**Title**
ESSAYS ON THE ECONOMICS OF CRIMINAL JUSTICE ADMINISTRATION



**Author**
GEORGIOU, GEORGIOS

**Publication Date**
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

# ESSAYS ON THE ECONOMICS OF CRIMINAL JUSTICE ADMINISTRATION

A dissertation submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECONOMICS

by

**Georgios Georgiou**

June 2012

The Dissertation of Georgios Georgiou
is approved:

Professor Donald Wittman, Chair

Associate Professor Carlos Dobkin

Associate Professor Justin Marion

Tyrus Miller
Vice Provost and Dean of Graduate Studies

# Contents

# List of Figures

# List of Tables

**Abstract**

ESSAYS ON THE ECONOMICS OF CRIMINAL JUSTICE ADMINISTRATION

by

Georgios Georgiou

In 1999, Washington State decided to allocate supervision based on offenders' risk characteristics. These risk characteristics were measured by the actuarial instrument "Level of Service Inventory - Revised." In analyzing the data, I discovered the unusual administration of the instrument by the local bureaucracy that resulted in many offenders being bumped to a higher supervision level. Using a regression discontinuity design, I uncover the mechanics of the bumping up process. After cleansing the instrument of the manipulation, I find that the manipulated instrument predicts time to recidivism better than the cleansed instrument. Moreover, using a regression discontinuity design I evaluate whether offenders of similar risk characteristics recidivate less if they receive a more intense supervision level. I find that offenders who received more supervision did not recidivate less. I also examined whether the authorities targeted certain groups of offenders more than others in the process of administering the instrument. Specifically, I tried to find out whether their behavior exhibits signs of what is known in the literature as "preference or prejudice-based discrimination" as opposed to "statistical discrimination." I did not find conclusive evidence that certain groups were bumped up more than others. Moreover, even if we accept that certain groups are bumped up more, this can be justified by their higher recidivism rates, which amounts to "statistical discrimination." Therefore, I found no evidence of unequal treatment in the administration of the risk assessment instrument. Finally, I found that I can significantly improve the predictive validity of the LSI-R instrument by having the items weighted. The results were found to be robust in a validation sample as well.

Στην μνήμη της γιαγιάς μου Αγγελικής

## Acknowledgments

I would like to thank my advisors, Professors Donald Wittman and Carlos Dobkin for their invaluable help all along this process.

Donald Wittman has been extremely supportive and helpful since the beginning of this effort. His clear, direct and unequivocal way of thinking always steered me in the right direction and helped me shape my own views on both research in general as well as this specific project. His elegant wit, pithy comments and delicate sense of humor have also been an endless source of bright new ideas. My gratitude to Donald Wittman cannot be overstated.

Carlos Dobkin has been the mastermind behind this project providing both the idea and the methodology for implementing it. With his profound theoretical and technical expertise as well as his approachable personality, he helped me put together this dissertation piece by piece. The truth is that this project would have never ended without his assistance. In fact, it would have never started. I thank him from the bottom of my heart.

David Kaun, a revered professor and generous philanthropist assisted me with editing a preliminary draft of this dissertation. For this and much more, I am most indebted to him.

I am also grateful to Elizabeth Drake and Laura Harmon of the Washington State Institute of Public Policy for providing me with the data sets used in this dissertation, and David Daniels of the Washington Department of Corrections for assisting me in understanding the state's community custody programs.

Finally, I benefited from comments and suggestions from my friends and colleagues: Julian Caballero, Jean Paul Rabanal and Sean Tanoos. I thank them all wholeheartedly.

Introduction

This dissertation is a study that evaluates a treatment program for offenders undertaken by the correctional authorities of Washington State between 1998 and 2008. The local bureaucracy designed and implemented a complex system of supervision allocation for released offenders that made use of advanced risk assessment techniques. The dissertation examines and analyzes several aspects of this system, spanning issues of criminal justice administration, implementation of treatment programs, discriminatory practices and risk prediction.

The first chapter of the dissertation presents the way a risk assessment instrument was implemented by Washington State authorities, leading to its manipulation. This manipulated instrument, though, is found to produce better prediction results than it would have produced in the absence of manipulation. Therefore, in contrast with current conventional wisdom that has associated bureaucratic mechanisms with inefficient and wasteful practices, it can be argued that Washington authorities acted with a considerable degree of efficiency and professionalism.

The second chapter of the dissertation assesses whether this method of supervision allocation that was based on an offender's risk level, was actually effective in reducing recidivism levels. To tackle this issue I used modern econometric techniques to create a valid control sample that would validate my results. I found that the method did not manage to have a statistically significant impact on recidivism.

The third chapter of the dissertation examines whether the authorities targeted certain groups of offenders more than others when they were making their decision to bump offenders. Specifically, I focused on demographic characteristics such as race, gender and age as well as on crime specific groups. I could not conclusively verify that certain groups were bumped up more than others. However, even in the event that such an unequal treatment is present, it can be justified by different recidivism rates across different groups. This is known in the literature as "statistical discrimination."

The fourth chapter of the dissertation builds on the already existing knowledge on actuarial risk assessment instruments. In particular, the instrument used in Washington State was an unweighted instrument, i.e. all its items were factored in equally (unit weights) without any special attention given to factors that are better predictors of the recidivism event. I showed that if weights are employed, the generated weighted instrument can outperform in predictive accuracy the unweighted one by a statistically significant margin. This result holds, even though the former instrument has significantly fewer items than the latter.

This dissertation describes the virtues and the shortcomings of a sophisticated system that administers offender supervision in the community. The four chapters shed light on several aspects of this system, assess its effectiveness and make suggestions for potential improvements. If any of the ideas presented herein turn out to be useful to policy makers even in the slightest way, the purpose of this work will be fulfilled.

# Part I

# First Part

# Chapter 1

# The Efficient Bureaucrat: Evidence from Washington State

## 1.1 Introduction

This study tells the story of how correctional authorities operated in Washington State between 1998 and 2008. The local bureaucracy designed and implemented a complex system of supervision allocation for released offenders that made use of advanced risk assessment techniques. In contrast with current conventional wisdom that has associated bureaucratic mechanisms with inefficient and wasteful practices, I show that Washington authorities acted with a considerable degree of professionalism in several respects.

Bureaucracies are usually regarded as slow and anachronistic organizations that serve mostly the interests of their members as well as their sponsors. The classic model for a budget maximizing bureaucracy has been proposed by Niskanen (1968, 1971). According to it, the sole purpose of bureaucrats is to maximize the amount of resources made available to them by their sponsors, i.e. the politicians. Variations of this model have been proposed to correct some of its rigidities. For example, Breton and Wintrobe (1975) incorporate control devices implemented by the politicians in order to monitor the performance of the bureaucrats, while Wyckoff (1990) builds on a "slack-maximizing" bureaucracy that tries to maximize the difference between a bureaucrat's revenue and the lowest production cost.

The correctional authorities in Washington State, part of bureaucratic machine that administers criminal justice in the state, were assigned with the task of determining the proper level of supervision an offender is to receive once he or she is released to the community after either serving time in prison or being sentenced straight to a community sentence. Supervision intensity was decided to depend on the risk level that each offender posed. Risk, in turn, is commonly measured by way of "actuarial" (also known as "mechanical," "algorithmic," or "statistical") instruments and these will be in the center of the present study.

According to Paul Meehl (1954, p. 15), a method is "actuarial" when "the prediction is arrived at by some straightforward application of an equation or table to the data." Alternatively, it can be said that these instruments are indices that assign a numerical or other value to the risk profile of an offender by using data that can be coded in a predetermined and standardized way. They are contrasted with clinical judgment of the risk, typically performed in a non-standardized fashion by professionals of the correctional system (clinicians) on the basis of their subjective evaluation of each individual case.[1]

The distinction between actuarial and clinical methods may be characterized as a special case of the rules versus standards debate, originating in the law and economics literature. Initially Ehrlich and Posner (1974) and later Kaplow (1992) use a cost minimizing approach to describe the choice between rules or standards, arguing that rules should be implemented where there exists much behavior governed by a law, whereas standards should be preferred where an activity is infrequent. In contrast, Friedman and Wickelgren (2012)

[1] According to proposed classifications of risk assessment methods (Bonta, 1996; Andrews and Bonta, 2006), clinical judgment is characterized as a "first generation assessment" whereas actuarial methods can be subdivided in "second", "third" and, more recently, "fourth generation assessments" depending on their characteristics. Second generation assessments just focus on measuring risk. Third generation assessments on top of risk also try to uncover the "needs" that each offender has and to direct the rehabilitation efforts to them which is why they are also known as "risk-needs assessments." Fourth generation assessments try to make use of the so called "responsivity principle" which claims that the administration of treatment programs should take into account the personality, ability, learning skills and other personal characteristics of each individual offender (Andrews and Bonta, 2006).

for example, take a more definite stance arguing that under a standards system, the part that desires the most efficient outcome will exert the most effort thus signaling to the judge the proper decision. The debate is ongoing and spills over in a number of disciplines.

Indeed, in the field of clinical psychology there exists a major debate on which of the two methods should be used for prediction making: actuarial (rules) or clinical (standards). On the basis of meta-analytic evidence from a wide range of social science studies, it is argued in the literature that actuarial methods are at least as good or better compared to clinical judgment (Meehl, 1954; Grove and Meehl, 1996; Grove, Zald, Lebow, Snitz, and Nelson, 2000). The same argument is made by proponents of actuarial instruments when we focus specifically on the issue of predicting the risk of reoffending (Quinsey, 1995; Bonta, Bogue, Crowley, and Motiuk, 2001). Some even go as far as to argue that actuarial instruments should in no way be complemented by clinical judgment, a situation referred to as "clinical overrides,"[2] although others take a more moderate stance and advise caution when such an override is to be made (Andrews and Bonta, 2006). On the other hand, proponents of clinical judgment note that such superiority of actuarial methods has not been proved yet and it is, therefore, premature to replace the former with the latter until a head to head comparison can be conducted (Litwack, 2001).[3] In a similar vein, research has shown that several risk instruments (including the one used in our study) do not perform better in predicting recidivism than arbitrary structured scales that are generated by combining randomly selected items of the original instruments (Kroner, Mills, and Reddon, 2005).

In Washington State, the authorities used the actuarial instrument called "Level of Service Inventory - Revised" or "LSI-R."[4] This instrument was originally developed in Canada in the late 1970s (Andrews and Bonta, 1995), and

[2]Quinsey, Harris, Rice, and Cornier (2006) bluntly note that: "Actuarial methods are too good and clinical judgment is too poor to risk contaminating the former with the latter."

[3]Although it should be noted that Harris, Rice, and Cormier (2002) responded with such a direct comparison and found similar results in favor of actuarial methods.

[4]There are several other actuarial instruments some of which have general applicability while others pertain to certain types of offenses. For example, the Wisconsin classification system, the General Statistical Information on Recidivism (GSIR) and the Self-Appraisal Questionnaire (SAQ) are general indices, while the Psychopathy Check List-Revised (PCL-R) and the Violence Risk Appraisal Guide (VRAG) are specifically tailored for the prediction of violence.

has since been increasingly popular in many jurisdictions across North America.[5] The state of Washington was using it from October 1998 until August 2008, when it switched to another instrument.

The LSI-R has been in use for several years, and its predictive validity has been established meta-analytically by several studies (Gendreau, Little, and Goggin, 1996; Andrews and Bonta, 1995; Gendreau, Goggin, and Smith, 2002).[6] At the same time, however, the importance of implementation integrity for actuarial instruments in general has also been highlighted by the literature. Follow-up training sessions, computerization, videotaping and revisiting of interview sessions, set up of quality assurance teams are some of the strategies proposed for enhancing quality in the administration of actuarial instruments (Bonta, Bogue, Crowley, and Motiuk, 2001). Specifically for the LSI-R, staff training, amount of time spent using the instrument (Flores, Lowenkamp, Holsinger, and Latessa, 2006), training on specific cultures represented in the offender population (Holsinger, Lowenkamp, and Latessa, 2006) are some of the intricacies that correctional professionals are advised to take into account.

However, even though in Washington all measures were taken for the proper administration of the LSI-R (Manchak, Skeem, and Douglas, 2008), upon graphing the distribution of the instrument, I obtained Figure 1.1. The three jumps observed are exactly on the points that separate the levels of supervision offered to offenders of different risk, as will be explained in detail below. I take this to be evidence of manipulation in the administration of the LSI-R instrument. Even though the literature has long ago identified practical problems that can arise when agencies use such instruments in order to assign supervision levels (Clear and Gallagher, 1983, 1985), to the best of my knowledge, a similar

[5]According to Manchak, Skeem, and Douglas (2008) it is the third most used instrument in the U.S. while Petersilia (2009) refers to it as the most popular instrument. On the other hand the instrument does not appear popular among forensic psychologists as 80% of a group of 64 had “no opinion” on its acceptability for evaluating violence risk (Lally, 2003).

[6]Similarly, it has been proved repeatedly valid in settings involving diverse subgroups of offender population, such as offenders in halfway houses (Bonta and Motiuk, 1985), female offenders (Coulson, Ilacqua, Nutbrown, Giulekas, and Cudjoe, 1996; Lowenkamp, Holsinger, and Latessa, 2001; Folsom and Atkinson, 2007), sex offenders (Simourd and Malcolm, 1998), long-term incarcerated offenders (Simourd, 2004; Manchak, Skeem, and Douglas, 2008) or different ethnic groups (Holsinger, Lowenkamp, and Latessa, 2003). Its inter-rater reliability has also been established (Andrews and Bonta, 1995; Lowenkamp, Holsinger, Brusman-Lovins, and Latessa, 2004).

case of instrument manipulation has not been reported.

Standard rules of using actuarial instruments dictate that the administrator is not supposed to tamper with their mechanics in order to influence the prediction outcome. Such an intervention negates the objective nature of the instrument and reintroduces the subjectivity that is concomitant with clinical judgment of each individual offender case.

In view of the above, the object of this Chapter is twofold: (a) to document the manipulation of the LSI-R instrument, uncover its mechanics and reconstruct the instrument by excluding the manipulated parts, and (b) evaluate whether the authorities' manipulation yielded better predictive results compared to the reconstructed corrected instrument.

The results in short are the following. I find compelling evidence that manipulation did indeed take place, and that it was focused on mostly subjective items of the LSI-R instrument. Moreover, I find that the manipulated instrument produced by the authorities outperforms in predictive accuracy the reconstructed corrected instrument.

The Chapter proceeds as follows. Section 1.2 gives a brief overview of the policy changes made in Washington's criminal justice system in the recent past. Section 1.3 describes the data sets used, their characteristics and the sources that made them available to me. Section 1.4 presents the theoretical background of the authorities' behavior. Section 1.5 gives an account of the empirical method used in this Chapter . Section 1.6 presents and discusses the results, and Section 1.7 concludes.

## 1.2 Background on Washington's Correctional Policies

In the past ten years, Washington State has been in the forefront of experimentation with new sentencing techniques. What is of more importance is the determination of the State's authorities to find out exactly which techniques work and which do not.[7] The considerable importance that Washington State

[7] It should not go unsaid that other states, such as California, are trying to replicate the Washington risk assessment methods (Turner, Hess, and Jannetta, 2009).

is attributing to the improvement of its sanctioning policies is also attested by the creation and operation of the Washington State Institute for Public Policy ("WSIPP"), which has followed and analyzed closely the effectiveness of the measures taken by the State's legislature.[8]

Washington State moved to determinate sentencing for juveniles in 1977 and 1981 for adults (Boerner and Lieb, 2001). In 1999 the Washington legislature passed the Offender Accountability Act ("OAA"), which set two goals for the relevant administrative authority, the Department of Corrections ("DOC"): (a) to classify all offenders using a research-based assessment tool, and (b) to use this information to allocate supervision and treatment resources.

The first goal set by the OAA was achieved in a timely fashion, as early as October 2000 in the form of the "Risk Management Identification" ("RMI") system (Aos, 2003). This was a mechanism that classified offenders based on two criteria: (a) their risk of reoffending, and (b) the harm that was done by them. The first criterion is forward looking, aiming to prevent the commission of new crimes by the same offender in the future, while the second is backward looking, trying to capture and measure the degree to which the offender has harmed the society at large.

The first criterion, the risk of reoffending, was addressed, as already noted, by using the LSI-R instrument, which was already in place prior to the enactment of the OAA. This index is measured by surveying offenders that are either imprisoned or are sentenced to serving time in the community. Offenders were interviewed at random intervals, but also close to the time of their release from prison or the beginning of their community sentence. The answers given by the offenders to the survey questions were corroborated by the use of documentation, where available, or by the use of other sources of information, such as the offender's employers, family and friends.

The LSI-R questionnaire consists of 54 items, divided in ten subcomponents, assigning scores from 0 to 54 to different offenders according to their

[8] Its reports have been of invaluable help in my effort to comprehend the State's criminal justice system and I inevitably drew on its expertise in a number of ways during my research for this study.

answers.[9] The 54-item survey comprises both static and dynamic risk factors.[10] Each of the items contributes either a 0 or a 1 to the offender's total score.[11] A 0 means that the relevant risk factor is not present, while a 1 means that it is. The LSI-R score of an offender is just the sum of the 1s recorded. Higher scores correspond to higher risk, lower scores to lower risk. In Appendix A, I present the list of the 54 items and the respective answers as they appear on the paper-based LSI-R questionnaire.[12]

Another set of questions addressed the second criterion, the harm done to society. Both criteria combined assigned offenders to four broad categories of the RMI system called RMA, RMB, RMC and RMD, the first corresponding to the highest risk and the last to the lowest risk.[13] Therefore, the score of an offender on the LSI-R index is not the only criterion for their risk classification on the RMI system. The seriousness of their offense also plays a role.

The second goal set by the OAA was achieved by dedicating more resources to higher risk offenders, RMA and RMB categories, and fewer resources to the RMC and RMD categories (Barnoski and Aos, 2003). The resources

[9] The subcomponents are (with number of items in parentheses): Criminal History (10), Education/Employment (10), Financial (2), Family/Marital (4), Accommodation (3), Leisure/Recreation (2), Companions (5), Alcohol/Drug Problems (9), Emotional/Personal (5), Attitudes/Orientation (4).

[10] "Static" are the factors that cannot change over time, such as criminal history, age, race, etc. "Dynamic" are risk factors, such as drug dependency, that can change over time through treatment or intervention. Therefore the LSI-R instrument qualifies as a "risk-needs" or "third generation" assessment, according to the aforementioned classification (Bonta, 1996; Andrews and Bonta, 2006) since it can identify the areas amenable to change through the use of rehabilitation techniques and it can measure the change effected by way of the dynamic factors.

[11] Most items take "Yes" or "No" answers which are counted as a 1 or a 0 respectively for the purposes of calculating the total score. Some items have a small scale from 0 to 3 but the responses are also converted to a binary, 0 or 1 form.

[12] Computer based LSI-R questionnaires are currently offered and used but their structure is identical to the paper-based form presented in Appendix A.

[13] The classification rules are the following (Aos, 2002; Aos and Barnoski, 2005): Offenders are RMA if their LSI-R score is 41 to 54 and they are convicted of a violent crime, or they are a very serious sex offender (Level III), or they are a dangerous mentally ill offender, or fulfill indicators of violent history. They are RMB if their LSI-R score is 41 to 54 and they are convicted of a non-violent crime, or their LSI-R score is 32 to 40 and they are convicted of a violent crime, or they are a serious sex offender (Level II), or they fulfill indicators of high level of needs. They are RMC if their LSI-R score is 24 to 40 and they have not been classified as RMA or RMB, or they are a less serious sex offender (Level I). Finally, they are RMD if their LSI-R score is 0 to 23 and they have not been classified as RMA, RMB or RMC. It should be noted that these thresholds are the ones chosen by the Washington State authorities and they are not the same as the ones suggested by the creators of the LSI-R (Andrews and Bonta, 1995).

refer to supervision intensity once offenders are again at risk to the community. Offenders who were imprisoned are at risk at the time of their release from prison, whereas offenders who were sentenced directly to the community are at risk immediately.

## 1.3 Data

The data used in this study have all been provided by WSIPP. Specifically, WSIPP made available to me data on the LSI-R score dating from January 12, 1999 until July 15, 2008.[14] The data consist of 437,335 unique scores for 110,421 individual offenders, that is many individual offenders have been evaluated by way of the LSI-R questionnaire multiple times. This data set, apart from the LSI-R score for individual offenders, also has information on the answers of offenders to each of the 54 items. Therefore, I can see the entries as recorded by the interviewers at the time of the administration of each LSI-R questionnaire. This is crucial for the analysis, since it can help identify the item(s) where the manipulation of the index might have occurred.

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 October 9, 2000 until July 15, 2008.[15] As with the LSI-R score, here too, many offenders were categorized multiple times.

Finally, WSIPP provided demographic, criminal history and recidivism data for offender cohorts from 1990 to 2004.[16] These data consist of 303,190 unique cases for 197,119 individual offenders. This last data set has been put together by WSIPP by combining and merging administrative data originating from the DOC and the Administrative Office of the Courts ("AOC").

To make use of the full extent of this information, and to enrich the

[14]As already mentioned, the Washington authorities discontinued the use of the LSI-R in August 2008.

[15]As already noted, the OAA came into force and the RMI started operating after the authorities had already started using the LSI-R score. This is why the earliest LSI-R interview predates the oldest RMI classification.

[16]Obviously, relevant for the present study are only entries for which I also have information on LSI-R and RMI. Therefore, I did not use demographic, criminal history and recidivism information for cohorts prior to the year 2000.

scope of the analysis, I merged the three data sets in the following way. The data set that contained information on the LSI-R scores was first merged with the data set that contained information on the RMI classification. The product of this first merge was further merged with the data set containing demographic and criminal history information. At that point, following WSIPP's precedent (Barnoski and Aos, 2003), I restricted my sample to cases where the LSI-R had been administered within 90 days from the day an offender became at risk to the community, i.e. either released from prison or sentenced directly to the community. If there were more than one LSI-R interviews within that period, the closest one to the release date was kept. 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. The earliest release date in the sample is July 14, 2000 and the most recent June 30, 2004. For the LSI-R interviews, the respective dates are October 9, 2000 and September 20, 2004.

Table 1.1 provides information on the composition of this data set with regard to gender, race and age of the subjects, as well as the nature of the most serious offense they are currently convicted of (seven broad categories) and the type of their sentence (imprisonment or community sentence). The sample consists mostly of male, white, adult (young and early middle age) offenders that serve sentences in the community, while almost one third of the offenses are drug related. Additionally, Table 1.1 presents 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.

As indicated above, Figure 1.1 gives the distribution of the LSI-R scores which would have been fairly normal had it not been for the easily noticeable large breaks between points 23 and 24, 31 and 32, and 40 and 41.[17] Again, 24 is the cut-off score for an offender to be upgraded from RMD to RMC, 32 is the cut-off point for an offender to be upgraded from RMC to RMB and 41 the cut-off for an upgrade from RMB to RMA or from RMC to RMB depending on

[17] An identification method for the problem of discontinuity in the density function of the running variable around the cut-off points has been proposed by McCrary (2008).

the crime committed.[18]

Figure 1.2 presents cumulative counts and percentage data for the way the LSI-R score translates to each of these four categories. We observe that even though passing the thresholds of the LSI-R score does not guarantee that an offender will be moved from one supervision category to another, nonetheless the probability that this will happen is affected sharply. For example, 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. Similarly, the probability that an offender is classified at the RMD category is about 73 percent if his or her score is 23, but falls to roughly 6 percent with a score of 24. Similar but less dramatic patterns can be observed on the other two thresholds.

This last observation suggests that the authorities follow closely the thresholds of the LSI-R grid when they are making risk classification decisions. This seems to explain their tendency to give extra points to offenders so that they exceed the thresholds and accordingly, with a considerable degree of certainty, be pushed to a higher level of supervision.

Finally, Table 1.2 presents recidivism information for the sample. Here, a recidivism event is any conviction for a felony or misdemeanor offense committed by an offender within a 36-month period after his or her being at-risk in the community (Drake, Aos, and Barnoski, 2010). A 12-month adjudication period is also allowed. Therefore technical violations of community custody conditions do not constitute recidivism events for my purposes. In Table 1.2 we notice that the sample has a general recidivism rate of almost 50 percent. The Table also indicates the rates for the specific types of recidivism both at the sample level and for each one of the four risk categories.

## 1.4 Theoretical Background

The argument in this Chapter is that the bureaucrats in Washington State were deciding to bump offenders to the next supervision category by adding

[18]The raw unprocessed data on LSI-R scores that I referred to above (437,335 unique scores) show similar discontinuities on the cut-off points.

extra points to their LSI-R scores. The question then is why they tampered with the instrument. Was it for a good reason or not? It can be argued that the authorities, by being in close contact with the offenders, have more information available to them, information that could not be possibly captured by a risk assessment instrument. Therefore, when they add points to an offender's score and upgrading him or her to a new category, it is because they believe that this new score and risk category reflects better the level of risk posed by the specific offender.

From a theoretical point of view, the authorities are maximizing the predictive accuracy of the instrument by tampering with it. Their concern is that the set up of the instrument would not properly reflect the risk level of a particular offender and therefore they intervene in order to correct what would otherwise be a misclassification of risk. The new score generated after their intervention (manipulation) is thought by them to be more consistent with the actual risk an offender poses. Therefore, the predictive accuracy of the LSI-R instruement is enhanced, even though the instrument has been manipulated.

The constraint the authorities are facing is that they have to make sure that the manipulation is subtle and not easily detectable or observable. To that end their intervention cannot affect items of the LSI-R questionnaire that are objective in nature as this would raise concerns on the part of the offenders. For example, the ten first items in the Criminal History section of the questionnaire cannot be subjected to manipulation. It would not be possible to report that an offender has three or more conviction on item 3, when in fact he or she has two. Therefore, the manipulation can target without raising suspicions only the items of the questionnaire that take more subjective responses. An inaccurate entry for those items would be difficult to uncover on the individual level and even harder to challenge. Only aggregate macroscopic analysis, such as the one undertaken in Section 1.5 below, could possibly unveil it.

In view of the above, the theoretical predictions that need to be verified below is both that the manipulation improved the predictive accuracy of the instrument and that the manipulated items where more subjective in nature. The empirical strategy presented below addresses both of these issues.

## 1.5 Empirical Methodology

### 1.5.1 Identification of manipulated items

The first step in identifying the manipulated items is to produce graphs that show the percentage of offenders that have answered each item affirmatively per LSI-R score, i.e. score 0, 1, 2, ..., 54.[19] Figure 1.3 presents, as an example, only the graph for item 1. The graphs for the remaining 53 items exhibit similar patterns and would not alter the analysis. As expected, as the score increases, so too does the percentage of offenders answering affirmatively on each item. For example, for offenders that have a score of 1, the percentage is very low for any given item. Conversely, for offenders that have a score of 54 (the maximum score) the percentage is 100 percent for all the items.

Based on Figure 1.1, my prior is that there should be some specific items that reveal where the manipulation took place. I suspect that if the scores were indeed manipulated, this would have been done in an organized rather than a random fashion. Therefore, for some items in Figure 1.3 and the other 53 graphs that are not presented here, there must be a discrete increase in the percentage of offenders that have a score of 24, 32 or 41. As it turns out, this is not entirely obvious in these graphs, and as a result I have to refine my analysis further.

The next step is to focus more on the set of offenders that have scores around the three thresholds in each item. As noted, it is expected that the percentage of offenders with a score of 24 who have answered affirmatively any of the 54 items should normally be higher than the percentage of offenders with a score of 23. This is the point where the solution can be given by the regression discontinuity design. By using this technique, I can unveil the items for which the difference in the two percentages is higher than what it would have been in the absence of manipulation.

To do this I use the micro data to run a binary response model, where the dependent variable is a dummy that takes the value 1 if the offender has an-

[19] The word "offenders" is used loosely here because, as indicated in the previous section, the data set is organized by offense since many offenders have committed multiple offenses. Maybe a more precise characterization of the percentage would be "percentage of interviewees" but I believe that use of this term in the text might lead to unnecessary confusion.

swered the item affirmatively and 0 otherwise ($yesonitem1_i, \ldots, yesonitem54_i$). The independent variable of interest is the regression discontinuity dummy that takes the value 1 if the offender has a score greater than or equal to the threshold I am examining and 0 otherwise. Therefore, when I am examining the jump at, say threshold 24, the regression discontinuity dummy has the form: $rd24_i = 1\{score_i \geq 24\}$. A linear polynomial of the risk score ($score_i$) concludes the list of the right hand-side variables. Therefore, for threshold 24, the first of the 54 regressions, one for each item, has the following form:

$$yesonitem1_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i. \quad (1.1)$$

The structure is identical for the remaining items. For example, the regression for the 54th item has the form:

$$yesonitem54_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i.$$

Similarly, when we try to identify the manipulated items for the jump at threshold 32, the first of the 54 regressions we run has the form:

$$yesonitem1_i = \beta_0 + \beta_1 rd32_i + \beta_2(score_i - 32) + \beta_3 rd32_i(score_i - 32) + \epsilon_i. \quad (1.2)$$

And finally, for threshold 41 the form of the first regression is:

$$yesonitem1_i = \beta_0 + \beta_1 rd41_i + \beta_2(score_i - 41) + \beta_3 rd41_i(score_i - 41) + \epsilon_i. \quad (1.3)$$

Following the literature on binary response models (Wooldridge, 2002, 2009), I use OLS to estimate the above regressions. Due to the simplicity of its implementation and interpretation, the use of OLS is generally recommended in this situation. To make up for the resulting heteroskedasticity I use robust standard errors in all the estimations. I cross-check the OLS results by also running separate probit and logit models for the above regressions.

The regressions are focused around the thresholds. Specifically, I use 5 points below each threshold and 6 points above it (including the threshold itself).[20] From the regressions of the micro data I obtain the predicted values of the percentages and I fit a linear line separately for the points below and above the threshold. These lines are superimposed on the actual aggregate percentage data.

This strategy produces 54 graphs for each of the three thresholds, a total of 162 graphs. Figure 1.4 presents here the three graphs, one for each of the three thresholds, that correspond to item 1. These graphs reproduce the aggregate data presented in Figure 1.3 but now the focus is around the thresholds and I superimpose the micro data applying the regression discontinuity design.

To be more precise, in this paragraph I attempt to link the numerical results obtained from the above regressions with their graphical representations, an example of which is Figure 1.4. Here I focus on the first threshold, 24, which is the largest of the three, but the analysis holds *mutatis mutandis* also for the other two thresholds. Looking at the first panel of Figure 1.4, I used the predicted values for scores 19-23 to create a counterfactual value of the percentage of offenders with score 24 that should have answered affirmatively a given item if there was no manipulation. This countefactual value is obtained by a linear extrapolation of the predicted values for scores 19-23. Then the regression discontinuity estimate, $\hat{\beta}_1$, measures the difference between this countefactual value for offenders with score 24 and the predicted value for the same offenders that is obtained from the regression. Graphically, this difference is the vertical distance between the two fitted lines in the three graphs of Figure 1.4. The numerical value of the regression discontinuity (RD) estimates and their standard errors (SE) are reported above each one of the three graphs.

The final step is to identify the items where the regression discontinuity estimates are the largest and the most statistically significant. To make the estimates for the different items comparable, I add one extra layer of analysis by dividing the RD estimates by the size of the constant of the respective regression, $\hat{\beta}_0$. Then the expression $\frac{\hat{\beta}_1}{\hat{\beta}_0}$ gives us a sense of the discontinuity in percentage terms and renders the size of the RD estimates comparable across regressions.

[20] Therefore 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.

### 1.5.2 Assessment of manipulation effectiveness

The issue of evaluating the effectiveness of the instrument manipulation by the authorities is dealt with by way of simple regression. The idea is that the authorities had a reason for bumping offenders up to the next supervision level, i.e. they thought they could outperform the predictive capacity of the instrument. The authorities basically reintroduced clinical judgment in the administration of the instrument, something that is strictly opposed by the literature, as indicated above.

The question is how it can be verified that the authorities improved the predictive capacity of the instrument by manipulating its items. The answer that I am proposing here is measuring how well the manipulated instrument explains the time that it takes for an offender to recidivate and compare it with the performance of the corrected instrument, the one that I generated after omitting the manipulated items. In other words, I am measuring and comparing the goodness of fit of a model where the dependent variable is the amount of time to recidivism and the independent variable is the manipulated or the corrected instrument. The specifications are as follows:

$$timetorecid_i = \beta_0 + \beta_1 manipulatedlsi_i + \epsilon_i, \tag{1.4}$$

and,

$$timetorecid_i = \beta_0 + \beta_1 correctedlsi_i + \epsilon_i, \tag{1.5}$$

where, $timetorecid_i$ is the amount of time to the recidivism event measured in days after being at-risk in the community and $manipulatedlsi_i$ and $correctedlsi_i$ are the manipulated LSI-R scores and the corrected LSI-R scores respectively. The measure of goodness of fit selected is a classic $R^2$ but also the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The analysis is refined by using different measures of recidivism, such as felony recidivism, violent felony recidivism, etc.

## 1.6 Results

### 1.6.1 Manipulated items

By running the 54 regressions specified above in equations (1.1) through (3.3) for each one of the three thresholds, I obtained the regression discontinuity estimates and their standard errors for each one of the 54 items.[21] I now use these econometric results to identify the items where the manipulation occurred.

#### Threshold 24

Figure 1.5 contains cumulative information on the RD estimates for all of the 54 items around threshold 24. In Panel (a) I present the size of the 54 RD estimates, subdivided based on whether they are statistically significant at the 5 percent level or not. Also Table 1.3 presents a cumulative account of these estimates. In column 1 we observe that for threshold 24 out of the 54 items, the number of positive and negative RD estimates is equal, both being 27. However, when we turn to the statistical significance of these estimates (column 2), naturally the positive estimates take the lead with 6 compared to 5 for the negative estimates. The highest positive estimate is over 0.05 and it is the one for item 54, the last item of the questionnaire. It is followed by the estimate for items 31, 53, 42 and 30.

In Panel (b) of Figure 1.5 I present again the same RD estimates subdivided by the degree of subjectivity of each of the 54 items. Specifically, I grouped the items in three categories:

**Subjective** These items are manipulable and the interviewer could tamper with them. An example is item 54: "Attitude poor toward supervision."

**Objective** Items that are not manipulable and take clear cut answers. An example is item 1: "Any prior adult convictions."

[21]As demonstrated in Appendix Figures B.1 and B.2, the marginal effects for the RD coefficients generated by both probit and logit regressions are for the most part extremely close to the coefficients obtained by regular OLS, justifying the use of the latter for the analysis that follows. Minor discrepancies in the coefficients' size for threshold 41 would not alter the conclusion on the manipulated items.

**Middle** Any items that don't fall into any of the previous two categories. An example is item 21: "Financial Problems."

Given this grouping, we observe in Panel (b) that most of the larger positive RD estimates fall either in the subjective or in the middle category. Specifically, items 30-31, in the Leisure/Recreation section of the questionnaire, and 53-54, in the Attitudes/Orientation section warrant our attention. Not only are they clearly subjective, but they also are large and significant, as we already saw in Panel (a). We should also note item 42 which is in the middle category. The objective items, which are mostly represented by the criminal history section of the questionnaire (the first 10 items), generally have small RD estimates, and are for the most part negative.

A last level of analysis but also a very important one for the purposes of identifying the manipulated items, is presenting the RD estimates divided by the respective constant for each regression. As indicated above, this gives us a picture of the jump in percentage terms. In Panel (c) we present this version of the RD estimates together with information on their statistical significance at the 5 percent level. Here the estimates for items 31, 53 and 54 continue to be among the highest and the only significant among the positive estimates. The size of the jump is almost 20 percent of the size of the constant for item 54 and close to 10 percent for items 31 and 53.

Given the above analysis, I believe that for the jump of the LSI-R scores on threshold 24, manipulation is generally occurring in the non-objective items of the questionnaire (30, 31, 42, 53, 54) and most notably in items 53 and 54. These two are highly subjective items and I suspect that as the interviewers were realizing that the offender needed more points to get past the 24 threshold, they were more likely to add some extra points to the score in these last few items of the questionnaire.

### Threshold 32

Figure 1.6 and Table 1.4 repeat this exercise for the jump on threshold 32, which as can be seen in Figure 1.1 is less pronounced than the jump on threshold 24. In column 1 of Table 1.4 we note that the negative RD estimates slightly outnumber the positive estimates in terms of their size, while in column 2, where we take into account statistical significance, the picture is completely balanced with 3 positive and 3 negative RD estimates. Panel (a) of Figure 1.6 presents these results graphically. Item 53 is again one of the three positive and statistically significant estimates having the second highest value after item 10.

Panel (b) gives the RD estimates based on the subjectivity of the items, showing more mixed signals, since item 10 is in the Criminal History section of the questionnaire and has the highest positive estimate. Other objective items, such as items 49 and 16 have high positive RD estimates, which was not expected. However, the second highest estimate does come from the subjective item 53.

Finally, Panel (c) gives the RD estimates divided by the constant. Here, among the 5 percent significant estimates, the estimate for item 10 retains its lead, while item 53 follows with a jump of around 5 percent of the size of the constant. The jumps here, however, are much less pronounced in percentage terms compared to threshold 24.

Based on these results, the manipulated items for the jump on threshold 32 are harder to identify. The preponderance of an objective item, item 10, blurs somewhat the picture. On the other hand, item 53 is again one of the manipulated items, which reinforces our conclusion for threshold 24 on the importance of the last items of the questionnaire for the bumping up strategy.

**Threshold 41**

Finally, Figure 1.7 and Table 1.5 repeat the exercise for the jump on threshold 41, which is also less pronounced than the jump on threshold 24. Here again, starting from column 1 of Table 1.5 we observe that there are 29 positive RD estimates and 25 negative ones, out of which 4 of the positive and 1 of the negative are significant. Not surprisingly, as seen in Panel (a) of Figure 1.7, the estimate for item 53 is once again one of the 4 positive and significant estimates.

Panel (b) presents the RD estimates based on the subjectivity of the item, and here we do not observe high RD estimates for objective items, as was the case for threshold 32. Items 50 and 45 are in the middle category, followed by item 53 in the subjective category. Most of the estimates for the objective items are actually negative. Panel (c) and the RD/Constant version of the estimates essentially repeat the results of Panel (a). The estimates for items 50, 45 and 53 are again the highest and significant at the 5 percent level.

The conclusion for threshold 41 is more in keeping with what was expected. We do not see large jumps for objective items, while item 53, which was identified as manipulated for the previous two thresholds is once again one of the manipulated items as shown in all three panels of Figure 1.7.

**Overall assessment and verification of analysis on manipulated items**

In view of the above I conclude that the items that were used by the authorities to push offenders over the thresholds were the following seven: 30, 31, 42, 45, 50, 53 and 54. This conclusion is based on either the pure size of the RD estimates or the RD/Constant estimates that are statistically significant at the 5 percent level.[22] Specifically, I found item 53 to be involved in the manipulation of all three thresholds. Items 30, 31, 42 and 54 were also used for threshold 24, while items 45 and 50 were also used for threshold 41.

Therefore, item 53 seems to be the item that is primarily responsible for the three discontinuities observed in Figure 1.1. I believe that the heavy use of this item in the manipulation process has to do with two factors: (a) it is an item that is close to the end of the questionnaire, and (b) it is a highly subjective item asking whether the offender has an attitude that is poor toward sentence.

Figures 1.8 through 1.10 provide the graphs that correspond to these seven items organized by the threshold that they correspond to. These are nine of the 162 graphs that I generated and which, as already noted, are not all reproduced here.[23]

This is a preliminary verdict for the manipulated items and I verify it by recalculating the LSI-R score without these seven items. Accordingly, the new index has a maximum score that is seven points lower.[24] Figure 1.11 presents

[22]In this conclusion I excluded items 10 and 16, which I found to be statistically significant for threshold 32, because they correspond to objective items.

[23]Note that item 53 appears for all three thresholds, so three times which is why the graphs are nine although they correspond to seven items.

[24]The three thresholds of the new index were calculated in a slightly more complicated way. First I found the average of each of the seven items around each of the three original thresholds, 24,32 and 41. Then I summed up these seven averages to get 3.46, 4.66 and 5.93

this new distribution of the LSI-R score. Here the jumps have disappeared and the overall distribution looks almost perfectly normal. I believe that this graph offers strong support of the fact that the items we identified as manipulated were actually the ones that were used for the bumping up process.

### 1.6.2 Assessment of manipulation effectiveness

Table 1.6 presents the results on various goodness of fit measures for equations (1.4) and (1.5). We notice that the manipulated scores fit the data on time to recidivism better than the corrected scores for most of the outcomes. This should be no surprise since the authorities must have had good reason to manipulate the data and these results serve as proof for the non-arbitrariness of their behavior.

Specifically, we notice that, with the exception of property felony recidivism, the goodness of fit measures for all other types of recidivism indicate that the manipulated instrument is a better predictor of time to recidivism compared to the corrected one. We also note that the differences in the measures are quite small. However, the fact that they are consistently, with one exception, in the same direction, shows a preponderance of the manipulated scores.

This result verifies that the local bureaucracy in Washington manipulated a risk assessment instrument in a way that it produced more accurate predictions. Moreover, this conclusion runs contrary to a general perception about the role of bureaucracies in modern government structures. The Washington authorities demonstrated that there is room for an efficient and innovative bureaucracy that not only implements but, occasionally, also improves standard procedures. Additionally, the authorities showed that a risk prediction instrument can produce better results if it is properly tampered with. This result is also in contrast with standard premises of the relevant literature on risk assessment.

---

respectively. Finally, I subtracted the sums from the respective thresholds to get the new thresholds. Rounded to the nearest integer the new thresholds were, 21, 27 and 35. The reason for adopting this approach is that I could not simply subtract seven points from each threshold since not all offenders had answered affirmatively the seven manipulated items. So what I needed to subtract had to depend on the sum of the seven averages around the three thresholds.

On the other hand, it should also be noted that the treatment program operated in Washington, according to which different supervision levels were assigned to offenders of different risk, did not manage to reduce recidivism.[25] Would it have been more rewarding for the authorities if supervision actually worked? I believe that it would. Does the fact that supervision does not work, diminish the contribution of the authorities, who, nevertheless, devised a better way to predict risk? I believe that it does not. Risk prediction is not related only to the specific treatment program. Risk prediction is a quintessential mechanism used by correctional authorities in order to administer a multitude of treatment programs. Understanding and predicting risk better is a benefit that accrues to all correctional interventions or programs, present or future that are based on risk assessment. Therefore, the work of the authorities in Washington may serve as an example to all correctional officers, proving that the efficient bureaucrat, even if endangered, is a species that still exists.

## 1.7 Conclusion

In this Chapter, I evaluated the implementation of a community supervision program for offenders in the state of Washington. In particular I analyzed the way that the risk assessment instrument "Level of Service Inventory - Revised" was used there in order to assign supervision levels to offenders who posed different risk of reoffending.

The unusual distribution of the scores produced by the instrument served as an initial indication of improper administration. The large jumps of the index exactly on the cut-off points that separated the supervision levels pointed to the fact that the authorities must have manipulated the instrument. Using the underlying data on which the instrument was built, I uncovered the way the manipulation was carried out and specifically, I pointed out the actual items of the index that were used in the manipulation process.

By using a regression discontinuity design, I found that seven of the 54 items of the LSI-R questionnaire were the ones mostly used in the process and I called them the "manipulated" items. Most of these items are subjective in

[25]This is shown in Chapter 2.

nature, although for a particular threshold I found evidence of manipulation in two items that are objective and should therefore normally be non-manipulable. After excluding the manipulated items from the calculation of the index, I recalculated all scores and generated a distribution that is much smoother, as it should have been, had there been no interference.

Additionally, I found that the manipulated instrument produced slightly better predictive results compared to the recalculated instrument that I generated by excluding the manipulated items. This result proves that the authorities acted in a subtle, yet efficient way and identified offenders whose higher risk would have otherwise gone unnoticed.

This analysis shows that bureaucracies are not necessarily egotistical, anachronistic and wasteful mechanisms, whose primary purpose is to prolong their existence by maximizing their budget. We found a case where professional and efficient bureaucrats, acting out of sincere devotion to the common good, bend the rules in order to ensure the proper level of post-release supervision for offenders. Even if it is an exception, the criminal agency of Washington State proves that unrefined generalizations and crude oversimplifications about bureaucracy are not warranted.

## 1.8 Figures and Tables

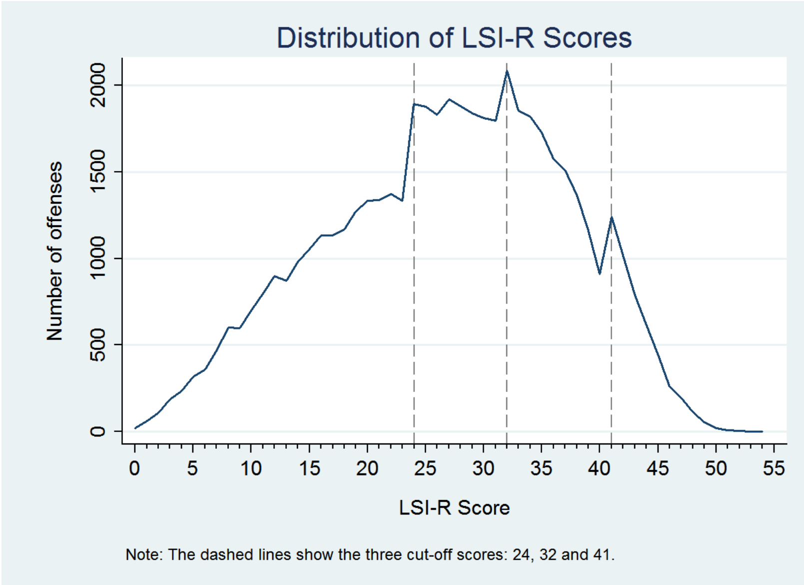


Figure 1.1: Distribution of LSI-R Scores. The LSI-R score was the risk instrument used by Washington State to measure the risk of offenders. It is a scale from 0 to 54, where higher numbers indicate higher risk.

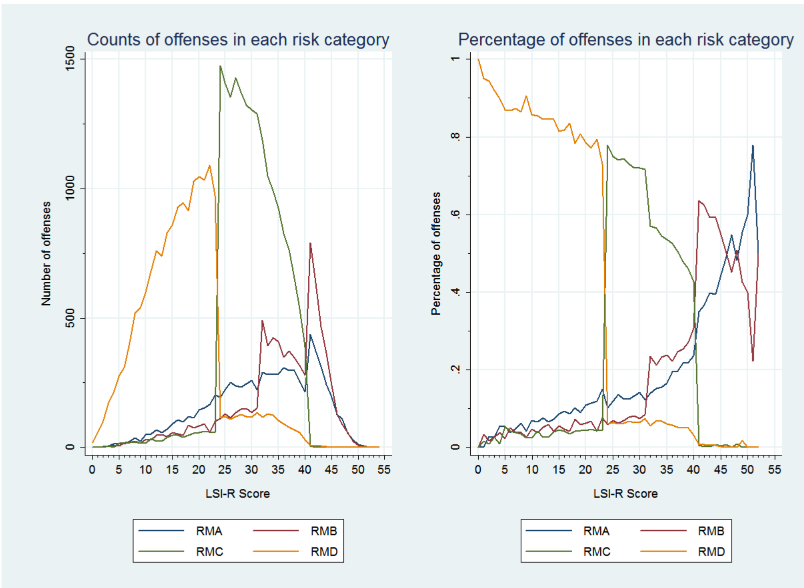


Figure 1.2: Counts and percentages of offenses in the four RMI categories per LSI-R score. The highest risk category is RMA and the lowest risk category is RMD. Offenders are assigned to them based on their LSI-R score.

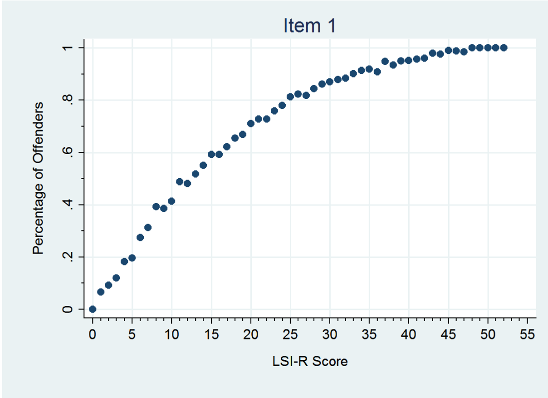


Figure 1.3: Percentage of offenders that answer affirmatively item 1 per LSI-R score. Item 1 is presented here as an example of the 54 items of the LSI-R. Obviously, as scores increase, the percentage of offenders that answer affirmatively increases and approaches 100 percent.

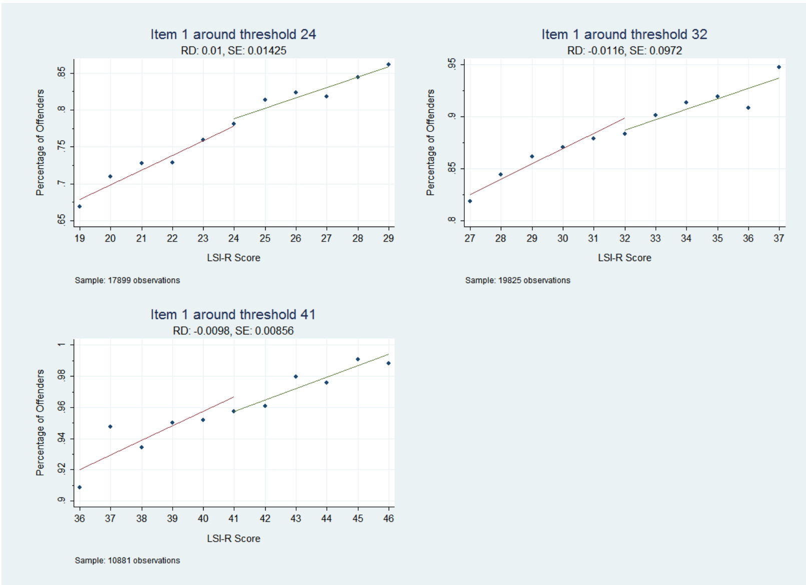


Figure 1.4: Three out of the 162 graphs generated by the regression discontinuity design. The three graphs correspond to Figure 1.3 but focus around the three thresholds. The dots in the scatter plots are the actual aggregate percentage data. From regressions of the micro data we obtain predicted values for the percentage data to which we fit a linear line separately for the points below and above each threshold. These lines are superimposed on the actual aggregate percentage data.

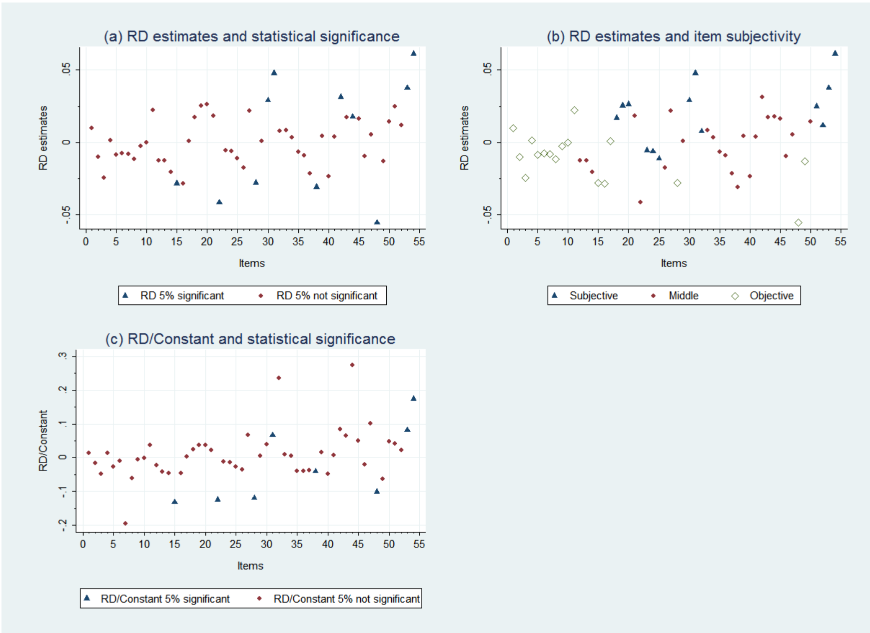


Figure 1.5: Panel (a) presents the Regression discontinuity (RD) estimates at threshold 24 for each one of the 54 items of the LSI-R questionnaire. Panel (b) presents the same estimates sub-categorized by the subjectivity of the item that they refer to. Panel (c) presents the RD estimates for each item divided by the constant of the respective regression. Item(s) identified as manipulated are: 30, 31, 42, 53, 54.

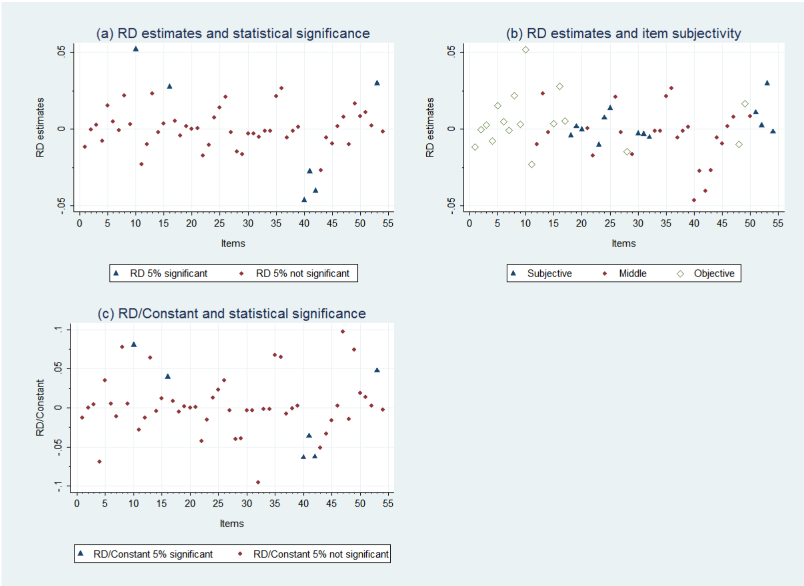


Figure 1.6: Cumulative data on RD estimates around threshold 32. Everything said on the note of Figure 1.5 applies here too but now with respect to threshold 32. Item(s) identified as manipulated are: 53.

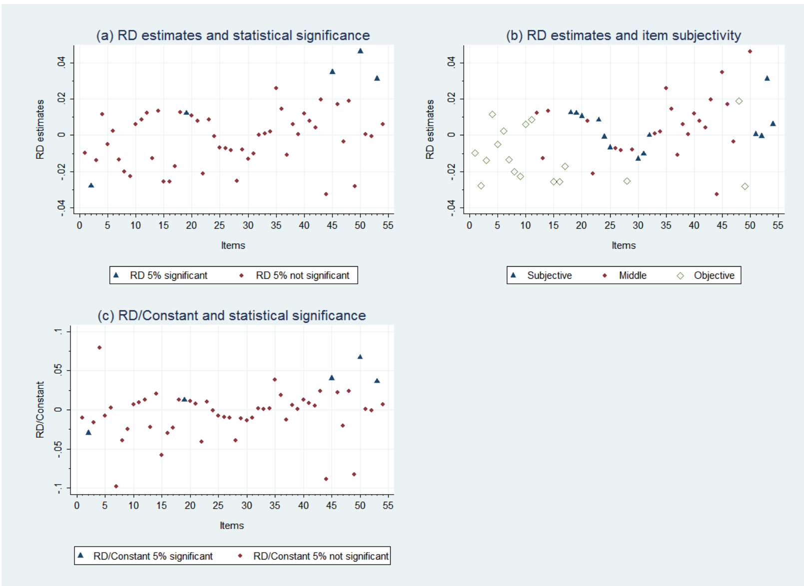


Figure 1.7: Cumulative data on RD estimates around threshold 41. Everything said on the note of Figure 1.5 applies here too but now with respect to threshold 41. Item(s) identified as manipulated are: 45, 50, 53.

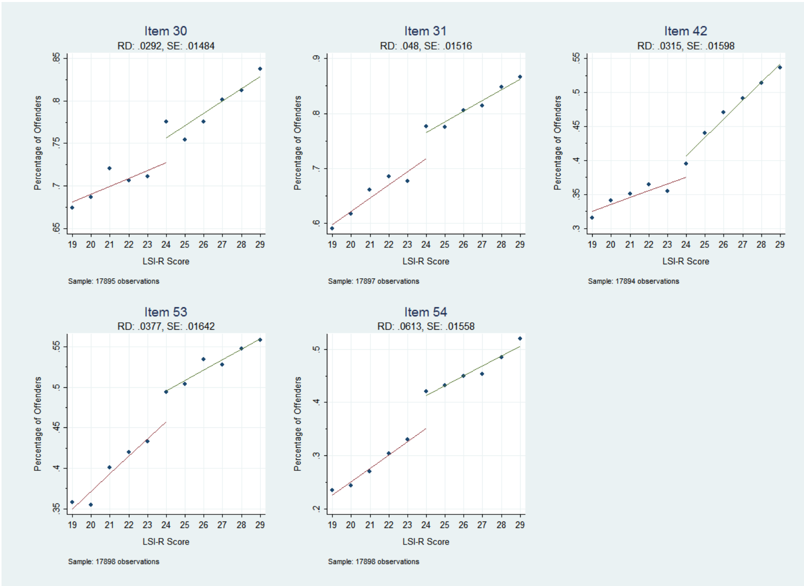


Figure 1.8: Items identified as manipulated for the discontinuity on threshold 24

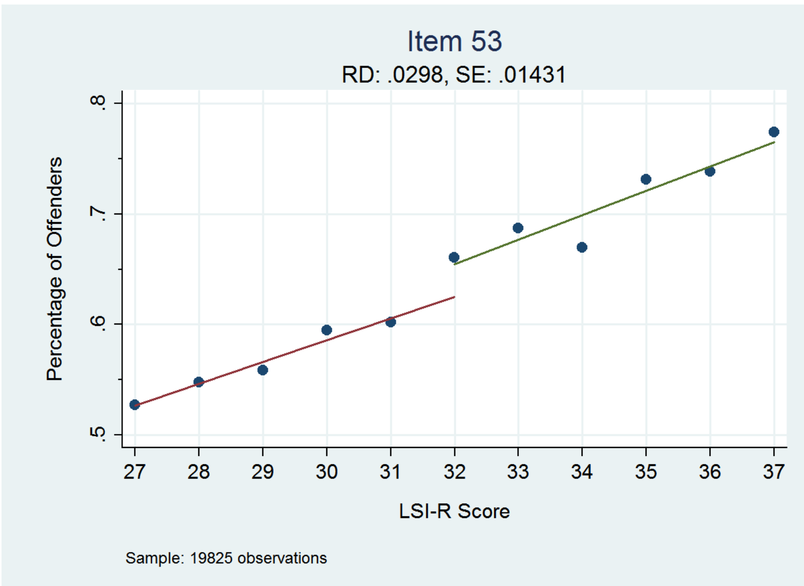


Figure 1.9: Item identified as manipulated for the discontinuity on threshold 32

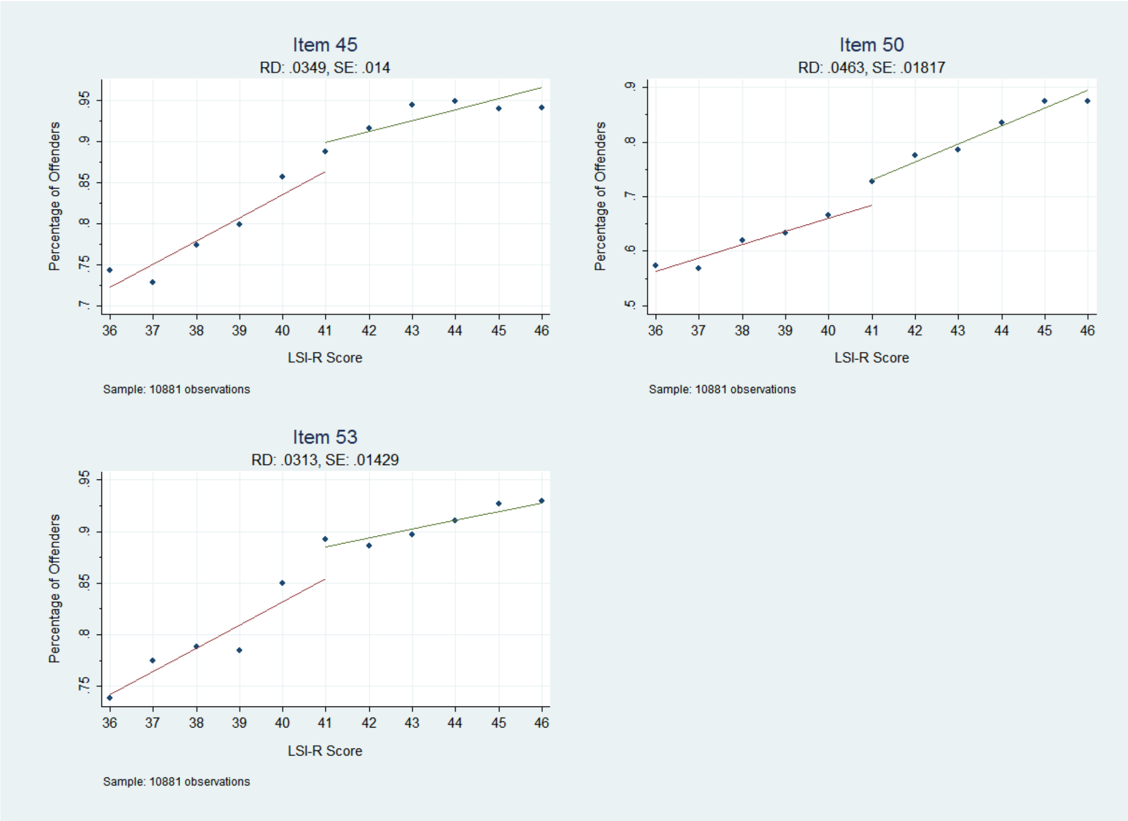


Figure 1.10: Items identified as manipulated for the discontinuity on threshold 41

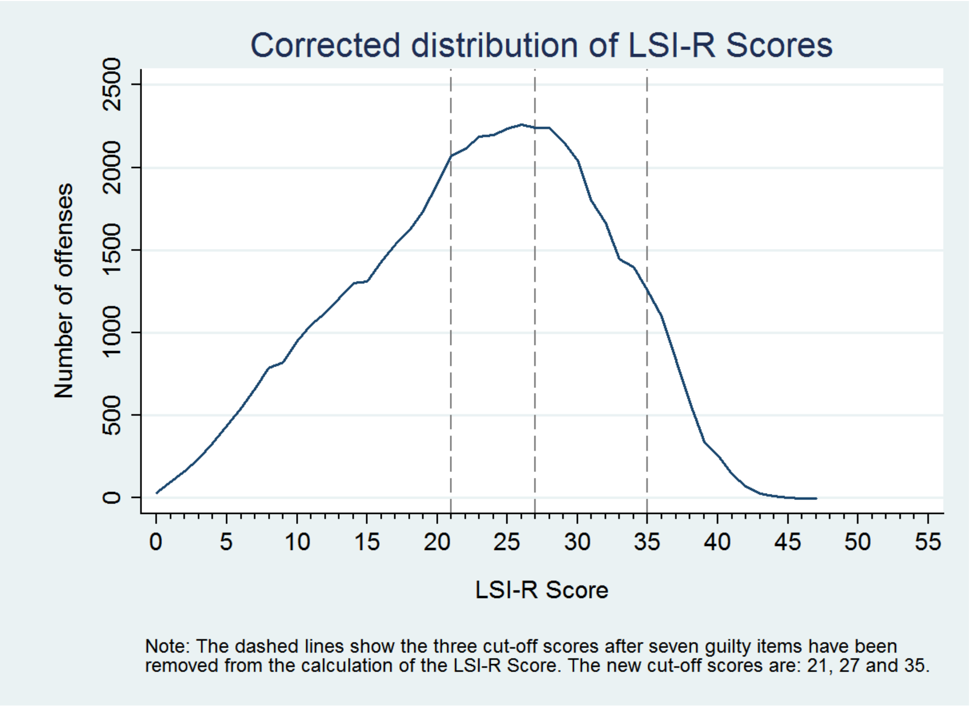


Figure 1.11: Distribution of LSI-R Scores after omitting the manipulated items. Without the seven manipulated items, the LSI-R index becomes a scale from 0 to 47.

Table 1.1: Composition of the data set in percentages

| 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 | 2.78 | 3.56 | 3.59 | 3.05 | 1.72 |
| Asian | 2.71 | 3.30 | 1.98 | 1.86 | 3.75 |
| Age 18-30 | 47.64 | 47.96 | 45.74 | 48.67 | 52.16 |
| Age 31-45 | 40.66 | 39.72 | 42.95 | 41.82 | 36.34 |
| Age over 45 | 11.59 | 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 |
| Observations | 51,957 | 7,902 | 8,126 | 19,008 | 16,831 |

Table 1.2: Recidivism 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 1.3: Manipulated items for the discontinuity on threshold 24

| Cumulative data on RD estimates for all the items based on Figure 1.5 | | | |
|---|---|---|---|
| | Number of estimates (1) | Number of estimates significant at least at the 5% level (2) | Number of RD/Constant estimates significant at least at the 5% level (3) |
| $RD > 0$ | 27 | 6 | 3 |
| $RD < 0$ | 27 | 5 | 5 |
| Total | 54 | 11 | 8 |

Table 1.4: Manipulated items for the discontinuity on threshold 32

| Cumulative data on RD estimates for all the items based on Figure 1.6 | | | |
|---|---|---|---|
| | Number of estimates (1) | Number of estimates significant at least at the 5% level (2) | Number of RD/Constant estimates significant at least at the 5% level (3) |
| $RD > 0$ | 25 | 3 | 3 |
| $RD < 0$ | 29 | 3 | 3 |
| Total | 54 | 6 | 6 |

Table 1.5: Manipulated items for the discontinuity on threshold 41

| Cumulative data RD estimates for all the items based on Figure 1.7 | | | |
|---|---|---|---|
| | Number of estimates (1) | Number of estimates significant at least at the 5% level (2) | Number of RD/Constant estimates significant at least at the 5% level (3) |
| $RD > 0$ | 29 | 4 | 4 |
| $RD < 0$ | 25 | 1 | 1 |
| Total | 54 | 5 | 5 |

Table 1.6: Goodness of fit for manipulated and corrected instruments

| | $R^2$ | | AIC | | BIC | |
|---|---|---|---|---|---|---|
| Time to event | Man. | Cor. | Man. | Cor. | Man. | Cor. |
| Recidivism | 0.0218 | 0.0202 | 351,954 | 351,995 | 351,970 | 352,011 |
| Felony Recidivism | 0.0337 | 0.0307 | 229,170 | 229,220 | 229,184 | 229,235 |
| Property Felony Recidivism | 0.0130 | 0.0151 | 104,680 | 104,664 | 104,694 | 104,678 |
| Drug Felony Recidivism | 0.0181 | 0.0126 | 111,817 | 111,860 | 111831 | 111,874 |
| Violent Felony Recidivism | 0.0107 | 0.0101 | 64,324 | 64,326 | 64,337 | 64,339 |
| Misdemeanor Recidivism | 0.0151 | 0.0142 | 325,409 | 325,429 | 325,425 | 325,445 |

*Note:* This Table shows different measures of goodness of fit for regressions where the dependent variable is time (counted in days) to a recidivism event. There is only one explanatory variable, either the manipulated or the corrected LSI-R instrument.

# Part II

# Second part

# Chapter 2

# Does increased supervision reduce recidivism?

## 2.1 Introduction

On December 31, 2009, there were 5,018,855 offenders on community supervision in the United States (Bureau of Justice Statistics, 2010). They were either sentenced directly to the community as probationers (4,203,967) or released from prison as parolees (819,308). This Chapter seeks to answer the following question: Does increasing the level of supervision reduce recidivism rates?

The issue has attracted considerable attention from the time of Martinson's "what works" study (1974) that included a special section on various treatment in the community programs. In the last decade, contributions from Petersilia (1999) and Travis (2005) generated a renewed interest in the topic but now focused more on prisoner reentry, the parole side of the community supervision spectrum. However, Petersilia (2009) has also noted the importance of "credible program evaluations," that will sift the programs that work from the ones that do not.

This study offers some new insight on the implementation process and the effectiveness of community supervision by evaluating empirically a treatment program that was initiated in the state of Washington in 1999 and is still in force. The program will be presented in detail in Section 1.2 below, but in essence what the Washington authorities did was to provide different levels of supervision to offenders that posed different levels of reoffending risk. Using a regression discontinuity design, I compare individuals that have similar risk characteristics, but received different levels of supervision due to the existence of cut-off scores separating offenders based on their risk. This approach provides quasi-experimental evidence on whether extra supervision is effective in reducing recidivism.

Previous work has yielded contradictory evidence on supervision effectiveness. Some studies have identified positive effects of supervision on recidivism. Solomon, Kachnowski, and Bhati (2005) showed that parole supervision managed to reduce rearrest rates of parolees compared to prisoners released unconditionally. Similarly, using self-reported data, MacKenzie, Browning, Skroban, and Smith (1999) found that probation resulted in reduced crime rates for probationers compared to the year prior to their arrest. On the other side of the spectrum, starting with Martinson (1974), probation and parole programs were found not to be particularly effective, and doubts were cast over programs that claimed they were. Some years later, a study by Petersilia and Turner (1991) showed that intensive supervision probation in California was no more effective than regular probation. An overall evaluation of intensive supervision and other programs (such as electronic monitoring and home confinement) by Cullen, Wright, and Applegate (1996) delivered similar results. Harris, Gingerich, and Whittaker (2004) showed that differential supervision of offenders, based on a case management classification system, did not produce the desired results in reducing recidivism either. Finally, a meta-analysis by Bonta, Rugge, Scott, Bourgon, and Yessine (2008) showed minimal effects of supervision on general recidivism, and even worse effects on violent recidivism. The authors go on to note that: "On the whole, community supervision does not appear to work very well" (p. 251).

One reason for these contradictory results and for most of these studies being inconclusive on supervision effectiveness is that they do no sufficiently control for simultaneity, that is offenders with more severe records are subjected to greater supervision. The purpose of this Chapter is to account for this simultaneity problem by using proper econometric techniques, thereby providing more convincing results.

The literature also stresses the importance of treatment program quality for ensuring the best possible results. Quality refers not only to the program characteristics that have to be consistent with the current state of the art (Harris, Gingerich, and Whittaker, 2004; Lowenkamp, Latessa, and Smith, 2006) but also to staff training and compliance with the program specifications (Dowden and Andrews, 2004). One of the most important requirements for a quality program is adhering to the so called "risk principle," according to which, higher risk offenders should receive more intensive treatment. It is therefore not surprising that Bonta, Bogue, Crowley, and Motiuk (2001, p. 228) note that "the assessment of offender risk is perhaps the single most important function of any correctional agency." Risk is commonly measured by way of "actuarial" (also known as "mechanical," "algorithmic," or "statistical") instruments and these will be in the center of the present study.

In Chapter 1, I showed that the authorities in Washington State engaged in a sophisticated system of instrument manipulation by adding extra points to the scores of offenders that they considered higher risks and ensuring increased supervision levels for them. In that Chapter I demonstrated that this practice generated instrument scores that could better predict time to recidivism compared to scores that were cleansed of all manipulation.

In this Chapter, I build on the results of Chapter 1 and use a regression discontinuity design to evaluate the effectiveness of the differential community supervision program based on several recidivism outcomes.[1] Additionally, I evaluate whether the program had an effect on the time span to the recidivism event.

The results in short are the following. I find sufficient evidence that the differential supervision program operated in Washington did not manage to reduce recidivism in a substantial or statistically significant way. Moreover, I did not find evidence that supervision had any effect on time to recidivism, i.e.

[1]Unfortunately the information available in the data sets does not allow me to explore other outcome measures such as employment or education as the literature suggests (Petersilia, 2004). This is one of the limitations of this study.

it did not prolong the amount of time it takes to recidivate.

The Chapter proceeds as follows. Section 2.2 presents the differential supervision program implemented in Washington State. Section 2.3 describes the data sets used. Section 2.4 gives an account of the empirical method used in this Chapter . Section 2.5 presents and discusses the results, and Section 2.6 concludes.

## 2.2 Differential Supervision in Washington

The correctional authorities in Washington State, part of a bureaucratic machine that administers criminal justice in the state, were assigned with the task of determining the proper level of supervision an offender is to receive once he or she is released to the community after either serving time in prison or being sentenced straight to a community sentence. To that end the authorities needed instructions with respect to the criteria that would determine supervision intensity for each offender. Legislation was passed in 1999 (Offender Accountability Act ("OAA")) that set those to be the following two: risk of reoffending and seriousness of the offense committed.

Reoffending risk was measured by way of the actuarial instrument "Level of Service Inventory - Revised" or "LSI-R." This is a popular risk assessment instrument and its qualities and mechanics have been analyzed in detail by numerous studies.[2] Briefly it can be said here that 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 offered to the offenders most importantly around the time that they will become at-risk to the community. Their answers are corroborated by official documentation or other methods, where possible.

The LSI-R builds into the system of supervision allocation ("Risk Management Identification" system or ("RMI")) provided for by the OAA. Three

[2]See e.g. Gendreau, Little, and Goggin (1996), Andrews and Bonta (1995), Gendreau, Goggin, and Smith (2002), and Chapter 1 herein.

cut-off scores across 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 passes the relevant cut-off score, he or she gets upgraded to the next supervision category. In Chapter 1 I showed how the manipulated LSI-R distribution depicted in Figure 2.1 can be transformed to the smoother distribution of Figure 2.2 by omitting the manipulated items of the questionnaire. Moreover, Figure 2.3 shows that crossing any of the three cut-off scores is concomitant with a considerable rise in the probability of being upgraded to the next supervision level.[3,4]

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

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 to the community. Existing community supervision legislation in Wash-

[3]The reason that there is a high probability and not a certainty is that the seriousness of the offense committed also plays a role in the decision of the authorities. However, as evidenced by Figure 1.2, the most important parameter the authorities follow is the LSI-R score.

[4]Table 2.1 provides further weak but corroborating evidence that there are certain dissimilarities between offenders around the three thresholds. Overall, Table 2.1 shows that for five out of the 45 regressions run, the characteristics are not comparable across the thresholds.

[5]The numbers were obtained by personal communication of WSIPP with DOC staff (Aos and Barnoski, 2005).

[6]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 budgeted hours switched from the 2002 numbers to the 2005 numbers. However, again further 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 approximation in the analysis.

ington was codified by House Bill 2718 in 2008, and subsequently modified by Senate Bill 5288 of 2009 and Senate Bill 5891 of 2011.

The length of community supervision ranges from less than a year to three years, depending on the nature of the offense committed.[7] The specific conditions of community supervision are determined initially by the court and subsequently by the DOC.[8] Additionally, DOC "Policies" further specify the operation of community supervision and the duties of supervision officers in a more detailed fashion.[9] Officers are required to develop an Offender Supervision Plan that outlines the objectives for each individual offender, the actions that need to be taken in that direction, reporting and contact requirements as well as any specific treatment programs that the offender may be subject to. Most importantly, the officers ensure that the Plan is followed by the offender.

As early as 2003, the officials of WSIPP had been proposing modifications in the LSI-R tool in order to improve its predictive power. Finally, in August 2008 the authorities decided to heed the proposals and moved to a radical overhaul of the instrument by excluding the dynamic factors and including only static factors in the new instrument (Barnoski and Drake, 2007; Drake and Barnoski, 2009). At that time the use of the LSI-R instrument was discontinued in Washington State.

The overall effectiveness of allocating supervision based on risk characteristics as ordained by the OAA was assessed by WSIPP in two of its reports, an interim and a final one (Aos and Barnoski, 2005; Drake, Aos, and Barnoski,

[7]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 that were sentenced to sentences of one year or less. However, I do not have information on the length of the community supervision term imposed on an 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 for which they were subject to supervision. According to the law, high risk offenders would be supervised for the whole three-year term, while low risk offenders for shorter terms. But since I don't have information on the exact length, I assume it to be the same for all offenders and spanning the entire three-year recidivism period. This means that the first stage estimates (relationship between LSI-R score and hours of supervision) produced in Section 2.5 below are understated in this respect.

[8]RCW 9.94A.703 & 704.

[9]E.g. Policies DOC 320.420 and 380.200 outline such duties, while the minimum contacts prescribed for each risk level are outlined in the Attachment to the latter Policy. As expected, the higher the risk category of the offender the more the contacts that the officer is required to make with him or her.

2010). Although the interim report (2005) presented optimistic preliminary results, the final report (2010) gave a much more ambiguous picture as to 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 Chapter is a valid way to sidestep this obstacle and obtain a quasi-experimental evaluation of the OAA's effectiveness.

## 2.3 Data

For this study I used the data set that I used in Chapter 1. This is a data set that was put together by combining three separate confidential data sets that were given to me by the "Washington State Institute of Public Policy." The first of them had information on LSI-R scores, the second on assignment of offenders to the four risk categories of the RMI system and the third on demographic, criminal history and recidivism characteristics.

Merging of these data sets produced a data set of 51,957 offenses committed by 47,154 individual offenders. The time span of the analysis is from July 2000 to September 2004. Table 2.2 presents descriptive statistics for general characteristics of the sample, such as gender, race, age, type of crime, type of sentence. Table 2.3 presents specific information on recidivism. For additional clarity this information is also broken down per risk category.

## 2.4 Empirical Methodology

In this study I use the results obtained in Chapter 1 with respect to the items of the LSI-R questionnaire that were manipulated. In that Chapter I 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 is smoother, since all manipulation has been eliminated, and thus allows me to make use of the similarity of individuals that have similar LSI-R scores. I can therefore proceed to my current task which is to assess how large, if any, is the effect of supervision on recidivism.

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

In the same vein as before, the answer that we are proposing is the regression discontinuity design. Having the offenders' risk scores ordered by way of the LSI-R instrument is a situation that lends itself to the use of this econometric technique. In particular, I compare individuals who have very similar scoring on the LSI-R index, but due to the artificial cut-off scores, they have been classified in different risk categories and therefore treated differently.[10] The regression discontinuity allows me to identify virtually identical individuals in terms of risk characteristics, and therefore reach a conclusion as to any possible causal effect of hours of supervision on recidivism rate.

The situation I am faced with is one where the risk instrument defines, not fully but still strongly, the risk category an offender is assigned to. This is known in the relevant literature (Imbens and Lemieux, 2008; Angrist and Pischke, 2009) as "fuzzy regression discontinuity." In Figure 1.2 I have already shown evidence that this situation falls into that category of regression discontinuity designs.

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

$$recidivism_i = \beta_0 + \beta_1 supervisionhours_i + \epsilon_i. \tag{2.1}$$

This again is a binary response model, where $recidivism_i$ is the outcome dummy variable that takes the value 1 if an offender has recidivated after he or she has

[10] Kuziemko (2007) used the same methodology in a study with offenders from the State of Georgia trying to answer the question if time served in prison has an effect on recidivism rates.

been at-risk to 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.[11] Then the estimate, $\hat{\beta}_1$, will measure the effect on recidivism of receiving more hours of supervision as a result of being assigned to a higher risk category.

The problem with this specification is that it is likely to result in estimates that are not consistent as 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, 2011). 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 (Angrist and Pischke, 2009).

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. For simplicity, I present here a generic form of the regression equation pertaining to any of the three thresholds, therefore the term *threshold* below refers to the numbers 24, 32 or 41 depending on the threshold the regression refers to. The exact specification is the following:

$$supervisionhours_i = \alpha_0+\alpha_1 rd_i+\alpha_2(score_i-threshold)+\alpha_3 rd_i(score_i-threshold)+u_i, \tag{2.2}$$

where again $rd_i$ is the regression discontinuity dummy defined as an indicator variable: $rd_i = 1\{score_i \geq threshold\}$. The estimate, $\hat{\alpha}_1$, measures the degree of association between an offender's scoring 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:

$$recidivism_i = \pi_0+\pi_1 rd_i+\pi_2(score_i-threshold)+\pi_3 rd_i(score_i-threshold)+v_i, \tag{2.3}$$

[11] I presented the relevant numbers of supervision hours in Section 2.2 above.

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 supervision on recidivism rates.

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 kinds of recidivism, such as felony or misdemeanor, violent or not, as well as on different types of crime, such as drug, property, etc. With regard to the right hand side variables, apart from the simple specifications presented above in equations (4.1)-(2.3), 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 and number of prior adult felony adjudications.

Apart from the issue of whether supervision had an effect on recidivism itself, it is also interesting to conduct a survival analysis on my sample, that is to verify whether supervision had an impact on the time it takes for a recidivism event to occur. The question is whether supervision prolonged the time an offender does not recidivate within the follow-up period, even if he or she recidivated eventually. To analyze this issue, I use the information I have on the number of days it takes an offender to recidivate since he or she became at-risk in the community, up to a maximum of three years. This question can be answered parametrically by using a Cox proportional hazard model. Again I repeat this analysis for multiple recidivism outcomes, such as felony, violent felony, property, etc.

## 2.5 Results

The short answer to the issue of whether the differential supervision program operated in Washington worked is No. The detailed answer is presented in Tables 2.4 and 2.5 and Figures 2.4 through 2.11 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 running equations (2.2) and (2.3), the first stage and the reduced form of the model respectively. The results for different recidivism outcomes are presented in Table 2.4 for each one of the three thresholds. Moreover, I use both the manipulated (first three columns) and the corrected LSI-R scores that I obtained after cleansing the instrument from the contaminated items (next three columns). The use of the manipulated scores, even though they don't allow me to capture individuals of similar risk characteristics around the cut-off points, is warranted by the fact that they represent the original risk decisions made by the authorities and therefore can serve as a benchmark.

Starting the analysis from the manipulated scores, the first section of Table 2.4 presents the RD estimates for the first stage of the model, i.e. the effect of the risk score on hours of supervision. For all three thresholds, the instrument (risk score) is strong and statistically significant. Indeed, offenders that cross all thresholds receive more hours of supervision. This effect is also captured graphically in panel (a) of Figure 2.4, and also in Figure 2.6, which focuses around the thresholds. The first and especially the third threshold entail a more significant increase in the hours of supervision compared to the middle threshold.

The second section of Table 2.4 reports the RD estimates for the reduced form, i.e. the effect of the risk score on the outcome variable. 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 is over 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 Figure 2.4 as well as in in Figure 2.7.

Finally, the third section of Table 2.4 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 basic result holds throughout the rest of our analysis, as we perform several robustness checks.

Turning to the corrected scores that are reported in the rightmost three columns of Table 2.4, the RD estimates for the first stage are now attenuated, as can also be seen in panel (a) of Figure 2.5 and in Figure 2.8 but they are still strongly significant. The attenuation can be attributed to the fact that after the recalculation of the score, the new thresholds do not perfectly capture the change in the amount of supervision. Appendix C offers a solution that addresses this issue. The reduced form RD estimates for general recidivism are still close to zero and non-significant (panel (b) of Figure 2.5 and Figure 2.9). For the other types of recidivism, except for three cases, the situation is generally similar. The IV estimates reveal again that, except for property felony recidivism and misdemeanor recidivism, there is no effect of hours of supervision on recidivism. For the former case, our results show that crossing the last threshold and getting more supervision reduces recidivism by 2.5 percent, while for the latter case, crossing the same threshold increases, surprisingly, recidivism by the same amount.[12]

In Table 2.5 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 a LSI-R score that is greater than *any* threshold. The observations that were used for each of the three threshold-specific regressions (five points below each threshold and six above) were stacked in a single data set for the manipulated scores, and 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. The difference in the first stage between manipulated and corrected scores observed in the threshold-specific regressions persists here too. We can see this graphically in the difference between panels (a) of Figure 2.10 and Figure 2.11. The IV estimates for both manipulated and corrected scores show again no statistically significant effect of hours of supervision on any type of recidivism. This is my preferred specification as it summarizes clearly and succinctly the absence of any effect of supervision on recidivism for any of the three thresholds.

Appendix C gives several robustness checks of the previous results.

[12]Given that I ran a large number of regressions for many different outcomes and found overall no effect of supervision on recidivism, the significance of these two conflicting results should not be overstated.

Specifically, Appendix Tables C.2 through C.4 indicate that our results are robust to the inclusion of control variables. The controls are powerful predictors of recidivism but they don't affect the size of the IV estimates. Finally, in Appendix Tables C.5 and C.6 I make up for the attenuated first stage in the corrected scores by removing observations that correspond to the two points around the thresholds. The first stage is indeed strengthened, but the IV estimates remain close to 0 and not significant.

One concern for the above analysis could be the so-called "surveillance effect." According to this argument, it might be the case that increased surveillance is actually reducing recidivism, but, on the other hand, this increased surveillance ends up uncovering many more 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 that these two effects counteract each other perfectly, 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.

Finally, Appendix Table C.7 presents the estimates from a Cox proportional hazard model. Again I used my simplified analysis where all the observations around *any* threshold are pooled in one data set. Moreover, I used both the manipulated and the corrected scores. The RD estimates reported in the Table show that there is essentially no difference in the probability to recidivate per unit of time if an offender is just over or just under the threshold. None of the estimates reported for various recidivism outcomes is statistically significant.

However, in spite of these negative results, it is interesting to reflect on whether there should be any difference on the size of the estimates between manipulated and corrected data and whether any such difference is observed in the data. Specifically, we know that the manipulation performed by the authorities pushed offenders up one risk category.[13] Therefore, looking at the manipulated

[13] This was shown in Chapter 1

data, we have (relatively) more risky offenders placed above the thresholds. When I corrected the data and readjusted the thresholds, these offenders were pushed back down, where they would have originally been placed had there been no manipulation. Therefore, looking at the corrected data, more risky offenders (relative to the manipulated data) are now just below the thresholds but their recidivism rates are high. On the other hand, fewer risky offenders (relative to the manipulated data) are now just above the thresholds.

In view of this description, when we use the corrected data we should see more benefits to supervision, i.e. less positive or more negative reduced form and IV estimates. Now, if the manipulated data show an effect of supervision that is barely over zero, the corrected data should show an effect that is even smaller or negative. Since, many of the manipulated data estimates are close to zero, sign reversals for the corrected data estimates should be observed (positive estimates in the manipulated data turn negative in the corrected).

Starting with the generic Table 2.5, we notice that the IV estimates for a general recidivism outcome is positive when we look at the manipulated scores (0.0030) but negative when we look at the corrected scores (-0.0033). The same pattern is observed for felony and violent felony recidivism, while for property and drug felony recidivism we observe a more negative number for the corrected scores. Our prediction is not verified only by misdemeanor recidivism. Overall, even if the numbers are very small and not significant, Table 2.5 confirms the prediction that the corrected data should show a more substantial effect of supervision on recidivism.

Moving to Table 2.4, we notice a similar pattern. The IV section of that Table shows us that 12 out of 18 estimates verify the prediction. In fact, the proportion is 9 out of 12 if we exclude misdemeanor recidivism, which is both less interesting and more scarce. Finally, Appendix Table C.7 repeats the same result for the Cox proportional hazard model with time to recidivism as the dependent variable.

These results show that there is a very mild (albeit non-significant) effect of supervision on recidivism that is masked by the manipulation of the scores and the bump up of more risky offenders. Even though this effect is weak, the overwhelming support of the data indicates that it should not go unnoticed.

## 2.6 Conclusion

In this Chapter, I assessed whether a treatment program that was implemented in Washington State and entailed differential supervision for offenders of different risk levels had any effects on recidivism probability. I also examined whether supervision at least slowed down the occurrence of reoffending after an offender has been put at-risk in the community.

I used a regression discontinuity design to assess the overall effectiveness of assigning different levels of supervision based on a risk instrument. In the analysis I used both the manipulated LSI-R scores, i.e. the scores as they were initially recorded, and the recalculated scores that I produced after omitting the contaminated items. By doing this, I tried to ensure that the individuals around the cut-off points are indeed comparable and, therefore, any outcome difference could be attributed to the different treatment.

The data, in all their forms, showed that the program did not succeed in reducing recidivism. Even though I made use of various recidivism outcomes and different specifications, I could not manage to verify any substantial or statistically significant reduction. Moreover, survival analysis techniques showed that supervision did not prolong the time to a recidivism event. The only shred of encouraging evidence that I could document was the difference in the size of the non-significant estimates observed between manipulated and corrected scores. This difference showed that there is a very small but non-significant effect of supervision on recidivism that has been concealed by the manipulation of the data.

No matter what the evidence of the literature on community supervision is, the reality remains that from 2000 to 2009, the number of offenders under community supervision increased from approximately 4.5 to 5 million. These offenders will be in the community whether community supervision works or not, since budget constraints prevent other more drastic forms of punishment. It is therefore important for research in the field to identify the programs that do work, as omitting to do so will be not only a minor scientific setback; it will be a most serious failure whose victims will be both the offenders themselves and the society at large.

## 2.7 Figures and Tables

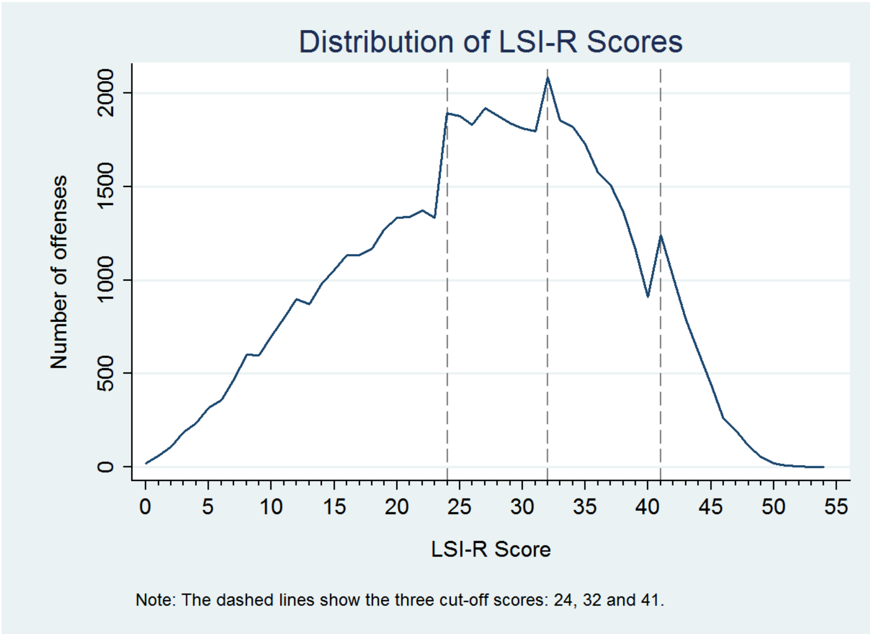


Figure 2.1: Distribution of LSI-R Scores. The LSI-R score was the risk instrument used by Washington State to measure the risk of offenders. It is a scale from 0 to 54, where higher numbers indicate higher risk.

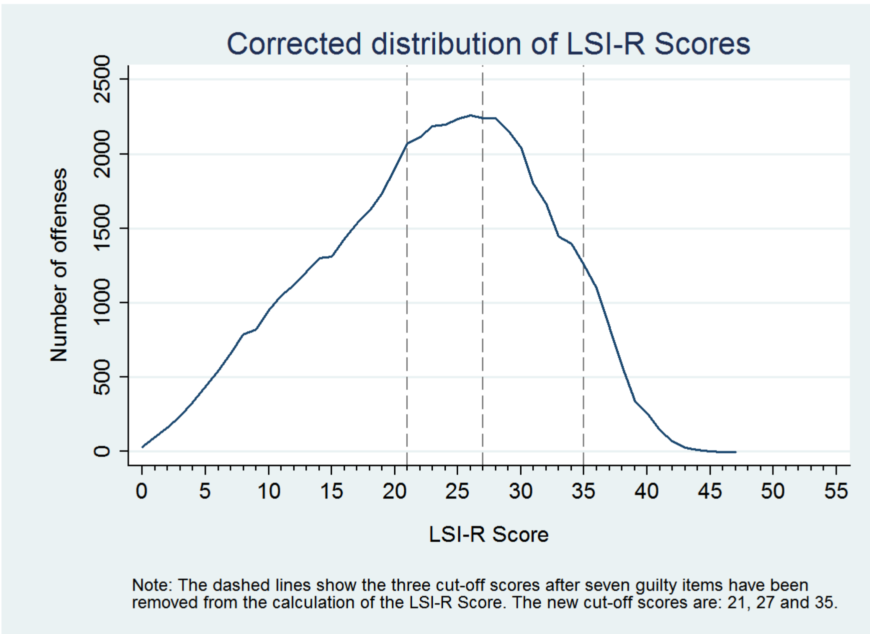


Figure 2.2: Distribution of LSI-R Scores after omitting the manipulated items. Without the seven manipulated items, the LSI-R index becomes a scale from 0 to 47.

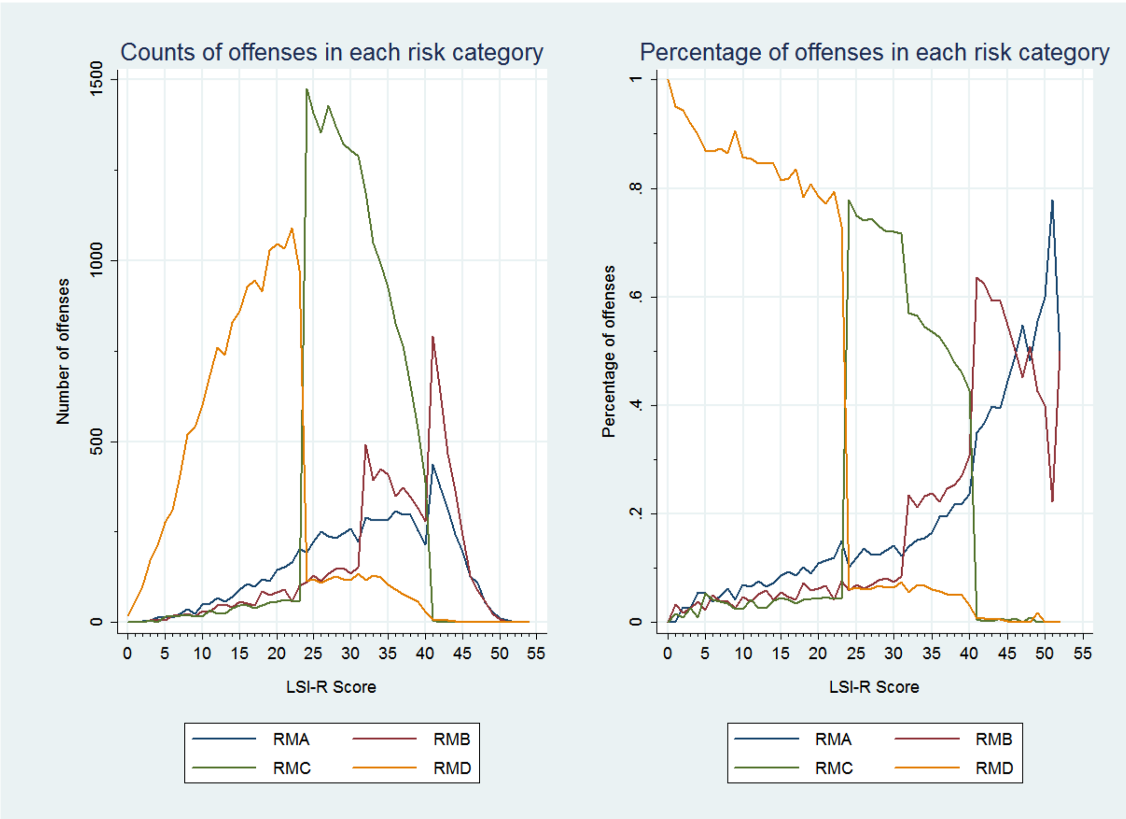


Figure 2.3: Counts and percentages of offenses in the four RMI categories per LSI-R score. The highest risk category is RMA and the lowest risk category is RMD. Offenders are assigned to them based on their LSI-R score.

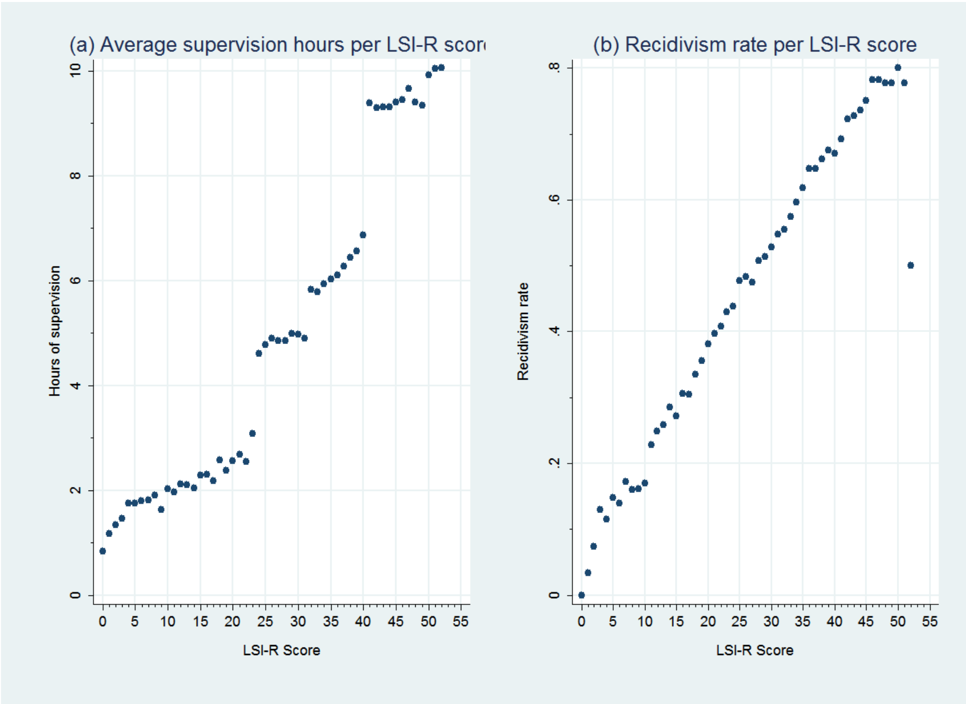


Figure 2.4: Average supervision hours and recidivism rate per LSI-R score using the manipulated data

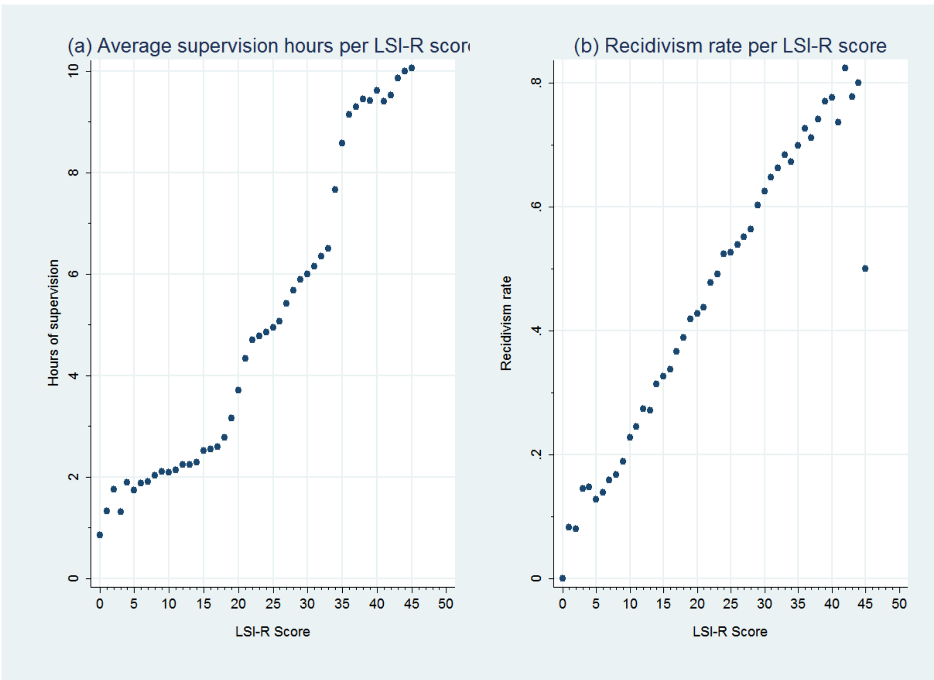


Figure 2.5: Average supervision hours and recidivism rate per LSI-R score using the corrected data

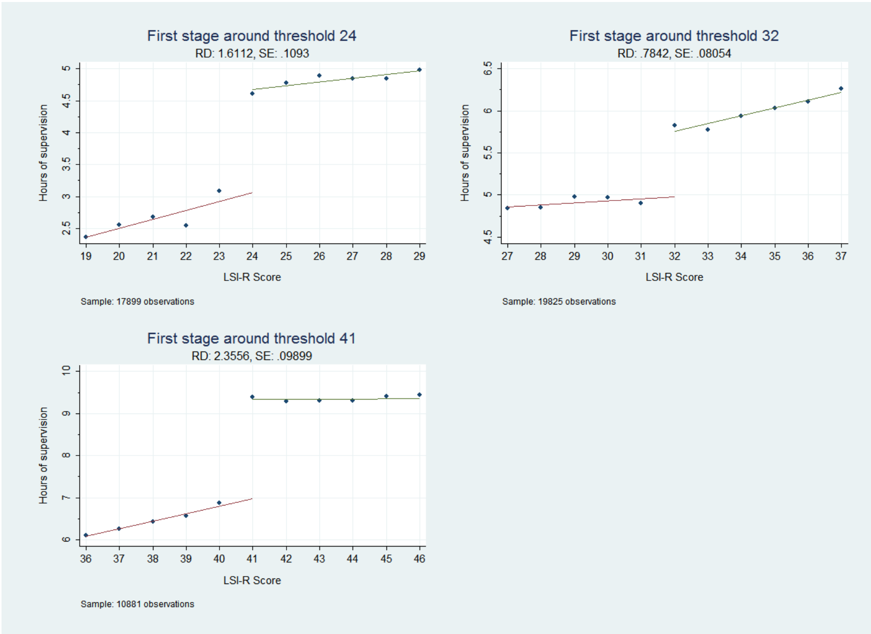


Figure 2.6: First stage for the three thresholds using the manipulated data. The Figure presents the effect of the LSI-R score on the hours of supervision for each one of the three thresholds.

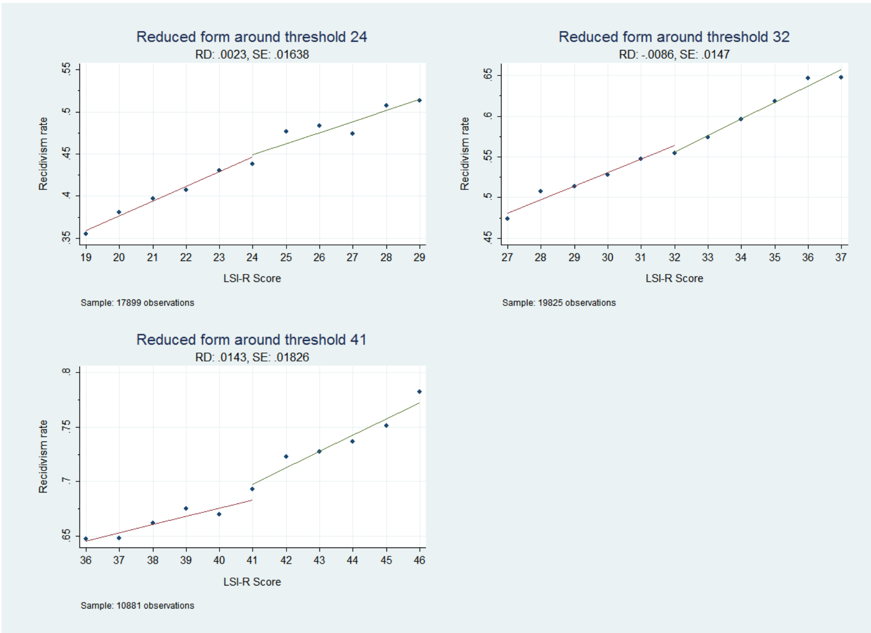


Figure 2.7: Reduced form for the three thresholds and “Recidivism” as the outcome variable using the manipulated data. The Figure presents the effect of the LSI-R score on recidivism for each one of the three thresholds.

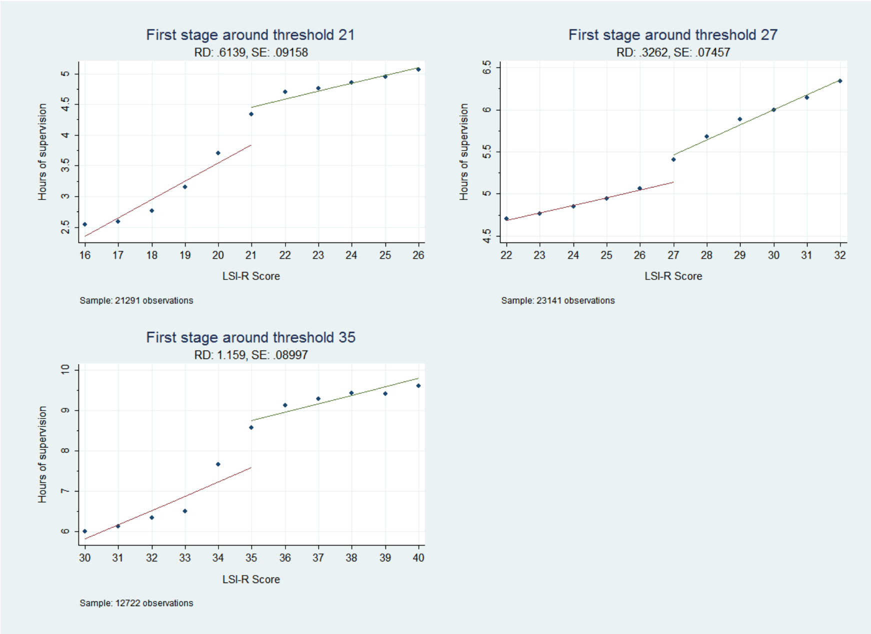


Figure 2.8: First stage for the three thresholds using the corrected data. The Figure presents the effect of the LSI-R score on the hours of supervision for each one of the three thresholds.

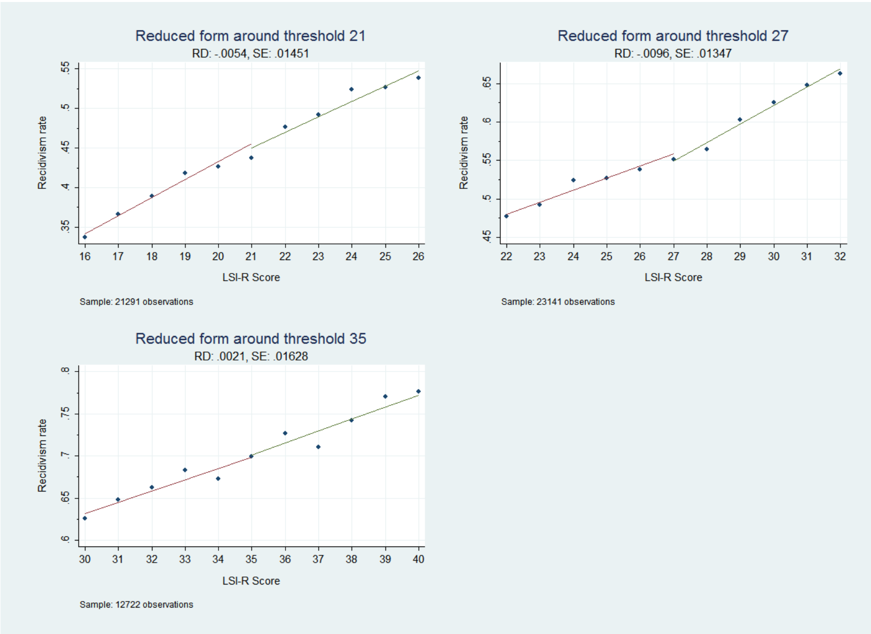


Figure 2.9: Reduced form for the three thresholds and "Recidivism" as the outcome variable using the corrected data. The Figure presents the effect of the LSI-R score on recidivism for each one of the three thresholds.

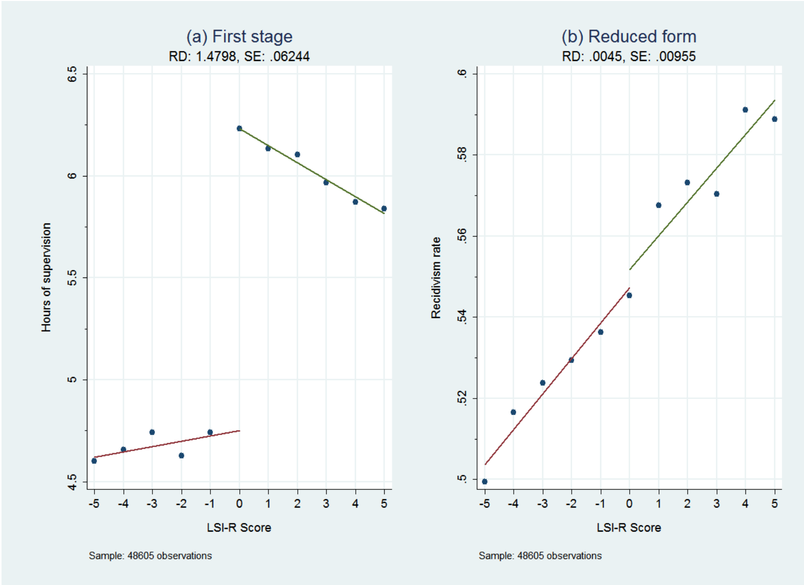


Figure 2.10: First stage and reduced form for the combined-threshold analysis using the manipulated data. For this Figure all the observations around the three thresholds have been pooled together to produce a single data set. In this manner, we can see the first stage and the reduced form combined and not split in three different figures.

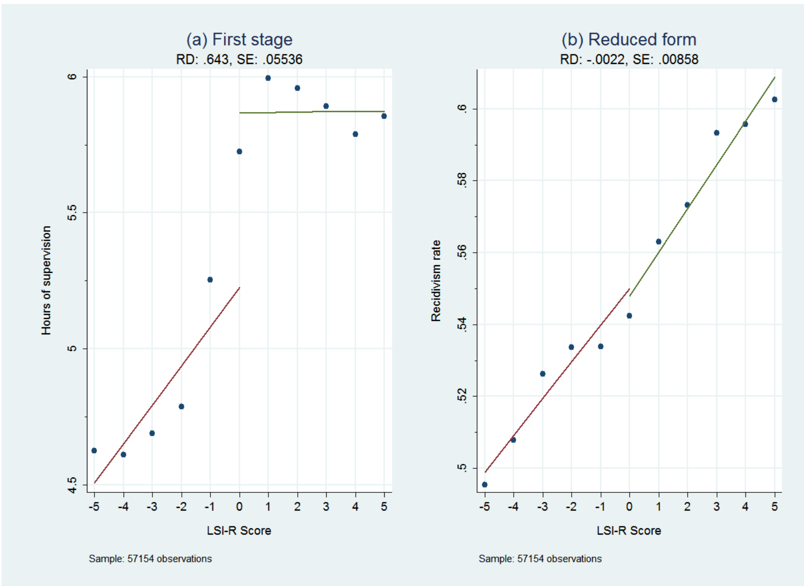


Figure 2.11: First stage and reduced form for the combined-threshold analysis using the corrected data. Everything said in the note to Figure 2.10 applies here too.

Table 2.1: Similarity of sample characteristics around thresholds

| 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 | -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] |
| Homicide | -0.0004 | 0.0015 | 0.0004 |
| | [0.0022] | [0.0014] | [0.0010] |
| Prison | -0.0095 | 0.0475*** | -0.0080 |
| | [0.0120] | [0.0116] | [0.0166] |

Standard errors in brackets

Note: This Table shows that the composition of the sample is generally similar around the three thresholds. 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 the RD 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 we use only scores that are five points below the respective threshold and six points above it (including the threshold itself).

Table 2.2: Composition of the data set in percentages

| 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 | 2.78 | 3.56 | 3.59 | 3.05 | 1.72 |
| Asian | 2.71 | 3.30 | 1.98 | 1.86 | 3.75 |
| Age 18-30 | 47.64 | 47.96 | 45.74 | 48.67 | 52.16 |
| Age 31-45 | 40.66 | 39.72 | 42.95 | 41.82 | 36.34 |
| Age over 45 | 11.59 | 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 |
| Observations | 51,957 | 7,902 | 8,126 | 19,008 | 16,831 |

Table 2.3: Recidivism 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 2.4: 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 | 17899 | 19825 | 10881 | 21291 | 23141 | 12722 |

Standard errors in brackets

*Note:* This Table presents the first stage, reduced form and IV estimates for each one 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 one of the three thresholds. 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 we use only scores that are five points below the respective threshold and six points above it (including the threshold itself).

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

Table 2.5: 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 | 48605 | 57154 |

Standard errors in brackets
*Note:* This Table is equivalent to Table 2.4 but now we combine the three thresholds in one. The estimates show the effect of increased hours of supervision if an offender had a LSI-R score that was greater than *any* threshold. The observations that were used for each of the three threshold-specific regressions (five points below each threshold and six above) were stacked in a single data set for the manipulated scores and another data set for the corrected scores. In the first column we 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.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

# Part III

# Third part

# Chapter 3

# Do correctional authorities treat all offenders equally? The case of a risk assessment instrument

## 3.1 Introduction

The issue of discriminating practices in the course of law enforcement procedures is not new. It has actually attracted considerable attention since Becker (1971) introduced economics into the field of discrimination. Moreover, in the last ten years the topic has been revisited repeatedly with a special focus on police searches. This Chapter examines the possibility of discriminatory behavior in the criminal justice system, notably, the system of assessing offenders' risk.

In Chapter 1, I showed that the risk instrument used by the Washington State's authorities to measure offenders' risk was subject to manipulation. The manipulation consisted in artificially inflating some offenders' risk score so that they receive more hours of supervision in the community than they would otherwise receive.

In this Chapter, I will answer the question whether this manipulation targeted some population groups more than others. Specifically, comparisons will be made between different gender, race or age groups as well as between groups of offenders that committed different crimes, such as drug versus non-drug offenders, property versus non-property offenders or offenders that were released from prison as opposed to those that went straight to the community.

However, to answer this question, it is not enough to compare the rates at which different groups appear to be subject to manipulation, something that is known in the literature as a "benchmarking test" (Persico and Todd, 2006). Alternative "outcomes-based tests" have been proposed in an effort to unravel the true motivation behind a seemingly unfair treatment.

Specifically, it is argued that, occasionally, even though a government policy violates fairness criteria, it meets other superior overarching objectives, in which case the violation is justified (Durlauf, 2006). The burden of proof lies with the proponent of the fairness violation. Such a consideration has to be taken into account when examining whether different offenders were treated differently. In the context of discriminatory behavior what needs to be determined is whether any potential inequality in the behavior of the authorities can be explained by some "legitimate" personal characteristic or can be attributed to characteristics that are not "properly relevant" (Arrow, 1973), such as race, age, etc. The former case is known as "statistical discrimination" whereas the latter as "preference or prejudice-based discrimination" (Antonovics and Knight, 2004; Dharmapala and Ross, 2003).

Following the literature initiated by Knowles, Persico, and Todd (2001), I try to identify whether any potential differential treatment of different groups can be attributed to inherent differences in their behavior, in which case the differential treatment can be sufficiently justified. In the case of this Chapter, such behavior that could possibly justify a different treatment can only be the recidivism rates observed for various groups. Therefore, recidivism is the outcome that the analysis will be based on and the objective is to verify whether what looks like discrimination is actually good practice attributable to the characteristics of the relevant group. If this turns out to be true, the possibility of discrimination can be safely ruled out.

The idea introduced by Knowles, Persico, and Todd (2001) is that hit rates, i.e. in their case police searches of motor vehicles that succeed in identifying contraband, should be equal across different groups. The reason is that if the hit rate for one group was higher, then the authors note "police would always search these motorists, who would in turn react by carrying contraband less often, until the returns to searching are equalized across groups" (p. 206). Therefore in their case equality of hit rates in equilibrium would serve as proof of absence of discrimination, even if certain groups are searched more than others.[1]

In other words, what at first sight might seem as a flagrant case of discrimination, a more in-depth analysis revealed to be good and efficient practice on the part of the authorities. Moreover, a low hit rate is taken to be evidence of discrimination against a specific group since the police can easily improve their general success rate by searching less the specific groups and moving their attention toward groups with higher hit rates (Sanga, 2009).

This literature was further refined and expanded in the years that followed. Hernández-Murillo and Knowles (2004) have extended the model by Knowles, Persico, and Todd (2001) to incorporate also identification methods that use aggregate rather than individual level data.[2] Anwar and Fang (2006) and Antonovics and Knight (2004) suggest officer race as a parameter of interest,[3] while Dharmapala and Ross (2003) introduce the possibility of multiple equilibria only one of which is addressed by the standard Knowles, Persico, and Todd (2001) model.

In this Chapter I use a regression discontinuity design to examine whether different groups were affected differently by the manipulation strategy followed by the authorities in Washington State. Further econometric techniques are used to determine differences in recidivism rates among the same groups. The result of my analysis in short is the following: although I find evidence that different groups were affected differently by the manipulation, I cannot say that this amounts to discrimination since these groups posed also different recidivism risks.

[1]It should be noted though that, for the purposes of this Chapter, it is difference in the hit rates rather than equality that serves as proof of non-discrimination. The reasoning will be laid out in detail in Section 3.4.

[2]For this Chapter, I use individual level data so I do not have to resort to this technique.

[3]Unfortunately, I do not have data on personnel characteristics so that I employ this method in my analysis.

The Chapter proceeds as follows. Section 3.2 gives an overview of the factual background that surrounds the use of a risk assessment instrument in Washington State and provides some information on the mechanics of the instrument's administration and how this fits into the state's post-release supervision system. Section 3.3 gives a brief description of the data sets used, their composition and their source. Section 3.4 outlines the econometric techniques that are employed in the Chapter . Section 3.5 presents and discusses the results, and Section 3.6 concludes.

## 3.2 Factual Background

Following the enactment of the Offender Accountability Act ("OAA") in 1999, Washington State started using the actuarial instrument called "Level of Service Inventory - Revised" or "LSI-R" to measure the risk that offenders released to the community pose and to allocate accordingly hours of supervision.[4] The OAA provided for the establishment of the "Risk Management Identification" ("RMI") system (Aos, 2003). According to this system four supervision categories were created that would largely reflect the offenders' risk of reoffending as measured by the risk assessment instrument, LSI-R.[5]

The LSI-R instrument was described in detail in Chapter 1. Here we can say that it is administered by way of a questionnaire consisting of 54 items that assigns to offenders scores ranging from 0 to 54. Each item represents a risk factor that if present contributes one point in the calculation of the score and if absent contributes zero points. The LSI-R score of an offender is simply the sum of the 1s recorded. Obviously, higher scores reflect higher risk and lower scores, lower risk. Along the 54-point scale there are three cut-off points that separate the four different supervision categories. If an offender's score exceeds a cut-off point, he/she gets upgraded to the next level of supervision.[6]

[4]The use of the LSI-R actually started prior to the OAA coming into force, as the authorities had anticipated its enactment (Barnoski and Aos, 2003; Aos and Barnoski, 2005)

[5]The supervision categories are called RMA, RMB, RMC and RMD, the first corresponding to the highest risk and the last to the lowest risk.

[6]It should be mentioned that according to the OAA the "harm done" by an offender also plays a role in his/her assignment to a supervision category (Aos, 2002; Aos and Barnoski, 2005). However, evidence suggests that crossing a cut-off point is strongly -albeit not perfectly-

The offenders that take the LSI-R questionnaire and are assigned to supervision categories are either released from prison or sentenced straight to the community. The questionnaire is administered at random intervals but also close to the time of release to the community so that the appropriate assignment can be made.

The problem that I came up with the administration of the LSI-R in Washington was that, even though I expected a normal-like distribution of scores across the population of offenders, with the bulk of offenders having scores in the middle of the scale and fewer offenders in the extremes, I obtained Figure 3.1. The three jumps observed are exactly on the cut off scores. I considered this to be evidence of manipulation which I documented in Chapter 1. Specifically, I identified the items of the questionnaire that were mostly used by the authorities to bump offenders up to the next supervision level. Then I removed these items, recalculated the instrument without them and obtained Figure 3.2. This served as proof that manipulation did indeed take place and was focused on particular items of the questionnaire. As indicated above, the objective of this Chapter is to show whether the offenders that were the subject of the manipulation and were bumped up to the next supervision level, were more likely to be part of a specific demographic or other group.

## 3.3 Data

For this study I used the data set that I used in Chapter 1. This is a data set that was put together by combining three separate confidential data sets that were given to me by the "Washington State Institute of Public Policy." The first of them had information on LSI-R scores, the second on assignment of offenders to the four risk categories of the RMI system and the third on demographic, criminal history and recidivism characteristics.

Merging of these data sets produced a data set of 51,957 offenses committed by 47,154 individual offenders. The time span of the analysis is from July 2000 to September 2004. Table 3.1 presents descriptive statistics for gen-

associated with being assigned to the immediately next supervision level (as shown in Chapter 1).

eral characteristics of the sample, such as gender, race, age, type of crime, type of sentence. Table 3.2 presents specific information on recidivism. For additional clarity this information is also broken down per risk category.

## 3.4 Empirical Methodology

The object of this Chapter is to examine whether the bumping up strategy undertaken by the authorities had the same impact across different groups of offenders. Specifically, my strategy for identifying differential treatment unfolds in two stages: (a) I want to see if different groups were equally likely to be bumped up, and (b) I want to see if any difference in the probability to be bumped up can be explained by a different recidivism probability, i.e. a different level of risk posed by the respective group.

### 3.4.1 First Task

#### Regression Discontinuity

To accomplish the first task, I use two approaches. The first is regression discontinuity. Specifically, I split the sample by characteristic (e.g. white offenders, Black offenders, male offenders, female offenders, etc. and obtain the distribution of the LSI-R score across the generated characteristic-specific group. For each group the distribution is presented as counts of offenses associated with a given LSI-R score for the offenders that committed them.

For example, my first pair to be compared is white against Black offenders. From the micro data, which is the data set of 51,957 offenses mentioned above, I find how many offenses committed by white offenders correspond to an LSI-R score of 0, how many to 1, and so on up to 54. I do the same for Black offenders. Having this information, I can generate a data set with aggregate data, specifically only 55 observations. Each observation corresponds to one of the 55 points of the LSI-R score (from 0 to 54). To each one of those scores I can associate the counts of offenses committed by white and Black offenders. These counts of offenses for each groups are the dependent variable in the model that is presented below. I perform the same task for all the other comparison pairs in my analysis (male vs. female, etc.).

The next step follows the logic that was used in Chapter 1. A regression discontinuity design is employed in order to identify the size of the manipulation for different groups. By focusing on the part of the distribution around each of the three thresholds, I obtain the regression discontinuity estimates for each of the groups and compare their size and significance. I use 5 points below each threshold and 6 above it. For example, for threshold 24, I use the scores 19-23 and 24-29 and by a linear extrapolation of the values of the counts for scores 19-23 I obtain a counterfactual value for score 24, which will then be compared with the actual value for score 24 by way of the regression discontinuity design.

The last step is to run a regression in which the dependent variable is the counts of offenses for each score around the relevant threshold ($CountsOffenses_i$) and the right hand side variable of interest is the regression discontinuity dummy. This variable takes the value 1 if the offender has a score greater than or equal to the threshold we are examining and 0 otherwise. For example, for threshold 24, the regression discontinuity dummy has the form: $rd24_i = 1\{score_i \geq 24\}$ (similarly for thresholds 32 and 41). A linear polynomial of the risk score concludes the list of right hand side variables. The regressions take the following form:

$$CountsOffenses_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i. \tag{3.1}$$

Similarly, for threshold 32, the regression equation has the form:

$$CountsOffenses_i = \beta_0 + \beta_1 rd32_i + \beta_2(score_i - 32) + \beta_3 rd32_i(score_i - 32) + \epsilon_i. \tag{3.2}$$

And finally, for threshold 41 the form is:

$$CountsOffenses_i = \beta_0 + \beta_1 rd41_i + \beta_2(score_i - 41) + \beta_3 rd41_i(score_i - 41) + \epsilon_i. \tag{3.3}$$

To be even more precise, I turn to the first pair to be compared, white vs. Black offenders. The two regressions for threshold 24 are the following:

$$CountsOffensesWhite_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i, \tag{3.4}$$

and,

$$CountsOffensesBlack_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i. \tag{3.5}$$

The estimation method is OLS. The coefficient of the regression discontinuity dummy, $\hat{\beta}_1$, will measure the difference between the line of best fit for scores 19-24 (but, as already mentioned, score 24 here is a counterfactual value that has been generated artificially by way of a linear extrapolation) and the line of best fit for scores 24-29. Differences in the size and significance of the RD estimates will indicate unequal treatment in the bumping up process.

However, since different groups appear in the sample with different frequencies, a comparison of the raw size of the RD estimates would be misleading as a sign of unequal treatment. For example, white offenders are 72 percent of the sample, while Black ones are only 15 percent. Therefore, one would naturally expect the RD estimate to be larger for the group of whites than for the group of Blacks. To deal with this problem, I divide the RD estimates by the constant of the respective regression, $\hat{\beta}_0$, to make them comparable across regressions. The expression $\frac{\hat{\beta}_1}{\hat{\beta}_0}$ shows us the size of the jump as a percentage of the constant of each regression. Therefore, now I am in a position to compare directly the jumps for different groups.

**Direct Comparison across Groups**

An alternative way to determine differences in the bumping up process as well as recidivism rates is to make use of the two distribution of LSI-R scores that I have at my disposal, namely the original manipulated scores and the corrected scores that I generated after removing the manipulated items.[7] The first manipulated distribution is a grid that runs from 0 to 54. The items that

[7]The process of uncovering the items that were subject to manipulation was described in detail in Chapter 1 and is not repeated here.

I identified as mostly used by the authorities were seven, therefore the new corrected distribution runs from 0 to 47.

To identify differential treatment I measure the number of offenses for each of the characteristic-specific subsamples that have been given the threshold scores. For example, I measure how many offenses committed by white offenders were associated with a score of 24 according to the manipulated scale, then do the same for offenses committed by Black offenders and then find the ratio of the two. Then repeat the process but using the respective threshold of the corrected scale. The ratio of Black to white is considered to be accurate in the corrected scale since all manipulation has been removed from it. If the difference between the ratio in the manipulated scale and the ratio in the corrected scale is statistically significant, I can conclude that the bump up was associated more with a given group. The difference in the ratios is measured again by way of the Pearson $\chi^2$ test, which is testing the null that the ratios are the same. For example, if the ratio of Black to white is higher in the manipulated scale compared to the corrected one, then I conclude that that the manipulation targeted Black offenders more.

### 3.4.2 Second Task

For my second task I need to determine if any potential inequality in the bumping up probability can be explained by a specific outcome that may justify the behavior of the authorities.[8] In the case of this Chapter, the only such outcome that I can use is each group's recidivism rate, since the offenders have already been convicted.

However, even though I follow the logic of the literature initiated by Knowles, Persico, and Todd (2001), I cannot apply directly here the rational choice model proposed.[9] The reason is that I do not believe that the offenders in my sample will adjust their recidivism behavior simply because they know that they are more likely to be bumped up to a higher supervision category.[10]

[8] See above in Section 3.1 for the logic and nature of outcomes-based tests.

[9] As indicated above, according to their model, if one demographic group yields a higher return to search, then the police will search these motorists exclusively. In turn, they best respond by lowering their criminal behavior thus equalizing the returns across groups.

[10] Supervision is mostly administered in an organized and predetermined fashion for certain

Therefore, in the case of this Chapter, I do not consider equality of recidivism rates as a sign of absence of bias, but rather the reverse. In other words, I conclude against discrimination if differences in recidivism rates go in the same direction as differences in the being bumped up probability.

I measure recidivism rates only for offenders located exactly on the three thresholds, 24, 32 and 41 by using all the observations for offenders that were given these scores.[11] The method I use to determine whether the recidivism rates across different groups are the same is the Pearson $\chi^2$ test. This allows me to test nonparametrically the null hypothesis that different groups recidivate at equal rates.

## 3.5 Results

Table 3.3 provides the results of the regressions following the specification in equation (3.1), using all the observations in the sample. I try to identify differences in the treatment of different demographic groups (race, gender and age groups) as well as differences between groups that committed different crimes (drug vs. non-drug, property vs. non-property) or received a different type of sentence (prison vs. community). The results on the size of the RD estimates are made comparable across different subsamples by dividing the RD estimates by the size of the constant in each subsample, thus obtaining a result in percentage terms. It is these adjusted estimates and their standard errors that are reported in Table 3.3. By way of example, in Figure 3.3 I also present graphically the jumps on threshold 24 for white and Black offenders, while in Figure 3.4 the jump for male and female offenders. Moreover, in Table 3.4 I present the results obtained by using the methodology presented in Section 3.4.1.

hours per week for each offender and cannot be compared with motor vehicle searches conducted by the police which are random and unpredictable. In other words, more intense supervision is not enough of an incentive to force offenders to alter their recidivism decisions.

[11]However, ideally I would like to measure recidivism rates not for all offenders that were given these scores but only for those offenders who were bumped up, i.e. they actually deserved a lower score. I identified those offenders by using both the original manipulated LSI-R scores and the corrected scores that I generated after removing the manipulated items. Specifically, I can identify those offenders that are given exactly the threshold score on the manipulated instrument but have a lower score on the corrected instrument. In Table D.1 of Appendix D, I performed such an analysis, which turns out to yield quite similar results to the ones obtained without any such filtering.

In Tables 3.6 and 3.8 I repeat this exercise using only observations that refer to specific offenses, drug offenses in Table 3.6 and assault offenses in Table 3.8. I do this only for the three demographic categories (gender, race and age). This approach allows me to control for the fact that different demographic groups might engage in different kinds of criminal activity. In such a case different bumping up patterns across demographic groups could be the result of the different nature of their criminal activity. By focusing on a specific crime, I significantly reduce the possibility of that occurring. I use drug and assault related crimes, the ones with the highest frequency in our sample, as shown in Table 1.1.

Finally in Tables 3.5, 3.7 and 3.9 I present the recidivism rates for offenders of different groups assigned exactly the three threshold scores, 24, 32 and 41. As noted in Section 3.4, this part of the analysis implements an outcomes-based test that could explain any differences observed across various groups in the bumping up process.

Before analyzing the results, it should be noted that, as shown in in Table 1.1, whites are the highest frequency group (72 percent of the sample), while Blacks come second (15 percent).[12] Table 3.3 shows that for threshold 24 Blacks have the greatest chance of being bumped up (50 percent) followed by whites (35 percent). This result is also presented graphically in Figure 3.3. Estimates for whites and Blacks are strongly significant. For threshold 32, Blacks continue to lead (21 percent) followed by whites with only a 14 percent chance. Finally, for threshold 41, whites have a 45 percent chance while Blacks now have a 73 percent chance.

Turning to Table 3.5, which will qualify the previous results, it is apparent that for thresholds 24 and 32, Blacks have consistently higher recidivism probability than whites. For threshold 41 their recidivism probability is still higher but the difference with the whites is marginally not significant at the 10 percent level. Therefore, with the exclusion of threshold 41, Table 3.5 gives a preliminary answer as to why, according to the regression discontinuity design,

[12] Offenders of Hispanic origin (7 percent of the sample) as well as Native American and Asian origin (both less than 3 percent of the sample) are very few and we have excluded them from the analysis as any obtained results would not be reliable.

Blacks have a higher probability of being bumped up than whites: it is because they are more likely to recidivate.[13]

Continuing to Tables 3.6 and 3.8, I get similar results as in Table 3.3. Excluding threshold 32 which is not significant for both groups, thresholds 24 and 41 show a clear preponderance of the Blacks in the probability of being bumped up, both for drug and assault offenses. That probability is two or three times higher depending on the threshold. Only for threshold 41 in the assault subsample is the estimate for Blacks not significant, although very high.

However, Tables 3.7 and 3.9 send more mixed signals. Both Tables indicate that even though for thresholds 24 and 32, Blacks continue to differ significantly from whites, threshold 41 shows the opposite picture.[14,15] For threshold 41 the difference in recidivism rates between Blacks and whites is not significant.

Continuing with gender groups, in Table 3.3 we observe that male offenders have a 38 percent chance of being bumped up across threshold 24 compared to almost 32 percent for female offenders.[16] This discrimination against men persists for threshold 32, where we observe that the being bumped up probability is about 16.5 percent and strongly significant for men while it is not even significant for women. For threshold 41 we notice that the situation is even more clear with the probability for men being 58 percent and very significant again while for women 30 percent and marginally significant at the 5 percent level.

Table 3.5 could explain again why male offenders were more likely to be bumped up than women, according to the regression discontinuity design. Men recidivate consistently more than women for all three thresholds. The difference

[13] As indicated above, in Table D.1 of Appendix D, I measure recidivism rates using only the offenders that were bumped up to the threshold scores. It is obvious that recidivism rates are quite similar with those presented in Table 3.5. Therefore, without loss of accuracy, I am using Table 3.5 as a source of information on recidivism rates, since the number of observations is much higher and the results more reliable.

[14] On the other hand, of course, threshold 32 does not present significant differences in the bumping up probabilities, as indicated above.

[15] However, it should be mentioned that the difference for threshold 32 in the assault case is marginally significant at the 10 percent level.

[16] The fact that male offenders constitute 77 percent of the sample rendering the two groups unequal in terms of their number is not a source of concern for Table 3.3 since we observe a sufficient number of women having the threshold scores. The problem resurfaces when the analysis focuses on specific crimes, notably assault.

is between 6 and 8 percent and statistically significant at the 5 percent level.

Tables 3.6 and 3.8 confirm this pattern. Both for drug and assault offenses, men seem to be discriminated against in the case of threshold 24 but more heavily in the case of threshold 41. Threshold 32 presents not significant estimates for both groups. Recidivism rates remain higher for men, although some times the difference is not significant.[17]

The final potentially discriminating demographic criterion is age at the time of the LSI-R interview.[18] For threshold 24 we note that the younger group has the lead with a 37 percent being bumped up probability followed by the middle group (32 percent). The order is the same for threshold 32, where the younger group has the highest chance (20 percent) followed by the middle group (12 percent). The order remains the same for threshold 41. The probability for young people is 69 percent followed by the middle group (35 percent). In Table 3.5 we notice that the younger group recidivates more followed by the middle group for thresholds 32 and 41, but less than the middle group for threshold 24.

Additionally, when we turn to Tables 3.6 and 3.8 we observe a similar pattern. Again the younger group has a significantly higher probability than the middle group of being bumped up, which seems to be justified by a higher recidivism rate.

Furthermore, for the pairs drug vs. non-drug and property vs. non-property crimes, we obtain a picture that does not offer much evidence in favor of differential bumping up of specific groups. In Table 3.3 we notice that the RD/Constant estimates are similar for both pairs, with the first and third threshold displaying similar percentages, while the second threshold shows a preponderance of the non-drug, non-property groups to be bumped up compared to the non-significant percentage of the drug and property groups respectively. In Table 3.5 we observe that for the first threshold the similar bumping up rates are reflected by similar recidivism rates. For threshold 41, the drug pair has a similar recidivism rate, while the property pair a significantly different rate with the property offenders recidivating more even though they are bumped up

[17]The results for assault in Table 3.9 should be read with caution given the limited number of women who commit assaults.

[18]Out of the three age groups, young, middle age and old, I have eliminated the older group because of lack of a sufficient number of observations.

less. For threshold 32, we note significantly different recidivism rates which, though, coincide with the direction of the bumping up only for the drug pair. For the property pair, the property offenders recidivate much more but we did not observe a statistically significant jump for them. Therefore, the results that we get for these two pairs of offenses do not seem to comply with some organized behavior of the authorities and cannot be explained by recidivism rates. The authorities did not systematically differentiate offenders based on these two crime types, drug and property crimes.

Finally, the prison vs. community pair does not support the hypothesis in favor of discrimination. In Table 3.3 we note that for thresholds 24 and 41, community offenders are bumped up relatively more than offenders that are being released from prison, while for threshold 32 we note the reverse with a large difference in favor of ex-prisoners. However, Table 3.5 shows that the recidivism rates of the two groups are very similar if not identical for all three thresholds. This would mean that for thresholds 24 and 41 community offenders were victims of discriminating behavior, while for threshold 32 the victims were ex-prisoners. Therefore, again we do not observe an organized behavior on the part of the authorities and, as a result, the differences in the bumping up probability cannot be attributed to discrimination.

Moreover, it must be mentioned that the method presented in Section 3.4.1 gives us more mixed signals about whether there was actually differential treatment in the bumping up process. In Table 3.4 we note that for virtually all six sections, the low and the middle threshold present conflicting results on whether the manipulated scores show evidence of targeting a specific group. Specifically, for the male/female, young/middle age, drug/non-drug, property/non-property ratios, what we observe for the low threshold is exactly the opposite of what we observe for the middle threshold. Additionally, for the high threshold and for all six sections, the difference between the two ratios is not significant. This picture indicates that the authorities did not target a specific demographic or crime category in contrast with what the regression discontinuity design tells us.

In view of these results and given my assumptions about how to interpret differences in recidivism rates, the following conclusions apply. When we use the regression discontinuity design, the preliminary result, which ignores differences in recidivism behavior across groups, indicates that for virtually all three thresholds, Blacks are more likely to be bumped up because they are more likely to recidivate.[19] This precludes racial discrimination. The situation is similar when we turn to gender differences. Men are bumped up more for all three thresholds but this is again explained by their higher recidivism rates. With respect to age differences, we observe a similar picture with younger offenders being bumped up more but again due to their higher tendency to recidivate.[20] Finally, we noted that the pairs drug vs. non-drug, property vs. non-property and prison vs. community did not provide us with evidence that the authorities targeted a specific type of crime or type of sentence.

When we turned to specific offenses, drug crimes and assault crimes, we noted that for race, gender and age differences we got a mixed picture that does not support the possibility of discriminating behavior on the part of the authorities. In fact, in the case of race groups e.g. we get conflicting evidence that would force us to believe that for one threshold there was discrimination against whites (threshold 32) while for another threshold there was discrimination against Blacks.[21]

Moreover, when I used the method of direct comparison across groups as presented in Table 3.4, I did not even find evidence in favor of differential bumping up in the first place. In summary, either I use regression discontinuity or the method of direct comparison, I cannot detect an organized attempt on the part of the authorities to target a particular demographic or criminal group, and I therefore exclude the possibility of discrimination.

[19]The marginally non-significant difference for threshold 41 should be noted at this point.

[20]The exception of threshold 24 should be noted here, which seems to defy the pattern observed, as older offenders are bumped up more.

[21]For threshold 32 we do not have higher probability of being bumped up for Blacks or whites, but Blacks continue to recidivate significantly more. This tells us that even though they should have been bumped up more, they actually were not. This would be an indication of discrimination against whites. For threshold 41 Blacks have a higher probability of being bumped up but they are not more likely to recidivate than whites. This would be an indication of discrimination against Blacks.

## 3.6 Conclusion

In this Chapter, I attempted to uncover whether the authorities in Washington State targeted certain demographic or other groups when they were making their decision to bump offenders up to a higher supervision level. By using a regression discontinuity design as well as other methods, I tested whether specific demographic or other groups were placed unusually more exactly on the cut-off points of a risk assessment instrument. Although the regression discontinuity design showed me that certain groups were more likely than others to be bumped up, other methods indicated that this was not the case.

However, even if certain groups were discriminated against, this result has to be qualified by taking into account the probability of recidivism by the respective groups. For most of the cases where unequal bumping up behavior was identified, it can be explained by similar unequal recidivism probability. Therefore, what a preliminary level of analysis would classify as discriminating behavior on the part of the authorities, turns out to be just exercising good sense.

The conclusion is that the artificial inflation of the LSI-R scores may not be explained by the use of discriminating logic on the part of the authorities. Higher recidivism rates is almost always behind a decision to inflate. It can be argued that this amounts to converting an actuarial instrument to a vehicle that camouflages clinical judgment decisions. However, it was not the object of this Chapter to substantiate such a statement. This Chapter just showed that discriminating or targeting logic based on criteria that are not "properly relevant" did not influence the bumping up decisions of Washington State's authorities.

## 3.7 Figures and Tables

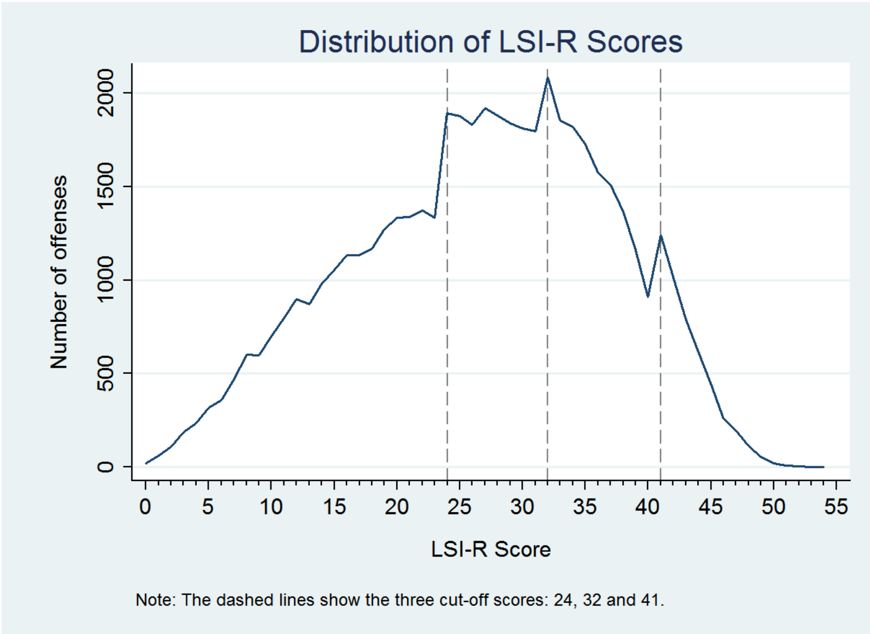


Figure 3.1: Distribution of LSI-R Scores. The LSI-R score was the risk instrument used by Washington State to measure the risk of offenders. It is a scale from 0 to 54, where higher numbers indicate higher risk.

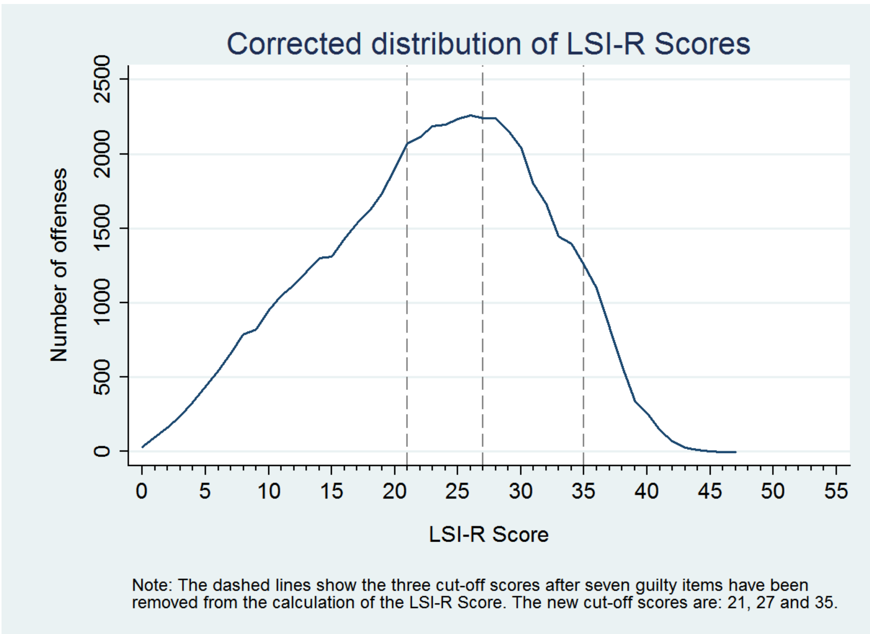


Figure 3.2: Distribution of LSI-R Scores after omitting the manipulated items. Without the seven manipulated items, the LSI-R index becomes a scale from 0 to 47.

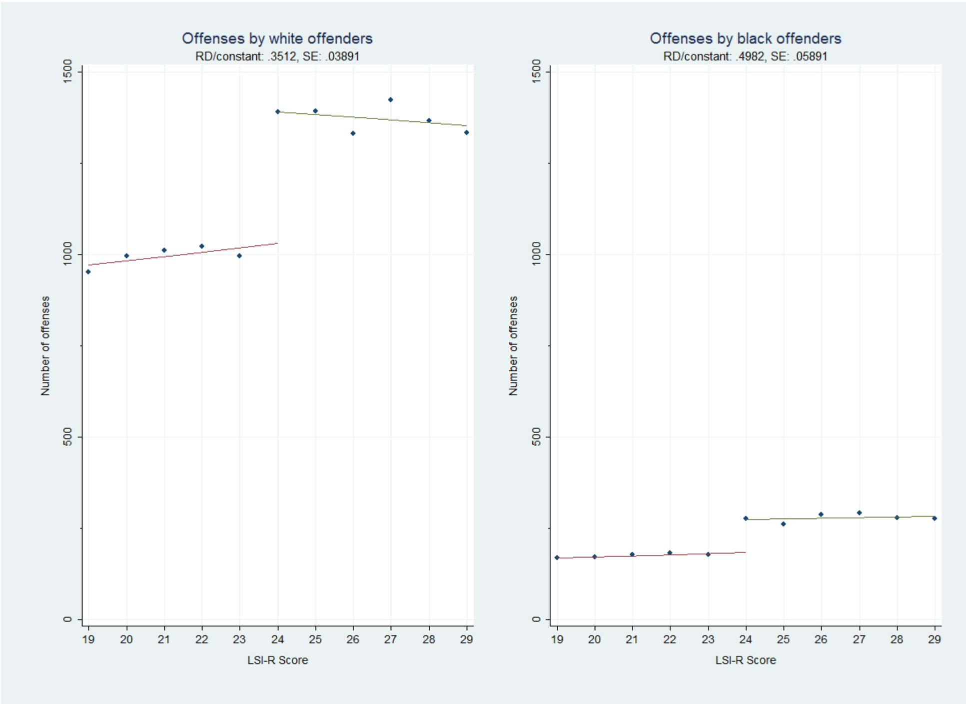


Figure 3.3: Jumps of white and Black offenders around threshold 24

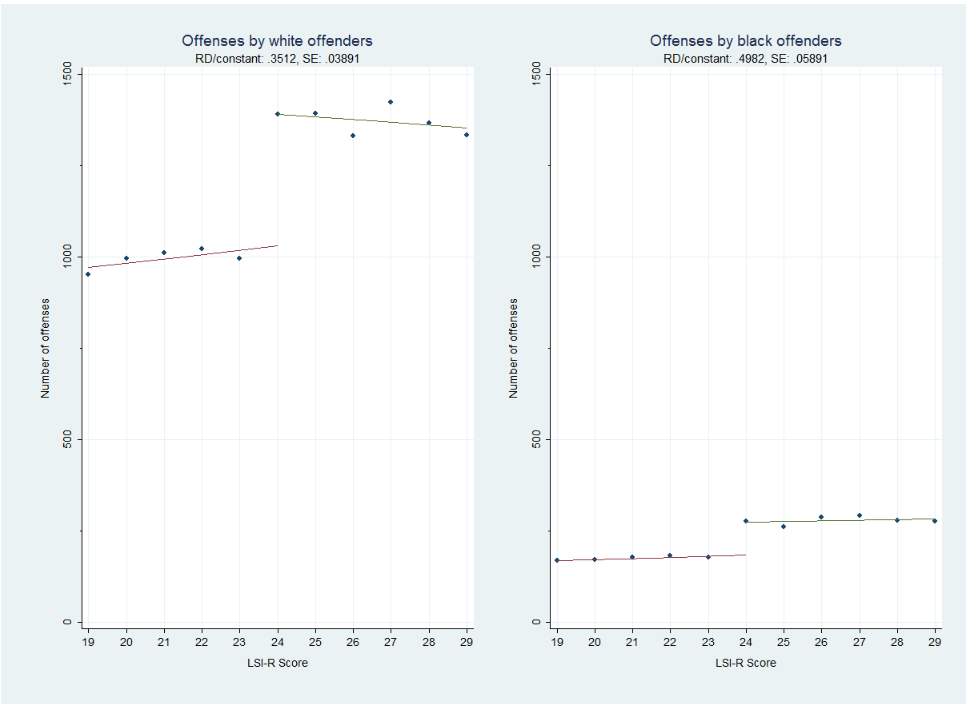


Figure 3.4: Jumps of male and female offenders around threshold 24

Table 3.1: Composition of the data set in percentages

| 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 | 2.78 | 3.56 | 3.59 | 3.05 | 1.72 |
| Asian | 2.71 | 3.30 | 1.98 | 1.86 | 3.75 |
| Age 18-30 | 47.64 | 47.96 | 45.74 | 48.67 | 52.16 |
| Age 31-45 | 40.66 | 39.72 | 42.95 | 41.82 | 36.34 |
| Age over 45 | 11.59 | 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 |
| Observations | 51,957 | 7,902 | 8,126 | 19,008 | 16,831 |

Table 3.2: Recidivism 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 3.3: Identifying differential treatment for different groups of offenders

| Size of RD/Constant estimates for all three thresholds | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.3512 | 0.1373 | 0.4458 |
| | (0.03891) | (0.0298) | (0.15481) |
| Black | 0.4982 | 0.2085 | 0.7317 |
| | (0.05891) | (0.05398) | (0.16761) |
| Male | 0.3808 | 0.1647 | 0.58 |
| | (0.03849) | (0.03809) | (0.1326) |
| Female | 0.3201 | 0.1339 | 0.307 |
| | (0.04384) | (0.10204) | (0.156) |
| Age 18-30 | 0.372 | 0.1967 | 0.6854 |
| | (0.0578) | (0.03333) | (0.16255) |
| Age 31-45 | 0.318 | 0.1179 | 0.3479 |
| | (0.0384) | (0.05243) | (0.16934) |
| Drug | 0.3823 | 0.0504 | 0.4923 |
| | (0.0544) | (0.0437) | (0.1542) |
| Non-Drug | 0.3593 | 0.2224 | 0.5255 |
| | (0.0394) | (0.0381) | (0.1475) |
| Property | 0.3222 | 0.0609 | 0.4945 |
| | (0.0745) | (0.0246) | (0.1393) |
| Non-Property | 0.3829 | 0.1953 | 0.5212 |
| | (0.0288) | (0.0396) | (0.1443) |
| Prison | 0.3093 | 0.4557 | 0.4136 |
| | (0.0779) | (0.0443) | (0.1256) |
| Community | 0.3791 | 0.0921 | 0.5480 |
| | (0.0380) | (0.0316) | (0.1456) |

Note: Standard errors in parentheses.

Table 3.4: Identifying differential treatment (Direct comparison across groups)

| Size of ratios between groups using manipulated (Man.) and corrected (Cor.) scores | | | |
|---|---|---|---|
| Group of offenders | Low Threshold | Middle Threshold | High Threshold |
| Man. (Black/White) | 0.143 | 0.175 | 0.301 |
| Cor. (Black/White) | 0.207 | 0.226 | 0.266 |
| p-value | 0.000 | 0.005 | 0.227 |
| Man. (Male/Female) | 4.458 | 1.633 | 3.809 |
| Cor. (Male/Female) | 3.519 | 3.913 | 3.621 |
| p-value | 0.000 | 0.000 | 0.639 |
| Man. (Young/MidAge) | 2.072 | 0.729 | 1.347 |
| Cor. (Young/MidAge) | 1.247 | 1.16 | 1.243 |
| p-value | 0.000 | 0.000 | 0.386 |
| Man. (Drug/Non-Drug) | 0.344 | 0.903 | 0.475 |
| Cor. (Drug/Non-Drug) | 0.528 | 0.515 | 0.48 |
| p-value | 0.000 | 0.000 | 0.921 |
| Man. (Property/Non-Property) | 0.752 | 0.257 | 0.432 |
| Cor. (Property/Non-Property) | 0.329 | 0.403 | 0.409 |
| p-value | 0.000 | 0.000 | 0.578 |
| Man. (Prison/Community) | 0.19 | 0.278 | 0.397 |
| Cor. (Prison/Community) | 0.283 | 0.326 | 0.425 |
| p-value | 0.000 | 0.034 | 0.483 |

Note: Last row shows p-values from Pearson $\chi^2$ tests of null hypothesis that the ratio of the first group to the second in each section and for a given column are equal. For example, the top left section says that the ratio of Black to White for offenders scoring on the low threshold was 0.14 using the manipulated data and 0.21 using the corrected data. The p-value that these ratios are the same is 0.000.

Table 3.5: Recidivism rates for offenders on the thresholds

| Recidivism rate (total number of observations in parenthesis) | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.41 (1,389) | 0.53 (1,478) | 0.68 (833) |
| Black | 0.57 (276) | 0.66 (342) | 0.73 (269) |
| p-value | 0.000 | 0.000 | 0.112 |
| Male | 0.45 (1,466) | 0.57 (1,638) | 0.71 (983) |
| Female | 0.39 (427) | 0.51 (447) | 0.63 (260) |
| p-value | 0.015 | 0.041 | 0.022 |
| Age 18-30 | 0.30 (1,235) | 0.61 (833) | 0.76 (543) |
| Age 31-45 | 0.46 (596) | 0.34 (1,142) | 0.67 (403) |
| p-value | 0.000 | 0.000 | 0.004 |
| Drug | 0.43 (616) | 0.51 (701) | 0.67 (417) |
| Non-Drug | 0.44 (1277) | 0.58 (1,384) | 0.70 (826) |
| p-value | 0.785 | 0.002 | 0.307 |
| Property | 0.47 (500) | 0.67 (517) | 0.76 (352) |
| Non-Property | 0.43 (1,393) | 0.52 (1,568) | 0.67 (891) |
| p-value | 0.073 | 0.000 | 0.001 |
| Prison | 0.44 (324) | 0.58 (466) | 0.70 (294) |
| Community | 0.44 (1,569) | 0.55 (1,619) | 0.69 (949) |
| p-value | 0.989 | 0.156 | 0.845 |

Note: Last row shows p-values from Pearson $\chi^2$ tests of null hypothesis that recidivism rates in each section of a given column are equal.

Table 3.6: Identifying differential treatment for drug related offenses

| Size of RD/Constant estimates for all three thresholds | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.36391 | 0.0744 | 0.3076 |
| | (0.04717) | (0.06491) | (0.149099) |
| Black | 0.6137 | 0.1080 | 0.9574 |
| | (0.21941) | (0.08991) | (0.08632) |
| Male | 0.3864 | 0.0356 | 0.5794 |
| | (0.052719) | (0.03114) | (0.176303) |
| Female | 0.3721 | 0.0856 | 0.3002 |
| | (0.08187) | (0.15249) | (0.15410) |
| Age 18-30 | 0.3745 | 0.0589 | 0.5117 |
| | (0.10366) | (0.08668) | (0.20103) |
| Age 31-45 | 0.3430 | 0.0533 | 0.3199 |
| | (0.04436) | (0.07356) | (0.15074) |

Note: Standard errors in parentheses.

Table 3.7: Recidivism rates by race for drug offenders on the thresholds

| Recidivism rate (total number of observations in parenthesis) | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.41 (452) | 0.48 (496) | 0.71 (109) |
| Black | 0.57 (98) | 0.62 (141) | 0.73 (263) |
| p-value | 0.003 | 0.002 | 0.236 |
| Male | 0.46 (443) | 0.54 (497) | 0.70 (300) |
| Female | 0.38 (173) | 0.43 (204) | 0.62 (117) |
| p-value | 0.071 | 0.009 | 0.112 |
| Age 18-30 | 0.47 (244) | 0.56 (239) | 0.72 (159) |
| Age 31-45 | 0.45 (289) | 0.51 (371) | 0.69 (210) |
| p-value | 0.630 | 0.193 | 0.517 |

Note: Last row shows p-values from Pearson $\chi^2$ tests of null hypothesis that recidivism rates in a given column are equal.

Table 3.8: Identifying differential treatment for assault related offenses

| Size of RD/Constant estimates for all three thresholds | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.2964 | 0.5426 | 0.4652 |
| | (0.06698) | (0.32794) | (0.21599) |
| Black | 0.9692 | 0.0904 | 1.4023 |
| | (0.41865) | (0.11584) | (1.37914) |
| Male | 0.4654 | 0.5779 | 0.8993 |
| | (0.06853) | (0.29826) | (0.33734) |
| Female | 0.3826 | 0.1310 | 0.0960 |
| | (0.35221) | (0.24113) | (0.14660) |
| Age 18-30 | 0.4915 | 0.5602 | 1.1876 |
| | (0.12460) | (0.30733) | (0.28744) |
| Age 31-45 | 0.2882 | 0.3399 | 0.4752 |
| | (0.21236) | (0.19314) | (0.46618) |

Note: Standard errors in parentheses.

Table 3.9: Recidivism rates by race for assault offenders on the thresholds

| Recidivism rate (total number of observations in parenthesis) | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.31 (168) | 0.47 (212) | 0.62 (93) |
| Black | 0.52 (52) | 0.59 (54) | 0.66 (47) |
| p-value | 0.006 | 0.099 | 0.677 |
| Male | 0.37 (226) | 0.51 (276) | 0.66 (137) |
| Female | 0.43 (42) | 0.38 (34) | 0.55 (20) |
| p-value | 0.486 | 0.157 | 0.317 |
| Age 18-30 | 0.41 (145) | 0.53 (144) | 0.72 (81) |
| Age 31-45 | 0.44 (78) | 0.52 (131) | 0.57 (63) |
| p-value | 0.675 | 0.795 | 0.071 |

Note: Last row shows p-values from Pearson $\chi^2$ tests of null hypothesis that recidivism rates in a given column are equal.

# Part IV

# Fourth part

# Chapter 4

# New risk assessment instrument based on data from Washington State

## 4.1 Introduction

Before and after Martinson (1974), officials in charge of designing and running treatment programs normally factor the offenders' risk of reoffending into the treatment choice and the possibility of successful rehabilitation. Such is the importance of risk assessment that is sometimes being regarded as the most important parameter of a correctional system (Bonta, Bogue, Crowley, and Motiuk, 2001). The goal of this Chapter is to build on current knowledge on risk instruments and produce an instrument that could be considered an improvement on an already existing one.

Specifically, following the results of Chapter 1 that found evidence of manipulation in the calculation of the risk instrument used in the state of Washington from 1998 to 2008, the object of the present study is to produce a version of the instrument that outperforms the prediction results of the previous instrument. Additionally, the study shows that this improved version of an old instrument produces results that are comparable, albeit slightly inferior, to the results produced by the instrument currently in use in Washington, which was tailored to the specific needs of the state's offender population.

Risk measuring instruments fall within the ambit of the so-called "actuarial" instruments. Actuarial predictions make use of mathematical models that transform factual observations to inputs (data) in a specific equation. The equation in turn provides a prediction of the outcome that is solely based on the data used (Meehl, 1954). An alternative method of arriving to predictions about an outcome is what is known as "clinical judgment," which differs fundamentally from actuarial instruments, in the sense that it does not rely on objective application of data to equations but on subjective assessment of each particular case by specialized professionals. These "clinicians" will pass their "judgment" based on their professional experience and scientific expertise.[1]

In the field of clinical psychology there exists a debate on which of the two methods should be used for prediction making: actuarial or clinical. On the basis of meta-analytic evidence from a wide range of social science studies, it is argued in the literature that actuarial methods are at least as good or better compared to clinical judgment (Meehl, 1954; Grove and Meehl, 1996; Grove, Zald, Lebow, Snitz, and Nelson, 2000). The same argument is made by proponents of actuarial instruments when we focus specifically on the issue of predicting the risk of reoffending (Quinsey, 1995; Bonta, Bogue, Crowley, and Motiuk, 2001). Some even go as far as to argue that actuarial instruments should in no way be complemented by clinical judgment, a situation referred to as "clinical overrides," although others take a more moderate stance and advise caution when such an override is to be made (Andrews and Bonta, 2006).[2] On the other hand, proponents of clinical judgment note that such superiority of actuarial methods has not been proved yet and it is, therefore, premature to replace the former with the latter until a head to head comparison can be conducted (Litwack, 2001).[3]

[1] A presentation of the different generation of prediction methods that have been used historically can be found in Bonta (1996) and Andrews and Bonta (2006). In these accounts clinical judgments is referred to a as a "first generation assessment" methods, while actuarial instruments under their different versions are classified as "second", "third" and, "fourth generation" methods.

[2] Quinsey, Harris, Rice, and Cornier (2006) bluntly note that: "Actuarial methods are too good and clinical judgment is too poor to risk contaminating the former with the latter."

[3] Although it should be noted that Harris, Rice, and Cornier (2002) responded with such a direct comparison and found similar results in favor of actuarial methods.

This Chapter does not intend to take a stance in this debate. Its objective is to touch on a more specific issue that the relevant literature on actuarial instruments and prediction making has dealt with, namely the mechanical set up of these instruments. Specifically, an important question is whether optimal weights should be employed for the items of an instrument or equal weights are good enough, an approach usually referred to in the literature as the Burgess method (Jones, 1996).

In the literature it is often suggested that equal weights or randomly scrambled weights perform sufficiently well, making the use of optimized coefficients unnecessary (Grove and Meehl, 1996). In fact it is argued that even such improper linear models can outperform clinical judgment, which should therefore be avoided (Dawes, 1979). It has even been proposed that the use of equal weights does not entail a serious loss of accuracy provided that there is a "sufficiently large number of intercorrelated predictor variables" (Wainer, 1976, p. 213). However, another strand of literature demonstrated that the loss in accuracy was actually twice as great as previously thought and, therefore, equal weighting schemes could not be considered a panacea, even though "they perform remarkably well" (Laughlin, 1978, p. 252). Wainer (1978, p. 272) conceded that his first conclusions were "overly simplistic" but maintained his position in favor of equal weights. Similar conclusions are reached by Einhorn and Hogarth (1975), while Parsons and Hulin (1982, p. 306) argue that equal weights are appropriate when "the variables are highly correlated and the sample sizes are small."

The risk assessment instrument employed from 1998 until 2008 in the state of Washington was the "Level of Service Inventory - Revised"or "LSI-R." This instrument was developed in Canada in the late 1970s (Andrews and Bonta, 1995) and it is still popular in several jurisdictions in North America and elsewhere. The predictive power of the LSI-R has been ascertained by several studies and has also been the subject of meta-analytic scrutiny (Gendreau, Little, and Goggin, 1996; Andrews and Bonta, 1995; Gendreau, Goggin, and Smith, 2002).[4]

[4]Its generalizability to diverse population groups, such as offenders in halfway houses (Bonta and Motiuk, 1985), female offenders (Coulson, Ilacqua, Nutbrown, Giulekas, and Cudjoe, 1996; Lowenkamp, Holsinger, and Latessa, 2001; Folsom and Atkinson, 2007), sex offenders (Simourd and Malcolm, 1998), long-term incarcerated offenders (Simourd, 2004; Manchak, Skeem, and

However, the LSI-R instrument is an unweighted scale in which all predictors are treated equally.

The goal of this Chapter is to show that introducing weights in the calculation of the LSI-R will significantly improve the predictive validity of the instrument. In other words, a weighted version of the LSI-R can perform much better in forecasting recidivism events compared to the unweighted version. The computational difficulty of using weights is far outweighed by the improvement in the accuracy of predictions. In any event, such complications can be easily overcome in the age of information technology. I, therefore, argue that a minor adjustment in the administration of the LSI-R, such as the use of weights, can have a substantial impact on its capacity to predict recidivism.

The Chapter proceeds as follows. Section 4.2 presents the methods of risk assessment and supervision allocation employed in Washington State. Section 4.3 describes the sources of the data and the way they are used in this Chapter . Section 4.4 gives an account of the empirical method used in this Chapter . Section 4.5 presents and discusses the results, and Section 4.6 concludes.

## 4.2 Background on Washington's Correctional Policies

The use of the LSI-R in Washington State began in 1998 (Barnoski and Aos, 2003; Aos and Barnoski, 2005). In 1999, the state's legislature enacted the Offender Accountability Act ("OAA"), which introduced the "Risk Management Identification" ("RMI") system (Aos, 2003). The RMI created a system of differential supervision for offenders released in the community that was largely based on their risk profile. More risky offenders would receive more supervision compared to less risky ones.[5] Four risk categories were created corresponding

Douglas, 2008) or different ethnic groups (Holsinger, Lowenkamp, and Latessa, 2003) has been proved repeatedly.

[5] According to the classification rules, the gravity of the offense, or the "harm done" by each offender, would also play a role in the assignment to the proper supervision category (Aos, 2002; Aos and Barnoski, 2005). However, the risk score turns out to be the most important factor.

to four different levels of supervision in terms of budgeted time and resources.[6] The LSI-R immediately became an integral part of the new system as the risk assessment instrument by which the risk of reoffending would be measured.

Offenders that were about to be released to the community either after having served time in prison or not (sentenced straight to the community) were subject to an interview in which the authorities made use of the LSI-R questionnaire.[7] The questionnaire comprises 54 items each of which represents a risk factor. As indicated above, all risk factors are weighted equally contributing either a 1 or a 0 to an offender's score, depending on whether his/her answer is affirmative or negative to the presence of the relevant factor.[8] An offender's score on the LSI-R grid is equal to the sum of the items answered affirmatively, i.e. the sum of the 1s, thus generating a scale from 0 to 54, where higher scores indicate higher risk and lower scores lower risk. A true replication of the LSI-R instrument is presented in Appendix A.

The 54 items of the questionnaire are divided in ten subcomponents.[9] An important feature of the questionnaire is that it contains both what are known as "static" and "dynamic" risk factors. Static are factors that do not change over time, such as criminal history, whereas dynamic are factors that can change over time and are thus possible targets of treatment programs, such as drug dependency. The advantages of including dynamic factors are twofold: (a) it allows the authorities to follow numerically the progress made by an offender over time, and (b) it can direct the authorities to focus their rehabilitative efforts to the areas of an offender's profile that require more attention.

Since 2003, the Washington State Institute of Public Policy ("WSIPP"), an organization charged with evaluating policies employed by the state's authorities, has been using its expertise in order to produce a new risk assessment

[6] The categories were called RMA, RMB, RMC and RMD, where the first one corresponded to high risk and the last one to low risk offenders.

[7] The answers provided by the offenders were corroborated by the authorities if that was possible. Ways of verification included documentation, contacts with family members or acquaintances of the offender, etc.

[8] For some items instead of a "Yes" or "No" answer, there is a small scale from 0 to 3 but again the answers translate directly to the 0 or 1 format.

[9] The subcomponents are (with number of items in parentheses): Criminal History (10), Education/Employment (10), Financial (2), Family/Marital (4), Accommodation (3), Leisure/Recreation (2), Companions (5), Alcohol/Drug Problems (9), Emotional/Personal (5), Attitudes/Orientation (4).

instrument tailored after the state's specific needs. This new instrument was finally introduced in August 2008, at which time the use of the LSI-R was discontinued (Drake and Barnoski, 2009).

The new instrument, known as the "static risk" instrument, contains only static risk factors and no dynamic ones and it has 26 instead of 54 items (Barnoski and Drake, 2007).[10] It does not follow the format of the LSI-R with the 0 or 1 answers but rather most items have small scales, e.g. 0-3 or 0-5, where a higher number reflects higher risk. Additionally, instead of just one score, the new instrument generates three scores that target specific recidivism types: (a) felony score, (b) property and violent score, and (c) violent score. Most importantly, the new instrument introduces an elementary version of weights for each specific item. These weights are different for each of the three aforementioned scores.[11]

Therefore, since 2008 Washington State authorities have been using an instrument that is employing a simple form of weighted risk factors that capture only a static picture of an offender's risk profile. In this Chapter, I show that a better accurately weighted version of the LSI-R instrument could achieve better levels of predictive validity compared to the unweighted version and quite similar levels of validity compared to the new static instrument.

## 4.3 Data

For this study I used the data set that I used in Chapter 1. This is a data set that was put together by combining three separate confidential data sets that were given to me by the "Washington State Institute of Public Policy" ("WSIPP"). The first of them had information on LSI-R scores, the second on assignment of offenders to the four risk categories of the RMI system and the third on demographic, criminal history and recidivism characteristics.

[10] It has the following six subcomponents (with number of items in parentheses): Demographics (2), Juvenile Record (4), Commitment to the Department of Corrections (1), Total Adult Felony Record (9), Total Adult Misdemeanor Record (9), Total Sentence/Supervision Violations (1).

[11] For example, the item "age at time of sentence" is weighted by being multiplied by 5 for the felony score, by 4 for the property and violent score and by 2 for the violent score (Barnoski and Drake, 2007).

Merging of these data sets produced a data set of 51,957 offenses committed by 47,154 individual offenders. The time span of the analysis is from July 2000 to September 2004. Table 4.1 presents descriptive statistics for general characteristics of the sample, such as gender, race, age, type of crime, type of sentence. Table 4.2 presents specific information on recidivism. For additional clarity this information is also broken down per risk category.

More specifically, in this Chapter I use a subset of this data set as my construction sample. The size of the construction sample is 10,000 randomly selected observations from the entire data set of 51,957 observations. I derive the weights for my weighted instrument from the observations in the construction sample. I cross check the results obtained by using a validation sample. The validation sample is 10,000 randomly selected observations from the 41,957 observations that were not used in the construction sample. Similarity of the results obtained in the construction sample with the results obtained in the validation sample serves as proof of the fact that the weighted instrument generated in this Chapter is generalizable to different populations of offenders.[12] The use of these two samples is a standard process in the relevant literature (Jones, 1996; Barnoski and Drake, 2007).

## 4.4 Empirical Methodology

I follow the methodology commonly used by the literature on evaluating the predictive validity of risk assessment instruments. That method is the Relative Operating Characteristic ("ROC") and it gives a measure of the "effect size" or accuracy of an instrument (Quinsey, Harris, Rice, and Cornier, 2006). Moreover, this method is the one chosen by WSIPP to evaluate the predictive validity of the new static instrument that was introduced in 2008. Therefore, the use of this method is warranted by the fact that it provides a common measure to compare my improved version of the LSI-R with the instrument currently in use in Washington.

The evaluation process begins in the construction sample by running a

[12]Although, a discrepancy in favor of the construction sample, known in the literature as *shrinkage* is to be expected.

regression of recidivism on all of the 54 items of the questionnaire. The specification is the following:

$$recidivism_i = \beta_0 + \beta_1 item1_i + \beta_2 item2_i + \ldots + \beta_{54} item54_i + \epsilon_i. \tag{4.1}$$

This is a binary response model and the estimation method is OLS. The dummy variable $recidivism_i$ is the outcome variable and takes the value 1 if an offender has recidivated after he or she has been at-risk to the community and 0 otherwise.[13] The recidivism event is defined as any conviction for a misdemeanor or felony offense that occurs within a 36-month period after being at-risk to the community (Drake, Aos, and Barnoski, 2010).[14] Additionally, a 12-month adjudication period is allowed. The variables $item1_i$ through $item54_i$ are the answers of the offenders to each one of the 54 items of the LSI-R questionnaire. As indicated above, all the items take 0 or 1 answers, therefore the 54 items are all dummy variables.

Using a stepwise regression and a 20% significance level, I find which of the 54 items of the LSI-R need to be part of the optimal specification.[15] The remaining items do not contribute in the predictive validity of the instrument and can thus be safely removed. After the elimination of the items, I obtain the coefficients for each one of them. These coefficients are the weights that are used for the construction of the "Weighted Instrument" ("WI").

Having obtained the weights, I construct the WI by multiplying the observations in the construction sample for each item by the respective weight as it was previously derived. Therefore, each 0 or 1 entry recorded for a specific item is multiplied by the weight that corresponds to it, yielding a product that is either 0 or equal to the weight. The WI is then the sum of these entries and it is rounded to the nearest integer.

After the creation of the WI, I proceed with the ROC. The ROC is

[13] An offender that has been incarcerated is at-risk in the community once he or she is released from prison, while an offender that is sentenced to a community sentence, is at-risk immediately.

[14] Technical violations of parole or probation conditions are not considered recidivism events.

[15] I started with a regression that included all 54 items. The item that was the least significant was removed (as long as its p-value was greater than 0.2). Then the regression was rerun and the next item was removed. This process continued until a specification was reached for which all the p-values for the remaining items were less than 0.2.

used to measure the predictive accuracy or validity of the new score and it serves as its graphical representation. The y-axis of a ROC graph represents the percentage of times that the instrument predicts accurately a recidivism event (hit rate) and the x-axis the percentage that it predicts falsely a recidivism event (false alarm rate). Each point of an instrument corresponds to a specific point on this x-y diagram, indicating a relationship between hit rate and false alarm rate. Therefore, the LSI-R with its 0-54 scale will produce 55 points on the x-y diagram that are then connected to form the ROC curve. Accordingly, the WI has a number of points that corresponds to its respective scale.

The diagonal is the locus on which an instrument predicts an event just as good as chance. The points of the new instrument on the diagram are then connected to form the ROC and I measure the area under it. Obviously an area of 0.5 corresponds to the diagonal. Any number higher than 0.5 and closer to 1 corresponds to an instrument with predictive validity greater than chance. The higher the number, the better the predictive validity.

I measure the area under the curve for the WI and then compare it with the respective areas obtained by using the original LSI-R as well as the new static instrument used by the authorities. Moreover, I cross-validate my results by using the same weights in the validation sample, generating similar ROCs in that sample and measuring the areas under them. The goal is to see if the results obtained in the construction sample are generalizable to any sample. The numbers obtained in the validation sample are the ones that are most reliable.

## 4.5 Results

The process described above resulted in three different versions of the WI depending on the outcome that we are interested in. Specifically, I generated distinct WIs for three different types of recidivism, namely general recidivism, felony recidivism and violent felony recidivism. This is consistent with the strategy of WSIPP (Barnoski and Drake, 2007), who produced different scores for different types of outcomes, as indicated above. However, I could only follow their categories for felony and violent felony recidivism as my data did not include a specific category for property/violent felony recidivism. Instead, I also introduced a general recidivism category, which includes felonies as well as misdemeanors. The fact that WSIPP have also produced their own ROC curves using the new static instrument, allows me to compare the WI not only with the LSI-R but also with that instrument that is specifically tailored for Washington State offenders.

The weights for the three different WI versions are presented in Table 4.3.[16] We notice that more items of the LSI-R are used to predict the general recidivism outcome (32 items), compared to 31 for felony recidivism and 24 for violent felony recidivism. The WI versions for recidivism and felony recidivism use mostly the same items, while the version for violent felony recidivism uses somewhat different items.

Specifically for the recidivism and felony recidivism versions, we note that items in the Criminal History and Education/Employment sections of the LSI-R questionnaire are relevant overall, with only a few items from each of these sections being removed. Moreover, these items are associated with the highest weights. Notably, item 9 that concerns violations during prior community supervision, has a weight of 8.4 for the recidivism outcome and 7.7 for the felony recidivism outcome. Item 3 dealing with the number of prior convictions is also worth noting as the weights associated with it are 8.2 and 7 respectively. These are extremely high number if they are compared with the unweighted LSI-R, where these items are weighted equally with all the other items.[17]

In contrast, the Leisure/Recreation with its two items seems to not be at all significant. Additionally, from the nine items of the Emotional/Personal and Attitudes/Orientations sections, only three are left in the calculations for the recidivism version and four for felony recidivism.

This shows that objective sections of the questionnaire, such as Criminal History and Education/Employment, that are clearly measurable and easily verifiable end up being important predictors of future recidivism. On the con-

[16] The weights presented have been multiplied by 100 so that the different version of the WI are scales of integer numbers similar to the LSI-R

[17] Furthermore, it must also be highlighted that several weights, namely 9 for recidivism and 10 for felony recidivism, are negative, i.e. they subtract off points from the scale rather than add.

trary, subjective sections, such as Leisure, Emotional or Attitudes seem to be much less helpful in predicting recidivism. Moreover, it should not go unnoticed that in Chapter 1, I showed that items 30, 31, 42, 45, 50, 53 and 54 have been subjected to manipulation, with the primary culprit being item 53. Five out of these seven items (all but items 50 and 54) are found now again to be *not* significant predictors in the calculations for recidivism and felony recidivism.

The results for the areas under the ROC curves generated by different instruments are presented in Figures 4.1-4.3 and cumulatively in Table 4.4. We notice that for all types of recidivism and whether we use the construction or the validation sample, the conclusion is that the WI clearly and statistically significantly outperforms the LSI-R. And this is true even though the three versions of the WI consist of much fewer items than the LSI-R. However, it should also be noted that the static instrument outperforms the WI for felony and violent felony recidivism outcomes.

More specifically and beginning our analysis with general recidivism and Figure 4.1, we notice that both in the construction and the validation sample, the WI predicts the outcome considerably better than the LSI-R. In the construction sample the area under the ROC curve for the WI is 0.7293 while for the LSI-R it is 0.6818. The difference is strongly statistically significant. The result is similar in the validation sample, with minor declines in both areas but still a difference that is strongly significant.

This pattern is repeated with felony recidivism in Figure 4.2. The only difference is that the numerical difference between the WI and the LSI-R becomes larger now both for the construction and the validation sample. Specifically, the WI retains similar levels for its areas under the ROC (around 0.72), while the LSI-R drops further to around 0.67.

Finally, with regards to violent felony recidivism, Figure 4.3 shows that the WI still outperforms the LSI-R and the difference becomes even larger. However, the significant drop of the area generated by the WI from 0.7221 in the construction sample to 0.6840 in the validation sample should be also noted. It can only be attributed to the fact that the predicted event is becoming increasingly rare as we move from general recidivism to more specific kinds. The difference with the LSI-R is still though large and significant.

Overall, we observe that a simple weighting of the items produces a predictive validity that is significantly better than the validity of the unweighted LSI-R instrument. This shows that fewer but weighted items can perform much better than more unweighted ones.

However, it should not go unsaid that the WI does not outperform in predictive validity the new static instrument currently used by Washington State authorities. As indicated in Table 4.4, which sums up areas under the ROC curve for different instruments, the new static instrument produces higher areas in all of its versions. Specifically, we note a higher area under the ROC curve for felony and violent felony recidivism ranging from 2.5 to 5 percentage points depending on the outcome and the sample observed.

The static instrument is better equipped to predict revidivism events in the state of Washington for several reasons. First, it should be noted that it was developed specifically for the offender population of that state, whereas the LSI-R, from which the WI is derived, has been developed for general use. Second, the static instrument makes use of important variables, such as age and gender, which the WI does not employ since they were never part of the LSI-R. The reason for non-inclusion of such predictors is my effort to keep the comparison between the WI and the LSI-R as clear as possible. Finally, as already indicated, the static instrument uses itself some form of weighting. Therefore, given the circumstances and the fact that the WI and the static instrument essentially use similar methods to improve the predictive validity of the LSI-R, the differences observed in the areas produced by these two instruments should not be overstated.

## 4.6 Conclusion

This Chapter examined the issue of whether the predictive accuracy of risk assessment instruments may be improved by using weights in their construction. Specifically, based on data originating from the state of Washington, I evaluated the predictive power of the index called "Level of Service Inventory - Revised" or "LSI-R," as was administered there from 2000 to 2004. The accuracy of that index was compared to a new weighted instrument, which I named "WI."

The WI was derived from the LSI-R by weighting all the items of the latter instrument and eliminating the ones that did not contribute sufficiently to the prediction.

I found that introducing weights in the calculation of the LSI-R improves significantly its predictive validity. I evaluated this validity by using the Receiver Operating Characteristic curve and measuring the area under it. I used three possible outcomes to be predicted, namely, recidivism, felony recidivism and violent felony recidivism. In all three cases the WI outperformed the LSI-R by a considerable and statistically significant margin. And that was the case, even though the WI used fewer items than the LSI-R, as several of them were eliminated in the process.

This result shows that the introduction of weights in actuarial instruments is an easy way of obtaining improved prediction results. The argument of simplicity that could be used in favor of unweighted instruments cannot be sustained in the age of computer technology. A simple spreadsheet software suffices for making use of a weighted instrument. In this respect, there is actually no extra time or effort needed for the calculation of a weighted instrument compared to the calculation of an unweighted one.

At the same time though, it should be noted that the static instrument, currently in use by Washington State authorities since 2008, performs even better than the WI. As already indicated however, this is actually an extra argument in favor of weighted instruments, since the static instrument employs a set of elementary weights itself.

The fact that there exists a simple and rather costless way to enhance the predictive power of actuarial instruments should not be confused with the issue of which prediction method is optimal, clinical or actuarial. This latter issue was not addressed by this Chapter and I can only refer to the relevant ongoing debate. It seems though that actuarial instruments have a lot to gain in this debate by employing techniques such as weighting. Weighting might no tip the balance but still, it might make a significant difference.

## 4.7 Figures and Tables

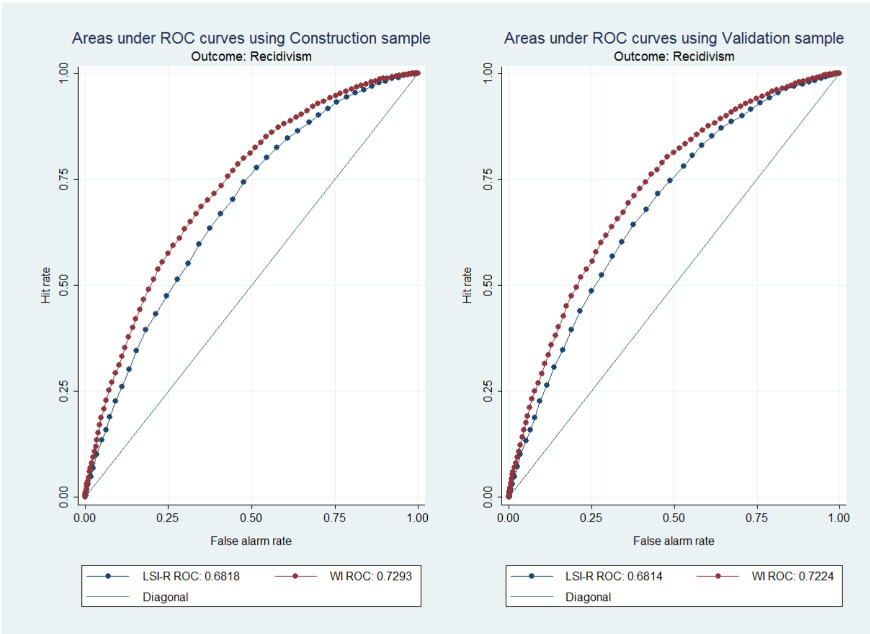


Figure 4.1: Areas under the ROC curve for LSI-R and WI predicting a recidivism event

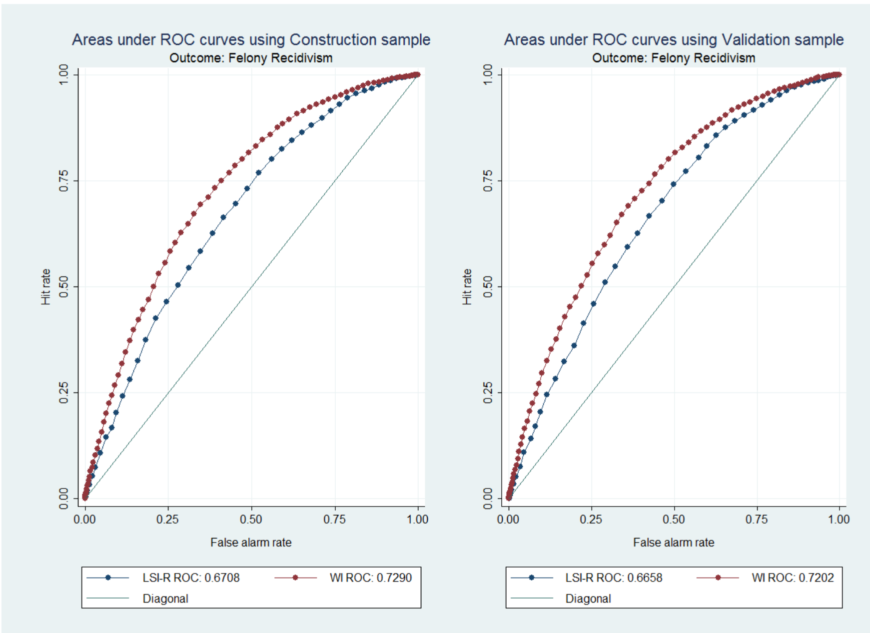


Figure 4.2: Areas under the ROC curve for LSI-R and WI predicting a felony recidivism event

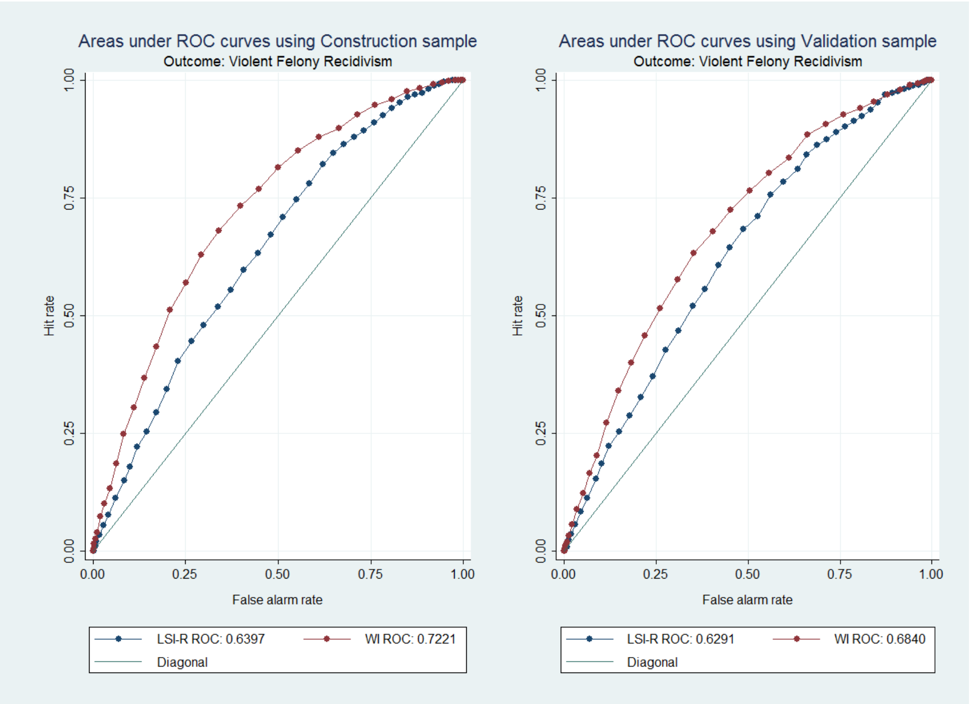


Figure 4.3: Areas under the ROC curve for LSI-R and WI predicting a violent felony recidivism event

Table 4.1: Composition of the data set in percentages

| 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 | 2.78 | 3.56 | 3.59 | 3.05 | 1.72 |
| Asian | 2.71 | 3.30 | 1.98 | 1.86 | 3.75 |
| Age 18-30 | 47.64 | 47.96 | 45.74 | 48.67 | 52.16 |
| Age 31-45 | 40.66 | 39.72 | 42.95 | 41.82 | 36.34 |
| Age over 45 | 11.59 | 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 |
| Observations | 51,957 | 7,902 | 8,126 | 19,008 | 16,831 |

Table 4.2: Recidivism 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 4.3: Weights obtained by linear regressions

| Item No. | Recidivism | Felony Recidivism | Violent Felony Recidivism |
|---|---|---|---|
| 1 | 3.8 | 2.9 | |
| 3 | 8.2 | 7 | 1.9 |
| 4 | | 2 | |
| 5 | 3.7 | 3.4 | 2.8 |
| 6 | 2.2 | 2 | |
| 7 | 4.9 | 6.9 | |
| 8 | 6.2 | 6.2 | 2.2 |
| 9 | 8.4 | 7.7 | 2.8 |
| 10 | 4.4 | | 4.2 |
| 11 | 3.8 | 5.2 | 1.4 |
| 12 | 5.1 | 4.6 | 1.4 |
| 13 | 6.7 | 5 | 3.5 |
| 14 | -2.3 | -3 | -2.8 |
| 15 | -3.9 | -2.4 | |
| 16 | 6.6 | 4.2 | |
| 17 | 3.5 | 3.3 | 2.9 |
| 18 | 3 | | |
| 22 | -2.7 | -4.9 | -3.1 |
| 24 | -2.3 | -2 | |
| 25 | | | -1.2 |
| 26 | 1.5 | | |
| 27 | | 2.7 | 1.7 |
| 28 | 5.6 | 4.3 | |
| 31 | | | 1.1 |
| 33 | 3.3 | | |
| 34 | 2.5 | 4.1 | |
| 35 | 3.8 | 3.7 | |
| 37 | -3.7 | -3.4 | |
| 38 | 2.6 | 2.9 | |
| 39 | -2.9 | -3.1 | 1.2 |
| 40 | 3 | 5 | |
| 41 | 3.8 | 2.6 | -1.6 |
| 42 | | | 1.5 |
| 44 | -3.1 | -3.1 | -3 |
| 45 | | | -1 |
| 46 | | | -1.4 |
| 48 | | | -0.9 |
| 49 | -2.6 | -3.6 | |
| 50 | -2.7 | -3 | 1.2 |
| 52 | | -1.5 | |
| 53 | | | 1 |
| 54 | 3.8 | 4.6 | 1.3 |

Note: All weights have been multiplied by 100 so that the WI is a scale of integer numbers similar to the LSI-R.

Table 4.4: Areas under the ROC curve using different instruments

| Outcome | Instrument | Construction Sample | Validation Sample |
|---|---|---|---|
| Recidivism | LSI-R | 0.6818 | 0.6814 |
| | WI | 0.7293 | 0.7224 |
| | p-value | 0.000 | 0.000 |
| Felony Recidivism | LSI-R | 0.6708 | 0.6658 |
| | WI | 0.7290 | 0.7202 |
| | p-value | 0.000 | 0.000 |
| | Static | 0.756 | 0.742 |
| Violent Felony Recidivism | LSI-R | 0.6397 | 0.6291 |
| | WI | 0.7221 | 0.6840 |
| | p-value | 0.000 | 0.000 |
| | Static | 0.745 | 0.732 |

Note: The p-value row originates from a $\chi^2$ test where the null hypothesis is that the area under the curve using the LSI-R and the WI are equal. For Felony Recidivism and Violent Felony Recidivism I also report the areas under the ROC curve generated by the new static instrument currently in use in Washington as reported by Barnoski and Drake (2007).

## Appendix A

# Supplemental material to Chapter 1: The LSI-R Questionnaire

Table A.1: Items of the LSI-R questionnaire

| Item No. | Item | Answer | | | |
|---|---|---|---|---|---|
| | **Criminal History** | | | | |
| 1 | Any prior adult convictions? Number_ | No | | Yes | |
| 2 | Two or more prior adult convictions? | No | | Yes | |
| 3 | Three or more prior convictions? | No | | Yes | |
| 4 | Three or more present offenses ? Number_ | No | | Yes | |
| 5 | Arrested under age sixteen? | No | | Yes | |
| 6 | Ever incarcerated upon conviction? | No | | Yes | |
| 7 | Escape history from correctional facility? | No | | Yes | |
| 8 | Ever punished for institutional misconduct? Number_ | No | | Yes | |
| 9 | Charge laid or probation/parole suspended during prior community supervision? | No | | Yes | |
| 10 | Official record of assault/violence? | No | | Yes | |
| | **Education/Employment** | | | | |
| 11 | Currently unemployed? | No | | Yes | |
| 12 | Frequently unemployed? | No | | Yes | |
| 13 | Never employed for a full year? | No | | Yes | |
| 14 | Ever fired? | No | | Yes | |
| 15 | Less than regular grade 10? | No | | Yes | |
| 16 | Less than regular grade 12? | No | | Yes | |
| 17 | Suspended or expelled at least once? | No | | Yes | |
| 18 | Participation/performance | 3 | 2 | 1 | 0 |
| 19 | Peer interactions | 3 | 2 | 1 | 0 |
| 20 | Authority interactions | 3 | 2 | 1 | 0 |
| | **Financial** | | | | |
| 21 | Problems | 3 | 2 | 1 | 0 |
| 22 | Reliance on social assistance | No | | Yes | |
| | **Family/Marital** | | | | |
| 23 | Dissatisfaction with marital or equivalent situation | 3 | 2 | 1 | 0 |
| 24 | Non-rewarding, parental | 3 | 2 | 1 | 0 |
| 25 | Non-rewarding, other relatives | 3 | 2 | 1 | 0 |
| 26 | Criminal family/spouse | No | | Yes | |
| | **Accommodation** | | | | |
| 27 | Unsatisfactory | 3 | 2 | 1 | 0 |
| 28 | Three or more address change last year | No | | Yes | |
| 29 | High-crime neighborhood | No | | Yes | |
| | **Leisure/Recreation** | | | | |
| 30 | Absence of a recent participation in an organized activity | No | | Yes | |
| 31 | Could make better use of time | 3 | 2 | 1 | 0 |
| | **Companions** | | | | |
| 32 | A social isolate | No | | Yes | |
| 33 | Some criminal acquaintances | No | | Yes | |
| 34 | Some criminal friends | No | | Yes | |
| 35 | Few anti-criminal acquaintances | No | | Yes | |
| 36 | Few anti-criminal friends | No | | Yes | |
| | **Alcohol/Drug problem** | | | | |
| 37 | Alcohol problem, ever | No | | Yes | |
| 38 | Drug problem, ever | No | | Yes | |
| 39 | Alcohol problem, currently | 3 | 2 | 1 | 0 |
| 40 | Drug problem, currently Specify type of drug:_ | 3 | 2 | 1 | 0 |
| 41 | Law violations | No | | Yes | |
| 42 | Marital/family | No | | Yes | |
| 43 | School/work | No | | Yes | |
| 44 | Medical | No | | Yes | |
| 45 | Other indicators Specify:_ | No | | Yes | |
| 46 | Moderate interference | No | | Yes | |
| 47 | Severe interference, active psychosis | No | | Yes | |
| 48 | Mental health treatment, past | No | | Yes | |
| 49 | Mental health treatment, present | No | | Yes | |
| 50 | Psychological assessment indicated Area:_ | No | | Yes | |
| | **Attitudes/Orientations** | | | | |
| 51 | Supportive of crime | 3 | 2 | 1 | 0 |
| 52 | Unfavorable toward convention | 3 | 2 | 1 | 0 |
| 53 | Poor, toward sentence | No | | Yes | |
| 54 | Poor, toward supervision | No | | Yes | |

Note: "Yes" or "No"answers are counted as 1 or 0 respectively for the calculation of the index. The total LSI-R score is the sum of all the 1s recorded. For items that have a "0-3" rating format, a 3 or a 2 corresponds to a "No" answer, while a 1 or a 0 corresponds to a "Yes" answer. The numbers are assigned according to the following scale (Andrews and Bonta, 1995):

**3** A satisfactory situation with no need for improvement

**2** A relatively satisfactory situation, with some room for improvement evident

**1** A relatively unsatisfactory situation with a need for improvement

**0** A very unsatisfactory situation with a very clear and strong need for improvement

# Appendix B

# Supplemental material to Chapter 1: Appendix of Figures

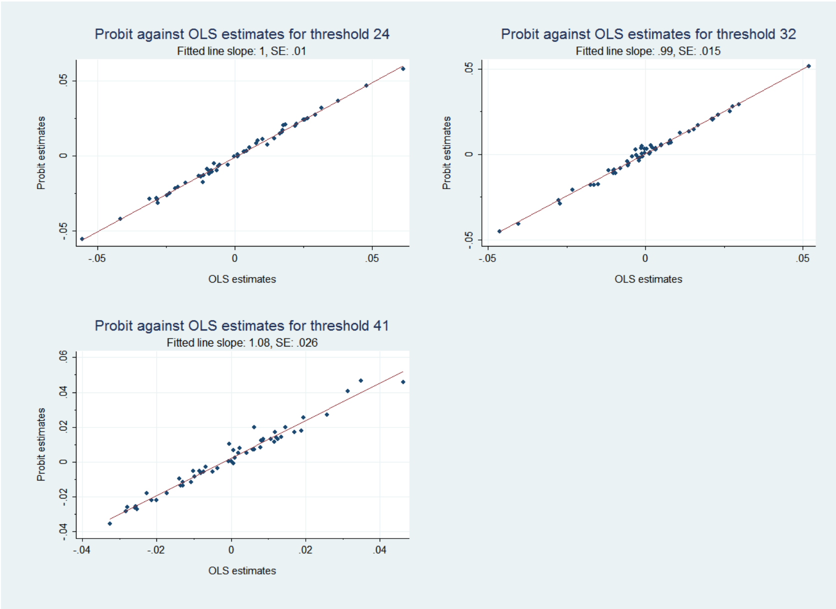


Figure B.1: Marginal effects of probit coefficients for the 54 items plotted against OLS coefficients for each one of the three thresholds with fitted lines

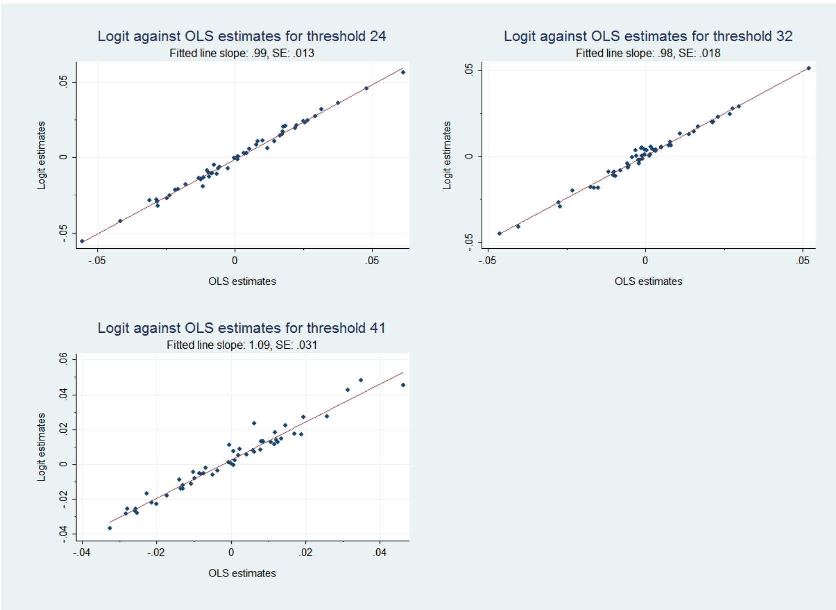


Figure B.2: Marginal effects of logit coefficients for the 54 items plotted against OLS coefficients for each one of the three thresholds with fitted lines

## Appendix C

# Supplemental material to Chapter 2: Appendix of Figures and Tables

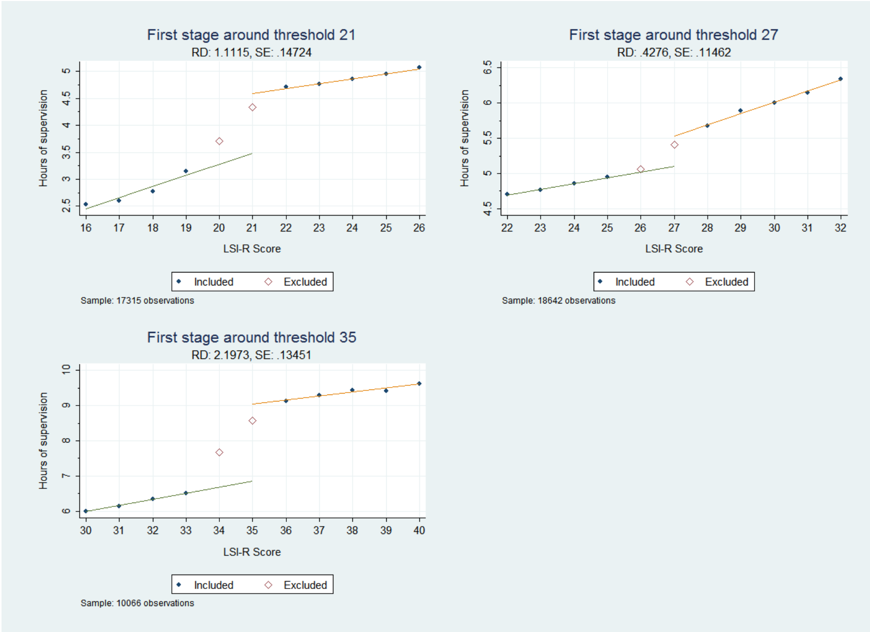


Figure C.1: First stage for the three thresholds excluding observations on the thresholds and one point prior to that using the corrected data

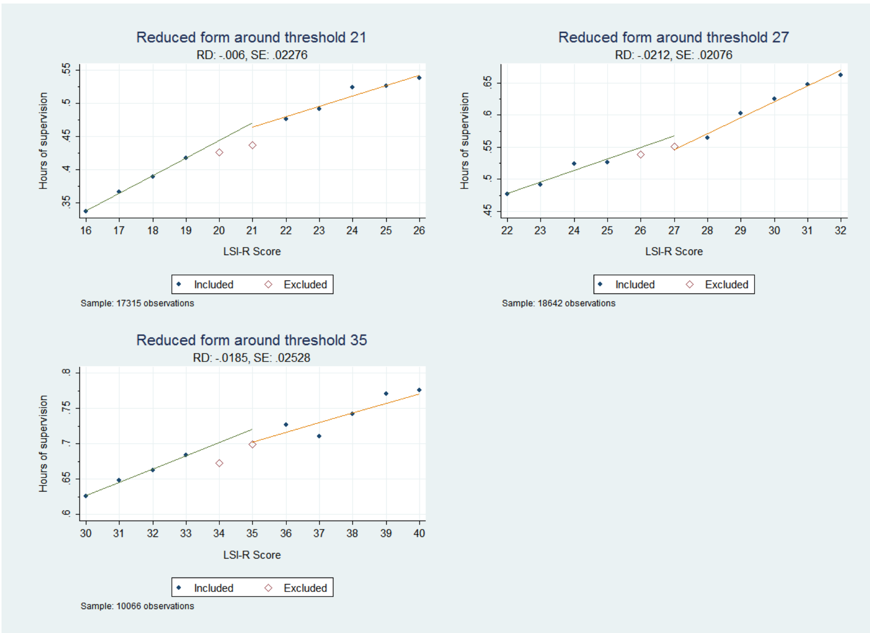


Figure C.2: Reduced form for the three thresholds and "Recidivism" as the outcome variable, excluding observations on the thresholds and one point prior to that using the corrected data.

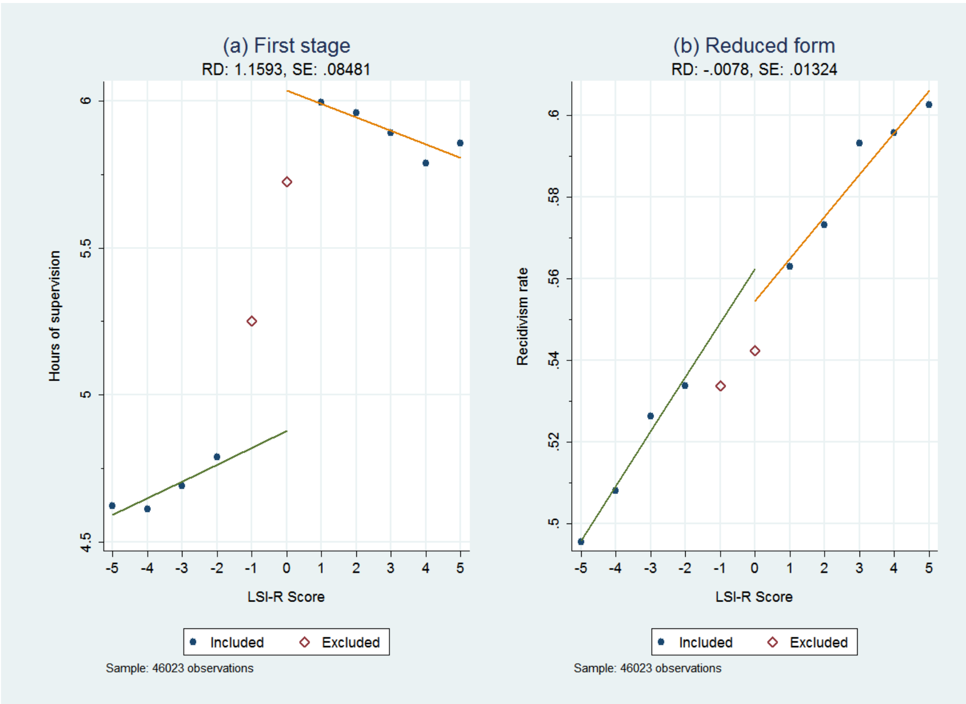


Figure C.3: First stage and reduced form with "Recidivism" as the outcome variable for the combined-threshold analysis, excluding observations on the thresholds and one point prior to that using the corrected data.

Table C.1: 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 | -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] |
| Homicide | 0.0017 | 0.0014 | 0.0018 |
| | [0.0020] | [0.0015] | [0.0010] |
| Prison | 0.0031 | 0.0048 | 0.0230 |
| | [0.0105] | [0.0107] | [0.0149] |

Standard errors in brackets
*Note:* This Table is similar to Table 2.1 of the body of the text but now we are using the corrected scores.

Table C.2: 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 | | 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 | 17899 | 17899 | 19825 | 19825 | 10881 | 10881 |

Standard errors in brackets

*Note:* This Table presents the IV estimates for each one 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 2.4 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 we use only scores that are five points below the respective threshold and six points above it (including the threshold itself).

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

Table C.3: 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 | | 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 | 21291 | 21291 | 23141 | 23141 | 12722 | 12722 |

Standard errors in brackets
*Note:* This Table is equivalent to Table C.2 but now we are using the corrected LSI-R scores. Everything said in the note of Table C.2 applies here too.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C.4: 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 | | 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 | 48605 | 48605 | 57154 | 57154 |

Standard errors in brackets

*Note:* This Table incorporates controls in the combined-threshold analysis presented in Table 2.5. 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 2.5 for the outcome variable "Recidivism." The columns labeled "Controls" include the control variables described in the note of Table C.2. Everything said in the note of Table 2.5 applies here too.

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

Table C.5: First Stage, Reduced Form and IV regressions using the corrected LSI-R scores excluding observations on the thresholds and one point prior to that

| 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 | 21291 | 17315 | 23141 | 18642 | 12722 | 10066 |

Standard errors in brackets
*Note:* In this Table we exclude from the regressions labeled (b) the observations on the thresholds and one point prior to that. In Figure 2.5 we noticed that these observations might have a confounding effect on the size of especially the first stage of our analysis. For each threshold the models labeled (a) repeat the estimates already reported in Table 2.4 for the outcome variables: Recidivism, Property Felony Recidivism and Misdemeanor Recidivism. The latter two types of recidivism are reported here because we found a statistically significant effect on them in Table 2.4. The models labeled (b) report the estimates following the exclusion of the said points. Everything said in the note of Table 2.4 applies here too.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table C.6: 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 to that

| 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 | 57154 | 46023 |

Standard errors in brackets

*Note:* This Table is equivalent to Table C.5 but now we have combined the three thresholds in one. The column labeled "Before exclusion" repeats the estimates already reported in Table 2.5 for the outcome variables: Recidivism, Property Felony Recidivism and Misdemeanor Recidivism.. The column labeled "After exclusion" implements the exclusion of the two points.

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

Table C.7: Cox proportional hazard estimates with combined thresholds

| | Manipulated | Corrected |
|---|---|---|
| Recidivism | 0.01 | -0.02 |
| | [0.026] | [0.024] |
| Felony Recidivism | 0.00 | -0.05 |
| | [0.031] | [0.028] |
| Property Felony Recidivism | -0.08 | -0.08 |
| | [0.051] | [0.046] |
| Drug Felony Recidivism | -0.01 | -0.04 |
| | [0.055] | [0.050] |
| Violent Felony Recidivism | 0.09 | -0.03 |
| | [0.060] | [0.053] |
| Misdemeanor Recidivism | 0.01 | 0.07 |
| | [0.048] | [0.043] |

Standard errors in brackets
*Note:* This Table shows estimates from a Cox proportional hazard regression model for various types of recidivism. All observations around thresholds are being used separately for the manipulated and the corrected scores. The estimates show the difference in recidivism probability per unit of time if an offender is just over a threshold compared to being just under it. For example, the first entry tells us that offenders that are over a threshold are 1 percent more likely to recidivate compared to those that are not per unit of time. None of the estimates is statistically significant. Apart from the RD estimates, a linear polynomial of the risk score concludes the right hand-side variables.

# Appendix D

# Supplemental material to Chapter 3: Appendix of Tables

Table D.1: Recidivism rates by race for offenders on the thresholds (using only bumped up offenders)

| Recidivism rate (total number of observations in parenthesis) | | | |
|---|---|---|---|
| Group of offenders | Threshold 24 | Threshold 32 | Threshold 41 |
| White | 0.35 (693) | 0.48 (392) | 0.65 (386) |
| Black | 0.56 (130) | 0.61 (82) | 0.77 (115) |
| p-value | 0.000 | 0.036 | 0.013 |
| Male | 0.40 (743) | 0.52 (437) | 0.69 (436) |
| Female | 0.35 (192) | 0.44 (106) | 0.66 (125) |
| p-value | 0.210 | 0.136 | 0.529 |
| Age 18-30 | 0.20 (804) | 0.56 (216) | 0.74 (236) |
| Age 31-45 | 0.41 (308) | 0.50 (172) | 0.68 (185) |
| p-value | 0.000 | 0.203 | 0.206 |
| Drug | 0.40 (301) | 0.47 (190) | 0.64 (204) |
| Non-Drug | 0.38 (634) | 0.53 (353) | 0.70 (357) |
| p-value | 0.652 | 0.237 | 0.156 |
| Property | 0.42 (226) | 0.63 (113) | 0.77 (152) |
| Non-Property | 0.38 (709) | 0.48 (430) | 0.65 (409) |
| p-value | 0.195 | 0.004 | 0.005 |
| Prison | 0.44 (149) | 0.59 (90) | 0.71 (116) |
| Community | 0.38 (786) | 0.49 (453) | 0.67 (445) |
| p-value | 0.135 | 0.094 | 0.472 |

*Note:* Last row shows p-values from Pearson $\chi^2$ tests of null hypothesis that recidivism rates in a given column are equal.

# Bibliography

ANDERSON, M., C. DOBKIN, AND T. GROSS (2011): "The effect of health insurance coverage on the use of medical services," *American Economic Journal: Economic Policy*, Forthcoming.

ANDREWS, D., AND J. BONTA (1995): *LSI-R User's Manual.* Multi-Health Systems Inc., Toronto, Ontario, Canada.

——— (2006): *The psychology of criminal conduct.* Anderson Publishing Co.

ANGRIST, J., AND J. PISCHKE (2009): *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

ANTONOVICS, K., AND B. KNIGHT (2004): "A new look at racial profiling: Evidence from the Boston Police Department," Discussion paper, National Bureau of Economic Research.

ANWAR, S., AND H. FANG (2006): "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review*, 96(1), 127–151.

AOS, S. (2002): "Washington's Offender's Accountability Act: An Evaluation of the Department of Corrections' Risk Management Identification System," Discussion paper, Washington State Institute of Public Policy, Olympia.

——— (2003): "Washington's Offender's Accountability Act: Update and Progress Report on the Act's Evaluation," Discussion paper, Washington State Institute of Public Policy, Olympia.

AOS, S., AND R. BARNOSKI (2005): "Washington's Offender's Accountability Act: A First Look at Outcomes," Discussion paper, Washington State Institute of Public Policy, Olympia.

ARROW, K. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by A. O., and R. A., pp. 3–33.

BARNOSKI, R., AND S. AOS (2003): "Washington's Offender's Accountability Act: An Analysis of the Department of Corrections' Risk Assessment," Discussion paper, Washington State Institute of Public Policy, Olympia.

BARNOSKI, R., AND E. DRAKE (2007): "Washington's Offender's Accountability Act: Department of Corrections' Static Risk Instrument," Discussion paper, Washington State Institute of Public Policy, Olympia.

BECKER, G. (1971): *The economics of discrimination.* University of Chicago Press.

BOERNER, D., AND R. LIEB (2001): "Sentencing Reform in the Other Washington," *Crime and Justice*, 28, 71–136.

BONTA, J. (1996): "Risk-needs assessment and treatment," in *Choosing correctional options that work: Defining the demand and evaluating the supply*, ed. by A. Harland, pp. 18–32. Sage Publications, Inc.

Bonta, J., B. Bogue, M. Crowley, and L. Motiuk (2001): "Implementing offender classification systems: Lessons learned," in *Offender rehabilitation in practice: Implementing and evaluating effective programs*, ed. by G. Bernfeld, D. Farrington, and A. Leschied, pp. 227–245. J. Wiley and Sons.

Bonta, J., and L. Motiuk (1985): "Utilization of an Interview-Based Classification Instrument," *Criminal Justice and Behavior*, 12(3), 333.

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.

Breton, A., and R. Wintrobe (1975): "The equilibrium size of a budget-maximizing bureau: A note on Niskanen's theory of bureaucracy," *The Journal of Political Economy*, 83(1), 195–207.

Bureau of Justice Statistics (2010): "Probation and Parole in the Unites States, 2009," NCJ 231674, Washington, D.C.: U.S. Department of Justice.

Clear, T., and K. Gallagher (1983): "Screening Devices in Probation and Parole," *Evaluation Review*, 7(2), 217.

——— (1985): "Probation and parole supervision: A review of current classification practices," *Crime & Delinquency*, 31(3), 423.

Coulson, G., G. Ilacqua, V. Nutbrown, D. Giulekas, and F. Cudjoe (1996): "Predictive utility of the LSI for incarcerated female offenders," *Criminal Justice and Behavior*, 23(3), 427.

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.

Dawes, R. (1979): "The robust beauty of improper linear models in decision making.," *American psychologist*, 34(7), 571.

Dharmapala, D., and S. Ross (2003): "Racial bias in motor vehicle searches: Additional theory and evidence," .

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.

Dowden, C., and D. Andrews (2004): "The importance of staff practice in delivering effective correctional treatment: A meta-analytic review of core correctional practice," *International Journal of Offender Therapy and comparative criminology*, 48(2), 203.

Drake, E., S. Aos, and R. Barnoski (2010): "Washington's Offender's Accountability Act: Final Report on Recidivism Outcomes," Discussion paper, Washington State Institute of Public Policy, Olympia.

Drake, E., and R. Barnoski (2009): "New Risk Instrument for Offenders Improves Classification Decisions," Discussion paper, Washington State Institute of Public Policy, Olympia.

Durlauf, S. (2006): "Assessing Racial Profiling," *The Economic Journal*, 116(515), F402–F426.

Ehrlich, I., and R. Posner (1974): "An economic analysis of legal rulemaking," *The Journal of Legal Studies*, 3(1), 257–286.

EINHORN, H., AND R. HOGARTH (1975): "Unit weighting schemes for decision making," *Organizational Behavior and Human Performance*, 13(2), 171–192.

FLORES, A., C. LOWENKAMP, A. HOLSINGER, AND E. LATESSA (2006): "Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity," *Journal of Criminal Justice*, 34(5), 523–529.

FOLSOM, J., AND J. ATKINSON (2007): "The generalizability of the LSI-R and the CAT to the prediction of recidivism in female offenders," *Criminal Justice and Behavior*, 34(8), 1044.

FRIEDMAN, E., AND A. WICKELGREN (2012): "A New Angle on Rules versus Standards," .

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.

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.

GROVE, W., AND P. MEEHL (1996): "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy.," *Psychology, Public Policy, and Law*, 2(2), 293.

GROVE, W., D. ZALD, B. LEBOW, B. SNITZ, AND C. NELSON (2000): "Clinical versus mechanical prediction: A meta-analysis.," *Psychological assessment*, 12(1), 19.

HARRIS, G., M. RICE, AND C. CORMIER (2002): "Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients," *Law and Human Behavior*, 26(4), 377–394.

HARRIS, P., R. GINGERICH, AND T. WHITTAKER (2004): "The âeffectivenessâ of differential supervision," *Crime & Delinquency*, 50(2), 235.

HERNÁNDEZ-MURILLO, R., AND J. KNOWLES (2004): "Racial profiling or racist policing? Bounds tests in aggregate data," *International Economic Review*, 45(3), 959–989.

HOLSINGER, A., C. LOWENKAMP, AND E. LATESSA (2003): "Ethnicity, gender, and the Level of Service Inventory-Revised," *Journal of Criminal Justice*, 31(4), 309–320.

——— (2006): "Exploring the validity of the Level of Service Inventory-Revised with Native American offenders," *Journal of Criminal Justice*, 34(3), 331–337.

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

JONES, P. (1996): "Risk prediction in criminal justice," in *Choosing correctional options that work: Defining the demand and evaluating the supply*, ed. by A. Harland, pp. 33–68. Sage Publications, Inc.

KAPLOW, L. (1992): "Rules versus standards: An economic analysis," *Duke Law Journal*, 42, 557.

KNOWLES, J., N. PERSICO, AND P. TODD (2001): "Racial bias in motor-vehicle searches: Theory and evidence," *Journal of Political Economy*, 109(1), 203–229.

KRONER, D., J. MILLS, AND J. REDDON (2005): "A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk," *International journal of law and psychiatry*, 28(4), 360–374.

Kuziemko, I. (2007): "Going off parole: How the elimination of discretionary prison release affects the social cost of crime," .

Lally, S. (2003): "What tests are acceptable for use in forensic evaluations? A survey of experts.," *Professional Psychology: Research and Practice*, 34(5), 491.

Laughlin, J. (1978): "Comment on" Estimating coefficients in linear models: It don't make no nevermind."," .

Litwack, T. (2001): "Actuarial versus clinical assessments of dangerousness.," *Psychology, Public Policy, and Law*, 7(2), 409.

Lowenkamp, C., A. Holsinger, L. Brusman-Lovins, and E. Latessa (2004): "Assessing the inter-rater agreement of the Level of Service Inventory-Revised," *Federal Probation*, 68, 34.

Lowenkamp, C., A. Holsinger, and E. Latessa (2001): "Risk/need assessment, offender classification, and the role of childhood abuse," *Criminal Justice and Behavior*, 28(5), 543.

Lowenkamp, C., E. Latessa, and P. Smith (2006): "Does correctional program quality really matter? The impact of adhering to the principles of effective intervention," *Criminology & Public Policy*, 5(3), 575–594.

MacKenzie, D., K. Browning, S. Skroban, and D. Smith (1999): "The impact of probation on the criminal activities of offenders," *Journal of Research in Crime and Delinquency*, 36(4), 423.

Manchak, S., J. Skeem, and K. Douglas (2008): "Utility of the Revised Level of Service Inventory (LSI-R) in predicting recidivism after long-term incarceration," *Law and human behavior*, 32(6), 477–488.

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.

Meehl, P. (1954): *Clinical versus Statistical prediction*. University of Minnesota Press, Minneapolis.

Niskanen, W. (1968): "The peculiar economics of bureaucracy," *The American Economic Review*, 58(2), 293–305.

——— (1971): "Bureaucracy and representative government," *New York City, Auldine.*

Parsons, C., and C. Hulin (1982): "Differentially weighting linear models in organizational research: A cross-validation comparison of four methods," *Organizational Behavior and Human Performance*, 30(3), 289–311.

Persico, N., and P. Todd (2006): "Generalising the Hit Rates Test for Racial Bias in Law Enforcement, With an Application to Vehicle Searches in Wichita," *The Economic Journal*, 116(515), F351–F367.

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

——— (2004): "What Works in Prisoner Reentry-Reviewing and Questioning the Evidence," *Fed. Probation*, 68, 4.

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

Petersilia, J., and S. Turner (1991): "An evaluation of intensive probation in California," *The Journal of Criminal Law and Criminology (1973-)*, 82(3), 610–658.

Quinsey, V. (1995): "The prediction and explanation of criminal violence," *International Journal of Law and Psychiatry*, 18(2), 117–128.

Quinsey, V., G. Harris, M. Rice, and C. Cornier (2006): *Violent Offenders: Appraising and Managing Risk*. American Psychological Association, Washington DC.

Sanga, S. (2009): "Reconsidering racial bias in motor vehicle searches: Theory and evidence," *Journal of Political Economy*, 117(6), 1155–1159.

Simourd, D. (2004): "Use of dynamic risk/need assessment instruments among long-term incarcerated offenders," *Criminal Justice and Behavior*, 31(3), 306.

Simourd, D., and P. Malcolm (1998): "Reliability and validity of the Level of Service Inventory-Revised among federally incarcerated sex offenders," *Journal of Interpersonal Violence*, 13(2), 261.

Solomon, A., V. Kachnowski, and A. Bhati (2005): "Does parole work? Analyzing the Impact of Post-Release Supervision on Rearrest Outcomes," Washington, DC: The Urban Institute.

Travis, J. (2005): *But they all come back: Facing the challenges of prisoner reentry*. Urban Inst Pr.

Turner, S., J. Hess, and J. Jannetta (2009): "Development of the California Static Risk Assessment Instrument," Discussion paper, Center for Evidence-Based Corrections, The University of California, Irvine.

Wainer, H. (1976): "Estimating coefficients in linear models: It don't make no nevermind.," *Psychological Bulletin*, 83(2), 213.

——— (1978): "On the sensitivity of regression and regressors.," *Psychological Bulletin*, 85(2), 267–273.

Wooldridge, J. (2002): *Econometric analysis of cross section and panel data*. The MIT press.

——— (2009): *Introductory econometrics: A modern approach*. South-Western Pub.

Wyckoff, P. (1990): "The simple analytics of slack-maximizing bureaucracy," *Public Choice*, 67(1), 35–47.