# The innovative bureaucrat: evidence from the correctional authorities in Washington State

Georgios Georgiou

National University of Singapore, Economics Department*†

**Abstract**

Bureaucracies are usually regarded as inefficient, wasteful mechanisms. Contrary to this deeply rooted perception of bureaucracy, this paper documents the case of the correctional authorities in Washington State, a bureaucracy that acted with a considerable degree of innovation and professionalism. Their task was to administer a risk assessment instrument that measured the level of risk posed by offenders by way of a numerical score. They used that score to identify the level of supervision offenders were to receive once released into the community. In analyzing the data, I discovered an unusual application of the instrument 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 and I generate an instrument that is cleansed of the manipulation. I find that the manipulated instrument predicts serious recidivism events better than the cleansed instrument, especially when these events involve high-risk offenders, thus providing evidence that the authorities had good reason to undertake the manipulation.

*JEL classification:* C26, K42

*Keywords:* bureaucracy, actuarial risk assessment instruments, recidivism

*National University of Singapore, Economics Department, Faculty of Arts and Social Sciences, AS2 #06-02, 1 Arts Link, Singapore 117570. E-mail: ecsgg@nus.edu.sg.

†I would like to thank my advisors Donald Wittman and Carlos Dobkin for their invaluable help. I am also grateful to Elizabeth Drake of the Washington State Institute for Public Policy for providing me with the data sets used in this paper. I benefited from comments and suggestions from Julian Caballero, David Kaun, Jean Paul Rabanal, Dean Scrimgeour, Sean Tanoos, seminar participants at Colgate University, and an anonymous referee. I thank them all wholeheartedly.

# 1 Introduction

Bureaucracies are usually regarded as slow and anachronistic organizations that primarily serve the interests of their members and their sponsors. This is why over the years the concept of bureaucracy has acquired a negative connotation in public opinion. The present study documents the case of a bureaucracy whose actions are in contrast to this conventional wisdom.

In the state of Washington, the legislature had designed a complex system of supervision for criminal offenders released into the community that made use of advanced risk assessment techniques. The state's correctional authorities, acting innovatively, implemented this system in a way that enhanced its capabilities and yielded improved results. The goal of this paper is to document the innovation and demonstrate the superior outcomes achieved.

The classic model of a budget-maximizing bureaucracy was proposed by Niskanen (1968, 1971). According to that model, the sole purpose of bureaucrats is to maximize the amount of resources made available to them by their sponsors, the politicians.[1] However, the literature has also identified conditions under which bureaucracies tend to not conform to this dismissive description. For example, a smaller bureau or administrative area allows for better monitoring and performance (Davis and Hayes, 1993). Similarly, competition from external agencies that offer services similar to those offered by a bureau also provides authorities with an incentive to increase efficiency (Niskanen, 1975; Duncombe, Miner, and Ruggiero, 1997). The innovative bureaucrats of the present study showcase one more example of efficient bureaucracy.

In Washington State, the job of the correctional authorities was to determine the proper level of supervision for an offender once he or she is released into the community after either serving time in prison or being sentenced directly into the community. Supervision intensity was assigned based on the risk of reoffending posed by each offender. The authorities deviated from the rules that governed the standardized procedure of allocating supervision, and in doing so produced superior outcomes.

Jurisdictions have to make a choice about how to measure the risk of reoffending. There are two major candidates: "actuarial" instruments (also known as mechanical, algorithmic, or statistical instruments) and clinical judgment.[2] The Washington State authorities chose the former method. Actuarial instruments 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. Such evaluations are unlike clinical judgment of risk, which is typically performed in an unstandardized fashion by professionals in the correctional system (clinicians) on the basis of their subjective evaluation of each case.[3]

[1] Or, in the words of Sir Humphrey Appleby, the chief bureaucrat in the BBC series, *Yes, Minister*, since the civil service can't measure success by way of profits, "we have to measure our success by the size of our staff and our budget. By definition, a big department is more successful than a small one."

[2] More recently, a blend of the two methods, called "structured professional (or clinical) judgment," has been developed, attempting to tap into the attributes of both methods (Webster, Hucker, and Haque, 2014).

[3] According to proposed classifications of risk assessment methods (Bonta, 1996; Andrews and Bonta, 2006),

In the field of clinical psychology there is a major debate over which method should be used to make predictions: 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 as or better than clinical judgment (Meehl, 1954; Grove and Meehl, 1996; Grove, Zald, Lebow, Snitz, and Nelson, 2000). The same argument is made with respect to predicting the risk of reoffending (Quinsey, 1995; Bonta, Bogue, Crowley, and Motiuk, 2001; Harris, Rice, and Cormier, 2002; Singh and Fazel, 2010).[4]

In Washington State, from October 1998 until August 2008, the authorities used the actuarial instrument Level of Service Inventory - Revised, or LSI-R.[5] This instrument was developed in Canada in the late 1970s (Andrews and Bonta, 1995), and has since become increasingly popular in many jurisdictions across North America.[6] The instrument's predictive validity has been established meta-analytically by several studies (Andrews and Bonta, 1995; Gendreau, Little, and Goggin, 1996; Gendreau, Goggin, and Smith, 2002). At the same time, the importance of maintaining implementation integrity for actuarial instruments in general has also been highlighted by the literature.[7]

However, even though in Washington all measures were taken to ensure the proper administration of the LSI-R (Manchak, Skeem, and Douglas, 2008), the distribution of the scores raises suspicions (see Fig. 1). The three jumps observed in offenders' scores are exactly at the cut-off points that separate the different levels of supervision, as will be explained in detail below. I take this to be evidence of manipulation in the administration of the LSI-R instrument, since the authorities know the cut-off scores.[8] Standard rules for using actuarial instruments dictate that

clinical judgment is characterized as a "first-generation assessment" whereas actuarial methods can be subdivided into "second-", "third-" and, more recently, "fourth-generation assessments," depending on their characteristics. Second-generation assessments just focus on measuring risk. Third-generation assessments also try to uncover the "needs" that each offender has and to direct the rehabilitation efforts to those needs, 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).

[4]With respect to violent reoffending in particular, Fazel, Singh, Doll, and Grann (2012) find that actuarial instruments have high predictive accuracy for low risk offenders but the evidence does not support the use of such instruments as sole determinants of risk prediction. In a similar vein, Kroner, Mills, and Reddon (2005) showed that several risk instruments (including the one used in this study) are not better at predicting recidivism than arbitrary structured scales that are generated by combining randomly selected items from the original instruments.

[5]There are several other actuarial instruments (Singh and Fazel (2010) mention that they examined 126 such instruments for their study), some of which have general applicability, and others that pertain to certain types of offenses.

[6]According to Manchak, Skeem, and Douglas (2008), the LSI-R is the third most used instrument in the U.S. and Petersilia (2009) refers to it as the most popular instrument. In contrast, the instrument does not appear popular among forensic psychologists, 80 percent of whom (in a group of 64) had "no opinion" about its acceptability for evaluating violence risk (Lally, 2003). Similarly, Singh, Grann, and Fazel (2011) report that the LSI-R performs worse than other well-known instruments when it comes to predicting violent recidivism.

[7]Follow-up training sessions, computerization, videotaping and revisiting interview sessions, and setting up quality-assurance teams are some of the strategies proposed for enhancing quality in the administration of actuarial instruments (Bonta, Bogue, Crowley, and Motiuk, 2001).

[8]The term "manipulation" carries a negative connotation and is used in this paper for lack of a better succinct

administrators should not tamper with their mechanics in order to influence the outcome. Such interventions negate the objectivity of the instrument and reintroduce the subjectivity of clinical judgment. Even though the literature long ago identified practical problems that can arise when agencies use such instruments to assign supervision levels (Clear and Gallagher, 1983, 1985), to the best of my knowledge, a similar case of instrument manipulation has not been reported.

In view of the above, the object of this paper can be stated more accurately as follows: *a*) to document the manipulation of the LSI-R instrument, uncover its mechanics by using a regression discontinuity design, and reconstruct the instrument by excluding the manipulated parts, and *b*) evaluate whether the authorities' manipulation yielded more accurate predictions than the reconstructed "corrected" instrument.

Briefly, I find compelling evidence that manipulation did indeed take place, and that it was focused primarily on subjective items in the LSI-R instrument. Moreover, I find that the manipulated instrument produced by the authorities outperforms the reconstructed corrected instrument in predicting serious recidivism events (violent felonies), especially when those events involved high-risk offenders.

It should also be stressed though, that the differential supervision program operated in Washington State did not actually manage to reduce recidivism rates (Georgiou, 2014). However, the evaluation and analysis of the risk assessment strategy that the authorities followed is useful for the administration of treatment programs that may prove to be more effective. In this respect, the improved technique that the authorities in Washington used could benefit all correctional programs, present or future, that are based on risk assessment.

From a policy perspective, these results favor the operation of bureaucracies in a modern state. When armed with autonomy and discretion, bureaucratic organizations are able to act innovatively and produce outcomes that outperform what would have been achieved if rules alone were followed.[9] The present study showed this in the context of the administration of a risk assessment instrument. Even if this is an isolated example, it indicates that law-makers should not be opposed to allocating discretionary powers to bureaucratic agencies.

It should be noted that one of the limitations of this study is that it cannot unveil the psychological underpinnings of the authorities' behavior. The literature has identified several possible ways to explain it. For example, Wilson (1980) considers risk aversion as an important parameter. The authorities might be afraid that a lower level of supervision might result in a bad outcome (e.g., an offender committing a crime) which would put them in the midst of a scandal. Alternatively, Maynard-Moody and Musheno (2003, p. 13) mention that another plausible motive is the

way to describe the authorities' intervention in the risk-assessment procedure. In fact, in this particular instance, the manipulation led to superior outcomes, thus putting a positive spin on the term.

[9]This recommendation squares with the findings of Kuziemko (2013) who demonstrated that parole board discretion has a positive impact on inmates' rehabilitation efforts and possibly also on recidivism outcomes.

"workers' beliefs about what is fair or the right thing to do." Between fear and fairness there are certainly many other possible motivations. However, identifying the one that best fits the actions of Washington State's bureaucrats is beyond the scope of this study.

Finally, it should also be highlighted that given the information available to me, I cannot say with certainty whether the decision to manipulate the scores in order to assign a higher supervision level was taken by high-ranking officials within the correctional authorities (managers—appointed politically or otherwise), or by individual lower-ranking officers acting independently. The fact that the manipulation was widespread and affected a multitude of LSI-R scores (as shown in Fig. 1), indicates that there may have been an organized and centralized directive, originating from managerial echelons, to implement the index with some latitude. On the other hand, it is also possible that different correctional officers across several sites, responding to the same range of incentives described in the previous paragraph, could have independently decided to manipulate the scores, leading to the large effect observed. However, I do not have sufficient information to determine the accuracy of either conjecture and this constitutes another limitation of the present study.

The paper proceeds as follows. Section 2 gives a brief overview of the policy changes made in Washington's criminal justice system in the recent past. Section 3 describes the data sets used, their characteristics, and the sources that made them available to me. Section 4 gives an account of the empirical method used in this study. Section 5 presents and discusses the results, and Section 6 concludes.

## 2 Background on Washington's correctional policies

In Washington State, the agency responsible for managing prison facilities and community supervision of offenders is the Department of Corrections (DOC), which employs approximately 8,100 employees. It supervises individuals residing in the state prisons and it makes sure that court-ordered supervision conditions are complied with by offenders residing in the community (Department of Corrections, 2014). Therefore, the authorities/bureaucrats that this study focuses on are DOC officials in charge of implementing the state's community supervision program.

Research and scientific support is provided by the Washington State Institute for Public Policy (WSIPP), a public research institute which evaluates public policy programs, such as the one described in the present study, and proposes new programs or modifications to existing ones, when appropriate. Ultimately, legislation on criminal justice matters is enacted by the legislature of the state. Such legislation may then be further specified by DOC policies, if operational details need to be made concrete.

In 1999 the Washington legislature passed the Offender Accountability Act (OAA), which set two goals for 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 they caused in the commission of their crime. 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 society at large.

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

The LSI-R questionnaire consists of 54 items, divided into ten subcomponents; a score of 0 to 54 was assigned to each offender based on his or her answers.[10] The 54-item survey comprises both static and dynamic risk factors.[11] Each item contributes either a 0 or a 1 to the offender's total score.[12] A 0 means that the relevant risk factor is not present, and a 1 means that it is. The LSI-R score of an offender is the sum of the 1s recorded. Higher scores correspond to higher risk, lower scores to lower risk. The Appendix presents the list of the 54 items as they appear on the LSI-R questionnaire.

Another set of questions, independent of the LSI-R, addressed the second criterion, the harm done to society. Both criteria combined assigned offenders to one of the four supervision categories in the risk management system: RMA (Risk Management A), RMB, RMC, and RMD, with RMA corresponding to the highest-risk offenders and RMD to the lowest-risk.[13] Therefore, the score of

[10] The subcomponents are (number of items in each subcomponent 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).

[11] "Static" factors are those that cannot change over time, such as criminal history, age, or race. "Dynamic" factors are those that can change over time through treatment or intervention (e.g., drug dependency). 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.

[12] Most items have a "Yes" or "No" answer; each answer is scored 1 or 0 respectively. Some questions can be answered on a scale from 0 to 3, but the responses are also converted to a binary 0 or 1 form by way of a simple conversion method, which can be found in the Appendix.

[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 were convicted of a violent crime, if they are a very serious sex offender (Level III), if they are dangerously mentally ill, or if they have other indicators of a violent history. They are RMB if their LSI-R score is 41 to 54 and they are convicted of a non-violent crime, if their LSI-R score is 32 to 40 and they are convicted of a violent crime, if they are a serious sex offender (Level II), or if they have other indicators of a high level of needs.

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 the higher-risk RMA and RMB categories, and fewer resources to the RMC and RMD categories (Barnoski and Aos, 2003). The resources refer to supervision intensity once offenders are again considered at-risk in 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.

These goals, as set by the legislature, were then implemented on the field by the correctional authorities. Their task was to follow the aforementioned rules in order to assign offenders to the proper supervision level. However, as noted, the authorities deviated from the rules in an attempt to improve the expected outcome.

In this sense, an analogy can be drawn between the behavior of the authorities and the decision making process of a rational actor according to standard microeconomic theory. 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. As a result, 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 instrument is enhanced, even though the instrument has been manipulated.

On the other hand, the authorities are also facing a constraint. The constraint 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. For example, items in the Criminal History section of the questionnaire are not good candidates for manipulation. It would not be possible to report that an offender has three or more convictions on item 3, when in fact he or she has two. Therefore, the manipulation can target without raising suspicions primarily 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 4 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.

They are RMC if their LSI-R score is 24 to 40 and they have not been classified as RMA or RMB, or if 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 set by the Washington State authorities and they are not the same as those suggested by the creators of the LSI-R (Andrews and Bonta, 1995).

# 3 Data

The data used in this study were provided by WSIPP. Specifically, WSIPP made available to me data on the LSI-R score dating from 1999 to 2008 (the use of the LSI-R was discontinued in August 2008). This data set, apart from the LSI-R score for individual offenders, also includes the answers the offenders gave to each of the 54 items on the questionnaire—that is, one can see the entries as recorded by the authorities at the time of the administration of each LSI-R questionnaire. This is crucial for the analysis, because it allowed me to identify the item(s) where the manipulation of the index might have occurred.

WSIPP also provided data on the classification of offenders into the four risk categories (RMA, RMB, RMC, and RMD) from 2000 to 2008. Last, WSIPP provided demographic, criminal history, and recidivism data for offender cohorts from 1990 to 2004. After merging the three data sets, following WSIPP's precedent (Barnoski and Aos, 2003), I restricted the sample to cases in which the LSI-R was administered within 90 days from the day an offender became at-risk in the community. If there was more than one LSI-R interview within that period, I used the one closest to the release date. The data set is organized by offense, and each offense corresponds to only one interview. The resulting data set contains 51,957 offenses committed by 47,154 individual offenders, spanning the period July 2000 to September 2004.

Table 1 shows the composition of the data set with regard to gender, race, and age of the subjects, as well as the nature of the most serious offense they had been convicted of (in seven broad categories) and the type of their sentence (imprisonment or community sentence). The sample consists mainly of white, male, adult (young and early middle age) offenders who served sentences in the community. Table 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, Fig. 1 shows the distribution of the LSI-R scores, which would have been fairly normal had it not been for the noticeable large breaks at scores of 24, 32, and 41.[14] Fig. 2 presents cumulative counts and percentage data for the way the LSI-R score translates to each of these four categories. Note that even though crossing any of the three thresholds does not guarantee that an offender will be moved from one supervision category to the next, 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

[14] As already indicated, 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 the crime committed. The raw unprocessed data on LSI-R scores that I received from WSIPP show similar discontinuities at the cut-off points.

roughly 6 percent with a score of 24. Similar but less dramatic patterns can be observed at the other two thresholds.

This last observation suggests that the authorities adhere closely to the thresholds of the LSI-R grid when they are making risk classification decisions. This seems to explain their tendency to assign extra points to offenders who are close to a threshold so that they cross it and accordingly, with a considerable degree of certainty, are pushed to a higher level of supervision.

Table 2 presents recidivism information for the sample. A recidivism event is any conviction for a felony or misdemeanor during a 36-month period after an offender has been at-risk in the community (Drake, Aos, and Barnoski, 2010). Technical violations of community custody conditions do not constitute recidivism events for the purposes of this study. A 12-month adjudication period is also allowed. Table 2 shows that the sample has a general recidivism rate of almost 50 percent. Apart from the general rate, more specific recidivism outcomes are also reported—namely misdemeanor, felony, property felony, drug felony, and violent felony recidivism. Finally, the table provides recidivism information for each of the specific risk categories. It is worth noting the high rate of violent felony recidivism among RMA offenders (16.38 percent), an observation that turns out to be important for the evaluation of the manipulation's effectiveness.

# 4 Empirical Methodology

## 4.1 Identification of manipulated items

To identify the items used most often by the authorities in manipulating the scores (the "manipulated items") one needs to see the distribution of offenders that answered each item affirmatively on the LSI-R questionnaire.[15] Fig. 3 presents, as an example, the graph for item 1. Graphs for the remaining 53 items exhibit a similar pattern and do not alter the analysis. As expected, as the score increases, so too does the percentage of offenders answering each item affirmatively. For example, for offenders with a score of 1, the percentage is very low for any given item. Likewise, for higher scores the percentage of those who answered affirmatively is higher for all the items.

My initial hypothesis was that it should have been obvious where (for which items in the list of 54) the manipulation of offenders' scores took place. If the scores were intentionally manipulated, this would likely have been done in an organized rather than a random fashion. Therefore, for some items there would have been a discrete increase in the percentage of offenders with a score of 24, 32, or 41 who answer affirmatively. However, Fig. 3 and the other 53 similar figures did not give a clear picture as to which items the authorities used to adjust the offenders' scores.

The solution is offered by the regression discontinuity design. By zooming in on the three thresholds for each item, the design can unveil details that are not obvious from the figures. As

[15] The expression "distribution of offenders" is used loosely here because, as indicated in Section 3, the data set is organized by offense, since some offenders committed multiple offenses.

noted, it is expected that the percentage of offenders with a score of 24 who have answered affirmatively any of the 54 items should be higher than the percentage of offenders with a score of 23. However, the regression discontinuity can identify the items for which the difference in the two percentages is greater than what would have been expected in the absence of manipulation.

To implement the design, I use the micro data (the "yes" or "no" answers provided by the offenders) to run a binary response model, where the dependent variable is a dummy that takes the value 1 if offender $i$ (where, $i = 1, \ldots, 51,957$) has answered an 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 (24, 32, or 41) and 0 otherwise ($rd_i = 1\{score_i \geq threshold\}$). 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:

$$yesonitem1_i = \beta_0 + \beta_1 rd24_i + \beta_2(score_i - 24) + \beta_3 rd24_i(score_i - 24) + \epsilon_i. \quad (1)$$

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, with respect to the manipulated items for the jump at threshold 32, the first of the 54 regressions has the form:

$$yesonitem1_i = \beta_0 + \beta_1 rd32_i + \beta_2(score_i - 32) + \beta_3 rd32_i(score_i - 32) + \epsilon_i. \quad (2)$$

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

$$yesonitem1_i = \beta_0 + \beta_1 rd41_i + \beta_2(score_i - 41) + \beta_3 rd41_i(score_i - 41) + \epsilon_i. \quad (3)$$

Following the literature on binary response models (Wooldridge, 2002; Angrist and Pischke, 2009), I use OLS to estimate the above regressions. Due to the fact that the explanatory variables are discrete and take on only limited values, the case for relying on OLS is even stronger 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 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).[16] From the regressions of the micro

[16] Therefore for threshold 24 I use observations in the range 19–29; for threshold 32 I use observations in the range

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. Fig. 4 presents the three graphs that correspond to item 1. These graphs reproduce the aggregate data presented in Fig. 3, but they focus around the thresholds. The lines of best fit for the micro data produced by the regression discontinuity design are superimposed on the three panels.

To be more precise, in this paragraph I explain how the numerical results obtained from the above regressions can be represented graphically. An example of such a representation is Fig. 4. I focus on the first panel, relating to threshold 24, but the analysis holds *mutatis mutandis* for the other two panels as well. Specifically, I used the predicted values for scores of 19–23 generated by Eq. (1) to create a counterfactual value of the percentage of offenders with score of 24 who should have answered a given item affirmatively if there was no manipulation. This counterfactual value is obtained by a linear extrapolation of the predicted values for scores of 19–23. Then the regression discontinuity estimate, $\hat{\beta}_1$, measures the difference between this counterfactual value for offenders with a score of 24 and the actual predicted value for the same offenders that is obtained from the regression of Eq. (1). Graphically, this difference is the vertical distance between the two fitted lines in the three panels of Fig. 4. The numerical value of the regression discontinuity (RD) estimates and their standard errors (SE) are reported above each graph.

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 me a sense of the discontinuity in percentage terms and renders the size of the RD estimates comparable across regressions.[17]

### 4.2 Assessment of manipulation effectiveness

To evaluate the effectiveness of the instrument manipulation by the Washington State authorities, one has to determine whether the manipulated LSI-R scores are good predictors of outcomes. Moreover, the present study generates "corrected" LSI-R scores—that is, scores cleansed of the manipulation by omitting the items in the LSI-R questionnaire that were mostly used by the authorities in the manipulation process. Therefore, to assess the effectiveness of the manipulation,

27–37; and for threshold 41 I use observations in the range 36–46.

[17]For example, for item 1 the regression for threshold 24 gives me an RD estimate of 0.01. The constant for the same regression is 0.78. Then the number $\frac{0.01}{0.78} = 0.013$ tells me that the jump at the threshold is 1.3 percent the size of the constant (the vertical distance between the lower (red) fitted line and the x-axis at the threshold—the regressions are centered on the thresholds). This percentage approach allows me to see how different items compare to each other. For a given RD estimate, a small constant would generate a large jump in percentage terms, while a large constant would generate a small jump in percentage terms.

one would need to show that the manipulated scores predict outcomes better than the corrected scores. In that case, the authorities would be in a position to justify their decision to manipulate the scores. The outcomes used in the present study are various types of recidivism events.

The method chosen is commonly employed by the literature in the field of predicting recidivism risk (Mossman, 1994), but more recently has also been used in economics (Schularick and Taylor, 2012; Caballero, 2012). It is the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUROC) gives a measure of the "effect size" or accuracy of an instrument (Quinsey, Harris, Rice, and Cormier, 2006). In this study's setting it represents the probability that a random pair of non-reoffending and reoffending offenders will be correctly ranked by an instrument in terms of their risk (Bamber, 1975; Hanley and McNeil, 1982).

The ROC curve is plotted by using each point of an instrument, such as the LSI-R, as a possible cut-off point, $c$. For example, if a score of 1 on the LSI-R grid were a cut-off point, all offenders with a score equal to or greater than 1 would be predicted to reoffend, and all offenders with a score lower than 1 (that is, zero) would be predicted to not reoffend. Following this classification, one can compute the true positive rate or "hit rate" (also known as $sensitivity$)—that is, how many of those predicted to reoffend actually did reoffend—as well as the false positive rate or "false alarm rate" (also known as $1 - specificity$)—that is, how many of those predicted to reoffend did not in fact reoffend. The point that corresponds to these two rates for $c = 1$ is then plotted on the graph. The same process is followed for all the scores on the LSI-R grid (0–54, plus a trivial point for scores greater than 54). Finally, the points are connected with straight lines to generate the ROC curve. Therefore, when all the scores on the LSI-R grid are used as cut-off points, the ROC curve has 56 points connected by straight lines.

An effective risk assessment instrument should predict recidivism events better than chance, and as a result, the ROC should lie above the 45-degree line. Moreover, a higher AUROC indicates higher predictive power for the instrument. Therefore, by measuring the AUROC using the manipulated LSI-R scores and then comparing it with the corrected scores, one can determine which of the two sets of scores is a better predictor of recidivism events.

# 5 Results and discussion

## 5.1 Identification of manipulated items

By running the 54 regressions specified in Eqs. (1)–(3) for each of the three thresholds, I obtained the regression discontinuity estimates and their standard errors for each of the 54 items. The marginal effects for the RD coefficients generated by both probit and logit regressions are extremely close to the coefficients obtained by regular OLS.[18] This justifies the use of the latter for the analysis

[18] The regression results of the nonlinear models have been omitted for simplicity.

that follows. In this section, I use these OLS results to identify the items where the manipulation occurred, but even if I had used the estimates from either of the nonlinear models, the manipulated items would have been the same.

#### 5.1.1 Threshold 24

Fig. 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. Note that there are 6 positive and 5 negative statistically significant RD estimates. The highest positive estimate is over 0.05, for item 54, the last item on the questionnaire. It is followed by the estimates for items 31, 53, 42, and 30.

In panel (b) of Fig. 5, I again present the RD estimates subdivided according to the degree of subjectivity of each of the 54 items. Specifically, I grouped the items in three categories: *a*) subjective—that is, items that are manipulable and that the interviewer could tamper with. An example is item 54: "Attitude poor toward supervision," *b*) objective—that is, items that are not manipulable and have specific, verifiable answers. An example is item 1: "Any prior adult convictions," and *c*) middle—that is, any items that do not fall into one of the previous two categories. An example is item 21: "Financial Problems."[19]

Given this grouping, panel (b) shows 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 items 53–54, in the Attitudes/Orientation section warrant our attention. Not only are they clearly subjective, but they also are large and significant, as already shown in panel (a). Note also 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.

A last level of analysis presents the RD estimates divided by the respective constant for each regression. As indicated above, this gives a picture of the jump in percentage terms. Panel (c) presents 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 ones among the positive estimates. The size of the jump is almost 20 percent 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 in LSI-R scores at threshold 24, manipulation is occurring for the non-objective items on the questionnaire (30, 31, 42, 53, and 54). Items 53 and 54, which are highly subjective, should be especially noted.

[19] Items 18–20, 23–25, 30–32, and 51–54 are classified as "subjective." Items 1–11, 15–17, 28, and 48–49 are classified as "objective." Items 12–14, 21–22, 26–27, 29, 33–47, and 50 are in the "middle" category.

#### 5.1.2 Threshold 32

Fig. 6 repeats this exercise for the jump at threshold 32, which as can be seen in Fig. 1 is less pronounced than the jump at threshold 24. Here, there are 3 positive and 3 negative statistically significant RD estimates. Item 53 is again one of the 3 positive estimates, having the second highest value after item 10.

Panel (b) gives the RD estimates based on the subjectivity of the items, and shows more mixed signals, since item 10, which is in the Criminal History section of the questionnaire, 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 a 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 the size of the constant. The jumps here, however, are much less pronounced in percentage terms than at threshold 24.

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

#### 5.1.3 Threshold 41

Finally, Fig. 7 repeats the exercise for the jump at threshold 41. Now, there are 4 positive and 1 negative RD estimate that are significant. Not surprisingly, as seen in panel (a) of Fig. 7, the estimate for item 53 is once again one of the 4 positive estimates.

Panel (b) presents the RD estimates based on the subjectivity of the item, and here high RD estimates for objective items are not observed, as was the case for threshold 32. Items 50 and 45 are in the middle category, followed by item 53 in the subjective category. 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. Large jumps for objective items are not found, 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 Fig. 7.

#### 5.1.4 Overall assessment and verification of analysis

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 size of the RD estimates or the RD/Constant estimates that are statistically

significant at the 5 percent level.[20] Specifically, I found item 53 to be involved in the manipulation of all three thresholds. Items 30, 31, 42, and 54 were used at threshold 24, and items 45 and 50 were used at threshold 41.

Therefore, item 53 seems to be the item that is primarily responsible for the three discontinuities observed in Fig. 1. I believe that the heavy use of this item in the manipulation process has to do with the fact that it is a highly subjective item, asking the offender to indicate whether he/she has a poor attitude about his/her sentence.

This is a preliminary verdict for the manipulated items, and I verify it by recalculating the LSI-R score after excluding these seven items from the calculation. Accordingly, the new index has a maximum score that is seven points lower.[21] Fig. 8 presents this new distribution of the LSI-R score. The jumps have disappeared and the overall distribution looks almost perfectly normal. I believe that this graph offers strong support for the fact that the items identified as manipulated were the ones used by the authorities to bump-up offenders' scores.

### 5.2 Assessment of manipulation effectiveness

Table 3 and Figs. 9 and 10 present the results on the issue of manipulation effectiveness. Table 3 shows the AUROCs for different types of recidivism and groups of offenders (models (1)–(5)) using either the manipulated or the corrected LSI-R scores. In addition, it gives the $p$-value for a $\chi^2$ test of the hypothesis that the two AUROCs for each of the five models are equal. The results present an interesting trend. As the type of recidivism becomes more serious, the manipulated scores become better predictors of recidivism than the corrected scores.

Starting the analysis from model (1) in Table 3, note that the corrected scores are better predictors of a recidivism event. Recidivism in this model is defined in the most generic way, including both felony and misdemeanor events. The AUROC for the corrected scores is 0.6890, while for the manipulated scores it is 0.6855. The difference, though small, is statistically significant.

Model (2) repeats the experiment, but this time the recidivism event includes only felonies. Here again, the AUROC for the corrected scores (0.6764) is greater than the AUROC for the manipulated scores (0.6727). The difference is again small but statistically significant. It should also be noted that the predictive power of both scores gradually falls as the predicted recidivism

[20] 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.

[21] The three thresholds of the new index were calculated in a slightly more complicated way. First I found the percentage of offenders that answered each of the seven items affirmatively. Included in this calculation were offenders with scores on both sides of each of the three original thresholds, 24, 32, and 41, as these bandwidths were defined in Section 4.1 above. Then I summed these seven percentages to get 3.46, 4.66, and 5.93 respectively. Finally, I subtracted the sums from the respective original thresholds to get the new thresholds. Rounded to the nearest integer, the new thresholds are 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 the seven manipulated items affirmatively. So what I needed to subtract depended on the sum of the seven averages around the three thresholds.

event becomes more specific—felonies in this model as opposed to felonies and misdemeanors for model (1).

The picture of the first two models is attenuated significantly in model (3), where recidivism is restricted to violent felonies. Now the AUROC for the two scores is almost identical, 0.6423 for the corrected scores and 0.6421 for the manipulated scores, a difference that is no longer statistically significant. Thus this model can be best described as a turning point, indicating a relative balance in the predictive power of the two sets of scores.

Model (4) limits the sample to high-risk offenders who have a score 41–46 in the manipulated LSI-R grid and received the highest level of supervision intensity, RMA. It is the group of high-risk offenders who would have been most likely to reoffend by committing a violent felony, the outcome variable for this model as well, since according to the classification rules, violent offenders with a score over 41 are assigned to the RMA category. Now, the manipulated scores become the better predictor of recidivism events with an AUROC of 0.5294, compared to 0.5181 for the corrected scores. Note, though, that this difference is not statistically significant at any conventional level ($p$-value = 0.1873) and that the predictive ability of both scores has decreased considerably. The latter observation is due to the fact that the instrument has been reduced to only seven thresholds, $c$, (for scores 41–46), which obviously decreases the ability of the instrument to make good predictions—at least compared to 56 thresholds, as was the case for models (1)–(3).

Finally, model (5) shows that if the sample of high-risk offenders is restricted further to the most populous category of violent offenders, namely those who committed assault (they constitute almost 14 percent of the entire sample as shown in Table 1), the predictive power of the manipulated scores increases further. Specifically, model (5) indicates that the AUROC for the manipulated scores is 0.5794, whereas for the corrected scores it is 0.5430. This difference is statistically significant at the 5 percent level. Therefore, when the sample is limited to violent recidivism events committed by high-risk offenders who have already committed a violent crime (assault), the manipulated scores outperform the corrected scores at a statistically significant level.

Thus Table 3 reveals the outcome of the subtle score manipulation undertaken by the Washington State bureaucrats, or, in other words, what the authorities achieved by manipulating the scores. Their manipulation generated scores that were more accurate predictors of serious recidivism events, especially when these events involved high-risk offenders. Serious events were violent felonies, and high-risk offenders were offenders with a score of 41–46 on the manipulated LSI-R instrument who received RMA supervision intensity. The effect is the most pronounced when the crime for which they were originally convicted was also violent (assault).

It should be highlighted, of course, that the corrected scores slightly outperform the manipulated scores for generic and less serious types of recidivism. This shows that the authorities were particularly successful in measuring recidivism risk and, accordingly, allocating the appropriate

supervision level to serious offenders.[22]

It is particularly interesting that the authorities did not target specific categories of offenders when they added points to offenders' scores. Georgiou (2014) showed that even though demographic, crime, and other characteristics are not entirely similar on both sides of the three thresholds of the manipulated scores, any dissimilarities that exist are isolated and do not amount to an intentional targeting of a specific group of offenders. In other words, the authorities did not add extra points just because an offender was, e.g., white or male or had committed a sex crime. However, Table 3 shows that the manipulated scores are particularly good at predicting violent recidivism offenses committed by offenders who committed assault, despite the fact that this group was not targeted by the authorities. This means that what prompted the authorities to manipulate an offender's score is unobservable and unmeasurable by standard variables available to the researcher. The authorities seem to have used a parameter or combination of parameters that only they could identify in the process of administering the LSI-R instrument, perhaps thanks to intuition or gut instinct.

The policy implications of these results are twofold. First, the authorities showed that a risk prediction instrument can produce better results for specific types of recidivism events (violent felonies committed by high-risk offenders in this case) if it is manipulated properly. This result contrasts with standard premises in the risk assessment literature, according to which actuarial instruments should not be tampered with.

Most important, 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 innovative bureaucracy that not only implements but, occasionally, also improves standardized procedures.

A standard criticism of bureaucrats is that they follow rules with rigid, unbending, almost religious devotion; that they refuse to take initiative because they are afraid of the responsibility a potential initiative may carry. The correctional authorities of Washington State are a case study of the reverse phenomenon. Not only did they take the initiative to manipulate the scores, but also their initiative led to an improved outcome, scores that predict more accurately serious recidivism events.

It should further be noted that the rules that Washington authorities bypassed (the proper administration of the actuarial instrument) were implemented in order to limit the authorities' discretion in making decisions (assigning supervision levels). By breaking the rules and exercising

[22]The strategy of the authorities could be open to criticism for requiring increased supervision of offenders who did not deserve it. However, from a welfare perspective, the benefit of preventing violent felonies through increased supervision would likely have outweighed the increased inconvenience of extra supervision allocated to cases that did not warrant it. As already indicated, though, despite the fact that the authorities went to great lengths to measure risk accurately and provide supervision accordingly, the differential supervision program operated in Washington did not succeed in reducing overall recidivism levels (Georgiou, 2014).

discretion, these authorities demonstrated that discretion combined with a standardized process can provide a better outcome than rules alone.

It is therefore not unreasonable to suggest that if legislatures give bureaucracies more authority, autonomy, and discretion in carrying out their responsibilities, improved outcomes may be achieved. Although the case presented here is isolated and limited in scope, it is highly suggestive in contradicting a deeply rooted prejudice about the function and effectiveness of bureaucracies.

# 6 Conclusion

In this paper I explore the workings of the correctional authorities in the state of Washington. In particular, I analyze the way that the authorities administered the risk assessment instrument LSI-R in order to assign supervision levels to offenders who posed different risk of reoffending.

The unusual distribution of the scores generated by the instrument served as an initial indication of improper administration. The large jumps exactly at the cut-off points that separated the supervision levels suggest that the authorities manipulated the instrument. Using the underlying data on which the instrument was built, I uncover how the manipulation was carried out and identify the specific items in the instrument that were used to manipulate the offenders' scores.

Using a regression discontinuity design, I find that seven of the 54 items on the LSI-R questionnaire were used most often. Most of these items are subjective in nature. After excluding the manipulated items from the calculation of the index, I recalculate all the scores and generate a distribution that is much smoother, as it should have been had there been no interference.

In addition, I find that the manipulated instrument produces better predictive results for serious recidivism cases, than the instrument as it is designed to be used. In manipulating the results the authorities acted in a creative way and identified offenders whose risk to society would have otherwise been treated more leniently than was appropriate.

This analysis shows that bureaucracies need not be wasteful and anachronistic mechanisms whose primary function is to prolong their existence by maximizing their budget. In this case, innovative bureaucrats bent the rules in order to ensure the proper level of post-release supervision for offenders. Even if it is an exception, the performance of the Washington State correctional authorities suggests that generalizations and oversimplifications about bureaucratic organizations are not always warranted.

# Appendix: The LSI-R questionnaire

| # | Item | # | Item |
|---|---|---|---|
| | **Criminal History** | | **Emotional/Personal** |
| 1 | Any prior adult convictions? Number_ | 46 | Moderate interference |
| 2 | Two or more prior adult convictions? | 47 | Severe interference, active psychosis |
| 3 | Three or more prior convictions? | 48 | Mental health treatment, past |
| 4 | Three or more present offenses ? Number_ | 49 | Mental health treatment, present |
| 5 | Arrested under age sixteen? | 50 | Psychological assessment indicated Area:_ |
| 6 | Ever incarcerated upon conviction? | | **Attitudes/Orientations** |
| 7 | Escape history from correctional facility? | 51 | Supportive of crime |
| 8 | Ever punished for institutional misconduct? Number_ | 52 | Unfavorable toward convention |
| 9 | Charge laid or probation/parole suspended during prior community supervision? | 53 | Poor, toward sentence |
| | | 54 | Poor, toward supervision |
| 10 | Official record of assault/violence? | | |
| | **Education/Employment** | | |
| 11 | Currently unemployed? | | |
| 12 | Frequently unemployed? | | |
| 13 | Never employed for a full year? | | |
| 14 | Ever fired? | | |
| 15 | Less than regular grade 10? | | |
| 16 | Less than regular grade 12? | | |
| 17 | Suspended or expelled at least once? | | |
| 18 | Participation/performance | | |
| 19 | Peer interactions | | |
| 20 | Authority interactions | | |
| | **Financial** | | |
| 21 | Problems | | |
| 22 | Reliance on social assistance | | |
| | **Family/Marital** | | |
| 23 | Dissatisfaction with marital or equivalent situation | | |
| 24 | Non-rewarding, parental | | |
| 25 | Non-rewarding, other relatives | | |
| 26 | Criminal family/spouse | | |
| | **Accommodation** | | |
| 27 | Unsatisfactory | | |
| 28 | Three or more address change last year | | |
| 29 | High-crime neighborhood | | |
| | **Leisure/Recreation** | | |
| 30 | Absence of a recent participation in an organized activity | | |
| 31 | Could make better use of time | | |
| | **Companions** | | |
| 32 | A social isolate | | |
| 33 | Some criminal acquaintances | | |
| 34 | Some criminal friends | | |
| 35 | Few anti-criminal acquaintances | | |
| 36 | Few anti-criminal friends | | |
| | **Alcohol/Drug problem** | | |
| 37 | Alcohol problem, ever | | |
| 38 | Drug problem, ever | | |
| 39 | Alcohol problem, currently | | |
| 40 | Drug problem, currently Specify type of drug:_ | | |
| 41 | Law violations | | |
| 42 | Marital/family | | |
| 43 | School/work | | |
| 44 | Medical | | |
| 45 | Other indicators Specify:_ | | |

*Note:* Most items take "Yes" or "No" answers, which are converted to 1 or 0 respectively for the calculation of the score. The total LSI-R score is the sum of all the 1s recorded. A few items follow a "0–3" rating scale, where a 3 or a 2 corresponds to "No," and a 1 or a 0 corresponds to "Yes" (Andrews and Bonta, 1995).

# References


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, Princeton, New Jersey, USA.

Aos, S. (2002): "Washington's Offender Accountability Act: An evaluation of the Department of Corrections' Risk Management Identification System," (Document No. 02-01-1201). Olympia: Washington State Institute for Public Policy.

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

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

Bamber, D. (1975): "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, 12(4), 387–415.

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

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., USA.

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.

Caballero, J. (2012): "Do surges in international capital inflows influence the likelihood of banking crises? Cross-country evidence on bonanzas in capital inflows and bonanza-boom-bust cycles," Inter-American Development Bank Working Paper Series No. IDB-WP-305.

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.

Davis, M. L., and K. Hayes (1993): "The demand for good government," *Review of Economics and Statistics*, 75(1), 148–152.

Department of Corrections (2014): "Strategic Action Plan, Transforming Washington State Corrections, 2015–2017 Biennium," P358 (9/2014), Washington State: Department of Corrections.

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

Duncombe, W., J. Miner, and J. Ruggiero (1997): "Empirical evaluation of bureaucratic models of inefficiency," *Public Choice*, 93(1-2), 1–18.

Fazel, S., J. P. Singh, H. Doll, and M. Grann (2012): "Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: systematic review and meta-analysis," *BMJ*, 345, e4692, doi: http://dx.doi.org/10.1136/bmj.e4692.

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

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

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

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.

Hanley, J. A., and B. J. McNeil (1982): "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve," *Radiology*, 143(1), 29–36.

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.

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. (2013): "How should inmates be released from prison? An assessment of parole versus fixed-sentence regimes," *Quarterly Journal of Economics*, 128(1), 371–424.

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

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.

Maynard-Moody, S. W., and M. C. Musheno (2003): *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service*. University of Michigan Press, USA.

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

Mossman, D. (1994): "Assessing predictions of violence: being accurate about accuracy," *Journal of Consulting and Clinical Psychology*, 62(4), 783.

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

——— (1971): *Bureaucracy and Representative Government*. New York City, Aldine.

——— (1975): "Bureaucrats and politicians," *Journal of Law and Economics*, 18(3), 617–643.

Petersilia, J. (2009): *When Prisoners Come Home: Parole and Prisoner Reentry*. Oxford University Press, USA.

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. Cormier (2006): *Violent Offenders: Appraising and Managing Risk*. American Psychological Association, Washington DC.

Schularick, M., and A. Taylor (2012): "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008," *American Economic Review*, 102(2), 1029–1061.

Singh, J. P., and S. Fazel (2010): "Forensic Risk Assessment: A Metareview," *Criminal Justice and Behavior*, 37(9), 965–988.

Singh, J. P., M. Grann, and S. Fazel (2011): "A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants," *Clinical Psychology Review*, 31(3), 499–513.

Webster, C. D., S. J. Hucker, and Q. Haque (2014): *Violence Risk-Assessment and Management: Advances Through Structured Professional Judgement and Sequential Redirections*. John Wiley and Sons, Ltd., West Sussex, UK.

Wilson, J. Q. (1980): "The politics of regulation," in *The Politics of Regulation*, ed. by J. Wilson, pp. 18–32. Basic Books, Inc., USA.

Wooldridge, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT press, Cambridge, Massachusetts.

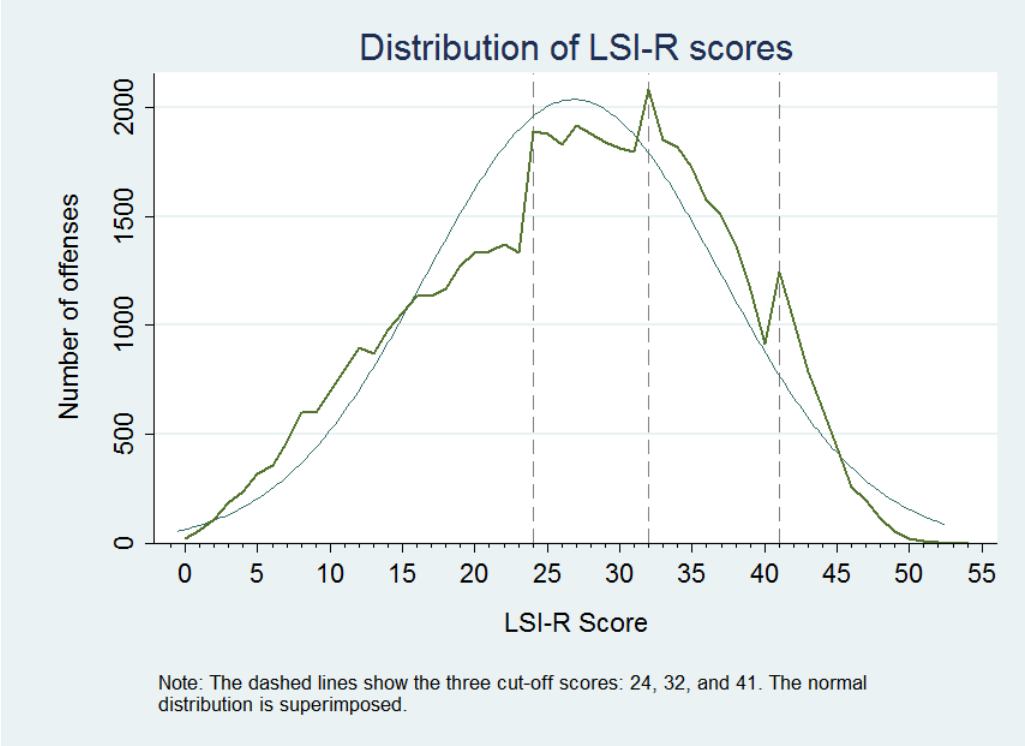


Figure 1: Distribution of LSI-R scores. The LSI-R score was the risk instrument used by Washington State to measure the risk posed by offenders. It is a scale from 0 to 54, where higher numbers indicate higher risk. Three cut-off scores separate four different levels of supervision intensity. If an offender's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

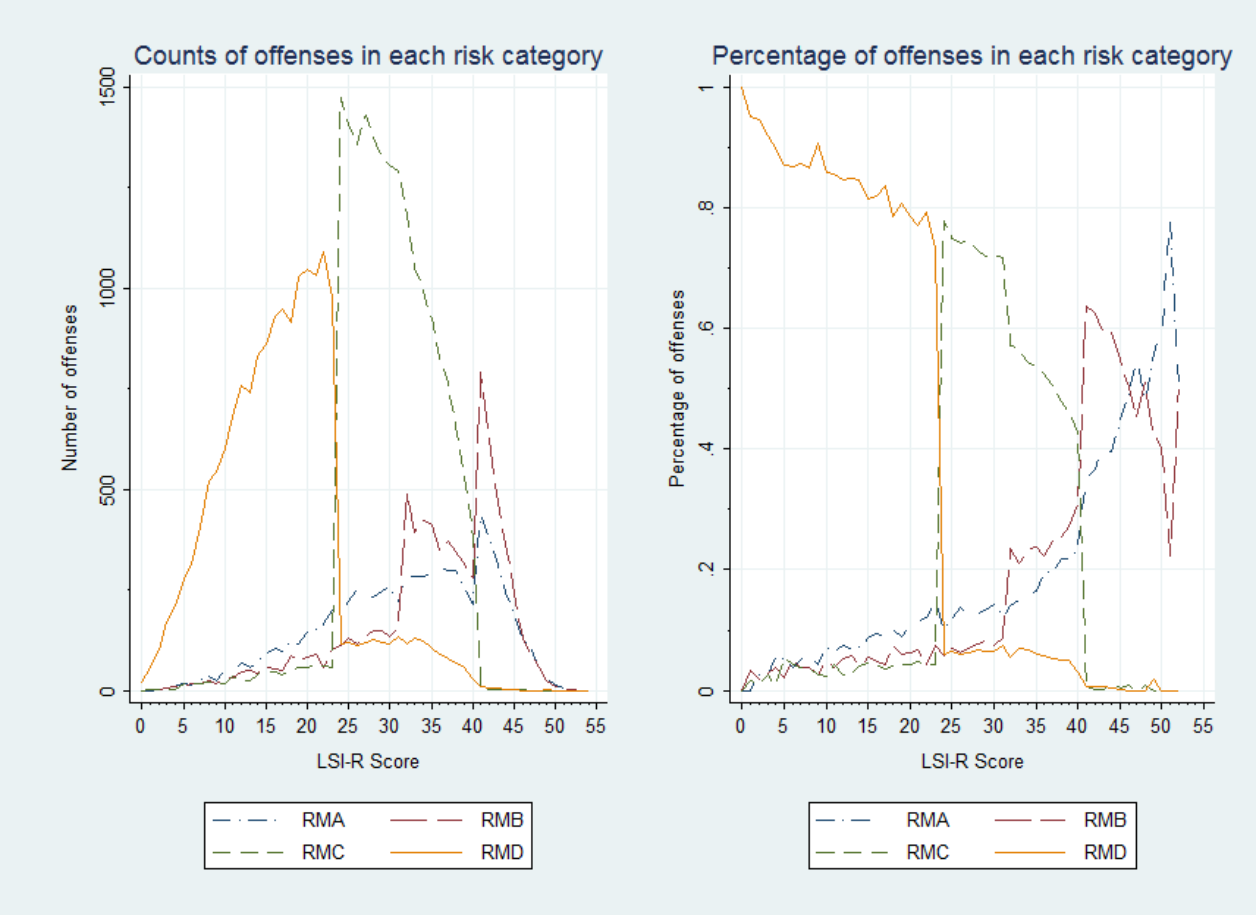


Figure 2: Counts and percentages of offenses in the four risk categories per LSI-R score. The highest-risk category is RMA and the lowest-risk category is RMD. Offenders are assigned to them based on their LSI-R score. This figure demonstrates that the three thresholds, 24, 32, and 41, are strongly binding when the authorities assign offenders to one of the four risk categories. For example, the probability that the authorities classify an offender at the low-risk RMD category is about 73 percent if his or her score is 23, but falls to roughly 6 percent if his or her score is 24. Conversely, the probability that an offender is placed at the RMC category is approximately 4 percent if his or her LSI-R score is 23, but increases to almost 78 percent if his or her score is 24.

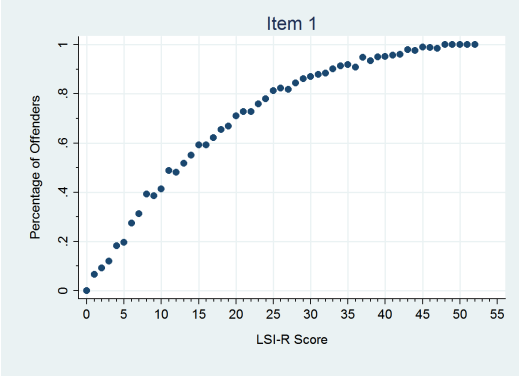


Figure 3: Percentage of offenders that answer affirmatively item 1 per LSI-R score. Item 1 is presented here as one example of the 54 items on the LSI-R questionnaire. As scores increase, the percentage of offenders that answer affirmatively each item increases and approaches 100 percent.

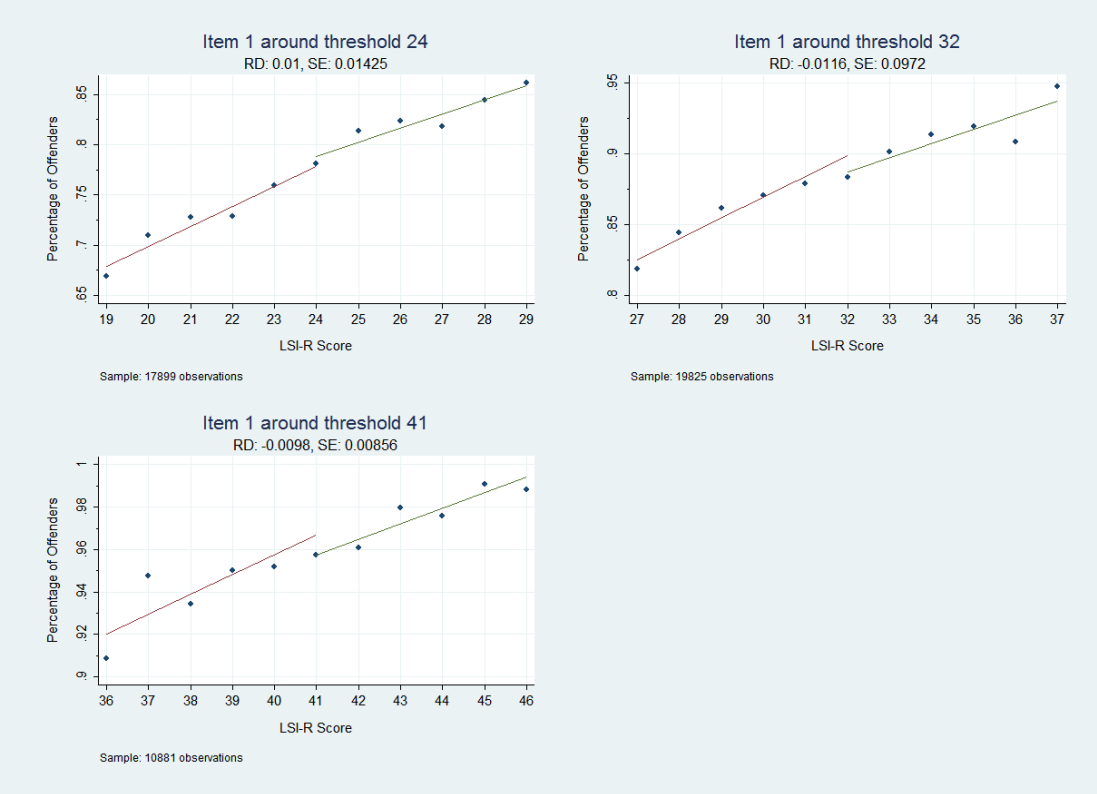


Figure 4: Three of the 162 figures (54 items times 3 thresholds) generated by the regression discontinuity design. The three figures correspond to Fig. 3 but focus around the three thresholds. The dots in the scatter plots are the actual aggregate percentage data. The two lines fit the micro data separately for the points below and above each threshold. These lines are superimposed on the aggregate percentage data. I use these graphs and the respective regressions to identify the items on the LSI-R questionnaire that were mostly used in the manipulation process.

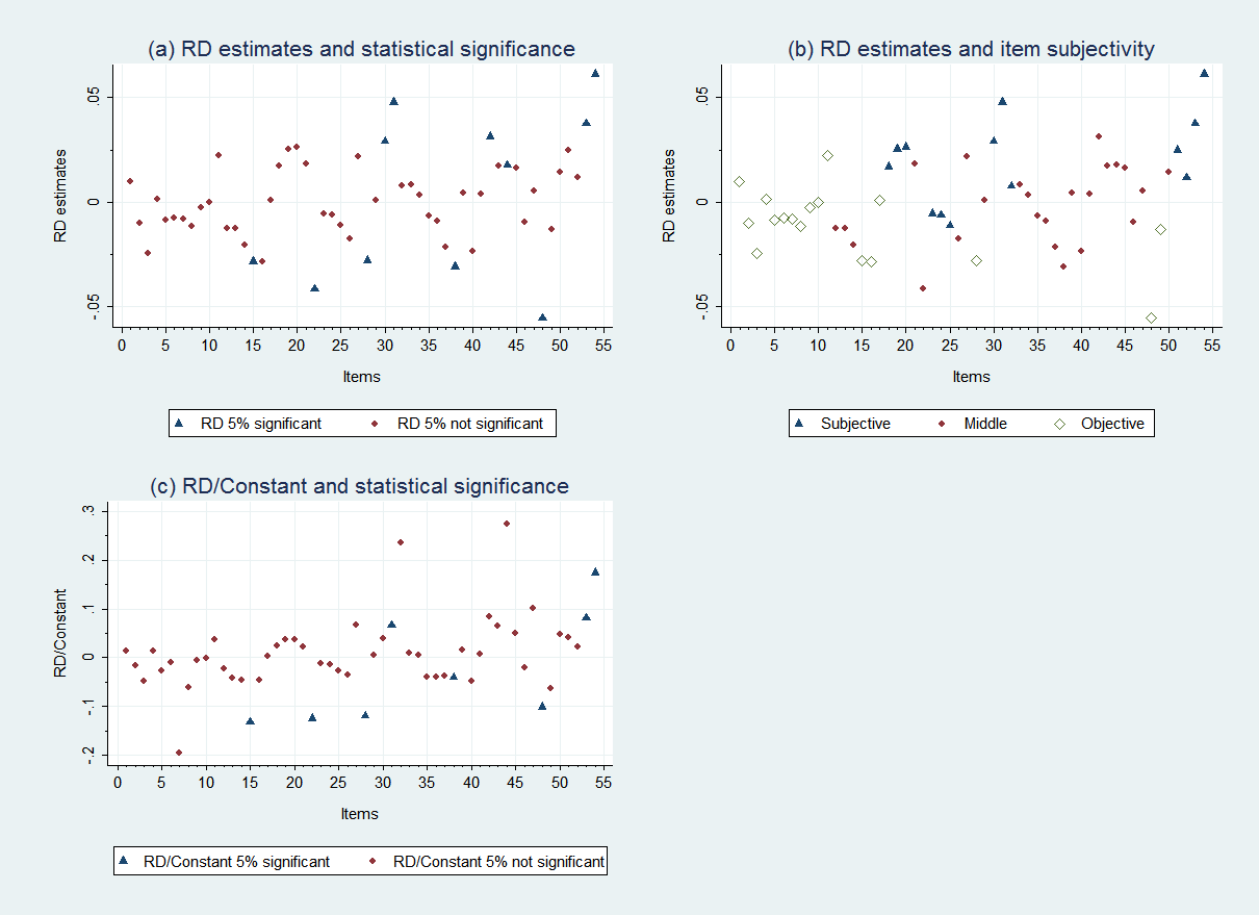


Figure 5: Panel (a) presents the regression discontinuity (RD) estimates at threshold 24 for each of the 54 items on the LSI-R questionnaire. Panel (b) presents the same estimates sub-categorized by the subjectivity of the item. Panel (c) presents the RD estimates for each item divided by the constant in the respective regression. Item(s) identified as manipulated are: 30, 31, 42, 53, and 54.

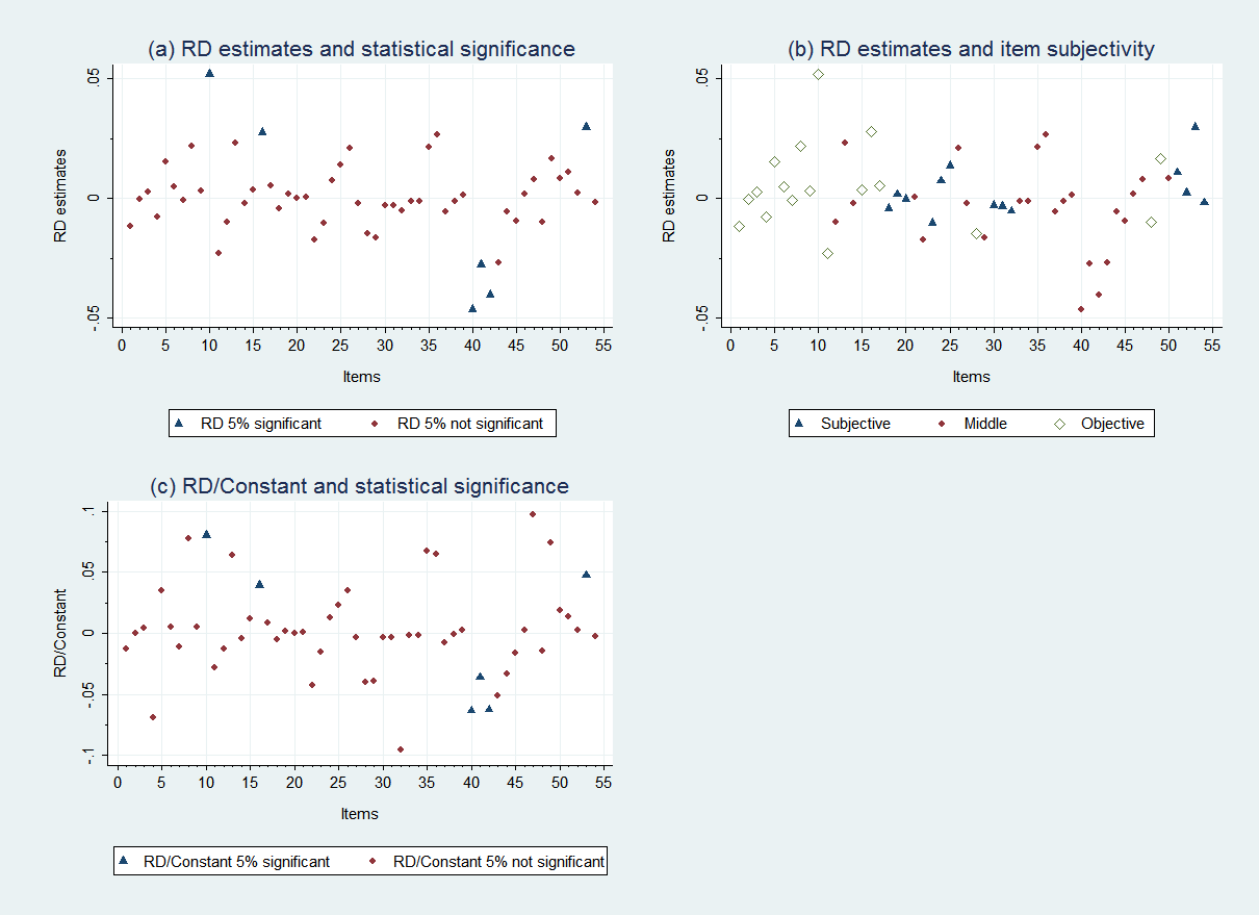


Figure 6: Cumulative data on RD estimates around threshold 32. Everything said in the note to Fig. 5 applies here, but now with respect to threshold 32. Item(s) identified as manipulated are: 53.

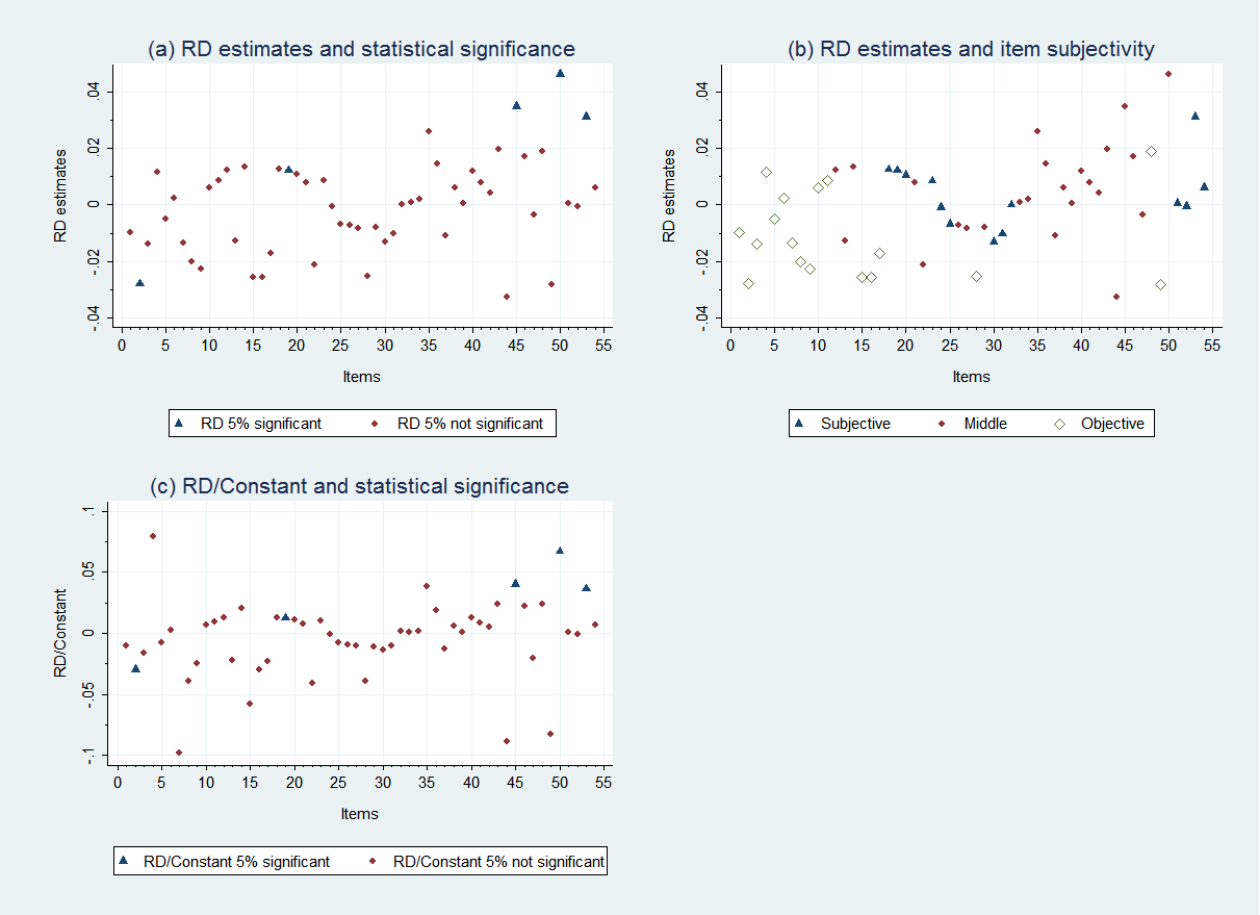


Figure 7: Cumulative data on RD estimates around threshold 41. Everything said in the note to Fig. 5 applies here, but now with respect to threshold 41. Item(s) identified as manipulated are: 45, 50, and 53.

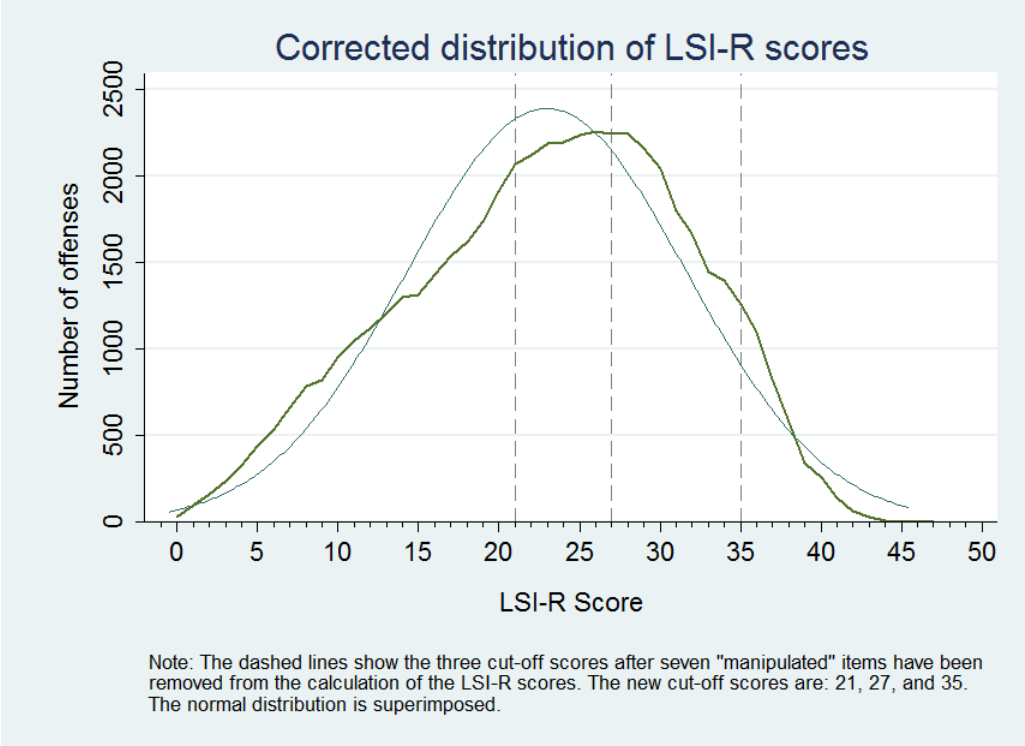


Figure 8: Distribution of LSI-R scores after omitting the manipulated items. Without the seven manipulated items, the LSI-R index becomes a scale from 0 to 47. Note that the discontinuities at the thresholds have disappeared and the scores exhibit a normal-like distribution.

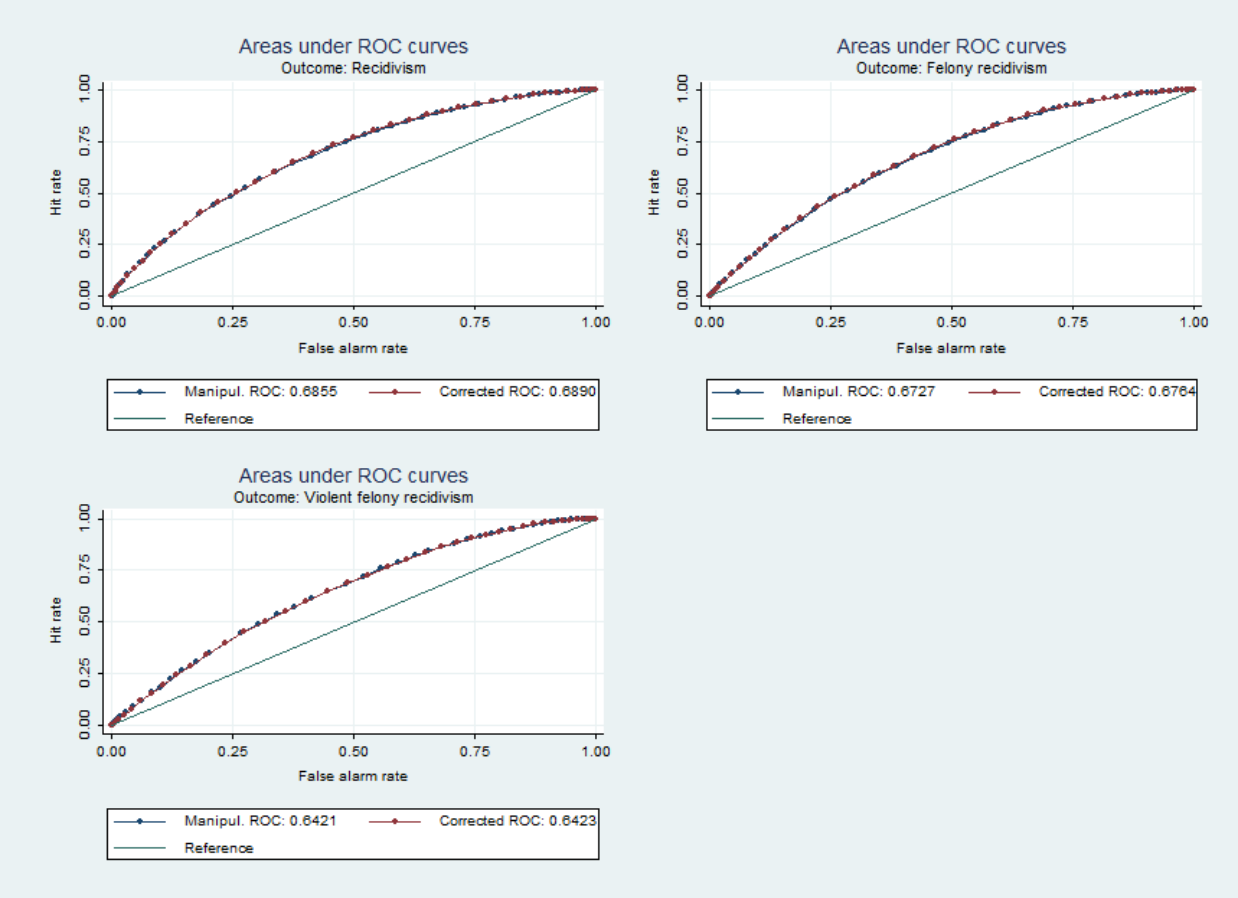


Figure 9: This figure shows the areas under the Receiver Operating Characteristic (ROC) curve for the types of recidivism where the corrected LSI-R scores have better predictive power than the manipulated scores. Measuring the area under the ROC curve is a method of assessing the predictive power of an instrument, such as the LSI-R. An area above 0.5 (depicted by the reference line in the three panels of the figure) signifies predictive power better than chance. The higher the area, the better the predictive power. In the figure, note that the corrected LSI-R scores outperform the manipulated ones in predicting general recidivism and felony recidivism. For violent felony recidivism, the difference in favor of the corrected scores is not statistically significant.

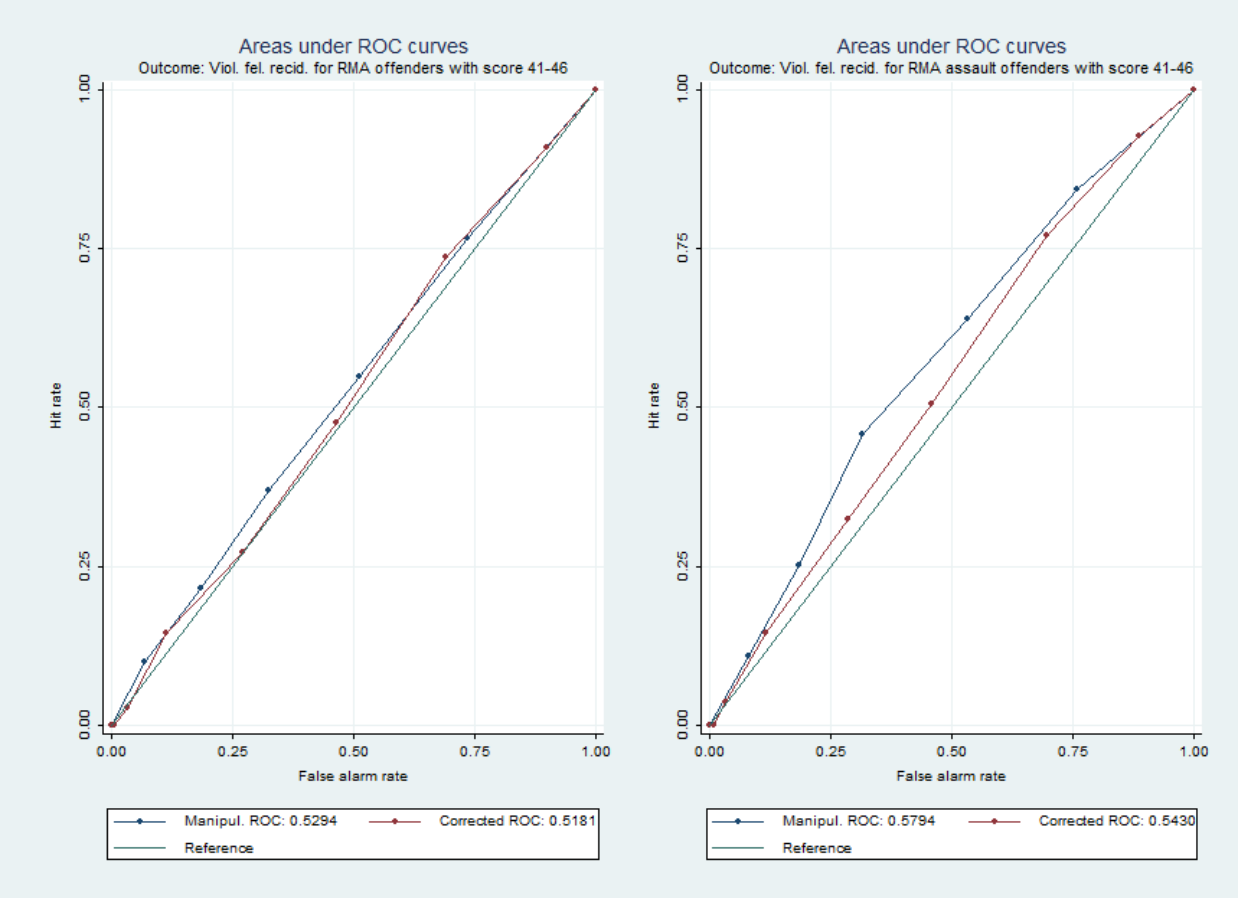


Figure 10: This figure shows the areas under the ROC curve for the types of recidivism where the manipulated LSI-R scores have better predictive power than the corrected scores. Note that the manipulated scores outperform the manipulated ones ($p$-value = 0.1873) for violent felony recidivism offenses committed by high-risk offenders (LSI-R score 41–46), who received the highest level of supervision, RMA. The difference is statistically significant in favor of the manipulated scores if the high-risk offenders were sentenced for assault offenses (right panel).

Table 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 American | 2.78 | 3.56 | 3.59 | 3.05 | 1.72 |
| Asian | 2.71 | 3.30 | 1.98 | 1.86 | 3.75 |
| Age 18–30 | 49.23 | 47.96 | 45.74 | 48.67 | 52.16 |
| Age 31–45 | 39.90 | 39.72 | 42.95 | 41.82 | 36.34 |
| Age over 45 | 10.66 | 11.92 | 10.97 | 9.33 | 11.41 |
| Drug crime | 31.74 | 12.98 | 26.23 | 39.71 | 34.24 |
| Assault | 13.76 | 30.85 | 17.79 | 8.26 | 9.98 |
| Property crime | 27.79 | 12.82 | 22.35 | 31.68 | 33.09 |
| Sex crime | 4.60 | 11.74 | 9.37 | 2.74 | 1.03 |
| Weapons | 2.82 | 2.18 | 2.60 | 3.21 | 2.79 |
| Robbery | 1.75 | 5.92 | 2.31 | 0.75 | 0.64 |
| Homicide | 0.31 | 0.80 | 0.50 | 0.12 | 0.21 |
| Community | 84.07 | 74.91 | 77.24 | 82.15 | 93.86 |
| Prison | 15.93 | 25.09 | 22.76 | 17.85 | 6.14 |
| Observations | 51,957 | 7,902 | 8,126 | 19,008 | 16,831 |

Table 2: Recidivism summary statistics in percentages

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

Table 3: Assessment of manipulation effectiveness: prediction of recidivism events

| | Outcome variable: various types of recidivism | | | | |
|---|---|---|---|---|---|
| LSI-R | Recidivism (1) | Felony recidivism (2) | Violent felony recidivism (3) | Violent felony recidivism for RMA offenders with score 41–46 (4) | Violent felony recidivism for RMA assault offenders with score 41–46 (5) |
| | Areas under the ROC curve | | | | |
| Manipul. | 0.6855 | 0.6727 | 0.6421 | 0.5294 | 0.5794 |
| | [0.0023] | [0.0024] | [0.0039] | [0.0171] | [0.0351] |
| Corrected | 0.6890 | 0.6764 | 0.6423 | 0.5181 | 0.5430 |
| | [0.0023] | [0.0024] | [0.0039] | [0.0169] | [0.0350] |
| | $p$-value for test that the two areas are equal | | | | |
| | 0.0000 | 0.0000 | 0.7370 | 0.1873 | 0.0278 |
| Observations | 51,957 | 51,957 | 51,957 | 1,686 | 356 |

Standard errors in brackets

*Note:* This table provides evidence on the superior effectiveness of the manipulated LSI-R scores in predicting certain types of recidivism events. The top part of the table presents the areas under the Receiver Operating Characteristic (ROC) curve for different outcomes. Each area is calculated by using either the manipulated or the corrected scores. A larger area indicates higher predictive ability. The outcome variable and the relevant sample is different for each model. Models (1)–(3) are self-explanatory. Model (4) includes only offenders who have scored 41–46 on the manipulated LSI-R instrument (high-risk offenders) and received RMA supervision intensity. Finally, model (5) includes only high-risk RMA offenders convicted of assault prior to their release. Note that as the recidivism event becomes more severe from model (1) to model (5), the predictive power of the manipulated instrument gradually outperforms the corrected instrument, indicating the ability of the authorities to identify possible serious recidivism cases. The second part of the table shows the $p$-value for a $\chi^2$ test of the hypothesis that the two areas (calculated using the manipulated and the corrected scores) for each model are equal. Note that this hypothesis is rejected for models (1) and (2), where the corrected scores outperform the manipulated, but also for model (5) where the manipulated scores outperform the corrected ones.