observed.

Given the large number of regressions and the different recidivism measures employed in the analysis, it would be most unlikely for these two effects to counteract each other perfectly for all the regressions, generating IV estimates that clearly show no effect of supervision on recidivism. Therefore I believe that the surveillance effect should not be a source of concern for the validity of my estimates.

The above result is undoubtedly discouraging for purposes of policymaking. A treatment method on which the post-release correctional system relies was found to be non-effective. If that is true, then the question arises: Why should society maintain a supervision program that does not work? The conclusion of the present line of research is that society would increase its welfare by reducing supervision intensity for those categories of offenders that were found not to be responsive to it.

As indicated in Section 4, regression discontinuity by construction can only identify treatment

[19] Since the mid-1980s (Petersilia and Turner, 1993) and until the present time, the violation of community custody conditions in Washington State may be sanctioned with up to 60 days' confinement per violation, when the condition violation is sanctioned by the court and 30 days' confinement, when the violation is sanctioned by the Department of Corrections (RCW 9.94A.633 and 2012 Washington State Adult Sentencing Guidelines Manual).

[20] Drake and Aos (2012) note that their study only measured the recidivism effect and not the "incapacitation" effect, that is the crimes in the community that have been prevented due to the confinement of the offenders. However, the small percentage of offenders under supervision that are confined for technical violations (4 percent on average) in combination with the short average confinement terms for those that end up being confined (30 days) indicate that the incapacitation effect would not substantially affect the estimates of the present study.

effects locally and cannot suggest general approval or disapproval of a program. For the purposes of my study, the non-responsive offenders were those who were just above the thresholds that separated each supervision level. For those offenders and only those, the study suggests that supervision intensity can be reduced without the risk of triggering higher recidivism rates. This finding does not apply to offenders with LSI-R scores far from the thresholds, such as extremely high-risk offenders. Such a policy conclusion, even though it is limited, can provide significant guidance to the competent authorities to rearrange the thresholds for the various supervision levels accordingly. Given the considerable cost of supervision programs, this would seem to be a welfare-enhancing move from a policy perspective.

6 Conclusion

In this paper I assess whether a treatment program that was implemented in Washington State and entailed different levels of supervision for offenders with different levels of assessed risk had any effect on the probability of recidivism. I use a large statewide sample of adult offenders and a regression discontinuity design to evaluate the overall effectiveness of assigning different levels of supervision based on a risk instrument, the LSI-R.

I find that the program succeeded in allocating different amounts of supervision intensity to offenders who posed different levels of risk. However, the data, in all their forms, show that the program did not succeed in reducing recidivism rates for the offenders who received more intensive supervision. Even though I made use of various recidivism types and different specifications, I could not verify any substantial or statistically significant recidivism reduction. Moreover, the result holds equally for high-risk and low-risk offenders.

The potential policy implication for correctional authorities is that a reduction of supervision intensity for specific groups of offenders would not cause an increase in their recidivism rates. However, it is worth noting that my results are local estimates of increased supervision around the three cut-off points of the LSI-R distribution and therefore do not apply to offenders who are far from these points (e.g., extremely high-risk offenders).

From 2000 to 2011, the number of offenders under community supervision increased from approximately 4.5 to 4.8 million. These offenders will be in the community whether community supervision works or not, since budget constraints prevent other more severe forms of punishment. It is therefore important for research in the field to identify the programs that do work, because not doing so will not be merely a minor scientific setback; it will be a serious failure whose victims will be both the offenders and society at large.

References

Anderson, M., C. Dobkin, and T. Gross (2012): "The effect of health insurance coverage on the use of medical services," *American Economic Journal: Economic Policy*, 4(1), 1–27.

Andrews, D., and J. Bonta (1995): *LSI-R user's manual.* Multi-Health Systems Inc., Toronto, Ontario, Canada.

Angrist, J., and J. Pischke (2009): *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press, Princeton, New Jersey, USA.

Angrist, J. D., and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly Journal of Economics*, 114(2), 533–575.

Aos, S. (2003): "Washington's Offender's Accountability Act: Update and progress report on the Act's evaluation," (Document No. 03-01-1201). Olympia: Washington State Institute for Public Policy.

Aos, S., and R. Barnoski (2005): "Washington's Offender's Accountability Act: A first look at outcomes," (Document No. 05-07-1202). Olympia: Washington State Institute for Public Policy.

Barnes, G. C., L. Ahlman, C. Gill, L. W. Sherman, E. Kurtz, and R. Malvestuto (2010): "Low-intensity community supervision for low-risk offenders: A randomized, controlled trial," *Journal of Experimental Criminology*, 6(2), 159–189.

Barnoski, R. (1997): "Standards for improving research effectiveness in adult and juvenile justice," (Document No. 97-12-1201). Olympia: Washington State Institute for Public Policy.

——— (2003): "Evaluation of Washington State's 1996 juvenile court program for high-risk, first-time offenders: Final report," (Document No. 03-04-1202). Olympia: Washington State Institute for Public Policy.

Barnoski, R., and S. Aos (2003): "Washington's Offender's Accountability Act: An analysis of the Department of Corrections' risk assessment," (Document No. 03-12-1201). Olympia: Washington State Institute for Public Policy.

Barnoski, R., and E. Drake (2007): "Washington's Offender's Accountability Act: Department of Corrections' static risk instrument," (Document No. 07-03-1201). Olympia: Washington State Institute for Public Policy.

Berk, R., G. Barnes, L. Ahlman, and E. Kurtz (2010): "When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment," *Journal of Experimental Criminology*, 6(2), 191–208.

Berk, R., and J. de Leeuw (1999): "An evaluation of California's inmate classification system using a generalized regression discontinuity design," *Journal of the American Statistical Association*, 94(448), 1045–1052.

Bonta, J., G. Bourgon, T. Rugge, T.-L. Scott, A. K. Yessine, L. Gutierrez, and J. Li (2011): "An experimental demonstration of training probation officers in evidence-based community supervision," *Criminal Justice and Behavior*, 38(11), 1127–1148.

Bonta, J., T. Rugge, T. Scott, G. Bourgon, and A. Yessine (2008): "Exploring the black box of community supervision," *Journal of Offender Rehabilitation*, 47(3), 248–270.

Bureau of Justice Statistics (2012): "Probation and parole in the Unites States, 2011," NCJ 239686, Washington, D.C.: U.S. Department of Justice.

Chen, K., and J. Shapiro (2004): "Does prison harden inmates? A discontinuity-based approach," Cowles Foundation Discussion Paper 1450.

Clear, T. R., and P. L. Hardyman (1990): "The new intensive supervision movement," *Crime & Delinquency*, 36(1), 42–60.

Cullen, F., J. P. Wright, and B. Applegate (1996): "Control in the community: The limits of reform?," in *Choosing correctional options that work: Defining the demand and evaluating the supply*, ed. by A. Harland, pp. 69–116. Sage Publications, Inc., USA.

Dobkin, C., R. Gil, and J. Marion (2010): "Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment," *Economics of Education Review*, 29(4), 566–575.

Drake, E. (2011): ""What works" in community supervision," (Document No. 11-12-1201). Olympia: Washington State Institute for Public Policy.

Drake, E., and S. Aos (2012): "Confinement for technical violations of community supervision: Is there an effect on felony recidivism?," (Document No. 12-07-1201). Olympia: Washington State Institute for Public Policy.

Drake, E., S. Aos, and R. Barnoski (2010): "Washington's Offender's Accountability Act: Final report on recidivism outcomes," (Document No. 10-01-1201). Olympia: Washington State Institute for Public Policy.

Drake, E., and R. Barnoski (2009): "New risk instrument for offenders improves classification decisions," (Document No. 09-03-1201). Olympia: Washington State Institute for Public Policy.

Erwin, B. S. (1986): "Turning up the heat on probationers in Georgia," *Federal Probation*, 50, 109.

Farrington, D. P., and B. C. Welsh (2005): "Randomized experiments in criminology: What have we learned in the last two decades?," *Journal of Experimental Criminology*, 1(1), 9–38.

Gendreau, P., C. Goggin, and P. Smith (2002): "Is the PCL-R really the "unparalleled" measure of offender risk?," *Criminal Justice and Behavior*, 29(4), 397–426.

Gendreau, P., T. Little, and C. Goggin (1996): "A meta-analysis of the predictors of adult offender recidivism: What works!," *Criminology*, 34(4), 575–608.

Georgiou, G. (2012): "Essays on the economics of criminal justice administration," Ph.D. thesis, University of California, Santa Cruz.

Gill, C. E. (2010): "The effects of sanction intensity on criminal conduct: A randomized low-intensity probation experiment," Ph.D. thesis, University of Pennsylvania.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69(1), 201–209.

Harris, P., R. Gingerich, and T. Whittaker (2004): "The effectiveness of differential supervision," *Crime & Delinquency*, 50(2), 235–271.

Hawken, A., and M. Kleiman (2009): "Managing drug involved probationers with swift and certain sanctions: Evaluating Hawaii's HOPE," (NCJ 229023). Washington, D.C.: National Institute of Justice.

Hjalmarsson, R. (2009): "Juvenile jails: A path to the straight and narrow or to hardened criminality?," *Journal of Law and Economics*, 52(4), 779–809.

Imbens, G., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142(2), 615–635.

Jacob, B., and L. Lefgren (2004): "Remedial education and student achievement: A regression-discontinuity analysis," *Review of Economics and Statistics*, 86(1), 226–244.

Jalbert, S. K., W. Rhodes, M. Kane, E. Clawson, B. Bogue, C. Flygare, R. Kling, and M. Guevara (2011): "A multi-site evaluation of reduced probation caseload size in an evidence-based practice setting," (NCJ 234596). Washington, D.C.: National Institute of Justice.

Kuziemko, I. (2007): "Going off parole: How the elimination of discretionary prison release affects the social cost of crime," National Bureau of Economic Research Working Paper 13380.

Lee, D., and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142(2), 655–674.

Lee, D., and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Lee, D., and J. McCrary (2005): "Crime, punishment, and myopia," National Bureau of Economic Research Working Paper 11491.

MacKenzie, D. L. (2002): "Reducing the criminal activities of known offenders and delinquents: Crime prevention in the courts and corrections," in *Evidence-Based Crime Prevention*, ed. by L. Sherman, D. Farrington, B. Welsh, and D. L. MacKenzie, pp. 330–404. Routledge, New York, NY, USA.

Martinson, R. (1974): "What works? -questions and answers about prison reform," *The Public Interest*, 35(1), 22–54.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142(2), 698–714.

Pearson, F. S. (1988): "Evaluation of New Jersey's intensive supervision program," *Crime & Delinquency*, 34(4), 437–448.

Petersilia, J. (1999): "Parole and prisoner reentry in the United States," *Crime and Justice*, 26, 479–529.

——— (2009): *When prisoners come home: Parole and prisoner reentry*. Oxford University Press, USA.

Petersilia, J., and S. Turner (1990): "Intensive supervision for high-risk probationers," *Santa Monica, CA: Rand*.

——— (1993): "Intensive probation and parole," *Crime and Justice*, 17, 281–335.

Petersilia, J., S. Turner, and E. P. Deschenes (1992): "Intensive supervision programs for drug offenders," in *Smart sentencing: The emergence of intermediate sanctions*, ed. by J. Byrne, A. Lurigio, and J. Petersilia, pp. 18–37. Sage Publications, Inc., USA.

Sherman, L. W., D. C. Gottfredson, D. L. MacKenzie, J. Eck, P. Reuter, and S. D. Bushway (1997): "Preventing crime: What works, what doesn't, what's promising," Washington, D.C.: United States Department of Justice, National Institute of Justice.

Taxman, F. S., C. Yancey, and J. E. Bilanin (2006): "Proactive community supervision in Maryland: Changing offender outcomes," Maryland Division of Parole and Probation.

Travis, J. (2005): *But they all come back: Facing the challenges of prisoner reentry.* The Urban Insitute, Washington, DC, USA.

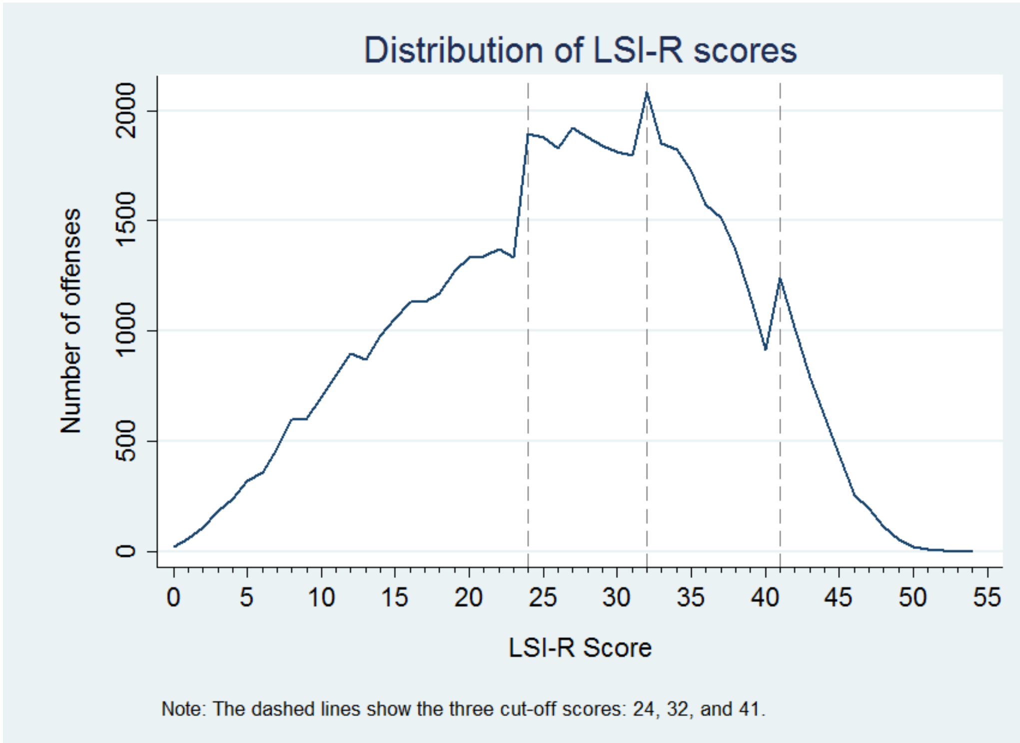

Figure 1: Distribution of LSI-R scores. The LSI-R score was the risk instrument used by Washington State to measure the risk of offenders. It is a scale from 0 to 54, where higher numbers indicate higher risk. Three cut-off scores separate four different levels of supervision intensity. If an offender's score reaches or passes one of the cut-off points, he or she is upgraded to a more intensive supervision level. The figure shows large jumps in the distribution of the scores that occur exactly on these three thresholds. This is an indication that the scores have been manipulated so that a certain number of offenders cross the thresholds and can therefore be assigned to a higher supervision level.

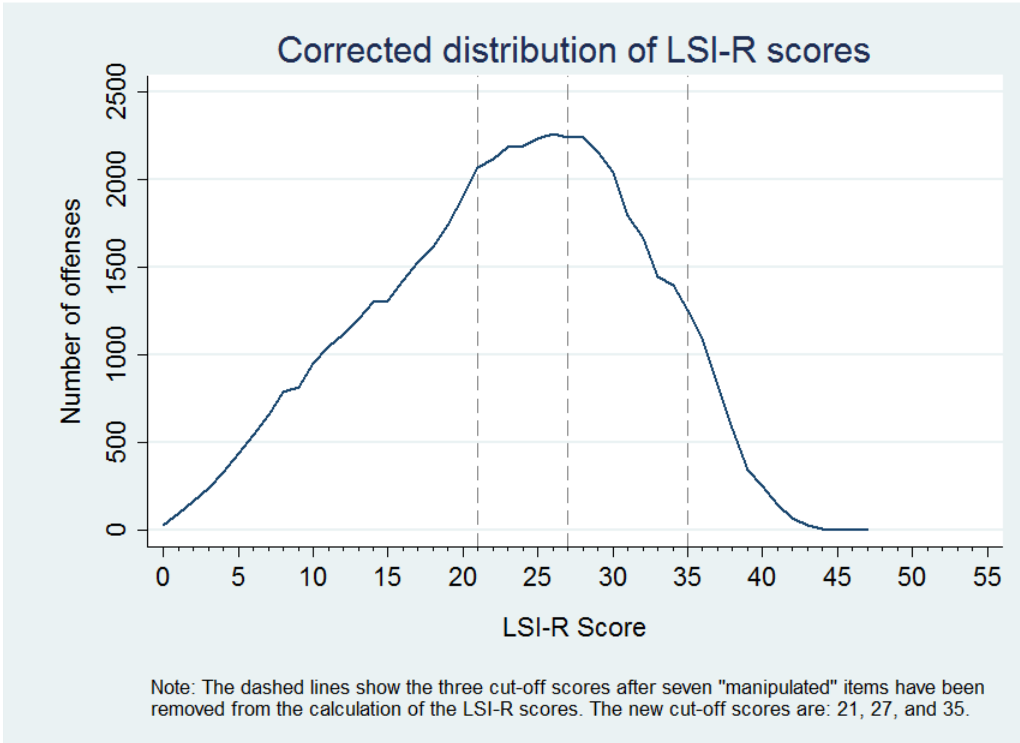

Figure 2: Distribution of LSI-R scores after omitting the seven “manipulated” items, as these were identified in Georgiou (2012). Without the seven manipulated items, the LSI-R index becomes a scale from 0 to 47. This “corrected” distribution of the scores does not exhibit discontinuities on the three thresholds, and it allows me to compare individuals with similar risk characteristics in the context of a regression discontinuity design.

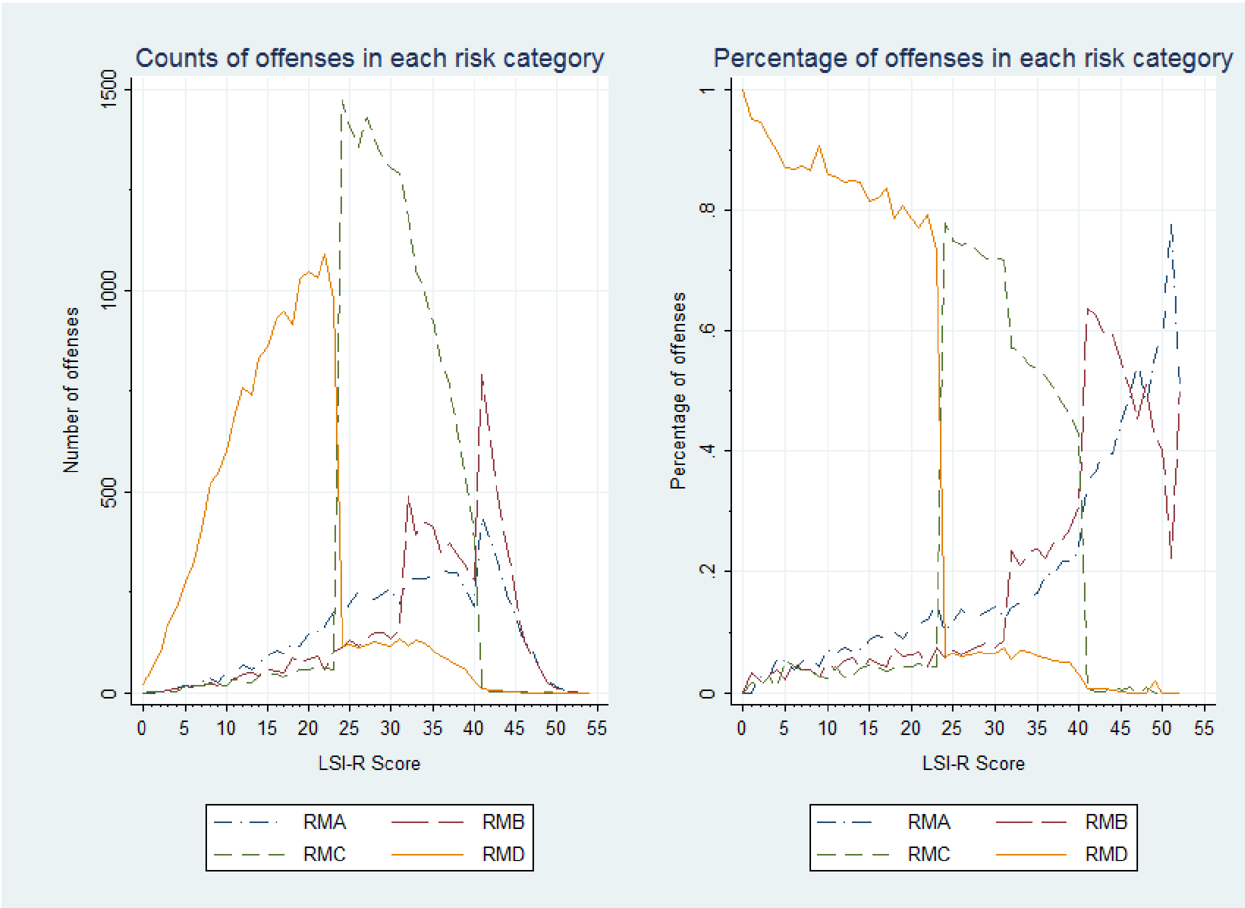

Figure 3: Counts and percentages of offenses in the four risk categories per LSI-R score. The highest-risk category is RMA and the lowest-risk category is RMD. Offenders are assigned to them based on their LSI-R score. This figure demonstrates that the three thresholds, 24, 32, and 41, are strongly binding when the authorities assign offenders to one of the four risk categories. For example, the probability that the authorities classify an offender at the low-risk RMD category is about 73 percent if his or her score is 23, but falls to roughly 6 percent if his or her score is 24. Conversely, the probability that an offender is placed at the RMC category is approximately 4 percent if his or her LSI-R score is 23, but increases to almost 78 percent if his or her score is 24.

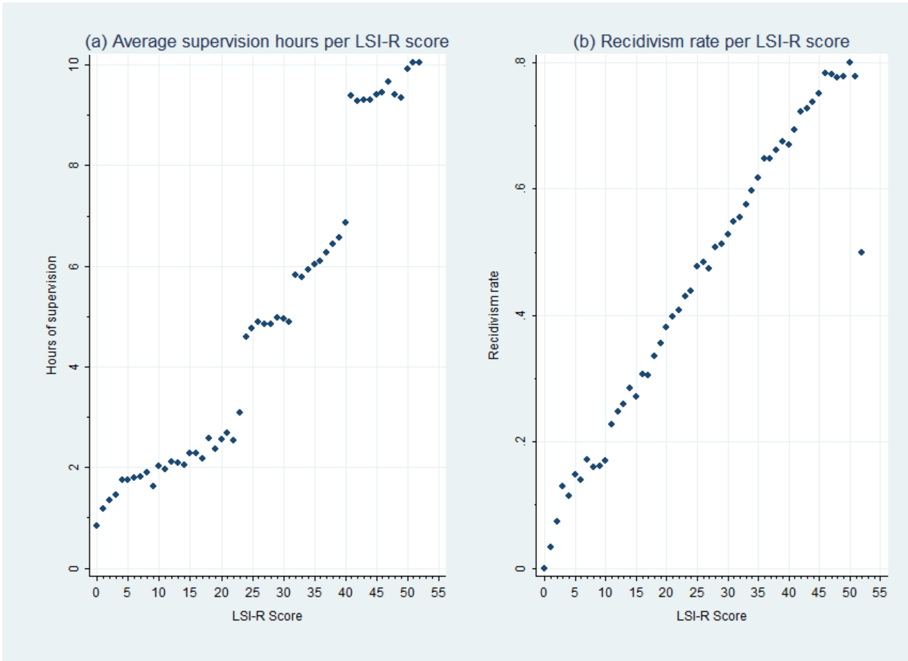

Figure 4: Average number of supervision hours allocated for each individual case (panel (a)) and recidivism rate (panel (b)) per LSI-R score using the manipulated data.

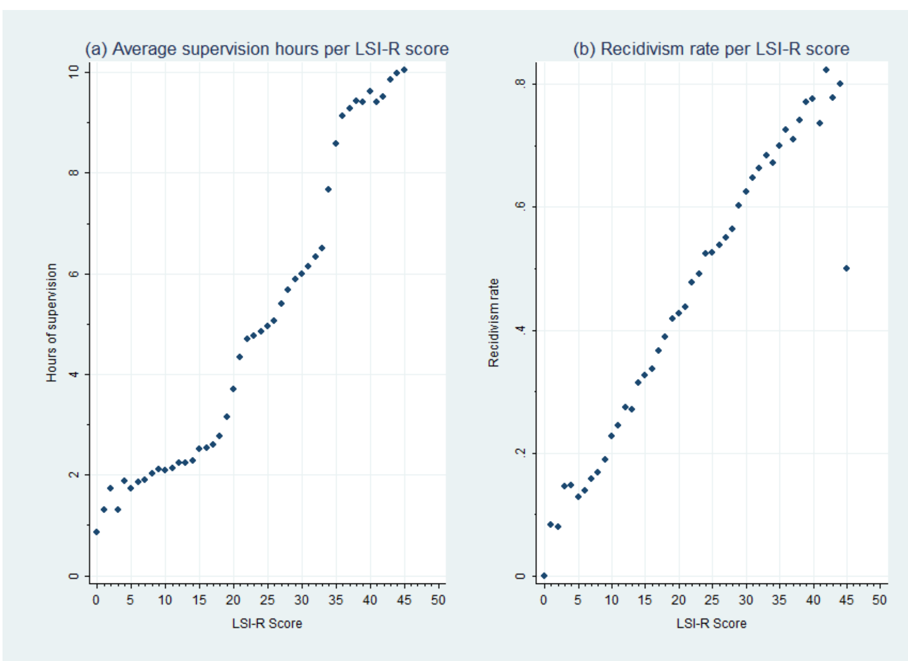

Figure 5: Average number of supervision hours allocated for each individual case (panel (a)) and recidivism rate (panel (b)) per LSI-R score using the corrected data.

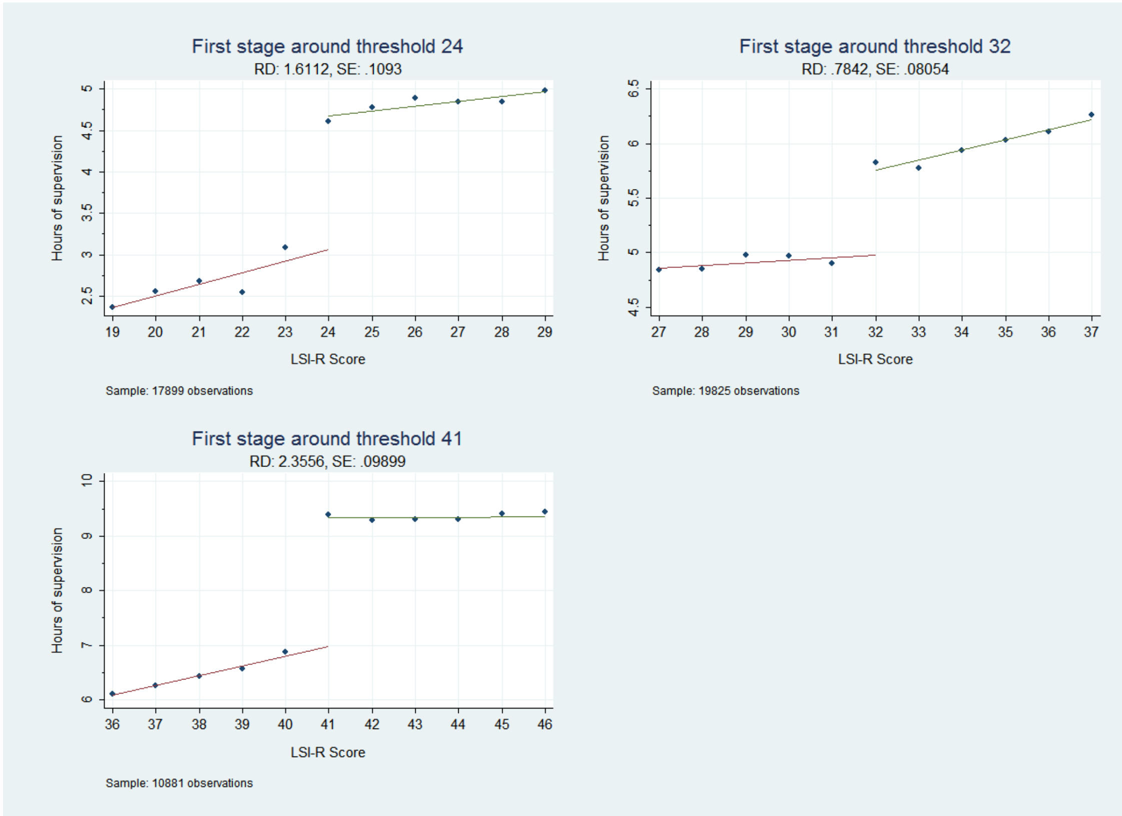

Figure 6: First stage for the three thresholds using the manipulated data. The figure presents the effect of the LSI-R score on the hours of supervision for each of the three thresholds. Each panel of the figure zooms in around each threshold in panel (a) of Fig. 4. For this and the following first-stage figures, the solid dots are the means of the hours of supervision for each level of the LSI-R score. The two regression lines fit the micro data for the same levels of the LSI-R score. For example, in the first panel, I used the predicted values for scores 19-23 to create a counterfactual value for the hours of supervision that offenders with a score of 24 would have received if 24 had not been the cut-off point. The regression discontinuity estimate measures the difference between this counterfactual value for offenders with a score of 24 and the actual predicted value. Graphically, this difference is the vertical distance between the two fitted lines in the three panels of the figure. The numerical value of the regression discontinuity (RD) estimates and their standard errors (SE) are reported above each graph.

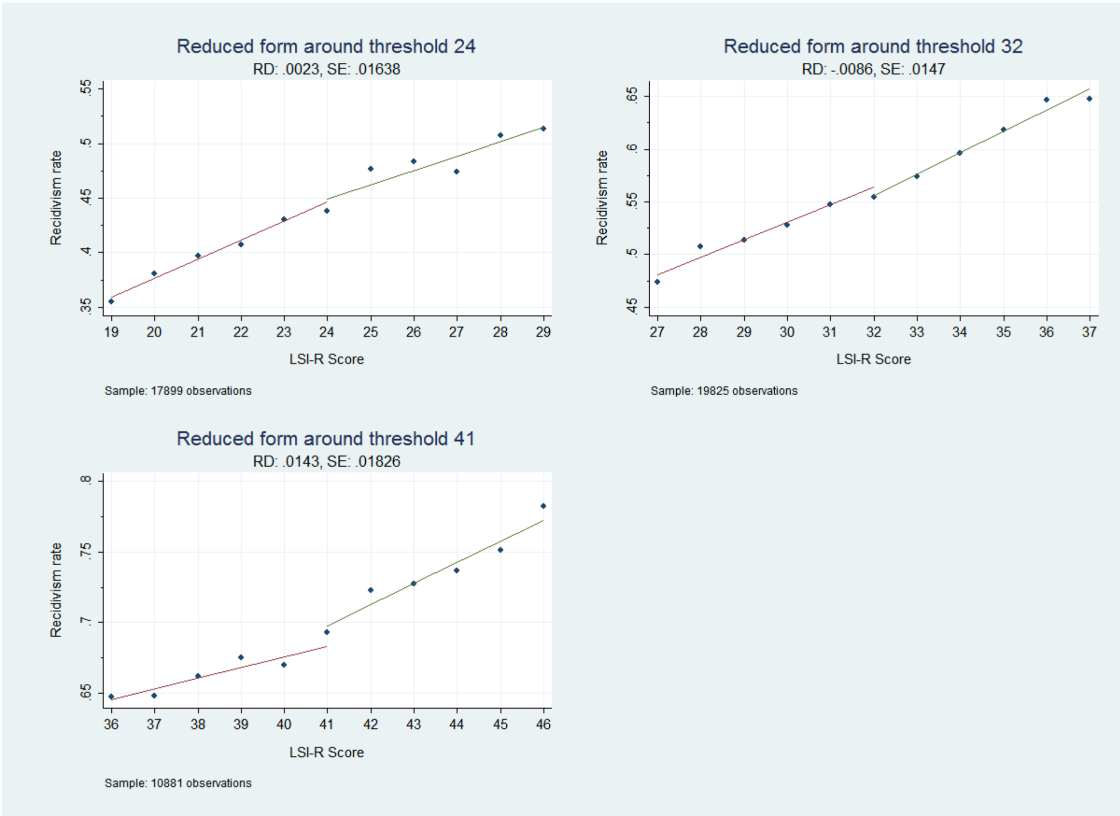

Figure 7: Reduced form for the three thresholds and "Recidivism" as the outcome variable using the manipulated data. The figure presents the effect of the LSI-R score on recidivism for each of the three thresholds. Each panel of the figure zooms in around each threshold in panel (b) of Fig. 4. This figure is equivalent to Fig. 6 but refers to the reduced form of the model. The construction method mentioned above was used here too. The numerical value of the regression discontinuity (RD) estimates and their standard errors (SE) are reported above each graph.

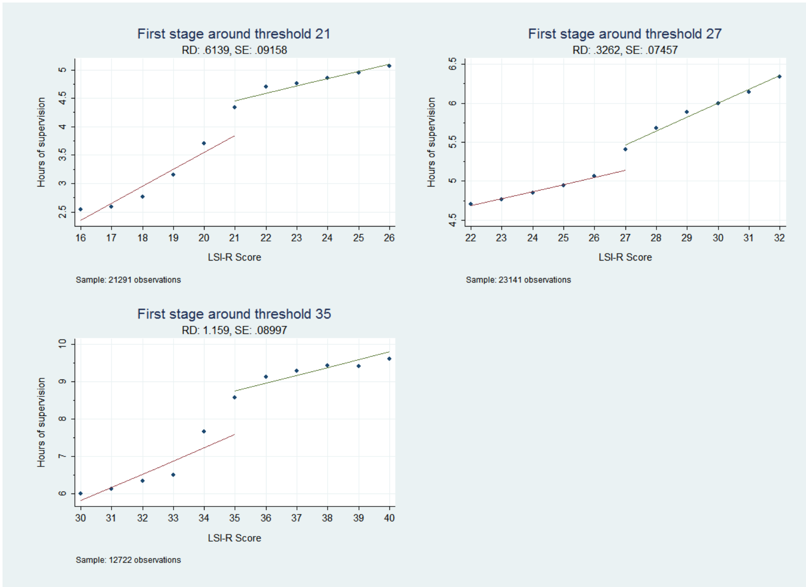

Figure 8: First stage for the three thresholds using the corrected data. The figure presents the effect of the LSI-R score on the hours of supervision for each of the three thresholds. Each panel of the figure zooms in around each threshold in panel (a) of Fig. 5.

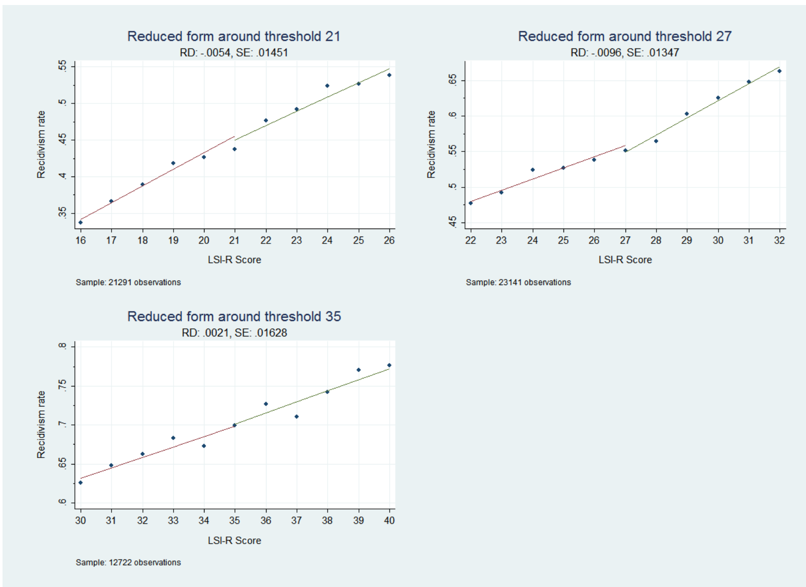

Figure 9: Reduced form for the three thresholds and "Recidivism" as the outcome variable using the corrected data. The figure presents the effect of the LSI-R score on recidivism for each of the three thresholds. Each panel of the figure zooms in around each threshold in panel (b) of Fig. 5.

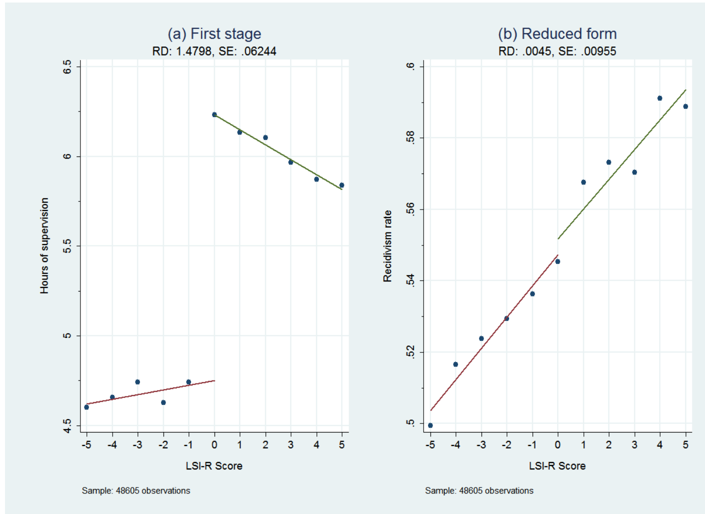

Figure 10: First stage and reduced form for the combined-threshold analysis using the manipulated data. For this figure all the observations around the three thresholds have been pooled together to produce a single data set. In this manner, one can see the first stage and the reduced form combined and not split among three different figures, as was the case with Figs. 6 and 7. Again the solid dots represent average hours of supervision and recidivism rates for every point around the thresholds, and the regression lines fit the respective micro data.

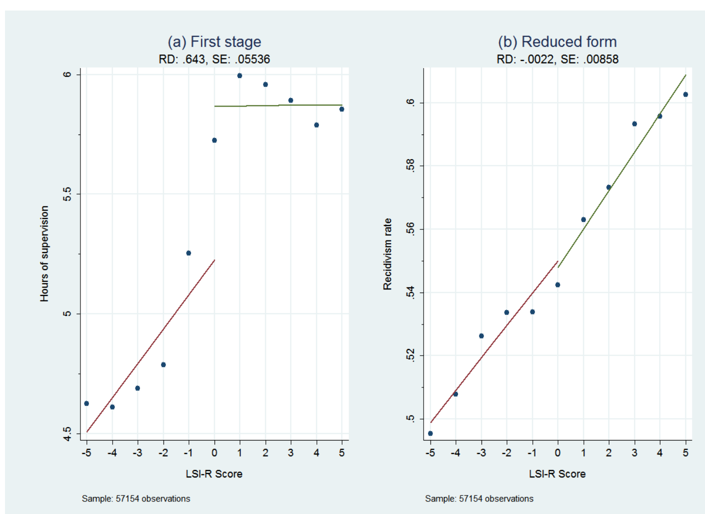

Figure 11: First stage and reduced form for the combined-threshold analysis using the corrected data. Everything said in the note to Fig. 10 applies here too. The two panels show the first stage and the reduced form combined and not split among three different figures, as was the case with Figs. 8 and 9.

Table 1: Composition of the data set in percentages (unless otherwise noted)

Characteristic	Full sample	RMA	RMB	RMC	RMD
Male	77.02	91.05	81.96	74.13	71.27
Female	22.98	8.95	18.04	25.87	28.73
White	72.33	62.72	69.30	74.81	75.54
Black	14.94	23.12	18.01	13.59	11.11
Hispanic	6.89	6.93	6.80	6.48	7.37
Native American	2.78	3.56	3.59	3.05	1.72
Asian	2.71	3.30	1.98	1.86	3.75
Age 18-30	49.23	47.96	45.74	48.67	52.16
Age 31-45	39.90	39.72	42.95	41.82	36.34
Age over 45	10.66	11.92	10.97	9.33	11.41
Drug crime	31.74	12.98	26.23	39.71	34.24
Assault	13.76	30.85	17.79	8.26	9.98
Property crime	27.79	12.82	22.35	31.68	33.09
Sex crime	4.60	11.74	9.37	2.74	1.03
Weapons	2.82	2.18	2.60	3.21	2.79
Robbery	1.75	5.92	2.31	0.75	0.64
Homicide	0.31	0.80	0.50	0.12	0.21
Community	84.07	74.91	77.24	82.15	93.86
Prison	15.93	25.09	22.76	17.85	6.14
Avg. hours of supervision (hours)	4.7	10.2	8.9	3.9	1
Observations	51,957	7,902	8,126	19,008	16,831

Table 2: Recidivism summary statistics in percentages

Recidivism type	Full sample	RMA	RMB	RMC	RMD
Recidivism	48.02	54.75	58.50	54.59	32.43
Felony Recidivism	33.14	37.52	41.58	39.00	20.35
Misdemeanor Recidivism	14.87	17.22	16.92	15.49	12.08
Property Felony Recidivism	12.21	10.25	14.48	15.56	8.23
Drug Felony Recidivism	10.33	8.52	12.54	13.08	7.00
Violent Felony Recidivism	9.30	16.38	12.73	9.17	4.45
Observations	51,957	7,902	8,126	19,008	16,831

Table 3: Similarity of sample characteristics around thresholds using the manipulated LSI-R scores

Characteristic	Threshold 24	Threshold 32	Threshold 41
Male	0.0072	0.0045	0.0266
	[0.0141]	[0.0123]	[0.0162]
White	-0.0084	-0.0129	-0.0312
	[0.0145]	[0.0132]	[0.0182]
Black	0.0126	0.0068	0.0308*
	[0.0114]	[0.0108]	[0.0154]
Hispanic	0.0014	0.0069	0.0042
	[0.0083]	[0.0074]	[0.0093]
Native American	-0.0016	-0.0033	-0.0033
	[0.0051]	[0.0050]	[0.0084]
Asian	-0.0022	0.0035	-0.0008
	[0.0055]	[0.0042]	[0.0046]
Age	0.3872	-0.3753	-0.5902
	[0.3462]	[0.2939]	[0.3696]
Drug crime	0.0037	-0.0341*	0.0021
	[0.0156]	[0.0142]	[0.0186]
Assault	0.0083	0.0327***	0.0055
	[0.0112]	[0.0097]	[0.0131]
Property crime	-0.0086	-0.0232	-0.0065
	[0.0147]	[0.0131]	[0.0175]
Sex crime	-0.0082	0.0097	0.0039
	[0.0074]	[0.0059]	[0.0080]
Weapons	-0.0045	0.0047	0.0087
	[0.0055]	[0.0050]	[0.0063]
Robbery	0.0030	0.0116**	0.0098
	[0.0046]	[0.0038]	[0.0055]
Prison	-0.0095	0.0475***	-0.0080
	[0.0120]	[0.0116]	[0.0166]

Standard errors in brackets

Note: This table provides weak but corroborating evidence that there are dissimilarities between offenders around the three thresholds. The table shows that for 5 out of the 42 regressions run the characteristics are not comparable across the thresholds. Four of the dissimilarities are around threshold 32, namely concerning drug crimes, assault, robbery, and offenders coming out of prison. The fifth concerns black offenders at threshold 41. The three thresholds, 24, 32 and 41, are scores of the LSI-R risk instrument and separate the four risk categories, RMA, RMB, RMC and RMD that an offender is assigned to, where RMA is the highest-risk category and RMD the lowest-risk one. The numbers shown correspond to the regression discontinuity (RD) estimates and their standard errors. The RD dummy takes the value 1 if an offender's LSI-R score is greater than or equal to the respective threshold and 0 otherwise. The outcome variable for each regression is a dummy variable that takes the value 1 if the characteristic is present and 0 otherwise. Therefore each RD estimate shows the change in the probability of having the characteristic if an offender's score is over the threshold. For the characteristic "age" the outcome variable is age in years, so the RD can be interpreted as change in age if an offender's score is over the threshold. A linear polynomial of the risk score is also included but the estimates for its terms are not reported here. For each regression I use only scores that are 5 points below the respective threshold and 6 points above it (including the threshold itself).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Similarity of sample characteristics around thresholds after correcting the scores

Characteristic	Threshold 21	Threshold 27	Threshold 35
Male	0.0121	0.0106	0.0042
	[0.0124]	[0.0113]	[0.0146]
White	0.0050	0.0052	0.0147
	[0.0129]	[0.0121]	[0.0164]
Black	0.0018	-0.0005	0.0079
	[0.0102]	[0.0099]	[0.0139]
Hispanic	-0.0018	-0.0017	0.0079
	[0.0072]	[0.0068]	[0.0086]
Native American	-0.0015	-0.0026	-0.0031
	[0.0044]	[0.0047]	[0.0073]
Asian	-0.0039	-0.0006	0.0014
	[0.0049]	[0.0037]	[0.0039]
Age	0.3670	0.2475	-0.3467
	[0.3110]	[0.2710]	[0.3288]
Drug crime	-0.0059	-0.0050	-0.0207
	[0.0139]	[0.0129]	[0.0167]
Assault	0.0112	0.0072	-0.0011
	[0.0100]	[0.0091]	[0.0118]
Property crime	-0.0029	-0.0019	-0.0141
	[0.0129]	[0.0120]	[0.0157]
Sex crime	-0.0064	0.0088	0.0129
	[0.0063]	[0.0055]	[0.0072]
Weapons	-0.0041	0.0006	0.0207***
	[0.0052]	[0.0045]	[0.0055]
Robbery	-0.0042	-0.0008	0.0050
	[0.0041]	[0.0037]	[0.0051]
Prison	0.0031	0.0048	0.0230
	[0.0105]	[0.0107]	[0.0149]

Standard errors in brackets

Note: This table is similar to Table 3, but here I am using the corrected scores. It shows that, after the removal of the manipulated items, there is only one characteristic, namely the weapons crime category at the highest threshold, for which offenders around the thresholds are not comparable.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: First-stage, reduced-form, and IV regressions using the manipulated and the corrected LSI-R scores

Outcome variable: various forms of Recidivism						
	Manipulated			Corrected		
	Threshold 24	Threshold 32	Threshold 41	Threshold 21	Threshold 27	Threshold 35
	First-Stage estimates			First-Stage estimates		
RD	1.6112***	0.7842***	2.3556***	0.6139***	0.3262***	1.1590***
	[0.1093]	[0.0805]	[0.0990]	[0.0916]	[0.0746]	[0.0900]
	Reduced-Form *RD* estimates			Reduced-Form *RD* estimates		
Recidivism	0.0023	-0.0086	0.0143	-0.0054	-0.0096	0.0021
	[0.0164]	[0.0147]	[0.0183]	[0.0145]	[0.0135]	[0.0163]
Felony Recidivism	-0.0011	-0.0039	-0.0014	-0.0147	-0.0109	-0.0276
	[0.0150]	[0.0143]	[0.0196]	[0.0132]	[0.0131]	[0.0175]
Property Felony Recidivism	-0.0053	-0.0196	-0.0046	-0.0147	0.0037	-0.0295*
	[0.0103]	[0.0105]	[0.0149]	[0.0090]	[0.0093]	[0.0135]
Drug Felony Recidivism	-0.0098	-0.0036	0.0196	0.0036	-0.0187*	0.0121
	[0.0096]	[0.0096]	[0.0141]	[0.0084]	[0.0089]	[0.0129]
Violent Felony Recidivism	0.0117	0.0148	-0.0120	-0.0078	-0.0006	-0.0090
	[0.0089]	[0.0089]	[0.0139]	[0.0080]	[0.0083]	[0.0124]
Misdemeanor Recidivism	0.0034	-0.0047	0.0157	0.0093	0.0013	0.0296*
	[0.0114]	[0.0108]	[0.0153]	[0.0102]	[0.0099]	[0.0136]
	IV $SupervisionHours$ estimates			IV $SupervisionHours$ estimates		
Recidivism	0.0014	-0.0109	0.0061	-0.0089	-0.0293	0.0018
	[0.0102]	[0.0188]	[0.0077]	[0.0236]	[0.0418]	[0.0140]
Felony Recidivism	-0.0007	-0.0049	-0.0006	-0.0239	-0.0334	-0.0238
	[0.0093]	[0.0183]	[0.0083]	[0.0217]	[0.0408]	[0.0152]
Property Felony Recidivism	-0.0033	-0.0250	-0.0019	-0.0239	0.0115	-0.0255*
	[0.0063]	[0.0134]	[0.0063]	[0.0148]	[0.0289]	[0.0117]
Drug Felony Recidivism	-0.0061	-0.0046	0.0083	0.0059	-0.0574	0.0105
	[0.0059]	[0.0122]	[0.0060]	[0.0137]	[0.0293]	[0.0111]
Violent Felony Recidivism	0.0072	0.0188	-0.0051	-0.0127	-0.0019	-0.0078
	[0.0055]	[0.0114]	[0.0059]	[0.0133]	[0.0256]	[0.0108]
Misdemeanor Recidivism	0.0021	-0.0060	0.0067	0.0151	0.0040	0.0256*
	[0.0071]	[0.0139]	[0.0065]	[0.0168]	[0.0304]	[0.0118]
Observations	17,899	19,825	10,881	21,291	23,141	12,722

Standard errors in brackets

Note: This table presents the first-stage, reduced-form and IV estimates for each of the three thresholds separately by using the original manipulated LSI-R scores in the first three columns and the corrected scores in the next three columns. The first section of this table reports the RD estimates for the first-stage regressions that capture the effect of the risk score on the hours of supervision for each threshold. The second section reports the RD estimates for the reduced-form regressions that capture the effect of the risk score on several recidivism outcomes for each threshold. The third section reports the IV estimates of the effect of supervision hours on several recidivism outcomes for each threshold. The regression equations for all thresholds include a regression discontinuity dummy that takes the value 1 if an offender's LSI-R score is greater than or equal to the respective threshold and 0 otherwise. A linear polynomial of the risk score is also included but the estimates for its terms are not reported here. For each regression I use only scores that are 5 points below the respective threshold and 6 points above it (including the threshold itself).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Combined-threshold analysis including first-Stage, reduced-form, and IV regressions using the manipulated and the corrected LSI-R scores

Outcome variable: various forms of Recidivism		
	Manipulated	Corrected
	First-Stage estimates	
RD	1.4798***	0.6426***
	[0.0624]	[0.0554]
	Reduced-Form *RD* estimates	
Recidivism	0.0045	-0.0022
	[0.0095]	[0.0086]
Felony Recidivism	0.0010	-0.0131
	[0.0093]	[0.0083]
Property Felony Recidivism	-0.0095	-0.0092
	[0.0067]	[0.0059]
Drug Felony Recidivism	-0.0006	-0.0037
	[0.0062]	[0.0056]
Violent Felony Recidivism	0.0092	-0.0034
	[0.0059]	[0.0053]
Misdemeanor Recidivism	0.0035	0.0109
	[0.0070]	[0.0063]
	IV $SupervisionHours$ estimates	
Recidivism	0.0030	-0.0033
	[0.0064]	[0.0134]
Felony Recidivism	0.0007	-0.0204
	[0.0063]	[0.0132]
Property Felony Recidivism	-0.0065	-0.0144
	[0.0045]	[0.0093]
Drug Felony Recidivism	-0.0004	-0.0058
	[0.0042]	[0.0087]
Violent Felony Recidivism	0.0062	-0.0053
	[0.0039]	[0.0083]
Misdemeanor Recidivism	0.0024	0.0170
	[0.0047]	[0.0098]
Observations	48,605	57,154

Standard errors in brackets

Note: This table is equivalent to Table 5, but here I combine the three thresholds in one. The estimates show the effect of increased hours of supervision if an offender had an LSI-R score that was greater than *any* threshold. The observations that were used for each of the three threshold-specific regressions (5 points below each threshold and 6 above) were pooled in a single data set for the manipulated scores and another data set for the corrected scores. In the first column I present the estimates for the manipulated scores and in the second column the estimates for the corrected scores. Now the regression discontinuity dummy takes the value 1 if the score is greater than or equal to *any* of the three thresholds.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix A

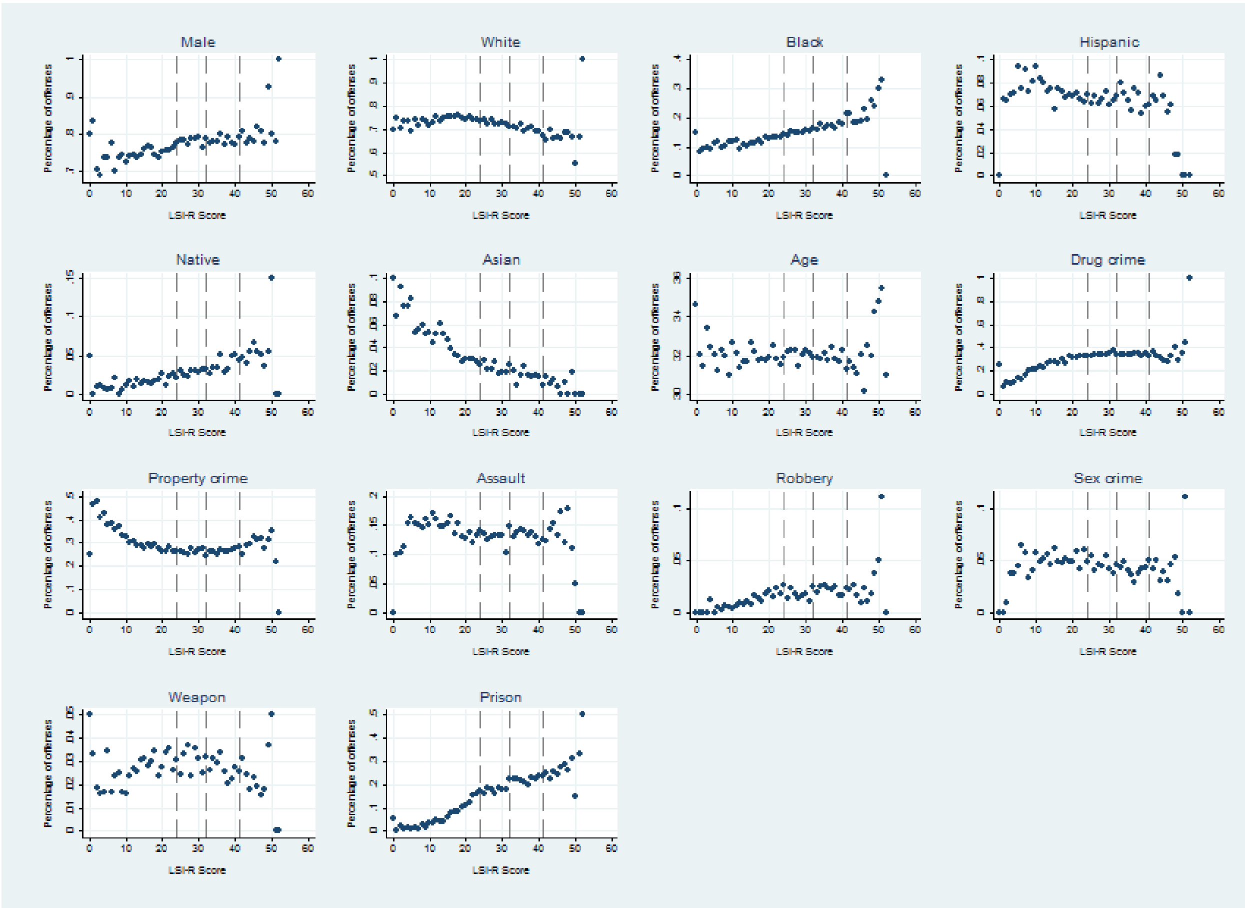

Figure 12: The figure shows the distributions of all the observed characteristics that are known to me, and it refers to the manipulated LSI-R distribution (Fig. 1). The three dashed lines in each panel correspond to the cut-off points of the manipulated distribution (24, 32, and 41). Regression lines around the cut-offs have been omitted for simplicity. However, the regression discontinuity coefficients are reported in Table 3 of the body of the text. As noted in that table, five out of the 42 RD coefficients are significant. Four of them are around threshold 32, namely concerning drug crimes, assault, robbery, and offenders coming out of prison. The fifth concerns black offenders at threshold 41. This observation provides weak but corroborating evidence that the characteristics across the thresholds in the manipulated LSI-R distribution were not entirely similar. At the same time, however, there is not enough evidence to support the claim that the authorities targeted specific crimes or other characteristics when they added points to, and thus manipulated, an offender's LSI-R score.

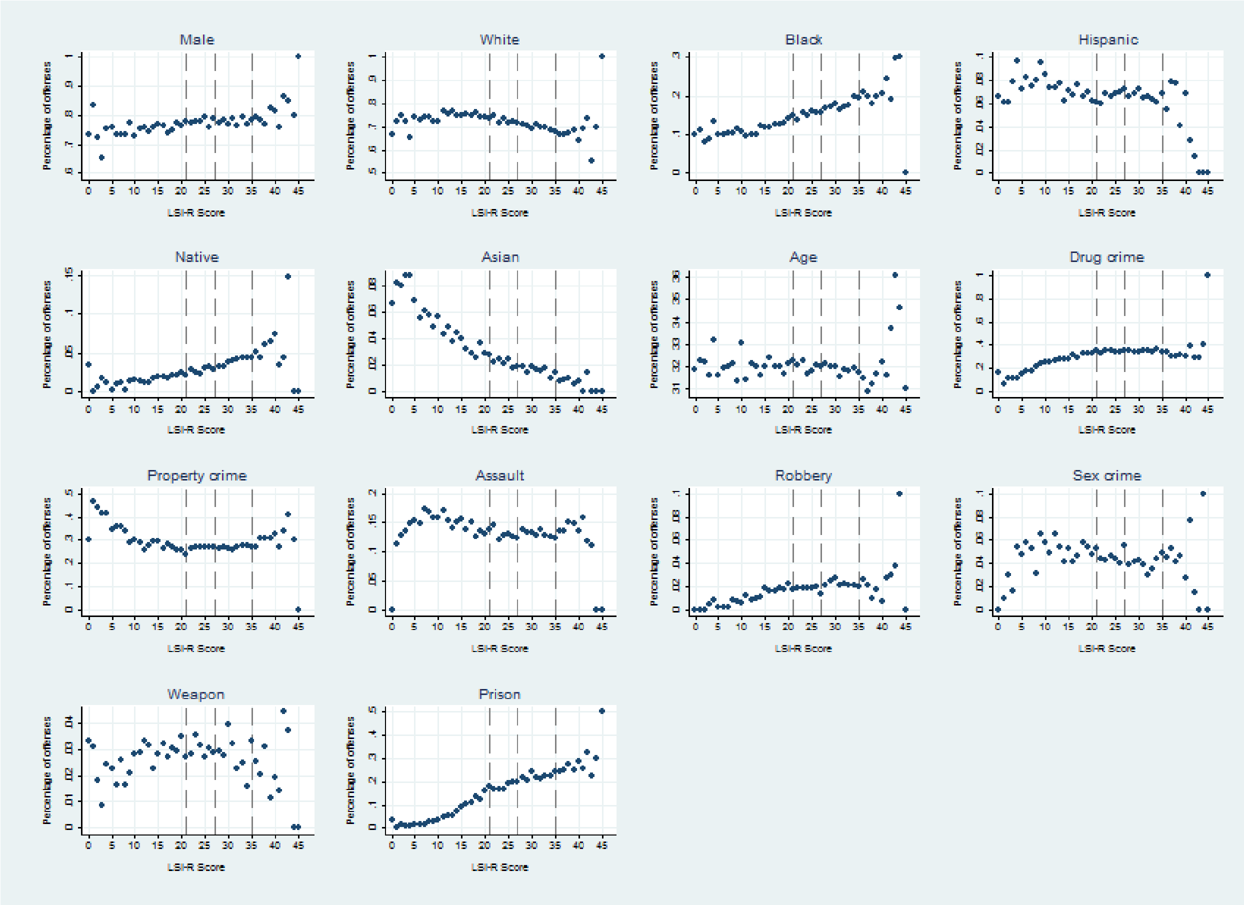

Figure 13: This figure is equivalent to Fig. 12 but it refers to the corrected LSI-R distribution (Fig. 2). The three dashed lines in each panel correspond to the cut-off points of the corrected distribution (21, 27, and 35). The regression discontinuity coefficients are reported in Table 4 of the body of the text. Note that after the correction of the scores there is only one dissimilarity, namely for crimes related to weapons around threshold 35, the high threshold of the corrected distribution.

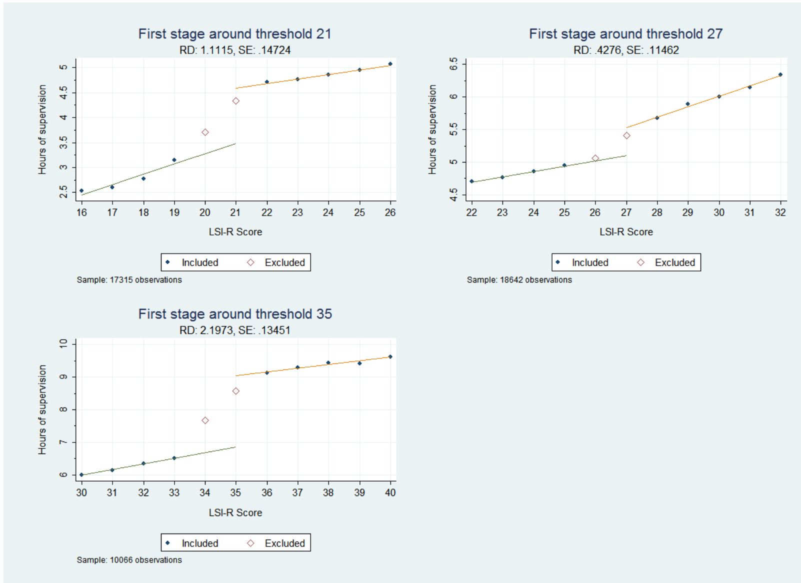

Figure 14: First stage for the three thresholds excluding observations on the thresholds and one point prior to that using the corrected data. The exclusion of these points offers a solution to the attenuated first stage observed when using the corrected data.

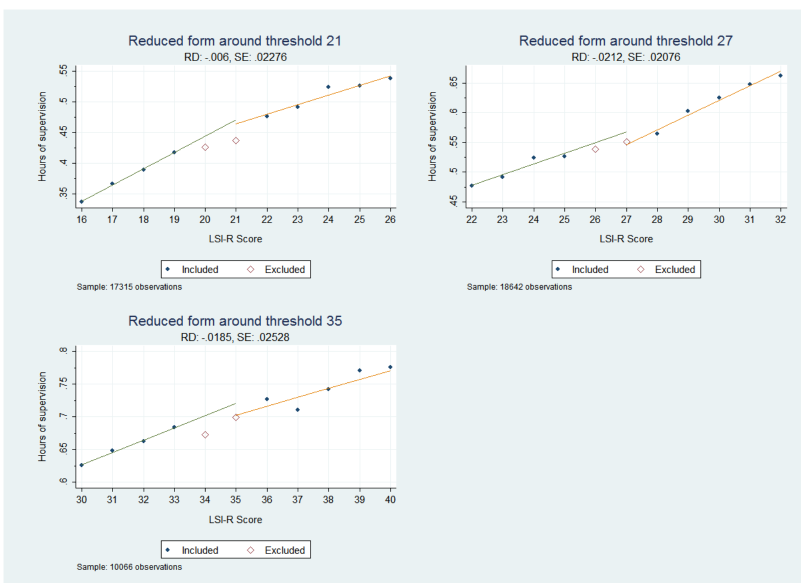

Figure 15: Reduced form for the three thresholds and "Recidivism" as the outcome variable, excluding observations on the thresholds and one point prior to that using the corrected data. This figure is the reduced-form equivalent of Fig. 14.

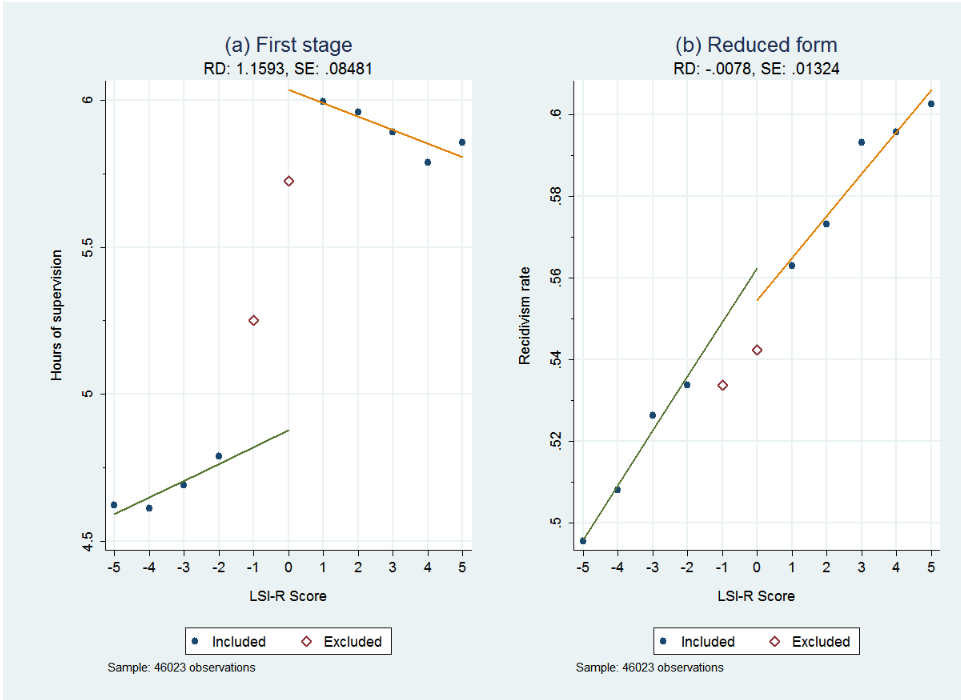

Figure 16: First stage and reduced form with "Recidivism" as the outcome variable for the combined-threshold analysis, excluding observations on the thresholds and one point prior to that using the corrected data.

Table 7: IV regressions using the manipulated LSI-R scores, including controls

Outcome variable: Recidivism						
	Threshold 24a	Threshold 24b	Threshold 32a	Threshold 32b	Threshold 41a	Threshold 41b
$SupervisionHours$	0.0014	0.0013	-0.0109	-0.0184	0.0061	0.0046
	[0.0102]	[0.0098]	[0.0188]	[0.0244]	[0.0077]	[0.0077]
Male		0.0976***		0.0932***		0.0541***
		[0.0098]		[0.0181]		[0.0124]
Age		-0.0070***		-0.0081***		-0.0088***
		[0.0004]		[0.0005]		[0.0005]
Black		0.0820***		0.0983***		0.0723***
		[0.0117]		[0.0192]		[0.0122]
Hispanic		0.0207		0.0259		0.0157
		[0.0145]		[0.0152]		[0.0181]
Native American		0.0634**		0.0812***		0.0951***
		[0.0228]		[0.0189]		[0.0198]
Asian		-0.0659**		-0.0172		0.0917**
		[0.0233]		[0.0296]		[0.0334]
Property Crime		0.0475***		0.0617***		0.0443***
		[0.0098]		[0.0121]		[0.0113]
Assault		-0.0568		0.0144		-0.0165
		[0.0298]		[0.0664]		[0.0196]
Homicide		-0.1783**		-0.0204		-0.2841*
		[0.0622]		[0.1094]		[0.1340]
Robbery		0.0573		0.1102		0.0168
		[0.0427]		[0.0962]		[0.0377]
Sex crime		-0.1753***		-0.0840		-0.1306***
		[0.0445]		[0.0922]		[0.0298]
Weapon		-0.0375		-0.0251		-0.0482
		[0.0222]		[0.0254]		[0.0293]
Prior Adult Fel. Adjs		0.0591***		0.0500***		0.0341***
		[0.0027]		[0.0022]		[0.0023]
Prison		0.0051		-0.0020		-0.0254*
		[0.0123]		[0.0170]		[0.0107]
Observations	17,899	17,899	19,825	19,825	10,881	10,881

Standard errors in brackets

Note: This table presents the IV estimates for each of the three thresholds separately by using the original manipulated LSI-R scores. For each threshold the models labeled (a) repeat the IV estimates already reported in Table 5 for the outcome variable "Recidivism." The models labeled (b) also incorporate control variables relating to gender, age, race, current most serious crime, number of prior adult felony adjudications, and type of sentence (released from prison or sentenced straight to the community). The benchmark category for gender is female, for race is white offenders, for current crime is drug crime and for type of sentence is community sentence. A linear polynomial of the risk score is also included in both (a) and (b) models but the estimates for its terms are not reported here. For each regression I use only scores that are 5 points below the respective threshold and 6 points above it (including the threshold itself).

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: IV regressions using the corrected LSI-R scores, including controls

Outcome variable: Recidivism						
	Threshold 21a	Threshold 21b	Threshold 27a	Threshold 27b	Threshold 35a	Threshold 35b
$SupervisionHours$	-0.0089	-0.0044	-0.0293	-0.0185	0.0018	0.0027
	[0.0236]	[0.0238]	[0.0418]	[0.0468]	[0.0140]	[0.0144]
Male		0.0966***		0.0982**		0.0585***
		[0.0152]		[0.0331]		[0.0148]
Age		-0.0068***		-0.0080***		-0.0088***
		[0.0005]		[0.0006]		[0.0005]
Black		0.0862***		0.1030**		0.0726***
		[0.0146]		[0.0330]		[0.0142]
Hispanic		0.0151		0.0210		0.0204
		[0.0133]		[0.0161]		[0.0177]
Native American		0.0481*		0.0608***		0.0892***
		[0.0210]		[0.0175]		[0.0185]
Asian		-0.0447		-0.0134		0.0762*
		[0.0257]		[0.0392]		[0.0335]
Property crime		0.0509***		0.0557***		0.0431***
		[0.0102]		[0.0166]		[0.0109]
Assault		-0.0326		0.0210		-0.0087
		[0.0671]		[0.1230]		[0.0297]
Homicide		-0.1765		-0.0302		-0.2834*
		[0.0956]		[0.1695]		[0.1304]
Robbery		0.0778		0.1162		0.0244
		[0.0807]		[0.1704]		[0.0527]
Sex crime		-0.1539		-0.0777		-0.1258**
		[0.1022]		[0.1757]		[0.0420]
Weapon		-0.0319		0.0050		-0.0214
		[0.0230]		[0.0318]		[0.0267]
Prior Adult Fel. Adjs		0.0596***		0.0487***		0.0338***
		[0.0027]		[0.0022]		[0.0023]
Prison		-0.0015		-0.0038		-0.0251*
		[0.0183]		[0.0283]		[0.0102]
Observations	21,291	21,291	23,141	23,141	12,722	12,722

Standard errors in brackets
Note: This table is equivalent to Table 7, but here I am using the corrected LSI-R scores. Everything said in the note to Table 7 applies here too.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: IV regressions from the combined-threshold analysis using the manipulated and the corrected LSI-R scores, including controls

Outcome variable: Recidivism				
	Manipulated		Corrected	
	No Controls	Controls	No Controls	Controls
$SupervisionHours$	0.0030	0.0017	-0.0033	0.0020
	[0.0064]	[0.0066]	[0.0064]	[0.0066]
Male		0.0816***		0.0810***
		[0.0067]		[0.0099]
Age		-0.0083***		-0.0082***
		[0.0002]		[0.0002]
Black		0.0819***		0.0840***
		[0.0075]		[0.0105]
Hispanic		0.0214*		0.0190*
		[0.0089]		[0.0086]
Native American		0.0995***		0.0867***
		[0.0123]		[0.0119]
Asian		-0.0438**		-0.0385*
		[0.0156]		[0.0152]
Property crime		0.0485***		0.0464***
		[0.0059]		[0.0061]
Assault		-0.0352		-0.0303
		[0.0188]		[0.0361]
Homicide		-0.1728***		-0.1875***
		[0.0447]		[0.0552]
Robbery		0.0470		0.0499
		[0.0283]		[0.0496]
Sex crime		-0.1614***		-0.1597**
		[0.0263]		[0.0515]
Weapon		-0.0423**		-0.0229
		[0.0136]		[0.0133]
Prior Adult Fel. Adjs		0.0564***		0.0558***
		[0.0020]		[0.0030]
Prison		-0.0089		-0.0108
		[0.0067]		[0.0090]
Observations	48,605	48,605	57,154	57,154

Standard errors in brackets

Note: This table incorporates controls in the combined-threshold analysis presented in Table 6. The first two columns present the results for the manipulated scores and the second two columns present the results for the corrected scores. The columns labeled "No Controls" repeat the IV estimates already reported in Table 6 for the outcome variable "Recidivism." The columns labeled "Controls" include the control variables described in the note to Table 7. Everything said in the note to Table 6 applies here too.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: First-stage, reduced-form and IV regressions using the corrected LSI-R scores, excluding observations on the thresholds and one point prior

Outcome variable: Recidivism, Property Felony Recidivism, and Misdemeanor Recidivism						
	Threshold 21a	Threshold 21b	Threshold 27a	Threshold 27b	Threshold 35a	Threshold 35b
			First-Stage estimates			
RD	0.6139***	1.1115***	0.3262***	0.4276***	1.1590***	2.1973***
	[0.0916]	[0.1472]	[0.0746]	[0.1146]	[0.0900]	[0.1345]
			Reduced-Form *RD* estimates			
Recidivism	-0.0054	-0.0060	-0.0096	-0.0212	0.0021	-0.0185
	[0.0145]	[0.0228]	[0.0135]	[0.0208]	[0.0163]	[0.0253]
Property Felony Recidivism	-0.0147	-0.0025	0.0037	-0.0003	-0.0295*	-0.0532*
	[0.0090]	[0.0140]	[0.0093]	[0.0144]	[0.0135]	[0.0216]
Misdemeanor Recidivism	0.0093	-0.0091	0.0013	-0.0216	0.0296*	0.0233
	[0.0102]	[0.0164]	[0.0099]	[0.0153]	[0.0136]	[0.0212]
			IV *SupervisionHours* estimates			
Recidivism	-0.0089	-0.0054	-0.0293	-0.0495	0.0018	-0.0084
	[0.0236]	[0.0204]	[0.0418]	[0.0503]	[0.0140]	[0.0115]
Property Felony Recidivism	-0.0239	-0.0023	0.0115	-0.0007	-0.0255*	-0.0242*
	[0.0148]	[0.0126]	[0.0289]	[0.0337]	[0.0117]	[0.0098]
Misdemeanor Recidivism	0.0151	-0.0082	0.0040	-0.0505	0.0256*	0.0106
	[0.0168]	[0.0148]	[0.0304]	[0.0386]	[0.0118]	[0.0096]
Observations	21,291	17,315	23,141	18,642	12,722	10,066

Standard errors in brackets

Note: In this table I exclude from the regressions labeled (b) the observations on the thresholds and one point prior to that. Fig. 5 indicated that these observations might have a confounding effect on the size of especially the first stage of the analysis. For each threshold the models labeled (a) repeat the estimates already reported in Table 5 for the outcome variables: Recidivism, Property Felony Recidivism, and Misdemeanor Recidivism. The latter two types of recidivism are reported here because I found a statistically significant effect for them in Table 5. The models labeled (b) report the estimates following the exclusion of the said points. Everything said in the note to Table 5 applies here too.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: First-stage, reduced-form and IV regressions from the combined-threshold analysis using the corrected LSI-R scores, excluding observations on the thresholds and one point prior

Outcome variable: Recidivism, Property Felony Recidivism, and Misdemeanor Recidivism		
	Before exclusion	After exclusion
	First-Stage estimates	
RD	0.6426***	1.1593***
	[0.0554]	[0.0848]
	Reduced-Form *RD* estimates	
Recidivism	-0.0022	-0.0078
	[0.0086]	[0.0132]
Property Felony Recidivism	-0.0092	-0.0106
	[0.0059]	[0.0092]
Misdemeanor Recidivism	0.0109	-0.0045
	[0.0063]	[0.0098]
	IV *SupervisionHours* estimates	
Recidivism	-0.0033	-0.0067
	[0.0134]	[0.0115]
Property Felony Recidivism	-0.0144	-0.0092
	[0.0093]	[0.0080]
Misdemeanor Recidivism	0.0170	-0.0039
	[0.0098]	[0.0084]
Observations	57,154	46,023

Standard errors in brackets
Note: This table is equivalent to Table 10, but here I have combined the three thresholds in one. The column labeled "Before exclusion" repeats the estimates already reported in Table 6 for the outcome variables: Recidivism, Property Felony Recidivism, and Misdemeanor Recidivism. The column labeled "After exclusion" implements the exclusion of the 2 points.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: IV estimates for different lengths of supervision

Outcome variable: Recidivism						
	Manipulated			Corrected		
	Threshold 24	Threshold 32	Threshold 41	Threshold 21	Threshold 27	Threshold 35
3 years (baseline)	0.0014	-0.0109	0.0061	-0.0089	-0.0293	0.0018
	[0.0102]	[0.0188]	[0.0077]	[0.0236]	[0.0418]	[0.0140]
2 years	0.0000	-0.0038	0.0010	-0.0180	-0.0476	-0.0031
	[0.0106]	[0.0032]	[0.0007]	[0.0258]	[0.0478]	[0.0146]
1.5 years	0.0024	-0.0042	0.0002	-0.0120	-0.0560	-0.0016
	[0.0107]	[0.0042]	[0.0005]	[0.0250]	[0.0584]	[0.0154]
1 year	0.0097	-0.0014	0.0024	-0.0048	-0.0772	-0.0065
	[0.0110]	[0.0034]	[0.0020]	[0.0250]	[0.0820]	[0.0154]
6 months	0.0149	0.0093	0.0014	-0.0073	-0.0684	-0.0114
	[0.0114]	[0.0051]	[0.0011]	[0.0283]	[0.0786]	[0.0176]

Standard errors in brackets

Note: The table provides robustness checks for different lengths of the supervision period. Due to lack of information about the number of months or years of supervision for each offender, the assumption was made that all offenders were supervised for a three-year period, which was the maximum period allowable by law. This assumption, however, could result in an understatement of the difference in the treatment levels since higher-risk offenders will be supervised for longer periods than lower-risk offenders. In this table I check that the three-year baseline result reported in Table 5 (and repeated in the first line of this table) holds when the length of supervision is successively reduced to shorter periods. For example, when the supervision period is two years, the recidivism event must have occurred within that period. The shorter the period, the more likely it is that all the observations in the relevant sample are actively under supervision. Note that the IV estimates for the shorter supervision periods do not show noticeable changes compared to the baseline specification of the body of the text.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: IV estimates for different population groups

Outcome variable: Recidivism						
	Manipulated			Corrected		
	Threshold 24	Threshold 32	Threshold 41	Threshold 21	Threshold 27	Threshold 35
Male	0.0027	-0.0122	0.0062	-0.0061	-0.0327	0.0058
	[0.0135]	[0.0182]	[0.0099]	[0.0297]	[0.0397]	[0.0171]
Female	-0.0029	0.0012	0.0029	-0.0265	-0.0216	-0.0097
	[0.0145]	[0.1034]	[0.0123]	[0.0395]	[0.6296]	[0.0245]
White	-0.0010	-0.0351	0.0113	-0.0014	-0.0446	0.0291
	[0.0114]	[0.0284]	[0.0086]	[0.0264]	[0.0823]	[0.0176]
Black	0.0161	0.0730	-0.0100	-0.0125	0.0168	-0.0931*
	[0.0303]	[0.0379]	[0.0323]	[0.0786]	[0.0530]	[0.0432]
Hispanic	-0.0095	0.0055	-0.0546	0.0147	-0.0563	-0.0493
	[0.0301]	[0.0314]	[0.0407]	[0.0800]	[0.0740]	[0.0465]
Age 18-30	0.0133	-0.0222	0.0051	0.0069	-0.0459	0.0165
	[0.0129]	[0.0234]	[0.0105]	[0.0319]	[0.0575]	[0.0194]
Age 31-45	0.0011	0.0195	0.0076	-0.0289	0.0247	-0.0101
	[0.0189]	[0.0358]	[0.0118]	[0.0412]	[0.0637]	[0.0222]
Age over 45	-0.0512	-0.0589	-0.0029	0.0205	-0.0594	-0.0302
	[0.0301]	[0.0610]	[0.0354]	[0.0745]	[0.2051]	[0.0510]
Drug	0.0054	-0.0304	0.0073	-0.0267	-0.0220	0.0151
	[0.0129]	[0.0512]	[0.0103]	[0.0567]	[0.1370]	[0.0193]
Assault	0.0407	-0.0548	0.0236	0.2594	-0.0345	-0.0051
	[0.1300]	[0.0382]	[0.0544]	[0.6422]	[0.0615]	[0.1630]
Property crime	-0.0273	0.0229	0.0063	-0.0154	-0.4427	-0.0153
	[0.0139]	[0.0538]	[0.0110]	[0.0327]	[0.4485]	[0.0212]
Prison	-0.0096	0.0334	-0.0134	0.0475	0.0292	-0.0480
	[0.0368]	[0.0354]	[0.0159]	[0.0734]	[0.0364]	[0.0330]
Community	0.0043	-0.0320	0.0119	-0.0166	-0.0980	0.0148
	[0.0104]	[0.0251]	[0.0088]	[0.0256]	[0.0908]	[0.0158]

Standard errors in brackets
Note: The table repeats the RD design for specific population groups in order to assess whether supervision had an impact on a particular subgroup of offenders. Each regression uses a sample that is limited to the specific population group, such as male, white, etc. Each one of the groups chosen comprised at least 5 percent of the whole sample. Only the IV estimates are reported for each group and each threshold, for both the manipulated and the corrected scores. Note that the IV estimates do not change substantially from the whole sample specification, and they remain insignificant. The only statistically significant case is the IV estimate for black offenders for the high threshold of the corrected scores.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: IV estimates for different windows around the thresholds

Outcome variable: Recidivism						
	Manipulated			Corrected		
	Threshold 24	Threshold 32	Threshold 41	Threshold 21	Threshold 27	Threshold 35
5 points (baseline)	0.0014	-0.0109	0.0061	-0.0089	-0.0293	0.0018
	[0.0102]	[0.0188]	[0.0077]	[0.0236]	[0.0418]	[0.0140]
4 points	0.0033	-0.0061	0.0056	-0.0129	-0.0281	0.0113
	[0.0116]	[0.0205]	[0.0089]	[0.0368]	[0.0505]	[0.0202]
3 points	0.0044	-0.0105	0.0084	-0.0270	0.0088	0.0371
	[0.0153]	[0.0221]	[0.0109]	[0.0738]	[0.0664]	[0.0406]
6 points	0.0001	0.0003	0.0027			
	[0.0085]	[0.0170]	[0.0069]			
7 points	-0.0027	0.0093	0.0001			
	[0.0079]	[0.0165]	[0.0063]			

Standard errors in brackets

Note: The table provides robustness checks for different ranges of the window around the various thresholds for the manipulated and corrected scores. The outcome variable is recidivism for all specifications and only the IV estimates are reported here. The baseline specification is 5 points around the thresholds. The first line of the table repeats the results from Table 5 of the body of the text. The next lines repeat the IV regressions, first shrinking the window to 4 and 3 points around the thresholds and then widening the window to 6 and 7 points. However, widening the window was only possible for the manipulated scores. Widening for the corrected scores was impossible because the window would overlap with the immediately next threshold. For example, a 6-point window above threshold 21 would end at score 27, which is the threshold for the next supervision category. Note that the IV estimates do not change substantially from the baseline specification of 5 points that was presented in the body of the text.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$